



Severe loss of precision in calculations of T -matrix integrals

W.R.C. Somerville, B. Auguie, E.C. Le Ru*

The MacDiarmid Institute for Advanced Materials and Nanotechnology, School of Chemical and Physical Sciences, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand

ARTICLE INFO

Article history:

Received 3 November 2011

Received in revised form

11 January 2012

Accepted 13 January 2012

Available online 21 January 2012

Keywords:

Light scattering

T -matrix

Extended Boundary Condition Method

Null-field method

Numerical cancellations

Spheroidal particle

ABSTRACT

A severe loss of precision is unravelled in the numerical calculation of surface integrals that appear in the Extended Boundary Condition Method (EBCM), to calculate the T -matrix elements of axisymmetric particles. We systematically study the occurrence of numerical cancellations for three basic particle shapes, namely cylinders, spheroids, and offset spheres, with typical sizes, aspect ratios and materials often studied as benchmark examples in the literature. The cancellations are evidenced both for spheroids and offset spheres, and are particularly pronounced in the latter case. The resulting loss of precision is independent from the commonly asserted problems of matrix inversion. We show that the origin of these severe cancellations can be further studied and understood by numerical investigations of the scaling of the integrands and integrals with respect to the particle size parameter. This allows us to develop a detailed mathematical proof of these cancellations. The results suggest that the EBCM method, in its usual formulation, suffers important numerical instabilities which reduce the domain of convergence for specific particle shapes that are commonly used for testing and benchmarking the method.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

From the vast array of numerical techniques available to rigorously solve Maxwell's equations [1], the Discrete Dipole Approximation [2], the Finite Differences method [3], and the T -matrix framework [4] have been used extensively to model the optical properties of nonspherical particles (the special case of spheres is very efficiently handled by Mie theory [5–7], with practically no disadvantage). Within its realm of applicability, the T -matrix approach is widely recognized for its elegant formulation of the scattering problem, which facilitates the development of semi-analytical procedures for efficient orientation averaging [8], and the treatment of multiple-scattering in ensembles of particles [9]. First introduced by Waterman [10], the T -matrix framework

considers the truncated series expansions of the incident, internal, and scattered fields in a basis of vector spherical wavefunctions (VSWFs). Such expansions read, for the incident and scattered fields, respectively [4],

$$\mathbf{E}_{\text{inc}}(\mathbf{r}) = E_0 \sum_{n,m} a_{nm} \mathbf{M}_{nm}^{(1)}(k_1 \mathbf{r}) + b_{nm} \mathbf{N}_{nm}^{(1)}(k_1 \mathbf{r}),$$

$$\mathbf{E}_{\text{sca}}(\mathbf{r}) = E_0 \sum_{n,m} p_{nm} \mathbf{M}_{nm}^{(3)}(k_1 \mathbf{r}) + q_{nm} \mathbf{N}_{nm}^{(3)}(k_1 \mathbf{r}),$$

where k_1 is the wavevector in the surrounding medium, \mathbf{r} denotes the position vector, $\mathbf{M}_{nm}^{(i)}, \mathbf{N}_{nm}^{(i)}$ are vector spherical harmonics, a_{nm}, b_{nm} are the known coefficients of the incident field, and p_{nm}, q_{nm} are the unknown coefficients of the scattered field. $n \geq 1$ and $|m| \leq n$ denote here the total and projected angular momentum numbers. Following the linearity of Maxwell's equations, the coefficients of these expansions can be linked in a linear relationship between incident and scattered field, written

* Corresponding author.

E-mail addresses: walter.somerville@vuw.ac.nz (W.R.C. Somerville), baptiste.auguie@vuw.ac.nz (B. Auguie), eric.leru@vuw.ac.nz (E.C. Le Ru).

in matrix form

$$\begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix} = \mathbf{T} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}.$$

The transition matrix (\mathbf{T}), so-called T -matrix, thus fully describes the scattering properties of the particle for an arbitrary incident beam, and formally provides an exact solution of Maxwell's equations. In practice, the infinite series expansions—and consequently the T -matrix—need to be truncated to a finite maximum order N , which (theoretically) dictates the final accuracy of the calculation.

Historically, the Extended Boundary Condition Method (EBCM) was introduced conjointly with the T -matrix framework by Waterman to calculate the matrix elements [10]. It has in fact become customary, if somewhat misleading, to see in the literature the generic appellation T -matrix being used in place of the more specific acronym EBCM. Although a variety of other techniques have been proposed to calculate the T -matrix of a given scatterer [11–16], the EBCM remains arguably the most efficient and elegant approach when it is applicable, and lends itself naturally to further analytical work.

Within the EBCM approach, the T -matrix is obtained as a matrix product [4]

$$\mathbf{T} = -\mathbf{Rg}\mathbf{Q}\mathbf{Q}^{-1}, \quad (1)$$

where the matrix \mathbf{Q} expresses the null-field condition relating the incident and internal fields, while $\mathbf{Rg}\mathbf{Q}$ describes the formation of the scattered field from the internal field. The matrix elements of $\mathbf{Rg}\mathbf{Q}$ and \mathbf{Q} are formally similar and can be expressed analytically as integrals of products of vector spherical harmonics over the particle surface. They differ only in the type of spherical Bessel function used: regular for $\mathbf{Rg}\mathbf{Q}$ and Hankel of the first kind for \mathbf{Q} . In practice, these integrals need to be evaluated numerically. For particles with symmetry of revolution, the surface integrals reduce to one-dimensional integrals with much simpler expressions (relatively speaking); the number of integrals to be evaluated is also reduced dramatically, as different m values are decoupled. For this reason, a predominant fraction of previous works, this one included, have focused on axisymmetric particles.

The EBCM has been widely applied to a variety of systems in both electromagnetic scattering (see [17] for a comprehensive review) and acoustic scattering [18]; the formulation is somewhat simplified in the latter case due to the scalar (as opposed to vectorial) nature of the problem. Despite these successes, it is also well-known that the method suffers, under some conditions, from serious numerical problems regarding convergence and loss of accuracy [19]. Several improvements have been reported on the commonly used implementation of Mishchenko [20], notably to extend the range of numerical convergence for large non-absorbing particles [21], and for strongly absorbing particles such as metals [22]. Most work on these aspects suggested that the crucial step in resolving issues of numerical stability lies in the inversion of the linear system (1). Increasing the numerical precision [23], or

improving the matrix inversion algorithm [24], was indeed found to extend the domain of convergence of the EBCM.

With a different perspective, Waterman, recently revisiting the foundations of the EBCM, noted the appearance of severe loss of precision in the calculation of the matrix elements themselves, *before* performing the inversion, clearly undermining the accuracy of any subsequent numerical operation [25,26]. Waterman also proposed some hints on where the origin of such loss of precision might lie, namely the presence of important cancellations in the calculation of the integrals. However, these findings and remarks remained without formal justification. Furthermore, Waterman's studies focused on two particular cases: acoustic scattering [26] (which is a scalar problem) and electromagnetic scattering by an infinite cylinder [25] (a 2D problem), thereby eluding the important case of electromagnetic scattering in 3D. These studies provide us with a starting point to further investigate important loss of precision in the T -matrix method for electromagnetic scattering by axisymmetric bodies, which is the subject of this work.

We here demonstrate in a systematic study that such severe loss of precision does occur in the computation of the \mathbf{Q} -matrix integrals, in particular for widely studied particle shapes that are generally considered as ideal case studies for their simple geometries. The underlying cancellations are particularly severe in the case of offset spheres, and to a lesser degree spheroids. They are also more prominent for small particles, but this may not be a problem if convergence is achieved for small N . Large particles are typically more problematic than smaller ones within the T -matrix framework, as more multipoles (i.e. a larger N) are required to describe the far-field scattering properties. While the far-field properties of small particles with moderate aspect ratios may reach a satisfying degree of convergence for moderate values of N , accurate near-field calculations or larger aspect ratios often require many more terms, and we show that term for term, the cancellations are more severe for small particles and may therefore compromise such calculations. A mathematical demonstration of the origin of such cancellations is provided, and its relation to the possible numerical loss of precision is discussed.

This work has important consequences for the application of the EBCM. Firstly, it highlights the fact that numerical/convergence issues in the T -matrix approach are not solely related to matrix inversion, but also to loss of precision in the computations of the integrals before inversion. Secondly, it suggests that more efficient implementations of the T -matrix method, avoiding or even making use of these cancellations, could be devised in the important case of spheroids. Thirdly, it shows that the convergence and accuracy of numerical implementations should not solely be tested on simple shapes like offset spheres and spheroids because of the shape-specific cancellations occurring.

2. Definitions and methods

2.1. Notations

We concern ourselves in this paper with particles having a symmetry of revolution, more specifically cylinders,

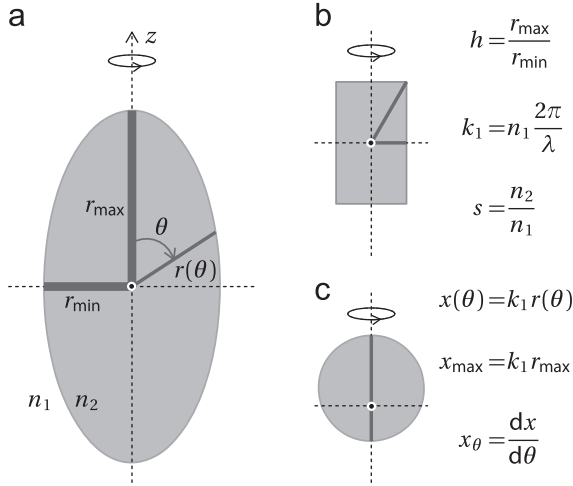


Fig. 1. Schematic illustration of the shapes and notations used to describe the scattering problem. Particles with the geometries considered in this paper include (a) prolate spheroid, (b) cylinder, (c) offset sphere, with common aspect ratio $h=2$. See Supplementary Information for the parametric equations used to define these geometries.

spheroids, and offset spheres (a spherical particle for which the T -matrix is evaluated from a point shifted from the geometrical center). Schematics of these shapes are presented in Fig. 1, together with the definition of various parameters of interest for the scattering problem. The refractive index of the embedding medium (n_1) and of the particle (n_2) are combined into a relative refractive index defined as the ratio $s = n_2/n_1$; we have used $n_1 = 1$ and $n_2 = 1.5 + 0.02i$ throughout this study, for consistency, but the results are general (we have for example also tested the case of silver particles in the visible). The incident wave is characterized by a wavevector $k_1 = 2\pi n_1/\lambda$, where λ is the wavelength in vacuum. The geometry is described in polar coordinates as $r(\theta)$; explicit expressions for the shapes of Fig. 1 are provided in Supplementary Information (S.I.). The particle aspect ratio is defined as $h = r_{\max}/r_{\min}$, where r_{\max} and r_{\min} are the maximum and minimum values of $r(\theta)$, respectively (note that this definition may differ slightly from other works in the case of a cylinder). Finally, the size parameter is defined as the maximum value of $x(\theta) = k_1 r(\theta)$ (again, this definition may differ from other works using, e.g., the radius of an equi-volume sphere).

2.2. The EBCM formulation

In order to further define our notations, we here provide a brief summary of the EBCM method applied to axisymmetric particles. We follow closely the formulation and notations of Mishchenko [4], except for the final expressions of the T -matrix integrals. Our group recently presented a more concise formulation [27] resulting from analytical simplifications of the integrals and alleviating some additional cancellations from the original EBCM equations. We therefore use these simplified expressions (summarized below) in this study since they are simpler to manipulate and more suited to the study of cancellations.

We will focus in the following on the Q -matrix. For axisymmetric particles (around the z -axis as in Fig. 1), the problem is decoupled between different values of m (angular momentum projection) and m can be viewed as an implicit fixed parameter. The matrix elements are then indexed by the total angular momentum only, denoted $n, k \geq |m|$ for row and column, respectively. We recently showed that the entire Q -matrix can be expressed with relatively simple expressions by computing the following six types of integrals (see Ref. [27] for full details):

$$K_{nk}^1 = \int_0^\pi d\theta m d_n d_k x_\theta \xi_n \psi'_k, \quad (2)$$

$$K_{nk}^2 = \int_0^\pi d\theta m d_n d_k x_\theta \xi'_n \psi_k, \quad (3)$$

$$L_{nk}^1 = \int_0^\pi d\theta \sin \theta x_\theta \tau_n d_k \xi_n \psi_k, \quad (4)$$

$$L_{nk}^2 = \int_0^\pi d\theta \sin \theta x_\theta d_n \tau_k \xi_n \psi_k, \quad (5)$$

$$L_{nk}^3 = \int_0^\pi d\theta \sin \theta d_k \psi'_k [x_\theta \tau_n \xi'_n - n(n+1) d_n \xi_n], \quad (6)$$

$$L_{nk}^4 = \int_0^\pi d\theta \sin \theta d_n \xi'_n [s x_\theta \tau_k \psi'_k - k(k+1) d_k \psi_k]. \quad (7)$$

In these and the rest of the paper, we have made the following simplification of notations:

$$\xi_n \equiv \xi_n(x(\theta)) \quad \text{and} \quad \xi'_n \equiv \left. \frac{d\xi_n(z)}{dz} \right|_{z=x(\theta)}, \quad (8)$$

$$\psi_n \equiv \psi_n(sx(\theta)) \quad \text{and} \quad \psi'_n \equiv \left. \frac{d\psi_n(z)}{dz} \right|_{z=sx(\theta)}, \quad (9)$$

$$\pi_n \equiv \pi_{mn}(\theta) \quad \text{and} \quad \tau_n \equiv \tau_{mn}(\theta), \quad (10)$$

$$d_n \equiv d_{0m}^n(\theta), \quad (11)$$

where the angular functions π_{mn} , τ_{mn} , d_{0m}^n are defined as in Ref. [4], and the Ricatti-Bessel and Hankel functions ξ, ψ are defined as in Ref. [28].

As is customary, the Q -matrix elements are grouped into a 2×2 block matrix $\mathbf{Q} = \begin{pmatrix} \mathbf{Q}^{11} & \mathbf{Q}^{12} \\ \mathbf{Q}^{21} & \mathbf{Q}^{22} \end{pmatrix}$, in relation to the two types of VSWFs used in the series expansions of the incident and scattered fields.

The Q -matrix elements are then given by [27]

$$Q_{nk}^{12} = A_n A_k \frac{s^2 - 1}{s} K_{nk}^1, \quad (12)$$

$$Q_{nk}^{21} = A_n A_k \frac{1 - s^2}{s} K_{nk}^2, \quad (13)$$

$$Q_{nk}^{11} = -i A_n A_k \left[-s L_{nk}^1 + L_{nk}^3 + \frac{L_{nk}^2 - L_{nk}^4}{s} \right], \quad (14)$$

$$Q_{nk}^{22} = -i A_n A_k \left[-L_{nk}^1 + \frac{L_{nk}^3}{s} + L_{nk}^2 - L_{nk}^4 \right], \quad (15)$$

where $A_n = \sqrt{(2n+1)/(2n(n+1))}$. For off-diagonal elements ($n \neq k$), the four L^i integrals are not linearly independent and we may therefore also use the following simplifications [27]:

$$Q_{nk}^{11} = \frac{iA_n A_k (s^2 - 1)/s}{n(n+1) - k(k+1)} \times [n(n+1)L_{nk}^2 - k(k+1)L_{nk}^1], \quad (16)$$

$$Q_{nk}^{22} = iA_n A_k (s^2 - 1)/s \left[L_{nk}^3 + \frac{sn(n+1)(L_{nk}^2 - L_{nk}^1)}{n(n+1) - k(k+1)} \right]. \quad (17)$$

It is also worth noting that for particles with mirror symmetry with respect to the xOy plane, we also have

$$K_{nk}^i = 0 \quad \text{if } n+k \text{ even}; \quad (18)$$

$$L_{nk}^i = 0 \quad \text{if } n+k \text{ odd}. \quad (19)$$

This is the case for example for the spheroid and cylinder, but not for the offset sphere. For particles that present this mirror symmetry, we may therefore study independently two decoupled sets of matrices. The first set considers the indices n, k both even for \mathbf{Q}^{11} , both odd for \mathbf{Q}^{22} , even/odd and odd/even in \mathbf{Q}^{12} and \mathbf{Q}^{21} , respectively. The complementary set contains the matrix elements with opposite conditions (odd/odd, even/even, odd/even, even/odd, respectively). In this paper we focus on the first set of matrices (in particular on \mathbf{Q}^{22} with odd/odd indices) for simplicity; our conclusions hold for either of them.

2.3. Numerical methods

Numerical computations were carried out both in double precision and in arbitrary precision. Further details on these two implementations are provided as Supplementary Information. For a rigorous evaluation of the numerical error in the T -matrix integrals, we proceeded as follows. The arbitrary-precision (AP) code was first used to compute the *exact* double-precision (DP) values of the integrals, independent of any cancellations that may occur. To this end, we first found the number of digits and quadrature points required to obtain exact DP results. This is achieved by checking that, to within double precision, none of the matrix elements change when either the precision (number of digits) or the number of quadrature points is increased. In this study we required that all of the (non-zero) values in both \mathbf{Q} and $\text{Rg}\mathbf{Q}$ had converged to within double precision (first 16 digits in agreement). As a rule of thumb, 100–300 quadrature points were typically sufficient to reach convergence (once this is determined, the same number of points is used in the DP implementation for any comparison), but up to 360 digits were necessary to obtain accurate DP results in some cases, which in itself reveals the presence of severe numerical cancellations. Once the AP results have converged within DP, they can be used as a benchmark to test the double precision implementation against.

One may express the number of digits of agreement $\alpha(A)$ between the quantity A^{DP} computed in double precision and its exact (within DP) value A^{AP} (assumed non-zero) as

computed in arbitrary precision as

$$\alpha(A) = -\log_{10} \left| \frac{A^{\text{DP}} - A^{\text{AP}}}{A^{\text{AP}}} \right|. \quad (20)$$

In general, the maximum number of digits in agreement possible is of the order of $\log_{10} \varepsilon$ where ε is the floating-point accuracy, which in double precision gives $\alpha_{\text{max}} \approx 16$. In instances of severe numerical errors, A^{DP} may even be different in magnitude from its correct value A^{AP} . In this case $\alpha \leq 0$ and $|\alpha|$ is then a measure of the order of magnitude of the error.

In the following we applied this methodology to the systematic study of loss of precision in specific cases.

3. Demonstration of severe cancellations and associated loss of precision

3.1. General considerations

In order to fix ideas, we here briefly describe what we mean by severe loss of precision in numerical computing. Loss of precision can in principle occur in a wide variety of contexts. The case that is most relevant to us here is the subtraction of two (or more) numbers of comparable magnitude but for which the difference is many orders of magnitude smaller. For example, using double-precision floating point arithmetic, one can perform (in any double-precision computing software) the following tests, chosen as illustrations of increasingly problematic numerical cancellations:

$$\begin{cases} a = 10^{10} + \pi \\ b = 10^{10} \end{cases} \Rightarrow a - b \stackrel{\text{DP}}{\approx} 3.141592025756836,$$

$$\begin{cases} a = 10^{40} + \pi \\ b = 10^{40} \end{cases} \Rightarrow a - b \stackrel{\text{DP}}{=} 0,$$

$$\begin{cases} a = 10^{40}(1/3) + \pi \\ b = 10^{40}(1 - 2/3) \end{cases} \Rightarrow a - b \stackrel{\text{DP}}{\approx} -6 \times 10^{23}.$$

In the first example, the DP result is only accurate up to the 6th decimal. We can precisely quantify the error using Eq. (20) since we know the exact result (π): $\alpha(a-b) = 6.7$. Increasing the exponent as in the second example, we then get complete loss of precision and obtain $a-b=0$ instead of π and $\alpha(a-b) = 0$, i.e. zero digits in agreement. The same problem can in fact be further compounded by additional rounding errors as in the third example where the DP result is more than 20 orders of magnitude wrong ($\alpha(a-b) \approx -23.3$).

Such problems arise whenever $|a-b| \ll |a+b|/2$, with complete loss of precision occurring when $2|a-b|/|a+b|$ is of the order or smaller than the floating point accuracy ($\approx 10^{-16}$ in double precision). This is a well-known problem in computing and there are in principle two ways around it. (i) One may perform the computation with increased precision, using for example arbitrary precision arithmetic packages. This does not really circumvent the loss of precision, but simply increases the precision such that, even after the loss, the remaining accuracy is acceptable. This approach, which we used in

this work for demonstration purposes, is unfortunately resource-intensive and in particular extremely inefficient in terms of computational speed. (ii) The other alternative is to find a way to remove analytically the terms causing the source of the cancellation (e.g. 10^{10} in the first example) and therefore compute the terms (a and b) and their difference ($a-b$) without the problematic terms (which should cancel out exactly anyway). This approach is obviously much better for problems where the origin of the cancellation can be identified.

Such cancellation problems may also arise when numerically evaluating integrals whose integrands are much larger in magnitude than the integral, as in the artificial example

$$I = \int_{0.01}^{0.1} \left[0.01 \frac{21}{t^{22}} - \frac{20}{t^{21}} \right] dt = 9 \times 10^{18}. \quad (21)$$

The integrand maximum value is $\approx 10^{42}$ (at $t=0.01$), more than 10^{20} times larger than the integral. Severe cancellations occur during the summation, and such an integral cannot reliably be computed numerically using standard methods. For example, the `quad` function in Matlab yields 1.1×10^{32} (i.e. an error characterized by $\alpha(I) \approx -13$). The problem mainly arises from a characteristic property of this integral: the integrand's magnitude (i.e. the envelope of its absolute value) varies widely across the range of integration (it is $\sim 10^{20}$ larger at $t=0.01$ than it is at $t=0.1$) and the integrand has an oscillating character, taking both large positive and negative values whose contributions cancel almost exactly. This is a situation that occurs, as we shall see, in the T -matrix integrals, though whether this results in cancellations and therefore problems for the calculation of the integrals is shape dependent.

3.2. Examples of cancellations in EBCM integrals for an offset sphere

In Fig. 2 we illustrate the problematic numerical integration of matrix elements for an offset sphere of

relative refractive index $s = 1.5 + 0.02i$, aspect ratio $h=2$, and size parameter $x_{\max} = 0.5$ (this is equivalent to a sphere of size parameter 0.375, but it is here offset by a third of its radius). The matrix elements of the Q^{22} matrix (for $m=1$) were calculated using Eq. (15) using either double precision (DP) or arbitrary precision (AP), making sure in the latter case that the result was exact at least to 16 digits (explicit values for the first column of the matrix are also given as a table in Supplementary Information). The color maps present the maximum modulus (magnitude) of the complex integrand (a) and of the DP integral (b) and exact (AP) integral (c). The discrepancy between the DP (b) and exact (c) integrals is further quantified by computing the error from Eq. (20), which is graphically represented as a color map in Fig. 3(d). This discrepancy is particularly visible in the bottom-left corner of the matrix (large n , small k), where α is negative (meaning that even the order of magnitude is wrong). For example, $Q_{19,1}^{22}$ is found to be $-9.32 \times 10^{16} - 2.42 \times 10^{16}i$ in DP, while its correct value is $1.97 \times 10^{-9} + 2.56 \times 10^{-10}i$, equivalent (from Eq. (20)) to an error $\alpha \approx -26$. This loss of precision is understandable given the maximum magnitude of the integrand, 1.93×10^{31} , compared to that of the integral. In fact, while both the integrand and the DP integral increase in value with increasing n (Fig. 2(a,b)), the correct result decreases (Fig. 2(c)). This behavior is suggestive of important cancellations occurring in the integration of some of the L_{nk}^i integrals when $n > k$, the problem becoming worse as $n-k$ increases. Such large errors will inevitably result in erroneous results upon inversion of the linear system (Eq. (1)) to get the T -matrix, and this will be shown explicitly later. We also note that, although T -matrix calculations for offset spheres are an artificial example with no apparent practical interest (since the results are more reliably and easily obtained from Mie theory), they are nevertheless often being used as a convenient test of the validity of the EBCM approach or of its implementation. This example suggests that this may not be a good practice.

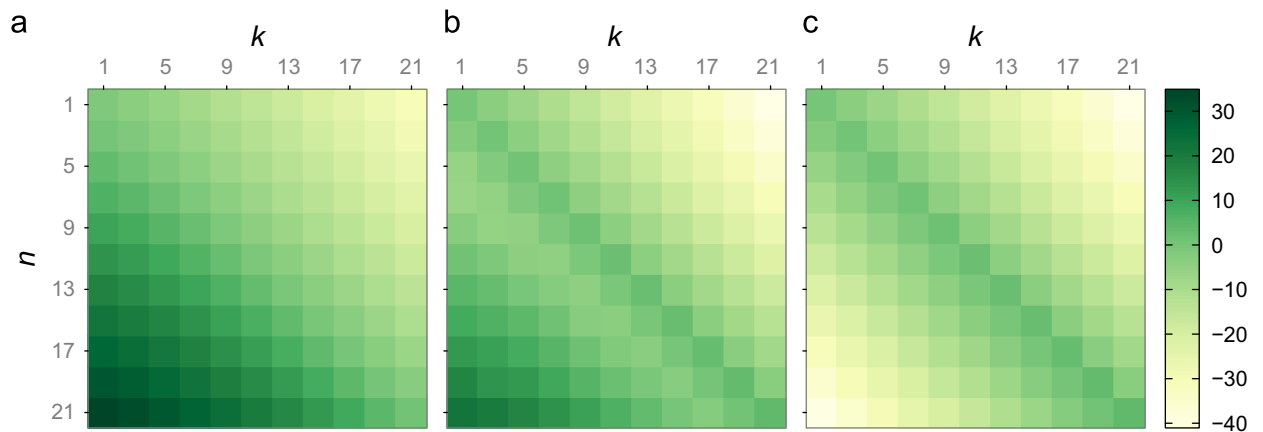


Fig. 2. Demonstration of cancellations and loss of precision in Q -matrix elements for an offset sphere, with $s = 1.5 + 0.02i$, $m=1$, $N=21$ and $h=2$. (a) Color maps showing the magnitude (maximum modulus) of the odd-odd integrands of Q^{22} indexed by their values of n and k in a base-10 logarithmic color scale. Also shown are the magnitude of the corresponding integrals when computed in double precision (b), and in arbitrary precision (c). The corresponding error between (b) and (c) is given as Fig. 3(d). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

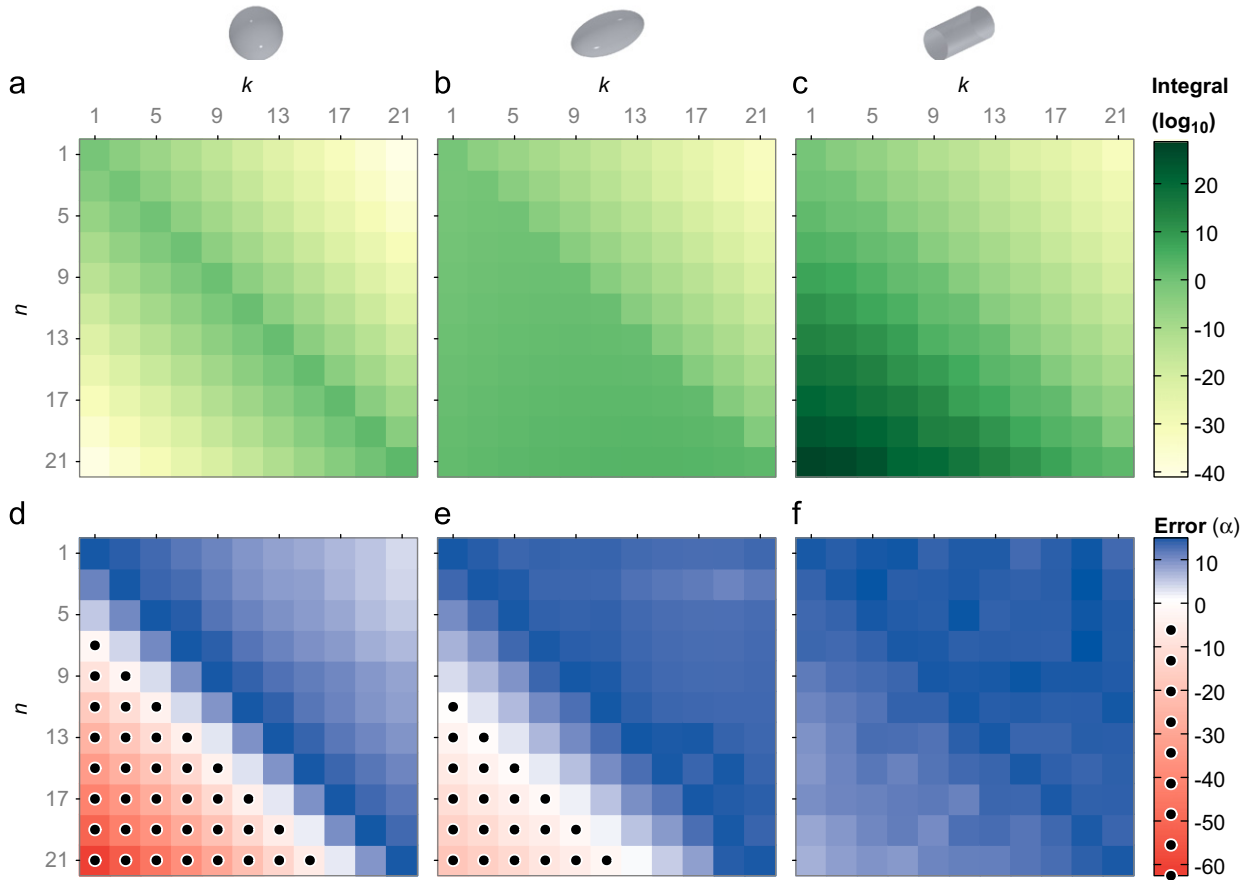


Fig. 3. Top row: map of the base-10 logarithm of the magnitude of the odd-odd integrals in \mathbf{Q}^{22} (computed in arbitrary precision and exact to double precision). From left to right, (a) for an offset sphere, (b) a prolate spheroid, (c) a cylinder. The integrand magnitude is similar in all three cases and follows that of the offset sphere shown in Fig. 2. Bottom row: (d–f) corresponding error represented as a color map of α (Eq. (20)), i.e. the number of digits of agreement for odd-odd elements of \mathbf{Q}^{22} between double precision and arbitrary precision results. Dots indicate that the order-of-magnitude is wrong ($\alpha < 0$). All results are shown for $s = 1.5 + 0.02i$, $m = 1$, $x_{\max} = 0.5$ and $h = 2$ (see Fig. 1 for definitions). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Shape and size dependence

As hinted at before, severe cancellations such as those evidenced in Fig. 2 are expected to be shape-dependent. The same study was therefore carried out for particles of different geometries, in particular spheroids and cylinders. The results are summarized in Fig. 3, where the offset sphere results are repeated for direct comparison. It is clear from these maps that spheroidal particles suffer from similar cancellations, albeit to a lesser extent, as those observed for the offset sphere. No significant error is evidenced however for cylindrical particles. A number of additional cases were also investigated using Chebyshev particles; the results obtained (not shown here) were similar in essence to those of the cylinder, indicating that the cylindrical particles are in fact more representative of the general case. The offset sphere and spheroids appear to be the only obvious cases where such cancellations occur. This observation is reminiscent of the suggestion in Refs. [25,26] that only so-called quadric surfaces exhibit this behavior for acoustic scattering. We also present in Table 1 numerical results for the three same particle




shapes in the case where $k = 1$, where the worst errors are observed. Two different size parameters are considered, $x_{\max} = 0.5$ and $x_{\max} = 5$. While the cylindrical shape shows again a good accuracy for the two sizes considered here, the spheroid, and more dramatically the offset sphere, both present an extreme loss of precision for moderate n , with a size parameter $x_{\max} = 0.5$. In the case of larger particles, we observe better agreement than for small particles in the Rayleigh regime, although the same loss of precision eventually occurs as n is increased. Finally, we also note that similar results are obtained when considering other values of m and other parts of the \mathbf{Q} -matrix, for example the even-even elements of \mathbf{Q}^{22} , the elements of \mathbf{Q}^{12} , \mathbf{Q}^{21} , \mathbf{Q}^{11} , or by directly examining the K_{nk}^i and L_{nk}^i matrix elements used in the computations. In contrast, no significant loss of precision was evidenced in the computation of the \mathbf{RgQ} matrix.

3.4. Consequences for the calculation of the T-matrix

As noted earlier, large errors in the magnitude of some of the matrix elements, even if they are isolated to a

Table 1

Number of correct digits in elements of Q_{nk}^{22} , for $k=1, m=1$. Negative values (in italics) indicate that the *order of magnitude is incorrect*. For all cases, the aspect ratio is fixed to $h=2$, the relative refractive index is $s = 1.5 + 0.02i$.

x_{\max}	n	 Offset sphere	 Spheroid	 Cylinder
0.5	1	14	14	14
	5	12	13	15
	9	-4	8	10
	19	-26	-3	8
5	1	14	14	14
	5	14	14	14
	9	13	14	14
	19	8	12	13

corner of the matrix, are very likely to result in further numerical errors upon solving the linear system (Eq. (1)) to obtain the T -matrix. We demonstrate this important consequence in Fig. 4 by examining the error in the resulting T -matrix in the case of a spheroid. We observe that the solution of the linear system in double-precision is substantially improved when the Q -matrix was computed reliably in arbitrary precision, and truncated to 16 digits. As an indicator, the condition number of the Q matrix was 2.4×10^{39} when computed in double-precision, and 8.1×10^4 in arbitrary-precision. The matrix $Rg\mathbf{Q}$, not suffering cancellations, does not affect the inversion. Those results were compared against the accurate computation of \mathbf{Q} and the inverse of the linear system in arbitrary-precision. This example demonstrates the importance of a robust integral evaluation, independently from the more commonly considered inversion problem, in the numerical accuracy of the T -matrix method. Indeed, the DP inversion results in an almost perfectly accurate T -matrix, provided that the exact Q -matrix (computed in AP) is used. In contrast, large errors appear in some part of the T -matrix for a full DP implementation. Such errors are likely to affect further computations of scattering properties from the T -matrix, although this will depend on whether the problematic matrix elements contribute or not to the scattering property under consideration. This is illustrated in Fig. 5 for the extinction coefficient Q_{Ext} . For an aspect ratio of $h=2$ (same as the one used in Fig. 4), the full DP calculation of Q_{Ext} converges to almost the exact results for $N=10$, but subsequently deteriorates (see Fig. 5(a)) as the errors in the T -matrix (Fig. 4(b)) start to contribute. This will however not affect our ability to compute Q_{Ext} to a high precision because (Fig. 4(b)) only depends on relatively small $n < 10$ in this case. Other physical properties, like the electric field in the vicinity of the particle, will however require much larger n to be computed accurately, and this will be prevented in a full DP implementation by the errors in the T -matrix resulting from the integral evaluations. Moreover for either larger particles, or particles with a larger aspect ratio, then larger n are also needed, and the errors in the T -matrix can become a problem even for far-field properties such as Q_{Ext} . This is illustrated in Fig. 5(b) with a higher aspect

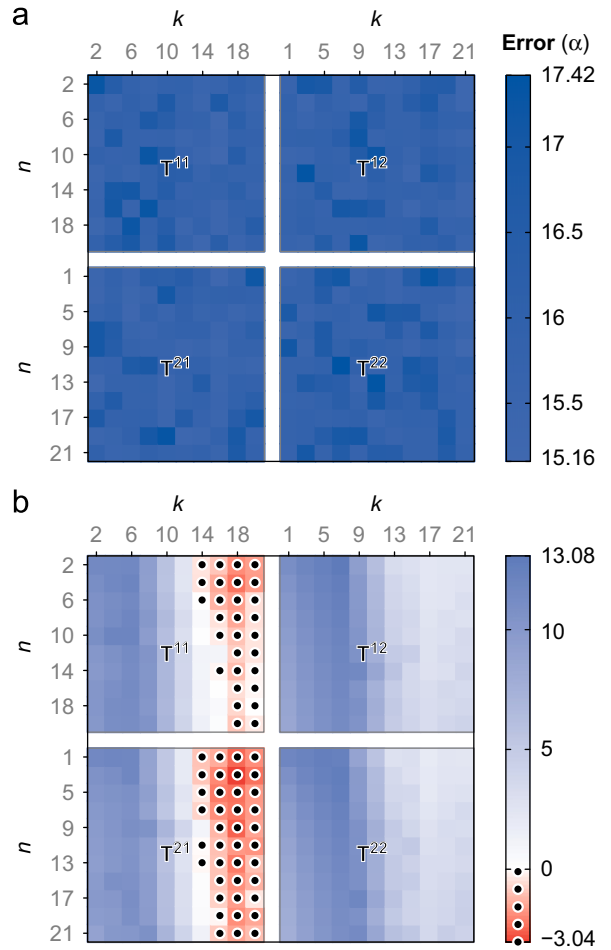


Fig. 4. Error in the T matrix of a spheroid with same parameters as in Fig. 3 after inversion in DP of the linear system (1), where the elements of $Rg\mathbf{Q}$ were computed in either AP (a), or DP (b). In both cases the inversion is computed in DP. The error is evaluated by performing all calculations, including the inversion step, in AP. The very good agreement in (a) compared to the large errors in (b) demonstrates that the problem lies here in the accurate computation of the \mathbf{Q} matrix, not in its inversion.

ratio, $h=20$. The computed Q_{Ext} starts to diverge (because of the T -matrix errors) as early as $N \geq 7$.

4. Origin and proof of the cancellations

A quick comparison between Fig. 2(a) and (c) provides a first hint of the origin of these cancellations: while the magnitude of the integrand for the offset sphere increases with n , the integral exhibits the opposite behavior. In order to reconcile this apparent contradiction, the following argument is proposed. First, a series expansion of the integrand is developed as a function of the angle-dependent size parameter. Upon individual examination of the dominant terms, we observe that a number of them can integrate to zero in specific cases that depend on n, k and the particle shape. To test this hypothesis, we have used a numerical approach to determine the scaling properties of the integrals. This practical analysis, though not

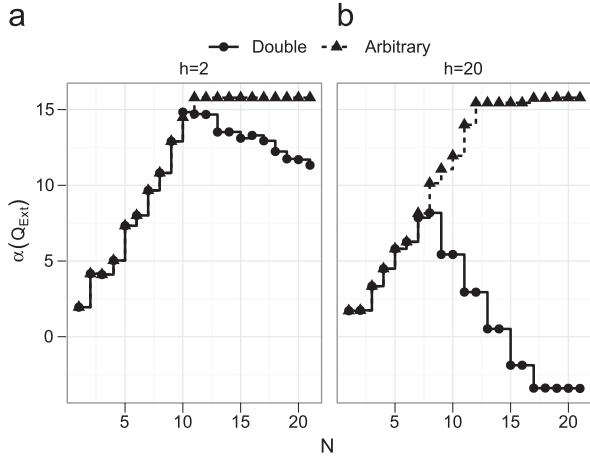


Fig. 5. Error in the extinction coefficient for light incident along the rotation axis of a spheroid as a function of N with $h=2$ and $h=20$ when \mathbf{Q} was computed in both arbitrary and double precision, as compared to the converged value for the arbitrary precision. Here $x_{\max}=0.5$ and $s=1.5+0.02i$. The arbitrary-precision results converge to a value (5.33×10^{-3} and 3.82×10^{-5} for $h=2$ and $h=20$, respectively), while the double-precision results approach that value and then diverge as the cancellations start to play a part.

constituting a formal proof, provides a simple and efficient way of testing for the presence and the extent of similar cancellations for any shapes. As such, it is worth presenting its principle and its main conclusions before getting to the more rigorous proof.

4.1. Numerical investigations

The aim of the method is to determine the scaling law for the integrals K_{nk}^1 and L_{nk}^1 (Eqs. (2)–(7)) for a given shape as a function of the size parameter, which we assume takes the form of a power law: $I \propto x_{\max}^p$, at least in the small size regime. The rationale behind this assumption is that all the integrands in Eqs. (2)–(7) follow such a power law (from the small argument expansion of the Bessel functions [28]), and one would therefore expect the power law with the same exponent for the integral, if cancellations do not occur. For example for the integrand of L_{nk}^1 , we have

$$x_{\theta} \xi_n(x) \psi_k(sx) \propto x_{\theta} x^{k-n+1} \propto x_{\max}^{k-n+2}. \quad (22)$$

This approximation is in fact good over a relatively large range of parameters, typically up to $x \approx n$, if $|s|$ is not too large. We should expect that the integral follow the same scaling law, unless this dominant term integrates exactly to zero.

The exponent of the power law for the integral can be obtained by computing the exact integral value (using arbitrary-precision) for different size parameters in the small size regime. Let us consider as an example the case of $L_{n,k}^1$, which is illustrated in Fig. 6. According to Eq. (22), the integral should scale, like the integrand, as x_{\max}^{k-n+2} , in the small size regime. The calculations of the integrals in arbitrary precision suggest that they do follow this scaling law for the cylinder, but for offset spheres and spheroids, this is only the case when $k > n$. If $n > k$ however, the

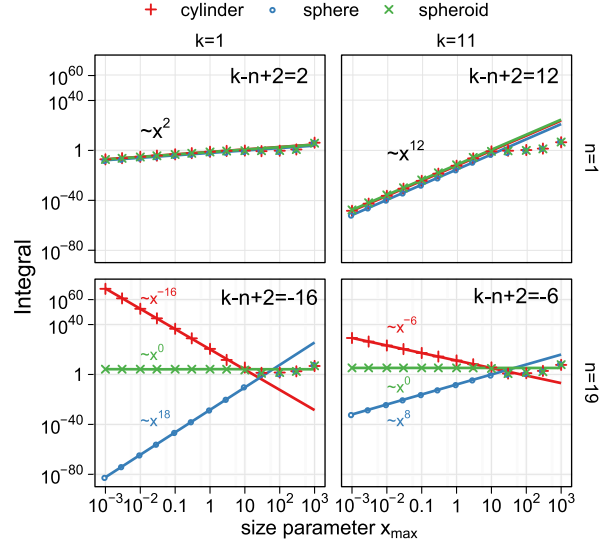


Fig. 6. Numerical determination of the scaling laws for the integral of L_{nk}^1 for different shapes. The difference in scaling factors (gradients) for the different shapes is evident below the diagonal (bottom row), while on or above the diagonal (top row) they are identical. The lines are a fit to the data in the small-size regime ($x_{\max} < 1$), and the exponents given are accurate to the second decimal place. In all cases the integrand's scaling was the same as that of the cylinder.

integral scales as x_{\max}^{n-k} for an offset sphere and as x_{\max}^0 (constant) for a spheroid. This clearly shows the presence of cancellations in these two cases, where the dominant terms must integrate to zero to modify the scaling law. More explicitly, we may write for the offset sphere the Laurent series of the radial part of the integral as

$$x_{\theta} \xi_n(x) \psi_k(sx) \tau_n d_k = \sum_{p=k-n+1}^{p=n-k-2} \gamma_p(s) x_{\theta} x^p \tau_n d_k + \sum_{p=n-k-1}^{\infty} \gamma_p(s) x_{\theta} x^p \tau_n d_k. \quad (23)$$

The fact that the integral is observed to scale as x_{\max}^{n-k} suggests that the first sum does not contribute at all and must therefore integrate to zero. In fact, we will demonstrate later an even stronger result: each individual term in this sum integrates to zero. A similar—albeit not as dramatic—situation arises for the spheroid where the sum up to $p=-2$ should integrate to zero. This explanation also reveals why no cancellation is observed when $k > n$, since the first sum only exists when $n > k+1$.

Such a scaling study was systematically carried out for all integrals (Eqs. (2)–(7)); the results are summarized in Table 2, showing that the cancellations are present in every integral for the offset sphere and the spheroid. For a general shape (illustrated here by cylinders and Chebyshev particles), the scaling is the same for both integrand and integrals, suggesting that such a cancellation mechanism does not operate. We note that a similar observation was made in Ref. [25] in the context of acoustic scattering, where it was postulated that for complete quadric surfaces (the analog of our spheroid here), the contribution of negative powers of x is zero. Our results generalize this observation to the more complex case of electromagnetic scattering, where the actual

Table 2

The scaling exponent (p) of integrands and integrals with respect to the size parameter x_{\max} , for different particle geometries, and $n > k$. For all particles except the Chebyshev, $s = 1.5 + 0.02i$, $h = 2$. In the case of the Chebyshev particle, the parameters were $s = 1.5 + 0.02i$, $\epsilon = 0.15$, and $n = 3$ and $n = 4$ were used. This notation follows that of Mugnai and Wiscombe [29].

Integral	Integrand	Offset sphere	Spheroid	Cylinder	Chebyshev
$K^1 \propto Q^{12}$	$k-n+1$	$n-k+1$	0	$k-n+1$	$k-n+1$
$K^2 \propto Q^{21}$	$k-n+1$	$n-k+1$	0	$k-n+1$	$k-n+1$
L^1	$k-n+2$	$n-k$	0	$k-n+2$	$k-n+2$
L^2	$k-n+2$	$n-k$	0	$k-n+2$	$k-n+2$
L^3	$k-n$	$n-k$	0	$k-n$	$k-n$
L^4	$k-n$	$n-k$	0	$k-n$	$k-n$
Q^{11}	$k-n+2$	$n-k$	0	$k-n+2$	$k-n+2$
Q^{22}	$k-n$	$n-k$	0	$k-n$	$k-n$

scaling varies slightly from one integral to another. Furthermore, in the case of offset spheres, we find that approximately the same number of terms of non-negative powers also make no contribution.

4.2. Analytical study of the cancellations

Using the numerical scaling study as a guide, a rigorous proof of these results was developed. Details are given in Appendix for the spheroid and in Supplementary Information for offset spheres. We here only sketch the key principles and conclusions. We start from Laurent series expansions of the radial part of the integrands, such as the one given in Eq. (23). It is then possible to prove that each individual term, denoted A_p , in the first sum integrates to zero when $x(\theta)$ describes an offset sphere. The same can be proved for a spheroid, although fewer terms integrate to zero (as predicted from the scaling study presented earlier). The central argument in these proofs is to express the term $x_\theta x(\theta)^p$ in a basis of Legendre polynomials $P_n^0(\cos(\theta))$ (which essentially means that they are polynomials of $\cos \theta$). Moreover, the angular functions $d_n(\theta)$ are proportional to associated Legendre functions (with order m) or in the case of $\tau_n(\theta)$ can be expressed as a sum of associated Legendre functions. As a result, A_p can be rewritten as a sum of products of three associated Legendre functions, one of which has order zero (i.e. is a Legendre polynomial). The key result to integrate these products is the use of Gaunt's formula [30], from which we use the following special case:

$$\int_0^\pi d\theta P_n^m(\cos \theta) P_k^m(\cos \theta) P_p^0(\cos \theta) \sin \theta = 0$$

$$\text{if } \begin{cases} |n-k| > p \\ \text{or} \\ n+k+p \text{ is odd.} \end{cases} \quad (24)$$

It is possible to show that the terms where this condition is satisfied are precisely the terms that were identified in the scaling study as making no contribution to the integral. A number of more subtle technical difficulties can be encountered in the formal derivation of the results

of Table 2. In Appendix A we present in details the complete proof in the case of a spheroid, while the detailed proof for offset-spheres is given in Supplementary Information. These proofs could be used as a guide to find other special shapes where such cancellations may occur.

4.3. Influence of other parameters

The results of Fig. 3 suggest that the cancellations, and the important resulting loss of precision identified in this work are more prominent for small size parameters. This is in fact naturally explained now that the origin of these cancellations has been identified. Most of the terms in the series expansion of the integrand that integrate to zero (and therefore cause the problems) are proportional to x^p , with p negative. Their magnitude therefore increases dramatically as the size parameter decreases, thereby causing more numerical problems for given n, k values. We should however note that since larger values of n, k are typically needed for larger particles, this requirement is also likely to result in cancellation problems for large particle size, at least for matrix elements with large n and small k .

A similar remark applies to particles with large aspect ratio. We carried out numerical tests of the effect of these cancellations on particles with larger aspect ratio and did not find any evidence that the cancellations were more significant for a given matrix element (given n and k). However, as for large size parameter, large aspect ratio require the computation of a larger matrix and the cancellations will therefore become increasingly problematic as N is increased, as already pointed out in Fig. 5.

Finally, the fact that the problematic terms are of the form x^p with p negative also explains why the right-upper part of the \mathbf{Q} matrix and the entire RgQ matrix do not suffer from such cancellations: such terms are simply absent from the Laurent expansion in those cases.

5. Conclusion and outlook

In conclusion, we first summarize the most important consequences of our demonstration of these severe cancellations. Firstly, it highlights a number of new features and counter-intuitive facts about the EBCM approach, notably: (i) numerical instabilities are not only due to problems in the inversion of the linear system of Eq. (1); severe loss of precision when computing the integrals (before inversion) are likely to contribute, possibly dominantly, to the problems in the case of offset spheres and spheroids; (ii) small-size particles, even with moderate aspect ratio are not immune from numerical problems; (iii) simple shapes such as offset spheres and spheroids are special cases. Numerical problems arising for these particular shapes may not be representative of those arising for other more general shapes. This needs to be borne in mind when testing convergence or accuracy of an EBCM implementation using these simple shapes. This remark leaves open the question of what should be the benchmark of choice for T -matrix calculations, since no other shapes admit a straightforward and rigorous solution

(using other methods) that could be used as a standard for comparison.

This work also opens up a number of potential new avenues for improving the accuracy and efficiency of the T -matrix/EBCM method, especially in the important case of spheroids. Our scaling study shows that all the spheroid integrals exhibit the same scaling factor. This suggests that the Q -matrix, if computed accurately, should be well-balanced and therefore easily invertible in double precision (this is precisely what Fig. 4 demonstrates), even for large matrices (for example for large aspect ratio). Arbitrary-precision computations could be used, as in this work, but are typically time-consuming. It would therefore be of great interest to devise a method to compute these integrals accurately and efficiently, e.g. taking into account, and removing, the terms causing cancellations. The identification of the origin of these cancellations, as presented here, is the first step toward such a goal. In the case of offset spheres it is not so critical to improve the method, as Mie theory readily provides the necessary results.

Acknowledgments

The authors are indebted to the Royal Society of New Zealand for support through a Marsden Grant (WRCS and ECLR) and Rutherford Discovery Fellowship (ECLR).

Appendix A. Formal proof of cancellations for spheroids

Preliminary results: We will use the Laurent expansion of the spherical Bessel functions. Although the actual coefficients of these expansions are not relevant to our proof, it is important to identify the highest order term in the series and highlight the fact that only every other term is present, at least up to some order in the case of ξ_n , namely

$$\psi_k(sx) = x^{k+1} \sum_{q=0}^{\infty} \alpha_{kq}(s)x^{2q}, \tag{A.1}$$

$$\xi_n(x) = x^{-n} \sum_{q=0}^{q=n} \beta_{np}x^{2q} + \mathcal{O}(x^{n+1}). \tag{A.2}$$

In these expressions, Greek letters are used to denote coefficients whose actual value are not relevant to the proof. We use the same convention in the rest of this appendix (with the obvious exceptions of the angular functions τ_n and π_n , and the radial functions ξ_n and ψ_k).

As explained in the main text, one key ingredient of the proof is the use of Gaunt's formula (Eq. (24)). We will more precisely show that some of the elements in the Laurent series of the form $x(\theta)^p$ can be expressed as polynomials of $\cos \theta$. We therefore define the notation $\mathcal{P}_N(\cos \theta)$ to denote a polynomial in $\cos \theta$ of degree N or less (the coefficients of which are not relevant to this proof). Then we use the fact that the set of Legendre polynomials $P_n^0(x)$ with $0 \leq n \leq N$ forms a basis for the vector space $\mathbb{R}_N[x]$ of real polynomials of degree N ,

therefore

$$\mathcal{P}_N(\cos \theta) = \sum_{v=0}^N \alpha_{vN} P_v^0(\cos(\theta)). \tag{A.3}$$

Gaunt's formula (Eq. (24)) therefore implies in particular that

$$\int_0^\pi d\theta d_n(\theta) d_k(\theta) \mathcal{P}_p(\cos \theta) \sin \theta = 0 \tag{A.4}$$

if $p \leq |n-k|-1$.

We will also need to use a similar expression, but involving τ_n rather than d_n . To find it, one can use the following property of the angular functions

$$\sin \theta \tau_n = \kappa_n d_{n+1} + \lambda_n d_{n-1}. \tag{A.5}$$

Using Eq. (A.4) on each term on the right hand side, we therefore find the equivalent for τ_n

$$\int_0^\pi d\theta \sin \theta \tau_n(\theta) d_k(\theta) \mathcal{P}_p(\cos \theta) \sin \theta = 0 \tag{A.6}$$

if $p \leq |n-k|-2$.

Note that we are looking here for a sufficient condition, but it may not be a necessary condition.

Spheroid-specific results: In this appendix we are concerned solely with a spheroid of revolution with semi-axes c along z and a along x, y , whose geometry is defined in Supplementary Information. We will in particular use the following (easily derived) relations:

$$\frac{x_\theta}{x^3} = \frac{a^2 - c^2}{k_1^2 a^2 c^2} \sin \theta \cos \theta, \tag{A.7}$$

$$\frac{1}{x^2} = \alpha + \beta \cos^2 \theta \tag{A.8}$$

$$\frac{\sin \theta}{x^2} = \gamma \sin \theta + \frac{x_\theta \cos \theta}{x^3}, \tag{A.9}$$

where α , β , and γ are constants.

From these, we also get for $q \geq 0$

$$\frac{x_\theta}{x^{2q+3}} = \cos \theta \sin \theta (\alpha + \beta \cos^2 \theta)^q = \sin \theta \mathcal{P}_{2q+1}(\cos \theta), \tag{A.10}$$

$$\frac{\sin \theta}{x^{2q+2}} = \sin \theta (\alpha + \beta \cos^2 \theta)^{q+1} = \sin \theta \mathcal{P}_{2q+2}(\cos \theta). \tag{A.11}$$

K^1 and K^2 integrals: We consider first the case of K_{nk}^1

$$K_{nk}^1 = \int_0^\pi d\theta m d_n d_k x_\theta \xi_n \psi'_k. \tag{A.12}$$

We also assume that $n+k$ is odd, as $K_{nk}^1 = 0$ if $n+k$ even. Both spherical Bessel functions can be expanded as a Laurent series as in Eqs. (A.1) and (A.2); in particular we have (for $n \geq k \geq 1$)

$$\xi_n \psi'_k = \frac{1}{x^{n-k}} \sum_{q=0}^{q=n-k} \alpha_{nkq}(s)x^{2q} + \mathcal{O}(x^{n-k+1}) \tag{A.13}$$

and the integral can therefore be rewritten as a sum

$$K_{nk}^1 = \sum_{\substack{p=n-k \\ p \text{ odd}}}^{p=n-k} v_{nkp}(s) K_{nkp} + \mathcal{O}(x_{\max}^{n-k+2}), \tag{A.14}$$

where

$$K_{nkp}^1 = \int_0^\pi d\theta x_\theta \chi(\theta)^p d_n(\theta) d_k(\theta). \quad (\text{A.15})$$

The scaling study (with an exponent zero for the power law of the integral) suggests that cancellations arise because a number of the leading K_{nkp}^1 terms are zero, more precisely: $K_{nkp}^1 = 0$ for $k-n \leq p \leq -3$, but $K_{nkp}^1 \neq 0$ for $p = -1$ (because $x_\theta \chi^{-1}$ scales as x_{\max}^0). We note that if $k-n > -3$, then no such cancellations occur, so only the bottom-left part of the matrix is subject to this problem. We therefore implicitly assume $n \geq k+3$ in the following.

Let us consider K_{nkp}^1 for p odd and $p \leq -3$ and define for convenience $q = -(p+3)/2$, which must be a non-negative integer. We then substitute Eq. (A.10) in Eq. (A.15) and using Eq. (A.4) (Gaunt's formula), we deduce that $K_{nkp}^1 = 0$ if $2q+1 \leq |n-k|-1$ or since $|n-k|$ is odd if $|n-k| \geq 2q+3$, equivalent to $k-n \leq p$. This is precisely what we set out to prove. It is also worth noting that this proof does not work for $p \geq -1$ (since q is then negative), as expected from the scaling study. Finally, the same arguments can be readily applied to the cancellations occurring in K_{nkp}^2 .

L^1 and L^2 integrals: We now consider the case of L_{nk}^1

$$L_{nk}^1 = \int_0^\pi d\theta \sin \theta \tau_n d_k x_\theta \zeta_n \psi_k. \quad (\text{A.16})$$

We also assume that $n+k$ is even, as $L_{nk}^1 = 0$ if $n+k$ is odd. As for K_{nkp}^1 , we can write

$$L_{nk}^1 = \sum_{p=k-1-n}^{p=n-k-1} \nu_{nkp}(s) L_{nkp}^1 + \mathcal{O}(x_{\max}^{n-k+1}), \quad (\text{A.17})$$

where

$$L_{nkp}^1 = \int_0^\pi d\theta x_\theta \chi(\theta)^p \sin \theta \tau_n(\theta) d_k(\theta). \quad (\text{A.18})$$

The scaling study (with an exponent zero for the power law of the integral) suggests that cancellations arise because $L_{nkp}^1 = 0$ for $k+1-n \leq p \leq -3$, but $L_{nkp}^1 \neq 0$ for $p = -1$. We note that if $k+1-n > -3$, then no such cancellations occur, so only the bottom-left part of the matrix is subject to this problem, as for K_{nkp}^1 . We therefore implicitly assume $n \geq k+4$ in the following.

Using the same reasoning as for K_{nkp}^1 , we consider $p \leq -3$ and set $q = -(p+3)/2$ (a non-negative integer), then substitute Eq. (A.10) in Eq. (A.18). Using Eq. (A.6) (our corollary to Gaunt's formula using τ_n), we deduce that $L_{nkp}^1 = 0$ if $2q+1 \leq |n-k|-2$, or since $|n-k|$ is even if $|n-k| \geq 2q+4$, equivalent to $k-n+1 \leq p$, as desired. Again the proof would fail for $p \geq -1$ as expected. The same proof can be applied to L_{nk}^2 .

L^3 and L^4 integrals: We now focus on the more complicated case of L_{nk}^3

$$L_{nk}^3 = \int_0^\pi d\theta \sin \theta d_k \psi'_k [x_\theta \tau_n \zeta'_n - n(n+1) d_n \zeta_n]. \quad (\text{A.19})$$

We here assume that $n+k$ is even, as $L_{nk}^3 = 0$ if $n+k$ is odd.

In this case, one is tempted to split the integral into a sum of two, each being in many respects similar to the cases we have treated so far. This approach works to some extent, demonstrating cancellation of all predicted terms in the expansion, *except for the dominant (lowest order)*

term. In fact, if one applies the scaling study to each of these two integrals, it is clear that no cancellation occur (which is because the highest order term does not integrate to zero). However, severe loss of precision occurs when summing these two integrals and the cancellations therefore do occur when evaluating L_{nk}^3 as one integral. This observation suggests that we need to consider the entire integrand to prove the cancellation of the dominant term. For this, we proceed as before; the spherical Bessel functions are expanded as Laurent series, but we isolate the highest order term and explicitly calculate its coefficients

$$L_{nk}^3 = \sum_{p=k-1-n}^{p=n+k-1} L_{nkp}^3 + \mathcal{O}(x_{\max}^{n+k+1}), \quad (\text{A.20})$$

where

$$L_{nkp}^3 = \int_0^\pi d\theta \sin \theta d_k [\alpha_{nkq}(s) x_\theta \chi^p \tau_n + \beta_{nkq}(s) x_\theta^{p+1} d_n], \quad (\text{A.21})$$

and more explicitly for $p = k-1-n$, we use the fact that for the highest order term $\zeta'_n \propto -n \zeta_n$ to get

$$L_{nkp}^3|_{p=k-1-n} = \int_0^\pi d\theta d_k \sin \theta \gamma_{nk}(s) \times [n x_\theta \chi^p \tau_n + n(n+1) x_\theta^{p+1} d_n]. \quad (\text{A.22})$$

Note that due to the absence of x_θ in the second term of these integrals, both terms scale as x_{\max}^{p+1} .

The scaling study (with an exponent zero for the power law of the integral) suggests that cancellations arise because $L_{nkp}^3 = 0$ for $k-1-n \leq p \leq -3$, but $L_{nkp}^3 \neq 0$ for $p = -1$. We note that if $k-1-n > -3$, then no such cancellations occur, so only the lower-left part of the matrix is subject to this problem, as before. We therefore implicitly assume $n \geq k+2$ in the following.

The first part of L_{nkp}^3 has the same form as L_{nkp}^1 and we have already shown that it is zero for $k-n+1 \leq p \leq -3$ (i.e. for all desired terms except the dominant one ($p = k-n-1$)). For the second part, we need to show that the integral

$$L_{nkp}^3 = \int_0^\pi d\theta x_\theta^{p+1} d_n d_k \sin \theta \quad (\text{A.23})$$

is zero. To this end, we define as before $q = -(p+3)/2$ ($q \geq 0$), which implies $-(p+1) = 2q+2$. We now substitute Eq. (A.11) in Eq. (A.23) and use Gaunt's formula (Eq. (A.4)) to deduce that $L_{nkp}^3 = 0$ if $2q+2 \leq |n-k|-1$, or since $|n-k|$ is even if $|n-k| \geq 2q+4$, equivalent to $k-n+1 \leq p$, which again corresponds to all desired terms except the dominant one ($p = k-n-1$).

To conclude the proof, we now need to prove that the dominant (lowest order) term (Eq. (A.22)) also integrates to zero. For this, we use Eq. (A.9) to re-arrange it as follows:

$$L_{nkp}^3|_{p=k-1-n} = \int_0^\pi d\theta n \gamma_{nk}(s) d_k x_\theta^{p+3} \times \left[\frac{x_\theta}{x_\theta^3} \tau_n \sin \theta + (n+1) \frac{x_\theta}{x_\theta^3} \cos \theta d_n + \gamma(n+1) \sin \theta d_n \right]. \quad (\text{A.24})$$

We then recognize the relation between angular functions

$$(n+1) \cos \theta d_n + \sin \theta \tau_n = \sqrt{(n+1)^2 - m^2} d_{n+1}, \quad (\text{A.25})$$

and obtain

$$L_{nk}^3 |_{p=k-1-n} = \eta_{nk}(s) \int_0^\pi d\theta n x_\theta x^p d_k d_{n+1} + \phi_{nk}(s) \int_0^\pi d\theta n x^{p+3} d_k d_n \sin \theta. \quad (\text{A.26})$$

The first part has the same form as $K_{n+1,k,p}^1$ and is in particular zero for $p=k-n-1$. The second part has the same form as $L_{n,k,p+2}^{32}$ and is therefore zero for $p=k-n-1$. This completes the proof that the dominant term does integrate to zero, and therefore the proof of the cancellations for L^3 . A similar proof can be applied to L^4 , although the results are also a direct consequence of the fact that L^4 can be expressed as a linear combination of L^1, L^2, L^3 [27].

Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:[10.1016/j.jqsrt.2012.01.007](https://doi.org/10.1016/j.jqsrt.2012.01.007).

References

- [1] Kahnert FM. Numerical methods in electromagnetic theory. *J Quant Spectrosc Radiat Transfer* 2003;79–80:775–824.
- [2] Draine BT, Flatau PJ. Discrete-dipole approximation for scattering calculations. *J Opt Soc Am A* 1994;11:1491–9.
- [3] Yee K. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans Antennas Propag* 1966;14:302–7.
- [4] Mishchenko MI, Travis LD, Lacis AA. Scattering, absorption and emission of light by small particles. 3rd ed. Cambridge: Cambridge University Press; 2002.
- [5] Mie G. Contributions to the optics of turbid media, particularly of colloidal metal solutions. *Ann Phys* 1908;25:377–445.
- [6] Bohren CF, Huffman DR. Absorption and scattering of light by small particles. New York: John Wiley & Sons Inc.; 1983.
- [7] Mishchenko MI, Travis LD. Gustav Mie and the evolving discipline of electromagnetic scattering by particles. *Bull Am Meteorol Soc* 2008;89:1853–61.
- [8] Mishchenko MI. Light scattering by randomly oriented axially symmetric particles. *J Opt Soc Am A* 1991;8:871–82.
- [9] Mackowski DW, Mishchenko MI. Calculation of the T matrix and the scattering matrix for ensembles of spheres. *J Opt Soc Am A* 1996;13:2266–78.
- [10] Waterman PC. Matrix formulation of electromagnetic scattering. *Proc IEEE* 1965;53:805–12.
- [11] Mackowski DW. Discrete dipole moment method for calculation of the T matrix for nonspherical particles. *J Opt Soc Am A* 2002;19:881–93.
- [12] Loke VLY, Nieminen TA, Parkin SJ, Heckenberg NR, Rubinsztein-Dunlop H. FDFD/ T -matrix hybrid method. *J Quant Spectrosc Radiat Transfer* 2007;106:274–84.
- [13] Doicu A, Wriedt T, Eremin YA. Light scattering by systems of particles: null-field method with discrete sources: theory and programs. In: Springer series in optical sciences, vol. 124. Berlin: Springer; 2006.
- [14] García de Abajo FJ, Howie A. Relativistic electron energy loss and electron-induced photon emission in inhomogeneous dielectrics. *Phys Rev Lett* 1998;80:5180–3.
- [15] Nieminen T, Rubinsztein-Dunlop H, Heckenberg N. Calculation of the T -matrix: general considerations and application of the point-matching method. *J Quant Spectrosc Radiat Transfer* 2003;79–80: 1019–29.
- [16] Schulz FM, Stamnes K, Stamnes JJ. Scattering of electromagnetic waves by spheroidal particles: a novel approach exploiting the T matrix computed in spheroidal coordinates. *Appl Opt* 1998;37: 7875–96.
- [17] Mishchenko MI, Videen G, Babenko VA, Khlebtsov NG, Wriedt T. T -matrix theory of electromagnetic scattering by particles and its applications: a comprehensive reference database. *J Quant Spectrosc Radiat Transfer* 2004;88:357–406.
- [18] Waterman PC. New formulation of acoustic scattering. *J Acoust Soc Am* 1969;45:1417–29.
- [19] Barber P. Resonance electromagnetic absorption by nonspherical dielectric objects. *IEEE Trans Microwave Theory Tech* 1977;25: 373–81.
- [20] Mishchenko MI, Travis LD. Capabilities and limitations of a current FORTRAN implementation of the T -matrix method for randomly oriented, rotationally symmetric scatterers. *J Quant Spectrosc Radiat Transfer* 1998;60:309–24.
- [21] Wielgaard DJ, Mishchenko MI, Macke A, Carlson BE. Improved T -matrix computations for large, nonabsorbing and weakly absorbing nonspherical particles and comparison with geometrical-optics approximation. *Appl Opt* 1997;36:4305–13.
- [22] Moroz A. Improvement of Mishchenko's T -matrix code for absorbing particles. *Appl Opt* 2005;44:3604–9.
- [23] Mishchenko M, Travis L. T -matrix computations of light scattering by large spheroidal particles. *Opt Commun* 1994;109:16–21.
- [24] Petrov D, Shkuratov Y, Videen G. Optimized matrix inversion technique for the T -matrix method. *Opt Lett* 2007;32:1168–70.
- [25] Waterman PC. The T -matrix revisited. *J Opt Soc Am A* 2007;24: 2257–67.
- [26] Waterman PC. T -matrix methods in acoustic scattering. *J Acoust Soc Am* 2009;125:42–51.
- [27] Somerville WRC, Auguie B, Le Ru EC. Simplified expressions of the T -matrix integrals for electromagnetic scattering. *Opt Lett* 2011;36: 3482–4.
- [28] Abramowitz M, Stegun IA, editors. Handbook of mathematical functions. New York: Dover; 1972.
- [29] Mugnai A, Wiscombe WJ. Scattering from nonspherical Chebyshev particles. I: cross sections, single-scattering albedo, asymmetry factor, and backscattered fraction. *Appl Opt* 1986;25:1235–44.
- [30] Gaunt JA. The triplets of helium. *Philos Trans R Soc London A* 1929;228:151–96.