Advancing numerics for the Casimir effect to experimentally relevant aspect ratios

Michael Hartmann[1], Gert-Ludwig Ingold[1] and Paulo A. Maia Neto[2]

[1] Institut für Physik, Universität Augsburg, 86135 Augsburg, Germany
[2] Instituto de Física, Universidade Federal do Rio de Janeiro, CP 68528, Rio de Janeiro RJ 21941-909, Brazil

E-mail: michael.hartmann@physik.uni-augsburg.de, gert.ingold@physik.uni-augsburg.de, pamn@if.ufrj.br

Abstract. Within the scattering theoretical approach, the Casimir force is obtained numerically by an evaluation of the round trip of an electromagnetic wave between the objects involved. Recently [Hartmann M *et al* 2017, *Phys. Rev. Lett.* **119** 043901] it was shown that a symmetrization of the scattering operator provides significant advantages for the numerical evaluation of the Casimir force in the experimentally relevant sphere-plane geometry. Here, we discuss in more detail how the symmetrization modifies the scattering matrix in the multipole basis and how computational time is reduced. As an application, we discuss how the Casimir force in the sphere-plane geometry deviates from the proximity force approximation as a function of the geometric parameters.

Submitted to: *Phys. Scr.*
Keywords: Casimir effect, electromagnetic scattering, sphere-plane geometry

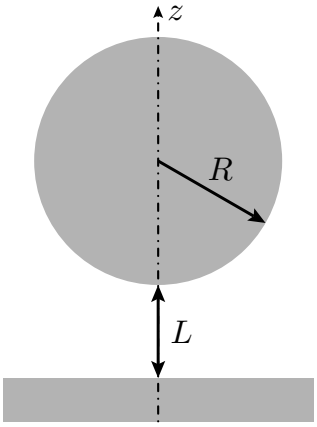

Figure 1. Geometry of the sphere-plane setup. The aspect ratio R/L depends on the sphere radius R and the smallest distance L between sphere and plane. The setup is axially symmetric about the z-axis.

1. Introduction

When Casimir first derived a force between two objects induced by quantum fluctuations of the electromagnetic vacuum [1], he considered a setup consisting of two infinitely extended, perfectly reflecting parallel plates and obtained a universal expression for the pressure pushing the plates towards each other. In a finite-size system, the Casimir force can thus be increased by increasing the size of the plates.

Instead of the plane-plane geometry, most modern experiments rather employ the sphere-plane geometry, thus avoiding the problem of misalignment. In experiments with a particularly large sphere radius, only a section of a sphere is used. The geometrical parameter characterizing the sphere-plane setup displayed in figure 1 is the aspect ratio R/L, i.e. the ratio between the sphere radius R and the minimal distance L between sphere and plane. Note that in figure 1 the sphere radius is chosen to be rather small. The corresponding aspect ratio of $4/3$ is more than two orders of magnitude smaller than common aspect ratios used in experiments. An intuitive idea of a typical aspect ratio is obtained by imagining the ratio of the earth radius and the typical cruising altitude of a commercial jet.

From an experimental point of view, the aspect ratio should be chosen large in order to obtain a sizeable force and to allow for precise measurements which are relevant in various respects. Precise Casimir force measurements are crucial to detect possible deviations from the gravitational interaction at submicrometer distances [2, 3] and thus to exclude or possibly support proposed mechanisms for a fifth fundamental interaction [4, 5]. Precise measurements render the results also sensitive to aspects of the experimental setup beyond the geometrical features, notably the material properties of sphere and plate [6]. This has led to the so-called Drude-plasma controversy. While the materials used in experiments clearly have a finite dc conductivity which can be accounted for within the Drude model, the plasma model with its infinite dc conductivity describes better most Casimir force measurements (for a recent experiment addressing this issue see Ref. [7]).

While on the experimental side, one aims for relatively large aspect ratios, on the theoretical side, the situation is a bit more complicated. The cases of extremely large aspect ratios and of very small aspect ratios can be treated rather easily while intermediate aspect ratios can be very demanding to cope with.

For the theoretical description of experiments, frequently the proximity force approximation (PFA) has been employed. It assumes that the Casimir force between a sphere and a plate can be decomposed into contributions from many small plane-plane segments which are integrated over. As the Casimir force is not additive, this approach can only be approximate. For perfect reflectors at zero temperature, the PFA yields the Casimir force in the sphere-plane geometry

$$F_{\mathrm{PFA}} = -\frac{\pi^3 \hbar c}{360} \frac{R}{L^3} \,. \tag{1}$$

The proximity force approximation can be generalized to account for arbitrary electromagnetic response of sphere and plate as well as arbitrary temperatures by using the corresponding expression for the force in the plane-plane geometry. Recently, it has been proven for the sphere-sphere geometry that the proximity force approximation yields the leading small-distance behaviour for arbitrary temperatures and materials [8]. In particular, the proximity force approximation for the sphere-plane geometry is obtained by taking one sphere radius to infinity. Furthermore, Ref. [8] has established a connection between the specular-reflection limit of Mie scattering and the proximity force approximation.

In the opposite limit of small aspect ratios, it is convenient to express the Casimir force in terms of

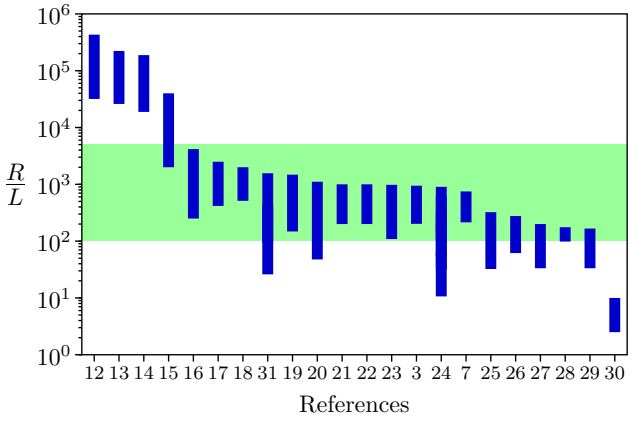

Figure 2. The aspect ratio R/L is shown for Casimir experiments involving a sphere or a spherical lens and a plane or another sphere. The range of aspect ratios, which became accessible through the numerical approach discussed here, is marked in green.

a multipole expansion. For very small aspect ratios, the dipole approximation allows for analytical results which have been used to analyse the appearance of a negative Casimir entropy [9, 10]. The number of multipole moments required in a numerical calculation of the Casimir force increases linearly with the aspect ratio. Since typically the numerical effort grows with the third power of the dimension, the numerical evaluation of the Casimir force has been restricted to aspect ratios $R/L \lesssim 100$ [11].

In figure 2, we show the aspect ratio R/L for recent experiments [3, 7, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29] measuring the Casimir force between a sphere and a plate. For large aspect ratios, the sphere radius is typically so large that the sphere is replaced by a sphere segment. In two of the experiments, measurements were performed on the sphere-sphere geometry [30, 31]. The smallest aspect ratio is reached by an experiment proposing extremely sensitive measurements of the Casimir force between two spheres by means of optical tweezers [30].

Figure 2 indicates that most experiments were out of reach of exact theoretical calculations until very recently. In order to cover a large number of them, it is necessary to numerically treat aspect ratios $R/L \lesssim 5000$. The extension to the aspect ratios marked in green in figure 2 became possible by symmetrizing the round-trip operator describing the scattering of electromagnetic waves between sphere and plate [32]. In the present paper, we will give a more detailed discussion of this approach. The very large aspect ratios not yet covered are rather well described by the proximity force approximation up to corrections typically smaller than one percent.

Recently, an alternative approach to the numerical computation of the Casimir force and force gradient has been proposed. Combining the exact result in the high-temperature limit for Drude metals with asymptotic techniques, a formula more accurate than the proximity force approximation has been derived [33, 34, 35]. This approach offers significant advantages with respect to numerical speed and simplicity as compared to our approach. However, it is not possible to make the error of this approximation arbitrarily small. Moreover, since in the high-temperature limit an exact result is only known for the Drude model, the extension of the approach to other models like perfect conductors or the plasma model requires non-trivial modifications [35].

In the following, we discuss how it becomes possible to advance into the green region of figure 2. We start in Section 2 by explaining as the central idea the symmetrization of the round-trip operator. Sections 3 and 4 provide technical details allowing to set up the required matrix elements. Numerical aspects related to hierarchical matrices and to the performance achieved by making use of them are given in Section 5. As a physical application, we discuss in Section 7 the size of the corrections to PFA as a function of the geometrical parameters in the sphere-plane geometry before presenting our conclusions in Section 8. Some more technical details are given in the appendices.

2. Symmetrization of the round-trip operator

The Casimir force for the setup originally considered by Casimir, i.e. two parallel perfectly reflecting plates, can be evaluated by summing the vacuum energy over all modes of the electromagnetic field [36]. For more general situations like the experimentally relevant sphere-plane geometry discussed here, the scattering approach has turned out to be very well suited [37].

Within the scattering approach to the Casimir effect in imaginary frequencies $\omega = \mathrm{i}\xi$, the free energy is expressed as a sum [38, 39]

$$\mathcal{F} = \frac{k_B T}{2} \sum_{n=-\infty}^{\infty} \log\det\left[1 - \mathcal{M}(|\xi_n|)\right] \tag{2}$$

over the Matsubara frequencies $\xi_n = 2\pi n k_B T/\hbar$. The round-trip operator

$$\mathcal{M} = \mathcal{R}_\mathrm{S}\mathcal{T}_\mathrm{SP}\mathcal{R}_\mathrm{P}\mathcal{T}_\mathrm{PS} \tag{3}$$

represents a complete round-trip of an electromagnetic wave between the sphere and the plane as indicated in figure 3a. The operator $\mathcal{T}_\mathrm{SP}$ describes a translation from the reference frame of the plane to that of the sphere, and vice versa for $\mathcal{T}_\mathrm{PS}$. $\mathcal{R}_\mathrm{S}$ denotes the reflection at the sphere, while $\mathcal{R}_\mathrm{P}$ denotes the reflection at the plane. The Matsubara sum (2) together with (3) holds even if sphere, plane or the medium in between are dissipative [40].

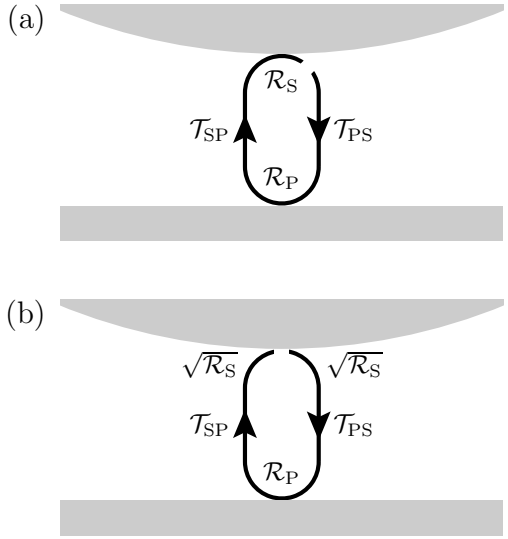

Figure 3. Graphical representation of the round-trip operator between sphere and plane in (a) its usual form (3) and (b) the symmetrized form (5). The reflection operators at plane and sphere are denoted by $\mathcal{R}_P$ and $\mathcal{R}_S$, respectively, while $\mathcal{T}_{PS}$ and $\mathcal{T}_{SP}$ denote the translation operators from sphere to plane and vice versa.

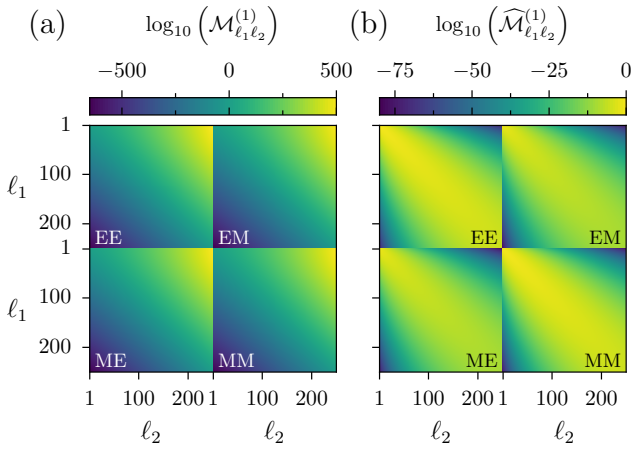

Figure 4. The logarithm of the matrix elements of (a) the non-symmetrized round-trip operator $\mathcal{M}^{(m)}$ and (b) the symmetrized round-trip operator $\widehat{\mathcal{M}}^{(m)}$ in the multipole basis is shown on a colour scale for $R/L = 50$, $\xi(L+R)/c = 1$, $m = 1$, and perfect reflectors. The four blocks correspond to the different sequences of polarizations during a round trip. While the non-symmetrized round-trip matrix is ill-conditioned, the matrix elements of the symmetrized round-trip operator take their maximum on the diagonal and decrease away from it.

We express the round-trip operator $\mathcal{M}$ in the multipole basis $|\ell, m, P\rangle$ described by the angular momentum quantum numbers ℓ and m ($\ell = 1, 2, \ldots$ and $|m| \leq \ell$). The polarization represents electric multipoles for $P = \mathrm{E}$ and magnetic multipoles for $P = \mathrm{M}$, respectively. Due to the rotational symmetry of the plane-sphere setup about the z-axis, the round-trip operator is diagonal in m and every block $\mathcal{M}^{(m)}$ yields an independent contribution to the free energy

$$\mathcal{F} = \frac{k_B T}{2} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \log\det\left(1 - \mathcal{M}^{(m)}(|\xi_n|)\right) . \quad (4)$$

The numerical problems associated with the definition (3) of the round-trip operator become clear from figure 4a where we depict the values of the matrix elements $\langle \ell_1, m, P_1 | \mathcal{M} | \ell_2, m, P_2 \rangle$ on a logarithmic colour scale. Even though the matrix elements depend on the parameters chosen here, $R/L = 50$, $\xi(L+R)/c = 1$, $m = 1$, and perfect conductors, the data shown are typical. Already for the relatively small aspect ratio used in figure 4a, the round-trip operator (3) clearly results in an ill-conditioned matrix with elements differing by hundreds of orders of magnitude. As a consequence, a fast and stable numerical evaluation of the determinant becomes extremely difficult. When the determinants in (4) are evaluated, the combination of very small and very large matrix elements can yield contributions of the order one. Small perturbations in the matrix elements may then cause large errors. Furthermore, common computer number formats cover a range of numbers from about 10^{-324} to 10^{308} [41] which is not sufficient to represent all matrix elements in figure 4a. Instead, one has to use number formats that cover a wider range of numbers, but are also significantly slower.

To overcome these problems, we make use of the fact that the round-trip operator is not uniquely defined. Instead of (3), we choose the symmetrized form for the round-trip operator

$$\widehat{\mathcal{M}} = \sqrt{\mathcal{R}_S}\mathcal{T}_{SP}\mathcal{R}_P\mathcal{T}_{PS}\sqrt{\mathcal{R}_S} . \quad (5)$$

as illustrated in figure 3b. The scattering operator at the sphere $\mathcal{R}_S$ is diagonal in the multipole basis and therefore the matrix square root exists. With this choice for the round-trip operator, the matrix elements take their maximum on the diagonal of each polarization block and decrease away from it, as can be seen in figure 4b. Here, the same parameters have been used as in figure 4a.

In figure 5, we depict a polarization-conserving block (left) and a polarization-mixing block (right) with the matrix elements taken from figure 4b, but now on a linear colour scale. This representation emphasizes the fact that a sizeable fraction of the matrix elements off the diagonal is numerically irrelevant. Furthermore, the contribution of the polarization-mixing blocks is significantly smaller than that of the polarization-conserving blocks.

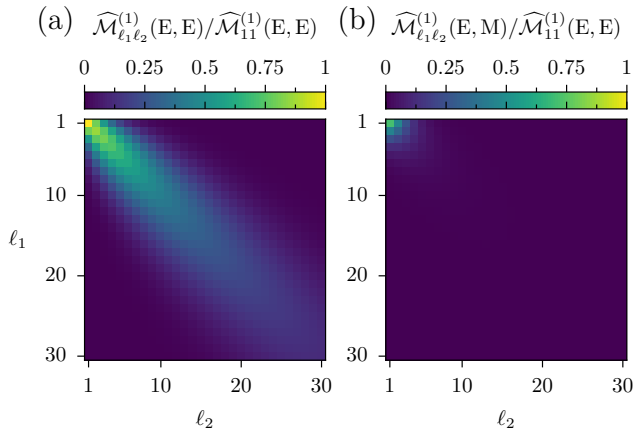

Figure 5. Details of the matrix elements for (a) the polarization maintaining block (E, E) and (b) the polarization mixing block (E, M) are shown for the same parameters as in figure 4, but now on a linear scale. All matrix elements are taken with respect to the largest matrix element in the round-trip matrix.

3. Reflection and translation operators

The reflection operator at the sphere is diagonal in the multipole basis

$$\mathcal{R}_{\mathrm{S}}|\ell,m,P\rangle = r^{(\mathrm{S})}_{\ell,P}|\ell,m,P\rangle\,. \tag{6}$$

The reflection coefficients

$$r^{(\mathrm{S})}_{\ell,E} = -a_\ell, \quad r^{(\mathrm{S})}_{\ell,M} = -b_\ell \tag{7}$$

are given by the Mie coefficients a_ℓ and b_ℓ, where the minus sign is a consequence of the definition of the Mie coefficients employed here [42]. For explicit expressions, we refer the reader to Appendix B.

While the multipole basis is well adapted for the scattering at the sphere, plane waves are better suited to describe the reflection at the plate and the translation between the reference frames of plate and sphere. A plane wave is characterized by the wave vector $\mathbf{K}$ and the polarization p. Given the special role of the z-axis as symmetry axis, we choose the basis vectors for transverse electric (TE) and transverse magnetic (TM) modes as

$$\hat{\varepsilon}_{\mathrm{TE}} = \frac{\hat{\mathbf{z}}\times\hat{\mathbf{K}}}{|\hat{\mathbf{z}}\times\hat{\mathbf{K}}|}, \quad \hat{\varepsilon}_{\mathrm{TM}} = \hat{\varepsilon}_{\mathrm{TE}}\times\hat{\mathbf{K}}\,. \tag{8}$$

Here, unit vectors are denoted by a hat. Within the Matsubara formalism, all quantities are expressed in terms of imaginary frequencies ξ. For the wave vector $\mathbf{K}$, we choose an imaginary z-component κ while the projection $\mathbf{k}$ onto the x-y-plane is kept real. The dispersion relation then reads

$$\xi^2 = c^2(\kappa^2 - |\mathbf{k}|^2) \tag{9}$$

with the speed of light c. As the frequency remains constant during a round trip, it is convenient to make use of the angular spectral representation [43]. The corresponding basis $\{|\mathbf{k},p,\phi\rangle\}$ is then labeled by the projection of the wave-vector $\mathbf{K}$ onto the x-y-plane, $\mathbf{k} = (k_x, k_y, 0)$, the polarization $p = \mathrm{TE}, \mathrm{TM}$, and the direction of propagation $\phi = \pm 1$ in $\pm z$-direction. The latter fixes the sign ambiguity when solving (9) for κ. We are therefore free to always choose κ as positive.

Plane waves propagating in $-z$-direction are reflected by the plate

$$\mathcal{R}_{\mathrm{P}}\,|\mathbf{k},p,-\rangle = r^{(\mathrm{P})}_p(\mathrm{i}\xi,k)\,|\mathbf{k},p,+\rangle \tag{10}$$

where the reflection coefficients $r^{(\mathrm{P})}_p$ are the Fresnel coefficients. Explicit expressions are given in Appendix C.

Finally, the translation operators are diagonal in the plane-wave basis

$$\mathcal{T}_{\mathrm{SP}}\,|\mathbf{k},p,+\rangle = \mathrm{e}^{-\kappa(L+R)}\,|\mathbf{k},p,+\rangle\,, \tag{11}$$

$$\mathcal{T}_{\mathrm{PS}}\,|\mathbf{k},p,-\rangle = \mathrm{e}^{-\kappa(L+R)}\,|\mathbf{k},p,-\rangle \tag{12}$$

with matrix elements given by exponential factors. Here, we have assumed vacuum between sphere and plane as we will do in the rest of the paper.

4. Matrix elements of the round-trip operator

Having discussed the individual building blocks in the previous section, we now combine them to obtain the matrix elements of the symmetrized round-trip matrix (5) in the multipole basis

$$\begin{aligned}\widehat{\mathcal{M}}^{(m)}_{\ell_1\ell_2}(P_1,P_2) &= \langle\ell_1,m,P_1|\widehat{\mathcal{M}}|\ell_2,m,P_2\rangle \\ &= \sqrt{r^{(\mathrm{S})}_{\ell_1,P_1}}\sqrt{r^{(\mathrm{S})}_{\ell_2,P_2}}\sum_p\int_0^\infty\frac{\mathrm{d}^2\mathbf{k}}{(2\pi)^2}r^{(\mathrm{P})}_p\mathrm{e}^{-2\kappa(L+R)} \\ &\quad\times\langle\ell_1,m,P_1|\mathbf{k},p,+\rangle\,\langle\mathbf{k},p,-|\ell_2,m,P_2\rangle\,.\end{aligned} \tag{13}$$

The last two factors arise due to a change from the multipole basis to the plane-wave basis and vice versa between the translation operators and the reflection operators at the sphere. Explicit expressions for these matrix elements are given in Ref. [44].

We organize the round-trip matrix in the form of a block matrix

$$\widehat{\mathcal{M}}^{(m)} = \begin{pmatrix}\widehat{\mathcal{M}}^{(m)}(\mathrm{E},\mathrm{E}) & \widehat{\mathcal{M}}^{(m)}(\mathrm{E},\mathrm{M}) \\ \widehat{\mathcal{M}}^{(m)}(\mathrm{M},\mathrm{E}) & \widehat{\mathcal{M}}^{(m)}(\mathrm{M},\mathrm{M})\end{pmatrix} \tag{14}$$

where the diagonal blocks correspond to matrix elements preserving polarization, and the off-diagonal blocks correspond to matrix elements with a change of polarization. Expressing the double integral in (13) in polar coordinates, the integral over the angle can be carried out. After a similarity transform of the round-trip matrix in order to remove phase factors, we finally obtain the matrix elements

$$\widehat{\mathcal{M}}^{(m)}_{\ell_1\ell_2}(\mathrm{M},\mathrm{M}) = \sqrt{|b_{\ell_1}b_{\ell_2}|}\left(A^{(m)}_{\ell_1\ell_2,\mathrm{TM}} - B^{(m)}_{\ell_1\ell_2,\mathrm{TE}}\right) \tag{15}$$

$$\widehat{\mathcal{M}}^{(m)}_{\ell_1\ell_2}(\mathrm{E},\mathrm{E}) = \sqrt{|a_{\ell_1}a_{\ell_2}|}\left(B^{(m)}_{\ell_1\ell_2,\mathrm{TM}} - A^{(m)}_{\ell_1\ell_2,\mathrm{TE}}\right) \tag{16}$$

$$\widehat{\mathcal{M}}^{(m)}_{\ell_1\ell_2}(\mathrm{M},\mathrm{E}) = \sqrt{|b_{\ell_1}a_{\ell_2}|}\left(C^{(m)}_{\ell_1\ell_2,\mathrm{TM}} - C^{(m)}_{\ell_2\ell_1,\mathrm{TE}}\right) \quad (17)$$

$$\widehat{\mathcal{M}}^{(m)}_{\ell_1\ell_2}(\mathrm{E},\mathrm{M}) = \widehat{\mathcal{M}}^{(m)}_{\ell_2\ell_1}(\mathrm{M},\mathrm{E})\,, \quad (18)$$

which depend on the integrals

$$A^{(m)}_{\ell_1\ell_2,p} = \int_1^\infty \mathrm{d}x f^{(m)}_{\ell_1,\ell_2,p}(x,-1)P^m_{\ell_1}(x)P^m_{\ell_2}(x) \quad (19)$$

$$B^{(m)}_{\ell_1\ell_2,p} = \int_1^\infty \mathrm{d}x f^{(m)}_{\ell_1,\ell_2,p}(x,1)P^{m\prime}_{\ell_1}(x)P^{m\prime}_{\ell_2}(x) \quad (20)$$

$$C^{(m)}_{\ell_1\ell_2,p} = \int_1^\infty \mathrm{d}x f^{(m)}_{\ell_1,\ell_2,p}(x,0)P^m_{\ell_1}(x)P^{m\prime}_{\ell_2}(x) \quad (21)$$

where

$$f^{(m)}_{\ell_1,\ell_2,p}(x,j) = m\Lambda^{(m)}_{\ell_1}\Lambda^{(m)}_{\ell_2}r^{(\mathrm{P})}_p\left(\mathrm{i}\xi,\frac{\xi}{c}\sqrt{x^2-1}\right) \times\left(\frac{x^2-1}{m}\right)^j \exp\left(-2\frac{\xi(L+R)}{c}x\right) \quad (22)$$

and

$$\Lambda^{(m)}_\ell = \sqrt{\frac{2\ell+1}{\ell(\ell+1)}\frac{(\ell-m)!}{(\ell+m)!}}\,. \quad (23)$$

Note that for the associated Legendre polynomials P^m_ℓ, we use an uncommon phase convention defined in Appendix A. The dimension of the matrices (15)–(18) is infinite. For a numerical evaluation, the vector space has to be truncated in the angular momentum.

The matrix elements for the round-trip operator $\mathcal{M}$ differ from those of the symmetrized round-trip operator $\widehat{\mathcal{M}}$ only with respect to the Mie coefficients. While for the symmetrized round-trip operator according to (15)–(18) the matrix elements are proportional to the square root of a product of Mie coefficients with different angular momenta, the matrix elements of $\mathcal{M}$ are proportional to one Mie coefficient with angular momentum ℓ_1, thus resulting in the numerical problems discussed above.

As the Fresnel coefficients are positive for $p = \mathrm{TM}$ and negative for $p = \mathrm{TE}$, the matrix elements of the round-trip operator $\widehat{\mathcal{M}}^{(m)}$ are positive. Also, as the integrals $A^{(m)}_{\ell_1\ell_2,p}$ and $B^{(m)}_{\ell_1\ell_2,p}$ are symmetric with respect to ℓ_1 and ℓ_2, the round-trip matrix is symmetric. Numerical tests suggest that the scattering matrix $1 - \widehat{\mathcal{M}}^{(m)}$ is diagonally dominant. Together with the positivity of the diagonal entries, it follows that the scattering matrix $1 - \widehat{\mathcal{M}}^{(m)}$ is positive definite. These properties ensure the stability of the numerical evaluation of the determinant. Small perturbations in the matrix elements thus only cause small changes in the value of the determinant [45].

5. Hierarchical matrices

The symmetrization allows to exploit further properties of the round-trip matrices. As the matrix elements decrease away from the diagonal, one might think that the dominant contribution to the determinant comes from matrix elements close to the diagonal. In fact, it turns out that the matrices $\widehat{\mathcal{M}}^{(m)}$ are hierarchical off-diagonal low-rank (HODLR) matrices [46]. This means that the round-trip matrices can be sub-divided into a hierarchy of rectangular blocks which can be approximated by low-rank matrices.

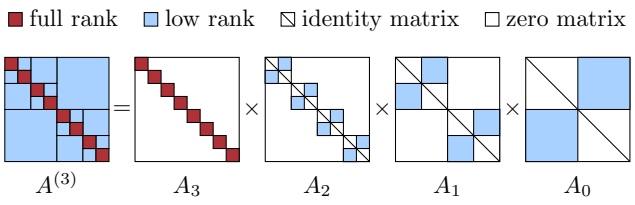

Figure 6. Graphical representation of the factorization (25) of a HODLR matrix for $n = 3$. Red blocks represent full-rank matrices while bright blue blocks correspond to low-rank matrices (after [46]).

A low-rank matrix M of dimension $N \times N$ can be efficiently approximated by

$$M \approx UV^T \quad (24)$$

where U and V are matrices of dimension $N \times p$ with $p \ll N$. The best rank p approximation of M can be obtained using the singular value decomposition [49]. Instead of a computationally expensive full singular-value decomposition, low-rank approximations can be computed using fast algorithms like adaptive cross approximation, pseudo-skeletal approximations, interpolatory decompositions or rank revealing QR and LU (see [46] and references therein).

A HODLR matrix A can be factored into $n + 1$ block-diagonal matrices

$$A \approx A^{(n)} = A_n A_{n-1} \dots A_0 \quad (25)$$

as sketched in figure 6 for $n = 3$. The matrix A_n consists of 2^n full-rank blocks around the diagonal while the other matrices A_{n-1} to A_0 represent low-rank updates to the identity. The error of this approximation can be made negligibly small by choosing appropriate ranks p.

The factorization (25) allows for a fast computation of the determinant $A^{(n)}$ for two reasons. Firstly, one can exploit the multiplicativity of determinants. Secondly, the block matrices appearing in A_0 to A_{n-1} are of the form

$$B = \begin{pmatrix} 1 & U_1V_1^T \\ U_2V_2^T & 1 \end{pmatrix}, \quad (26)$$

requiring at first sight the evaluation of the determinant of an $N \times N$-matrix according to

$$\det(B) = \det\left(1 - U_1V_1^TU_2V_2^T\right)\,. \quad (27)$$

However, exploiting Sylvester's determinant identity

$$\det(1 + UV) = \det(1 + VU)\,, \quad (28)$$

we obtain

$$\det(B) = \det\left(1 - V_2^T U_1 V_1^T U_2\right) . \tag{29}$$

It is thus sufficient to evaluate the determinant of a $p \times p$-matrix, resulting in a significant speed-up.

Instead of the free energy $\mathcal{F}$, Casimir experiments measure either the force $F = -\partial\mathcal{F}/\partial L$ or the force gradient $F' = -\partial^2\mathcal{F}/\partial L^2$. The scattering formula for the force does not involve determinants, but traces of products of matrices

$$\frac{\partial}{\partial L} \log\det(A) = \mathrm{tr}\left(A^{-1}A'\right), \tag{30}$$

where $(A')_{jk} = \partial A_{jk}/\partial L$.

The factorization (25) also allows for a fast evaluation of (30). Firstly, the matrix product of two factorized matrices can be computed efficiently. Secondly, the inverse of A is given as a product of matrices

$$A^{-1} \approx \left(A^{(n)}\right)^{-1} = A_0^{-1} A_1^{-1} \dots A_n^{-1}, \tag{31}$$

and each matrix A_j can be inverted independently.

The inverse of the block matrices B defined in (26) is given by

$$B^{-1} = \begin{pmatrix} 1 + U_1 V_1^T W U_2 V_2^T & -U_1 V_1^T W \\ -W U_2 V_2^T & W \end{pmatrix}, \tag{32}$$

with the $N/2 \times N/2$-matrix

$$W = (1 - U_2 V_2^T U_1 V_1^T)^{-1} . \tag{33}$$

Using the Woodbury matrix identity in the form

$$(1 + UV)^{-1} = 1 - U\,(1 + VU)^{-1}\,V, \tag{34}$$

(33) can be reexpressed as

$$W = 1 + U_2\left(1 - V_2^T U_1 V_1^T U_2\right)^{-1} V_2^T U_1 V_1^T . \tag{35}$$

Consequently, the evaluation of A^{-1} can be reduced to the inversion of $p \times p$-matrices, thus allowing for an efficient evaluation of the force. A similar approach can be used to compute the force gradient F'.

The numerical library used in our calculation and further discussed in the following section, unfortunately has not yet implemented these ideas. Therefore, we determine the force and the force gradient numerically by computing the first and second derivative of the free energy using a symmetric finite difference formula of order 8 [47]. The numerical error can be controlled by variation of the step size.

From comparisons with the analytical result in the high-temperature case for Drude metals [48], we find as a rule of thumb that the numerical error of the force is typically twice as large as the numerical error for the free energy while it is increased by a factor of 10 for the force gradient. With typical relative errors for the free energy of 10^{-7}, we thus still achieve numerical errors of 10^{-6} for the force gradient.

6. Numerical complexity

In order to assess the numerical advantages of the approach discussed above, we compute the determinants of the scattering matrices either by means of a Cholesky decomposition or the implementation [50] of the algorithm for HODLR matrices described in [46]. The Cholesky decomposition factorizes a symmetric positive-definite matrix into the product of a triangular matrix and its transpose allowing for a simple computation of the determinant. The factorization requires $\mathcal{O}(N^3)$ of time for an $N \times N$ matrix and is about twice as fast as an LU decomposition. The computation of determinants using the HODLR approach takes $\mathcal{O}(p^2 N \log^2 N)$ steps where, depending on the nature of the problem, p may be a function of N.

In figure 7 we compare the average time to compute $\log\det\left(1 - \widehat{\mathcal{M}}^{(m)}(\xi)\right)$ depending on the aspect ratio using the HODLR approach and a Cholesky decomposition. We specifically choose $m = 1$, $\xi = c/(L + R)$, and perfect reflectors, but other parameters yield similar results. For aspect ratios $R/L \lesssim 100$, both algorithms take about the same time. For larger aspect ratios $R/L \gtrsim 200$, the computational time using the Cholesky decomposition scales as $\propto (R/L)^{2.56}$. This is faster than the theoretical complexity $\mathcal{O}((R/L)^3)$ of the Cholesky decomposition, because our numerical implementation saves and reuses intermediate values so that the computational time per matrix element is not constant. In contrast, the HODLR algorithm becomes significantly faster for $R/L \gtrsim 200$ and the computational time scales only as $\propto (R/L)^{1.31}$. For the largest aspect ratio displayed in figure 7, $R/L = 2000$, we find a speed-up by a factor 33. At even larger aspect ratios, the time required by the Cholesky decomposition becomes prohibitively long.

7. Corrections to the PFA for the force

As an application of our numerical approach, we consider the corrections to the Casimir force

$$F = -\frac{\partial\mathcal{F}}{\partial L} \tag{36}$$

beyond the proximity force approximation. The geometrical parameters of the sphere-plane geometry R and L (cf. figure 1) for which the relative correction $1 - F/F_{\mathrm{PFA}}$ of the force F with respect to its PFA value F_{PFA} takes the values $0.25\,\%, 0.5\,\%$, and $1\,\%$ can be read off from the corresponding symbols shown in figure 8. The numerical data have been determined for gold surfaces at a finite temperature of 300 K. According to Appendix B and Appendix C, we need the dielectric function $\epsilon(\mathrm{i}\xi)$ for gold on the imaginary axis which can be obtained from tabulated data [53] by means of a procedure explained in Ref. [54]. In view of

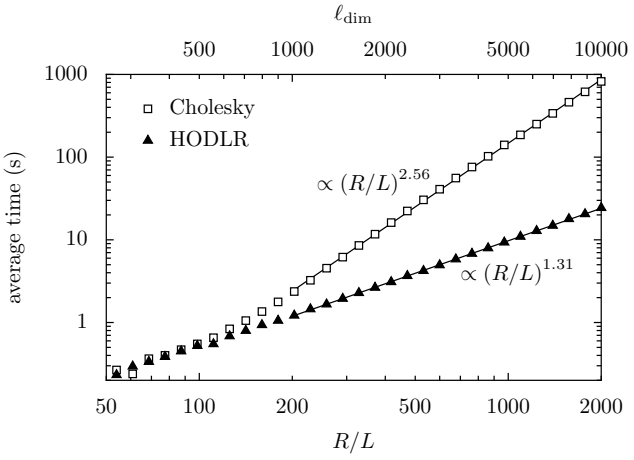

Figure 7. Average runtime to compute $\log\det\left(1-\widehat{\mathcal{M}}^{(m)}(\xi)\right)$ as a function of the aspect ratio R/L for $m=1$, $\xi=c/(L+R)$, and perfect reflectors. The angular momentum is truncated at $\ell_{\mathrm{dim}}=5R/L$, yielding a scattering matrix of dimension $2\ell_{\mathrm{dim}}\times 2\ell_{\mathrm{dim}}$. Squares and triangles correspond to a computation using Cholesky decomposition and the HODLR algorithm [50], respectively. The lines represent fits and correspond to the asymptotic scaling of the two algorithms. The computations were carried out on an Intel Core i7 with 3.4 GHz. For the Cholesky decomposition LAPACK [51] in combination with ATLAS [52] was used.

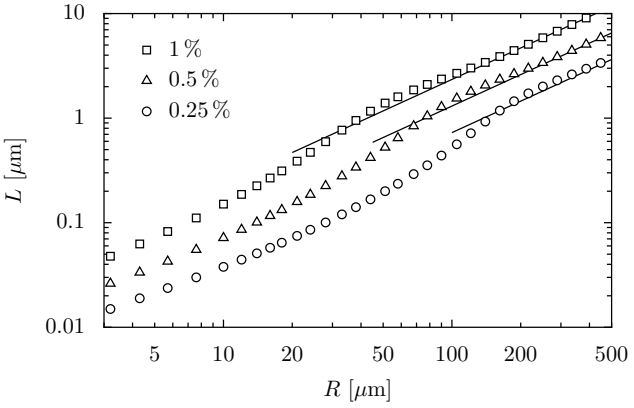

Figure 8. The correction $1-F/F_{\mathrm{PFA}}$ of the force with respect to the corresponding PFA result is displayed as a function of the radius R of the sphere and its distance L from the plane. The numerical data depicted by symbols corresponding to different values of the correction have been obtained for gold at $T=300\,\mathrm{K}$ as explained in the text. The solid lines indicate the corresponding high-temperature result according to Ref. [48].

the still ongoing debate on how to correctly treat the zero-frequency contribution to the Matsubara sum, we note that here we use the Drude prescription where the transverse electric mode does not contribute. For the plasma prescription, the corresponding curves for a given value of the correction with respect to PFA lie below the ones shown in figure 8, i.e. at larger aspect ratios [32].

For a given value of the correction to the PFA force, the data in figure 8 display three different regimes. In the upper right corner, i.e. for large values of L and R, the thermal wavelength $\lambda_{\mathrm{T}}=\hbar c/k_{\mathrm{B}}T$ can be taken to be small, thus indicating the high-temperature regime. In contrast, the low-temperature behaviour is found in the opposite corner, i.e. the lower left part. The two regimes are joined by a transition region visible in the middle of figure 8.

In the high-temperature regime, only the term with $n=0$ in the Matsubara sum (2) contributes and thus only the zero-frequency limit of the dielectric function is relevant. This special case within the Drude model allows for an analytical solution for the free energy in the sphere-plane setup [48]. As R and L are the only length scales remaining in the high-temperature limit, a fixed value of the correction to the PFA force yields a straight line of slope 1 in the presentation of figure 8. The high-temperature limits for the chosen values of the correction are indicated by solid lines. As expected, our numerical data approach the high-temperature limit for large sphere radii and large distances between sphere and plane. We note that with decreasing sphere radius, the distance L required for a given value of the correction to the PFA force exceeds the distance obtained within the high-temperature limit. However, decreasing the radius further, the distance falls below the prediction of the high-temperature limit.

8. Conclusions

The standard approach to calculating Casimir forces and free energies within the scattering theory in the multipole basis has been plagued with ill-conditioned round-trip matrices resulting in various numerical problems. We have shown that these problems can be eliminated by a symmetrization of the round-trip operator which physically amounts to splitting the round-trip of an electromagnetic wave between the two scattering objects in the middle of the reflection at one of these objects. The symmetrization of the round-trip operator allows to significantly reduce computational time and thus to perform calculations in the experimentally relevant regime of aspect ratios in the sphere-plane geometry. Furthermore, numerical errors become controllable.

As large aspect ratios are now numerically accessible, it has become possible to assess the quality of the proximity force approximation by determining the deviations from the exact result (cf. figure 8). This is particularly relevant as the analysis of experimental results so far has relied on the proximity force approximation. Specifically, precise numerical results are of importance in the ongoing Drude-plasma

debate. As an example, with the numerical approach detailed here it has become possible recently [32] to demonstrate that the experimental bounds for the Casimir force gradient found in Ref. [24] are violated both for the Drude and the plasma prescription, even though the violation for the latter is found to be significantly larger.

Acknowledgments

This paper is dedicated to Wolfgang Schleich on the occasion of his 60$^{\text{th}}$ birthday. Particularly during our yearly Augsburg-Ulm meetings, we could admire his sagacious physical reasoning and draw inspiration from it.

We thank A. Canaguier-Durand, R. Guérout, A. Lambrecht, and S. Reynaud for discussions and D. Dalvit for providing numerical data for the dielectric function of gold at imaginary frequencies. The authors acknowledge support from CAPES and DAAD through the PROBRAL collaboration program. P. A. M. N. also thanks CNPq and FAPERJ for partial financial support.

Appendix A. Associated Legendre polynomials

We define the associated Legendre polynomials as [55]

$$P_\ell^m(x) = (x^2-1)^{m/2}\frac{\mathrm{d}^m}{\mathrm{d}x^m}P_\ell(x) \tag{A.1}$$

where $P_\ell(x)$ denotes the ordinary Legendre polynomial of degree ℓ. The phase convention employed here differs from the common choice usually made in physics: We omit the Condon-Shortley phase, and change the sign in the first term. With this choice the associated Legendre polynomials are real and non-negative functions for $x \geq 1$.

Appendix B. Mie coefficients

The Mie coefficients are given by [42]

$$a_\ell(\mathrm{i}x) = (-1)^\ell \frac{\pi}{2}\frac{n^2 s_\ell^{(\mathrm{a})} - s_\ell^{(\mathrm{b})}}{n^2 s_\ell^{(\mathrm{c})} + s_\ell^{(\mathrm{d})}}, \tag{B.1}$$

$$b_\ell(\mathrm{i}x) = (-1)^{\ell+1} \frac{\pi}{2}\frac{s_\ell^{(\mathrm{b})} - s_\ell^{(\mathrm{a})}}{s_\ell^{(\mathrm{c})} + s_\ell^{(\mathrm{d})}}, \tag{B.2}$$

where

$$s_\ell^{(\mathrm{a})}(x) = I_{\ell+\frac{1}{2}}(nx)\left[xI_{\ell-\frac{1}{2}}(x) - \ell I_{\ell+\frac{1}{2}}(x)\right], \tag{B.3}$$

$$s_\ell^{(\mathrm{b})}(x) = I_{\ell+\frac{1}{2}}(x)\left[nxI_{\ell-\frac{1}{2}}(nx) - \ell I_{\ell+\frac{1}{2}}(nx)\right], \tag{B.4}$$

$$s_\ell^{(\mathrm{c})}(x) = I_{\ell+\frac{1}{2}}(nx)\left[xK_{\ell-\frac{1}{2}}(x) + \ell K_{\ell+\frac{1}{2}}(x)\right], \tag{B.5}$$

$$s_\ell^{(\mathrm{d})}(x) = K_{\ell+\frac{1}{2}}(x)\left[nxI_{\ell-\frac{1}{2}}(nx) - \ell I_{\ell+\frac{1}{2}}(nx)\right]. \tag{B.6}$$

The Mie coefficients are evaluated at $x = \xi R/c$ and $n = \sqrt{\epsilon(\mathrm{i}\xi)}$ denotes the refractive index. The dielectric function $\epsilon(\mathrm{i}\xi)$ is evaluated at imaginary frequencies. The coefficients $s_\ell^{(\mathrm{a})}$, $s_\ell^{(\mathrm{b})}$, $s_\ell^{(\mathrm{c})}$, $s_\ell^{(\mathrm{d})}$, and the numerators in (B.1) and (B.2) are positive independently of ℓ and ξ.

Appendix C. Fresnel coefficients

The Fresnel coefficients are given by

$$r_{\mathrm{TE}}^{(\mathrm{P})}(\mathrm{i}\xi, k) = \frac{c\kappa - \sqrt{c^2\kappa^2 + \xi^2\left[\epsilon(\mathrm{i}\xi) - 1\right]}}{c\kappa + \sqrt{c^2\kappa^2 + \xi^2\left[\epsilon(\mathrm{i}\xi) - 1\right]}}, \tag{C.1}$$

$$r_{\mathrm{TM}}^{(\mathrm{P})}(\mathrm{i}\xi, k) = \frac{\epsilon(\mathrm{i}\xi)c\kappa - \sqrt{c^2\kappa^2 + \xi^2\left[\epsilon(\mathrm{i}\xi) - 1\right]}}{\epsilon(\mathrm{i}\xi)c\kappa + \sqrt{c^2\kappa^2 + \xi^2\left[\epsilon(\mathrm{i}\xi) - 1\right]}}. \tag{C.2}$$

For perfect reflectors they simplify to $r_{\mathrm{TM}}^{(\mathrm{P})} = 1$ and $r_{\mathrm{TE}}^{(\mathrm{P})} = -1$.

References

[1] Casimir H B G 1948 *Proc. K. Ned. Akad. Wet.* **51** 793
[2] Decca R S, Fischbach E, Klimchitskaya G L, Krause D E, López D and Mostepanenko V M 2003 *Phys. Rev.* D **68** 116003
[3] Decca R S, López D, Fischbach E, Klimchitskaya G L, Krause D E and Mostepanenko V M 2007 *Phys. Rev.* D **75** 077101
[4] Fischbach E and Talmadge C 1992 *Nature* **356** 207
[5] Antoniadis I *et al* 2011 *C. R. Physique* **12** 755
[6] Klimchitskaya G L, Mohideen U and Mostepanenko V M 2009 *Rev. Mod. Phys.* **81** 1827
[7] Bimonte G, López D and Decca R S 2016 *Phys. Rev.* B **93** 184434
[8] Spreng B, Hartmann M, Henning V, Maia Neto P A and Ingold G-L 2018 Proximity force approximation and specular reflection: Application of the WKB limit of Mie scattering to the Casimir effect arXiv:1803.02254
[9] Ingold G-L, Umrath S, Hartmann M, Guérout R, Lambrecht A, Reynaud S and Milton K A 2015 *Phys. Rev.* E **91** 033203
[10] Umrath S, Hartmann M, Ingold G-L and Maia Neto P A 2015 *Phys. Rev.* E **92** 042125
[11] Canaguier-Durand A 2011 Ph.D. thesis (Paris: Université Pierre et Marie Curie)
[12] Masuda M and Sasaki M 2009 *Phys. Rev. Lett.* **102** 171101
[13] Sushkov A O, Kim W J, Dalvit D A R and Lamoreaux S K 2011 *Nat. Phys.* **7** 230
[14] Lamoreaux S K 1997 *Phys. Rev. Lett.* **78** 5
[15] Garcia-Sanchez D, Fong K Y, Bhaskaran H, Lamoreaux S and Tang H X 2012 *Phys. Rev. Lett.* **109** 027202
[16] van Zwol P J, Palasantzas G, van de Schootbrugge M and De Hosson J Th M 2008 *Appl. Phys. Lett.* **92** 054101
[17] van Zwol P J, Palasantzas G and De Hosson J Th M 2008 *Phys. Rev.* B **77** 075412
[18] de Man S, Heeck K, Wijngaarden R J and Iannuzzi D 2009 *Phys. Rev. Lett.* **103** 040402
[19] Decca R S, López D, Fischbach E and Krause D E 2003 *Phys. Rev. Lett.* **91** 050402
[20] Chan H B, Aksyuk V A, Kleiman R N, Bishop D J and Capasso F 2001 *Science* **291** 1941
[21] Munday J N, Capasso F, Parsegian V A and Bezrukov S M 2008 *Phys. Rev.* A **78** 032109
[22] Munday J N, Capasso F and Parsegian V A 2009 *Nature* **457** 170

[23] Mohideen U and Roy A 1998 *Phys. Rev. Lett.* **81** 4549
[24] Krause D E, Decca R S, López D and Fischbach E 2007 *Phys. Rev. Lett.* **98** 050403
[25] Elzbieciak-Wodka M, Popescu M N, Ruiz-Cabello F J M, Trefalt G, Maroni P and Borkovec M 2014 *J. Chem. Phys.* **140** 104906
[26] Banishev A A, Klimchitskaya G L, Mostepanenko V M and Mohideen U 2013 *Phys. Rev. Lett.* **110** 137401
[27] Jourdan G, Lambrecht A, Comin F and Chevrier J 2009 *EPL* **85** 31001
[28] Chang C-C, Banishev A A, Castillo-Garza R, Klimchitskaya G L, Mostepanenko V M and Mohideen U 2012 *Phys. Rev.* B **85** 165443
[29] Torricelli G, Pirozhenko I, Thornton S, Lambrecht A and Binns C 2011 *EPL* **93** 51001
[30] Ether Jr. D S *et al* 2015 *EPL* **112** 44001
[31] Garrett J L, Somers D A T and Munday J N 2018 *Phys. Rev. Lett.* **120** 040401
[32] Hartmann M, Ingold G-L and Maia Neto P A 2017 *Phys. Rev. Lett.* **119** 043901
[33] Bimonte G 2017 *EPL* **118** 20002
[34] Bimonte G 2018 *Phys. Rev. D* **97** 085011
[35] Bimonte G 2018 *arXiv:1804.07182*
[36] Schleich W 2001 *Quantum Optics in Phase Space* (Berlin: Wiley-VCH)
[37] For an elementary introduction, see Ingold G-L and Lambrecht A 2015 *Am. J. Phys.* **83** 156
[38] Lambrecht A, Maia Neto P A and Reynaud S 2006 *New J. Phys.* **8** 243
[39] Emig T, Graham N, Jaffe R L and Kardar M 2007 *Phys. Rev. Lett.* **99** 170403
[40] Guérout R, Ingold G-L, Lambrecht A and Reynaud S 2018 *Symmetry* **10** 37
[41] IEEE Computer Society 2008 *IEEE Standard for Floating-Point Arithmetic: IEEE Std 754-2008*
[42] Bohren F C and Huffman D R 1983 *Absorption and Scattering of Light by Small Particles* (Weinheim: Wiley-VCH) ch. 4.
[43] Nieto-Vesperinas M 2006 *Scattering and diffraction in physical optics* (Singapore: World Scientific)
[44] Messina R, Maia Neto P A, Guizal B and Antezza M 2015 *Phys. Rev.* A **92** 062504
[45] Dailey M, Dopico F M and Ye Q 2014 *SIAM J. Matrix Anal. Appl.* **35** 1303
[46] Ambikasaran S and Darve E 2013 *J. Sci. Comput.* **57** 477
[47] Fornberg B 1988 *Math. Comp.* **51** 699
[48] Bimonte M and Emig T 2012 *Phys. Rev. Lett.* **109** 160403
[49] Eckart C and Young G 1936 *Psychometrika* **1** 211
[50] Ambikasaran S 2013 *A fast direct solver for dense linear systems*, `https://github.com/sivaramambikasaran/HODLR`
[51] Anderson E *et al* 1999 *LAPACK Users' Guide* 3rd edition (Philadelphia: Society for Industrial and Applied Mathematics)
[52] Whaley R C and Petitet A 2005 *Softw. Pract. Exper.* **35** 101
[53] Palik E (ed) 1998 *Handbook of Optical Constants of Solids* vol 1 (New York: Academic Press)
[54] Lambrecht A and Reynaud S 2000 *Eur. Phys. J. D* **8** 309
[55] Zhang S and Jin J 1996 *Computation of Special Functions* (New York: Wiley)