Truncation strategy for the series expressions in the advanced ENZ-theory of diffraction integrals

S. van Haver
S[&]T Experts Pool (STEP)
P.O. Box 608, 2600 AP Delft, The Netherlands, and
Optics Research Group, Faculty of Applied Sciences,
Technical University Delft,
Van der Waalsweg 8,
2618 CH Delft, The Netherlands.
E-mail svenvanhaver@gmail.com

A.J.E.M. Janssen
Department of Mathematics and Computer Science,
Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands.
E-mail a.j.e.m.janssen@tue.nl

Abstract.

The advanced ENZ-theory of diffraction integrals, as published recently in J. Europ. Opt. Soc. Rap. Public. **8**, 13044 (2013), presents the diffraction integrals per Zernike term in the form of doubly infinite series. These double series involve, aside from an overall azimuthal factor, the products of Jinc functions Jinc_h for the radial dependence and structural quantities c_t that depend on the optical parameters of the optical system (such as NA and refractive indices) and the defocus value. The products in the double series have coefficients that are related to Clebsch-Gordan coefficients and that depend on the order h of the Jinc function and the index t of the structural quantity, as well as on the azimuthal order m and degree n of the involved Zernike term Z_n^m. The structural quantities themselves are also given in the

form of doubly infinite series, the terms of which are products of Zernike coefficients a_l, pertaining to an algebraic function containing the optical parameters, and Zernike coefficients b_k, pertaining to a focal factor, and these products have coefficients that are again related to Clebsch-Gordan coefficients. Finally, the a_l, are also given in the form of an infinite series. In this paper, we give truncation rules for the various infinite series depending on required accuracy. In particular, we make precise the following rule-of-thumb for truncation of the double series per Zernike term: For a given value of the radial variable r and the defocus parameter f, it is enough to include in the double series
– all Jinc functions of order h less than H,
– all structural quantities with index t less than T,
where H is somewhat larger than $2\pi r$ and T is somewhat larger than $\frac{1}{2}\,|f|$. We present of this rule both a global version, which can be used for all Zernike terms at the same time, and a dedicated version, in which the H and T take into account order and degree of the involved Zernike term.

Keywords.
Diffraction integral, advanced ENZ-theory, double series, Jinc functions, structural quantities, Debye asymptotics of Bessel functions.

1 Introduction and overview

The advanced ENZ-theory of diffraction integrals, as presented in [1], aims at the computation of the Debye approximation of the Rayleigh integral for the optical point-spread functions of radially symmetric optical systems that range from as basic as having low NA and small defocus value to advanced high-NA systems, with vector fields and polarization, that are meant for imaging of extended objects into a multilayer structure. As in the classical Nijboer-Zernike theory, the generalized pupil function is developed into a series of Zernike terms. This gives rise to diffraction integrals per Zernike term that are expressed in [1] as doubly infinite series

$$I = I_n^m = \sum_{h,t} A_{2t,n,h}^{0mm} (-1)^{\frac{h-m}{2}} \, c_t \, \frac{J_{h+1}(2\pi r)}{2\pi r} \ . \tag{1}$$

In Eq. (1), m and n are the azimuthal order and degree of the involved Zernike term Z_n^m, the $c_t = c_t(OS, f)$ are the Zernike coefficients of the radially symmetric front factor composed of an algebraic factor comprising the parameters of the optical system and a factor comprising the defocus parameter f, the $J_{h+1}(2\pi r)/2\pi r$ are Jinc functions whose order h has the same parity as m with argument $2\pi r$ where r is the value of the radial parameter, and the A are to Clebsch-Gordan coefficients related numbers. In [1], Eq. (59), there occurs a slightly more general expression, in which the vectorial nature and polarization conditions are accounted for, leading to 5 series expressions involving an integer j, $|j| = 0, 1, 2$, of which Eq. (1) is the case $j = 0$. We shall not consider this generalization, since for truncation matters all these 5 cases behave the same. Furthermore, in the low-NA, small-defocus case, where a scalar treatment is allowed, the only required diffraction integral is the one with $j = 0$.

The A-coefficients in the double series in Eq. (1) have attractive properties with respect to their size and the set of h, t for which they are non-vanishing. The main effort in getting truncation rules goes therefore into bounding Jinc functions Jinc_h and structural quantities c_t. The Jinc functions are directly given in terms of Bessel functions while the structural quantities involve products of spherical Bessel and Hankel functions evaluated at $f/2$ and $f/2v_0$, respectively, where v_0, $0 < v_0 < 1$, is a quantity determined by the optical system. Now it is a fact that (spherical) Bessel functions, considered as a function of the order, are of constant magnitude as long as the order is less than the value of the argument. Beyond this point a super exponential decay as a function of order takes place. The situation for the structural quantities is somewhat complicated by the occurrence of the Hankel functions (causing decay to slow down to exponential for t beyond $|f|/2v_0$). These observations are basic to the approach taken in this paper and lead to the general rule-of-thumb that it suffices to include in Eq. (1) all terms h, t with $0 \leq h \leq H$, $0 \leq t \leq T$ in which H is slightly larger than $2\pi r$ and T is slightly larger than $|f|/2$. It is the aim of this paper to give a more precise meaning to this rule-of-thumb, in which the required absolute accuracy is included. Furthermore, by taking advantage of the (m,n)-dependent support properties of the A-coefficients, it is possible to formulate a truncation rule per Zernike term Z_n^m that achieves a particular accuracy with substantially less terms than when the general rule were used.

We shall do this in all detail for the diffraction integral $I = I_{VM}$ of [1], Sec. 8, which is meant for systems with high NA, vector fields and magnification. Explicitly, I assumes the form

$$I = I_{VM} = I_{n,VM}^m = \int_0^1 a(\rho)\, f(\rho)\, p(\rho)\, b(\rho)\, \rho\, d\rho \ , \tag{2}$$

where

$$a(\rho) = \frac{(1 - s_0^2\rho^2)^{1/2} + (1 - s_{0,M}^2\rho^2)^{1/2}}{(1 - s_0^2\rho^2)^{1/4}\,(1 - s_{0,M}^2\rho^2)^{3/4}} \ , \tag{3}$$

$$f(\rho) = \exp\left[\frac{if}{u_0}\,(1 - \sqrt{1 - s_0^2\rho^2})\right] \ , \tag{4}$$

$$p(\rho) = R_n^{|m|}(\rho) \ , \qquad b(\rho) = J_m(2\pi r\rho) \ , \tag{5}$$

are the algebraic, focal, polynomial and Bessel function factor, respectively. Here s_0 is the NA in image space, $s_{0,M}$ is built from the refractive indices in image and object space and the magnification factor in object space according to [1], Eq. (31), and $u_0 = 1 - \sqrt{1 - s_0^2}$.

The I_{VM}-case is with respect to truncation issues quite representative for all diffraction integrals considered in [1], except for the case of I_{VMML} in [1], Sec. 9, with backward propagating waves in a layer of the multilayer structure in image space. The I_{VM}-case is also general enough to illustrate the various intricacies that come with the computation of the Zernike coefficients c_t, the structural quantities, of the front factor $a(\rho)\, f(\rho)$, see [1], Sec. 4, requiring truncation rules as well.

In Sec. 2 we consider rules for the truncation of the double series in Eq. (1) for the I_{VM}-case for which we use bounds on the Jinc functions and on the structural quantities that follow from Debye's asymptotics for Bessel functions. In Sec. 3 we consider the truncation issues associated with the computation of the structural quantities. In Sec. 4 the whole computation scheme and the truncation rules are summarized. In Sec. 5 we illustrate the performance of the truncation rules by plotting actually achieved accuracy and computation times against required accuracy. In Sec. 6 we present our conclusions. In Appendix A we present basic properties of φ-functions that arise in bounding the (spherical) Bessel and Hankel functions using Debye's asymptotics. The results of Appendix A are used in Appendix B and C where we develop bounds on Jinc functions and structural quantities. In Appendix D we present some proofs concerning the validity of the truncation rules. In Appendix E we present a number of results containing the computation and asymptotics for the Zernike coefficients of the algebraic factors that occur in the I_{VM}-case.

2 Truncation rules for the double series for I_{VM}

2.1 Double series for I_{VM} and truncation strategy

We have

$$I_{VM} = \sum_{h,t} A^{0mm}_{2t,n,h} (-1)^{\frac{h-m}{2}} \, c_t \, \frac{J_{h+1}(2\pi r)}{2\pi r} \tag{6}$$

as in Eq. (1), where c_t are the Zernike coefficients of the front factor $a(\rho)\, f(\rho)$, with $a(\rho)$ and $f(\rho)$ as in Eqs. (3–4) so that

$$\frac{(1-s_0^2\rho^2)^{1/2} + (1-s_{0,M}^2\rho^2)^{1/2}}{(1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{3/4}} \exp\left[\frac{if}{u_0}\,(1-\sqrt{1-s_0^2\rho^2})\right]$$

$$= \sum_{t=0}^{\infty} c_t \, R^0_{2t}(\rho) \ . \tag{7}$$

Our approach to get truncation rules for the double series uses the following observations. The coefficients A are all non-negative and bounded by 1 and satisfy other boundedness properties such as

$$\sum_h A^{0mm}_{2t,n,h} = 1 = \sum_t \frac{2t+1}{h+1} \, A^{0mm}_{2t,n,h} \ . \tag{8}$$

In Subsec. 2.2 we give bounds on the Jinc functions $J_{h+1}(2\pi r)/2\pi r$ and the coefficients c_t that show rapid decay after $h = 2\pi r$ and $t = \frac{1}{2}\,|f|$, respectively. For values of absolute accuracy ε that are relevant in the optical practice, the double series in Eq. (6) is truncated at values $h = H$ and $t = T$ where both the Jinc functions and the coefficients have reached their plunge ranges. Accordingly, the absolute truncation error in approximating I_{VM} in Eq. (6) by

$$\sum_{h+1\le H,t\le T} A^{0mm}_{2t,n,h}(-1)^{\frac{h-m}{2}}\, c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \tag{9}$$

is safely bounded by

$$\max_{(h,2t)\in S^m_n\,;\,h+1>H\,\mathrm{or}\,t>T} \left| c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \right| , \tag{10}$$

where S^m_n is the set of all h, t such that $A^{0mm}_{2t,n,h} \neq 0$.

In the general truncation rule, the dependence on n and m of the supporting set S^m_n is totally ignored and the functions bounding Jinc_{h+1} and c_t are replaced by simple functions allowing convenient determination of set points H and T for which

$$\max_{h+1>H\,\mathrm{or}\,t>T} \left| c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \right| \tag{11}$$

is below a specified $\varepsilon > 0$.

In the dedicated rule, we use a more careful approximation of the bounding functions, and we include explicitly the supporting set S^m_n. It thus appears that an inspection of the product of the approximated bounding functions along the boundary $\partial\, S^m_n$ of the supporting set in the $(h, 2t)$-plane produces numbers $H = H^m_n$ and $T = T^m_n$ such that the quantity in Eq. (10) is below a specified $\varepsilon > 0$.

2.2 Bounding Jinc functions and structural quantities

We let for $c > 0$ and $x \ge 0$

$$\varphi(x\,;\,c) = \begin{cases} 0 & , \quad 0 \le x \le c \ , \\ x\,\mathrm{arccosh}(x/c) - c\,\sqrt{(x/c)^2-1} & , \quad x \ge c \quad , \end{cases} \tag{12}$$

where $\mathrm{arccosh}(y) = \ln(y + \sqrt{y^2-1})$. In Appendix B, the following is shown. Let $r > 0$, and set

$$R = \max\Big(\frac{1}{2\pi}\,, r\Big) \ . \tag{13}$$

Then

$$\left| \frac{J_{h+1}(2\pi r)}{2\pi r} \right| \le \frac{1}{2\pi^2\, R\sqrt{R}}\, \exp(-\varphi(h+1\,;\,2\pi R)) \ . \tag{14}$$

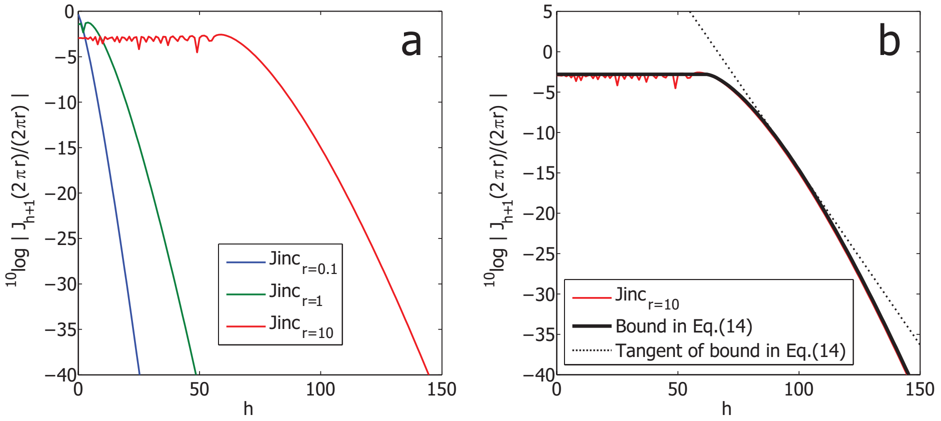

Figure 1: (a) Plot of $\log_{10}|J_{h+1}(2\pi r)/2\pi r|$ as a function of $h = 0,\ 1, \cdots,\ 150$ for the case $r = 0.1$ (blue), 1 (green), 10 (red). (b) Plot of $\log_{10}|J_{h+1}(2\pi r)/2\pi r|$ as a function of $h = 0,\ 1, \cdots,\ 150$ case $r = 10$ (red), together with the $\log_{10}$ of the bound at the right-hand side of Eq. (14) (solid black) and the tangent line (dashed) corresponding to the right-hand side of Eq. (20).

The bound in Eq. (14) is valid for all $h \geq 0$, except for a small range of h's near $2\pi r$ with $r \to \infty$. In fact, Eq. (14) is valid for all $r \geq 0$ and $h \leq 2$, it is valid within a factor of 2 for all $r \geq 0$ and all $h \leq 175$, it is valid within a factor of 4 for all $r \geq 0$ and all $h \leq 11194$, and so on. Of course, we also have the general bound $|J_{h+1}(2\pi r)/2\pi r| \leq \frac{1}{2}$.

In Figure 1a, we show $\log_{10}|J_{h+1}(2\pi r)/2\pi r|$ as a function of h, $0 \leq h \leq 150$, for $r = 0.1,\ 1$ and 10, respectively. It can be seen that there is rapid decay from $h + 1 = 2\pi r = 0.63,\ 6.28$ and 62.83, respectively onwards. For the case that $r = R = 10$, we have plotted in Figure 1b both $\log_{10}|J_{h+1}(2\pi r)/2\pi r|$ and the bound $\log_{10}[\exp\{-\varphi(h+1; 2\pi R)\}/2\pi^2 R\sqrt{R}\,]$, see Eq. (14). The (asymptotic) maximum of $\log_{10}|J_{h+1}(2\pi r)/2\pi r|$ can be found from Appendix B and equals -2.5609, assumed at $h = 58.67$ when $r = 10$. At this point h, the upper bound $\log_{10}[1/2\pi^2 R\sqrt{R}] = -2.7953$ is slightly lower than the asymptotic maximum. We have also shown in Fig. 1b the linear function $\log_{10}[\exp\{-(h + 1 - 2\pi R \sinh(1))\}/(2\pi^2 R\sqrt{R})\,] = 28.8387 - 0.4343h$ which is a tangent line of the bounding function, see Subsec. 2.3.

For the structural quantities c_t a similar result holds. In Appendix C the following is shown. let f be a real number, and set

$$g = \max(1, |f|)\ . \tag{15}$$

Then

$$|c_t| \leq 4w_0\, a_0\, \exp(-\varphi(t\,;\, g/2) + \varphi(t\,;\, g/2v_0))\ , \tag{16}$$

where

$$a_0 = 2\int\limits_0^1 a(\rho)\, \sqrt{1 - s_0^2\rho^2}\,\rho\, d\rho \tag{17}$$

is the R_0^0-coefficient of $A(\rho) = a(\rho)\, \sqrt{1 - s_0^2\rho^2}$, and

$$w_0 = \frac{1}{1 + \sqrt{1 - s_0^2}}\ , \qquad v_0 = \frac{1 - \sqrt{1 - s_0^2}}{1 + \sqrt{1 - s_0^2}}\ . \tag{18}$$

Here it has been assumed that $s_0 \geq s_{0,M}$. In the case that $s_{0,M} > s_0$, we should replace s_0 in Eqs. (17-18) by $s_{0,M}$ and change the right-hand side of Eq. (16) accordingly. The value of a_0 is in almost all cases well approximated by

$$A(\tfrac{1}{2}\,\sqrt{2}) \quad \text{or} \quad \tfrac{1}{6}\,A(0) + \tfrac{2}{3}\,A(\tfrac{1}{2}\,\sqrt{2}) + \tfrac{1}{6}\,A(1) \tag{19}$$

(midpoint rule or Simpson rule for integration over $x = \rho^2$). The bound in Eq. (16) is shown in Appendix C using a somewhat heuristic approach so as to arrive at manageable expressions. As with the bound in Eq. (14) there are small exceptional ranges of t near $\frac{1}{2}\,g$ and $g \to \infty$, where Eq. (16) holds safe for a factor that grows to infinity very slowly as $g \to \infty$.

In Figure 2a, we show $|c_t|$ as a function of t, $0 \leq t \leq 150$, for $f = 150$, $s_0 = 0.95$ and $s_{0,M} = 0.50$, with $j = 0,\ 1,\ 2$ determining the precise form of the algebraic function in the vectorial setting according to [1], Eq. (30). It can be seen that the graphs for these three cases are qualitatively the same, except for an overall amplitude factor that is related to the R_0^0-coefficient a_0 of $a(\rho)\sqrt{1 - s_0^2\rho^2}$. There is rapid decay from $t = \frac{1}{2}f = 75$ onwards. For the case $j = 0$, we have plotted in Figure 2b both $\log_{10}|c_t|$ and the bound $\log_{10}[4w_0a_0 \exp\,(-\varphi(t; g/2) + \varphi(t; g/2v_0))]$, see Eq. (16). The (asymptotic) maximum of $\log_{10}|c_t|$ occurs somewhat before $t = 75$ and exceeds the value $\log_{10}[4w_0a_0]$ obtained from the bounding function somewhat. We also show in Figure 2b the linear function $\log_{10}[4w_0a_0 \exp\,(\frac{1}{2}g \sinh(\gamma_0) - \gamma_0 t)] = 23.1718 - 0.2806t$, where $\gamma_0 = \ln\,(1/v_0) = 0.6461$, which is a tangent line of the bounding function, see Subsec. 2.4.

In Figure 3, we show the graph of v_0, as given in Eq. (18), against s_0, $0 \leq s_0 \leq 1$. The asymptotic decay of c_t is Cv_0^t, and so there is rapid decay of c_t for all s_0 until $s_0 = 0.95$ (with $v_0 = 0.5241$), and even cases like $s_0 = 0.99$ are still practicable.

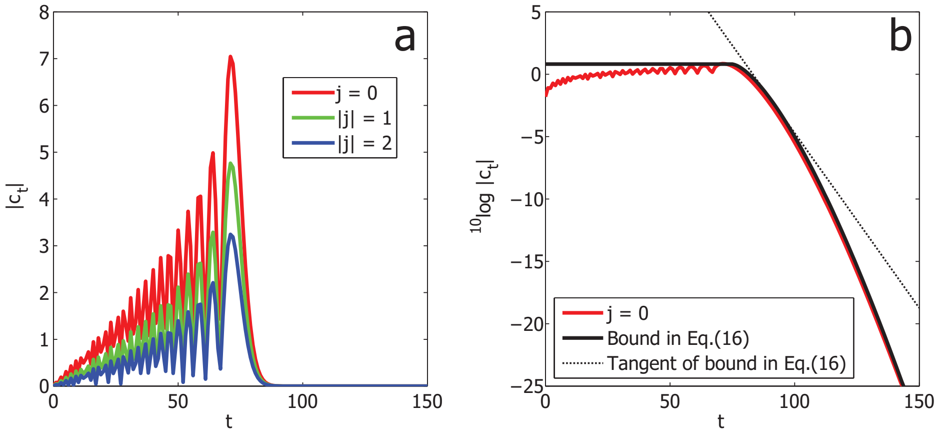

Figure 2: (a) Plot of $\log_{10}|c_t|$ as a function of $t = 0,\ 1, \cdots,\ 150$, for $f = 150$, $s_0 = 0.95$, $s_{0,M} = 0.50$, where c_t are the Zernike coefficients of the front factors that occur in accordance with [1], Eq. (30) for $|j| = 0$ (red), 1 (green), 2 (blue) and of which c_t in Eq. (7) gives the case $|j| = 0$. (b) Plot of $\log_{10}|c_t|$ as in (a) for the case $|j| = 0$ (red), together with the $\log_{10}$ of the bound at the right-hand side of Eq. (16) (solid black) and the tangent line (dashed) corresponding to the right-hand side of Eq.(21).

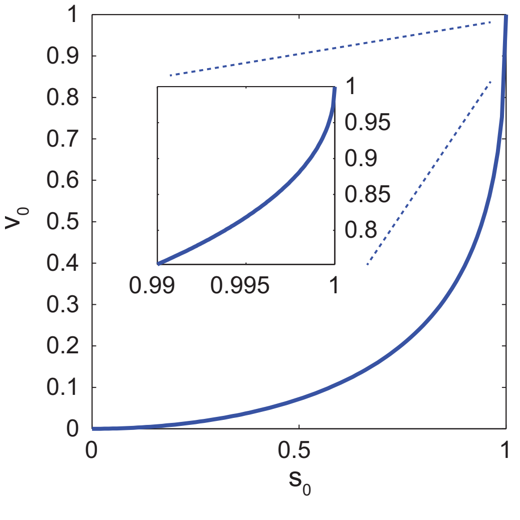

Figure 3: Graph of $v_0 = (1 - \sqrt{1 - s_0^2})/(1 + \sqrt{1 - s_0^2})$ as a function of s_0, $0 \leq s_0 \leq 1$.

2.3 General truncation rule

In Appendix A the functions $\varphi(h+1\,;\,2\pi R)$ and $\varphi(t\,;\,g/2)-\varphi(t\,;\,g/2v_0)$ are bounded from below by piecewise linear functions according to

$$\varphi(h+1\,;\,2\pi R) \geq \max(0, h+1-2\pi R\sinh(1)) \;, \tag{20}$$

and

$$\varphi(t\,;\,g/2)-\varphi(t\,;\,g/2v_0) \geq \max(0, \gamma t - \tfrac{1}{2}\,g\sinh(\gamma)) \;, \tag{21}$$

where

$$\gamma = \min(1, \ln(1/v_0)) \;, \tag{22}$$

respectively. This leads to the following general truncation rule: Let $0<\varepsilon<1$, and let

$$B = \max\Big(0, \ln\big(\frac{2w_0a_0}{\pi^2\,\varepsilon\,R\sqrt{R}}\big)\Big) \;. \tag{23}$$

Then the quantity in Eq. (11) is less than ε when

$$T = T^{\mathrm{gen}} = \frac{1}{\gamma}\,B + \tfrac{1}{2}\,g\,\frac{\sinh(\gamma)}{\gamma} \;, \qquad H = H^{\mathrm{gen}} = B + 2\pi R\sinh(1) \;. \tag{24}$$

See Appendix D for a proof.

By observing that we can write T and H in Eq. (24) as

$$T = \tfrac{1}{2}\,g + \frac{1}{\gamma}\,B + \tfrac{1}{2}\,g\,\frac{\sinh(\gamma)-\gamma}{\gamma} \;, \qquad H = 2\pi R + B + 2\pi R(\sinh(1)-1) \;, \tag{25}$$

where for $0<\gamma\leq 1$

$$0 < \frac{\sinh(\gamma)-\gamma}{\gamma} \leq \sinh(1)-1 = 0.1752 \;, \tag{26}$$

we have given precision to the rule-of-thumb that the truncation points should be chosen somewhat larger than $\frac{1}{2}|f|$ and $2\pi r$, respectively.

2.4 Dedicated truncation rule

We now present a truncation rule that takes into account the (n,m)-dependence of the supporting set S_n^m of the A's in Eq. (6). We also use better approximations for the functions $\varphi(h+1\,;\,2\pi R)$ and $\varphi(t\,;\,g/2)-\varphi(t\,;\,g/2v_0)$ on the left-hand sides of Eqs. (20–21). Thus we consider

$$F(h,t) = \varphi(h+1\,;\,2\pi R) + \varphi(t\,;\,g/2, g/2v_0) \;, \tag{27}$$

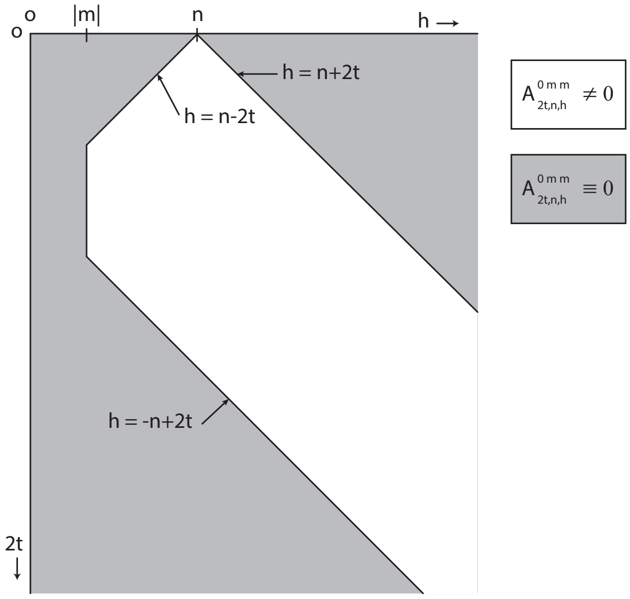

Figure 4: For given integers n and m with $n - |m|$ even and non-negative, the unshaded set $h \geq |m|$, $|h - n| \leq 2t \leq h + n$ contains all points $(h, 2t)$ with non-negative integer h and t such that $A^{0mm}_{2t,n,h} \neq 0$.

where

$$\varphi(t\,;\,g/2, g/2v_0) = \begin{cases} \varphi(t\,;\,g/2) & , \quad 0 \leq t \leq \frac{1}{2}\,g\cosh(\gamma_0)\ , \\ \gamma_0 t - \frac{1}{2}\,g\sinh(\gamma_0)\ , & \quad t \geq \frac{1}{2}\,g\cosh(\gamma_0) \qquad , \end{cases} \tag{28}$$

with $\gamma_0 = \ln(1/v_0)$. The function $\varphi(t\,;\,g/2, g/2v_0)$ is the largest convex function bounding $\varphi(t\,;\,g/2) - \varphi(t\,;\,g/2v_0)$, which is convex in $t \leq g/2$ but concave in $t \geq g/2v_0$, from below. The function $\varphi(h+1\,;\,2\pi R)$ is convex in $h \geq 0$. See Appendix A.

In Figure 4 we depict, for given n and m such that $n - |m|$ is even an non-negative, the set S^m_n in the $(h, 2t)$-plane that contains all non-zero coefficients $A^{0mm}_{2t,n,h}$ (S^m_n is the convex hull of those points $(h, 2t)$). The boundary $\partial\, S^m_n$ of S^m_n consists of 4 line segments I, II, III, IV in accordance with the conditions, see [1], Sec. 5,

$$h \geq |m|\ , \qquad |h - n| \leq 2t \leq h + n\ . \tag{29}$$

We consider the function $F(h, t)$ of Eq. (27) along $\partial\, S^m_n$ with continuous $t \geq 0$, $h \geq 0$. We have that $F(h, t)$ is non-negative and increasing and convex in both h and t, and

$$\left| c_t \,\frac{J_{h+1}(2\pi r)}{2\pi r} \right| \leq \frac{2 w_0 a_0}{\pi^2 \, R \sqrt{R}} \, \exp(-F(h,t)) \ . \tag{30}$$

We let B as in Subsec. 2.3, and we let

$$M = \min \{ F(h,t) \,|\, (h,2t) \in \partial \, S_n^m , \ h+1 \leq H^{\rm gen} , \ t \leq T^{\rm gen} \} \tag{31}$$

with $H^{\rm gen}$ and $T^{\rm gen}$ from Subsec 2.3. From the monotonicity and convexity properties of F, we then get, see Appendix D,

- when $M > B$, we have that

$$\max_{(h,2t) \in S_n^m} \left| c_t \,\frac{J_{h+1}(2\pi r)}{2\pi r} \right| < \varepsilon \ , \tag{32}$$

- when $M \leq B$, there are two points $(h_1, 2t_1)$ and $(h_2, 2t_2) \in \partial \, S_n^m$ such that for any $(h,2t) \in S_n^m$

$$h \geq \max(h_1, h_2) \ \text{ or } \ t \geq \max(t_1, t_2) \Rightarrow F(h,t) \geq B \ . \tag{33}$$

The dedicated truncation rule becomes then as follows. Determine M in Eq. (31). When $M > B$, we set $H = H_n^m = 1$, $T = T_n^m = 0$. When $M \leq B$, we search the boundary $\partial \, S_n^m$, as long as contained in the box $h+1 \leq H^{\rm gen}$ & $t \leq T^{\rm gen}$, for the two points $(h_1, 2t_1)$ and $(h_2, 2t_2)$ satisfying Eq. (33), and we set $H = H_n^m = \max(h_1, h_2) - 1$, $T = T_n^m = \max(t_1, t_2)$. With H and T defined this way, we have that the quantity in Eq. (10) is less than ε.

By the monotonicity and convexity properties of F, the minimum M of F along $\partial \, S_n^m$ is assumed on edge II. Hence, it is sufficient to inspect F along this edge to find M.

The actual variables h, t are non-negative integer, and this should be accounted for. We intersect $\partial \, S_n^m$ with the box $(h, 2t)$, $h \leq \hat{H} - 1$ or $t \leq \hat{T}$, where $\hat{H} - 1$ is the smallest integer of same parity as n with $\hat{H} \geq H^{\rm gen}$ and $\hat{T}$ is the smallest integer with $\hat{T} \geq T^{\rm gen}$. In case that we find 0 or 1 point $(h, 2t)$ in the intersection, the inspection is a trivial matter. In the case that we find two intersection points, we let the inspection start at the point with largest value of h and lowest values of $2t$, and we end the inspection at or before the point with lowest value of h and largest value of $2t$, following the boundary curve counterclockwise with points $(h, 2t)$, integer h and t and h same parity as n.

3 Computation of structural quantities and truncation issues

3.1 Series expressions for structural quantities

We consider in this section computation of the Zernike coefficients of the front factor $a(\rho)\,f(\rho)$, with $a(\rho)$ and $f(\rho)$ given in Eqs. (3–4). We make a slight variation of the approach in [1], Sec. 4 and 8, in that we write

$$a(\rho)\,\sqrt{1-s_0^2\rho^2}=\sum_{l=0}^{\infty} a_l\,R_{2l}^0(\rho)\ , \tag{34}$$

$$f(\rho)/\sqrt{1-s_0^2\rho^2}=\sum_{k=0}^{\infty} b_k\,R_{2k}^0(\rho)\ , \tag{35}$$

and we use linearization coefficients A to write

$$a(\rho)\,f(\rho)=\sum_{t=0}^{\infty} c_t\,R_{2t}^0(\rho)\ , \tag{36}$$

where

$$c_t=\sum_{l,k=0}^{\infty} A_{2l,2k,2t}^{000}\,a_l\,b_k\ . \tag{37}$$

The reason for moving a factor $\sqrt{1-s_0^2\rho^2}$ from the focal factor $f(\rho)$ to the algebraic factor $a(\rho)$ is the fact that this yields the most convenient expression for the expansion coefficients b_k, viz.

$$b_k=\frac{1}{iu_0}\,\exp\,[if/u_0]\,(2k+1)\,j_k(f/2)\,h_k^{(2)}(f/2v_0)\ . \tag{38}$$

Here j_k and $h_k^{(2)}$ are the spherical Bessel and Hankel functions of order k, given as

$$j_k(z) \;=\; \sqrt{\frac{\pi}{2z}}\,J_{k+1/2}(z)\ , \tag{39}$$

$$\begin{aligned} h_k(z) &= j_k(z)-i\,y_k(z)\\ &= \sqrt{\frac{\pi}{2z}}\,(J_{k+1/2}(z)-i\,Y_{k+1/2}(z))\\ &= \sqrt{\frac{\pi}{2z}}\,H_{k+1/2}^{(2)}(z)\ , \end{aligned} \tag{40}$$

with J_ν, Y_ν and $H_\nu^{(2)}$ the Bessel function of first, second and third kind (Hankel function) and of order ν, see [2], Ch. 10. The quantities b_k can be computed, via Eqs. (39–40) using MatLab routines, efficiently at any desired accuracy.

As to the coefficients a_l, we first write, see Eq. (3),

$$\begin{aligned} a(\rho)\,\sqrt{1-s_0^2\rho^2} &= (1-s_0^2\rho^2)^{3/4}\,(1-s_{0,M}^2\rho^2)^{-3/4} \\ &\quad + (1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{-1/4}\ . \end{aligned} \tag{41}$$

Next, either term on the right-hand side of Eq. (41) is developed into a power series

$$a_{\alpha\beta}(\rho) = (1-s_\alpha^2\rho^2)^\alpha\,(1-s_\beta^2\rho^2)^\beta = \sum_{N=0}^{\infty} r_N\rho^{2N}\ , \tag{42}$$

where the coefficients r_N are computed recursively according to [1], Eqs. (37–39) and [1], Eq. (106). Finally, the Zernike coefficients $a_{l,\alpha\beta}$ are computed from r_N according to

$$a_{l,\alpha\beta} = \sum_{N=l}^{\infty} b_N(l)\,r_N\ , \qquad l = 0, 1, \ldots\ , \tag{43}$$

with $b_N(l)$ given explicitly and computed recursively in [1], Eqs. (41–44).

3.2 Truncation and accuracy issues

The accuracy by which the c_t must be computed is dictated by the absolute accuracy ε in the truncation analysis of Sec. 2 that involves the products of c_t's and Jinc functions $J_{h+1}(2\pi r)/2\pi r$ as in Eqs. (10–11). Now $|J_{h+1}(z)/z| \le 1/2$ for $z \ge 0$. Hence, when c_t is computed with absolute accuracy ε, and the truncation rules of Subsecs. 2.3–2.4 are used with $\varepsilon/2$ instead of ε, a final absolute accuracy better than ε results.

Next, given integers $L, K > 0$, the absolute error due to approximating c_t of Eq. (37) by

$$c_{t,LK} = \sum_{l=0}^{L}\sum_{k=0}^{K} A_{2l,2k,2t}^{000}\,a_l\,b_k \tag{44}$$

is, as in Eqs. (9–10), safely bounded by

$$\max_{l>L\,\mathrm{or}\,k>K} |a_l b_k|\ . \tag{45}$$

Now there are the bounds

$$|a_l| \le \tfrac{16}{3}\,, \quad |b_k| \le 4\;, \qquad l,k = 0,1,\ldots\;. \tag{46}$$

The second bound in Eq. (46) follows from Appendix C, Eq. (C18), while the first bound is obtained by considering in Appendix E, Eq. (E1) the worst case $l = 0$ with $s_0 = 0$ and $s_{0,M}$ close to 1. Hence, when $\varepsilon \in (0,1)$, we have that the quantity in Eq. (45) is less than ε when L and K are such that

$$l > L \Rightarrow |a_l| < \tfrac{1}{4}\,\varepsilon \quad \& \quad k > K \Rightarrow |b_k| < \tfrac{3}{16}\,\varepsilon\;. \tag{47}$$

According to Appendix C we have

$$|b_k| \le 4\,\exp(-\varphi(k\,;\,g/2) + \varphi(k\,;\,g/2v_0))\;, \tag{48}$$

and this is less than $\frac{3}{16}\,\varepsilon$ when

$$k > \frac{1}{\gamma}\,\max\Bigl(0, \ln\Bigl(\frac{64}{3\varepsilon}\Bigr)\Bigr) + \tfrac{1}{2}\,g\,\frac{\sinh(\gamma)}{\gamma}\;, \tag{49}$$

with γ as in Eq. (22).

The quantities b_k are computed using Eq. (38), involving the spherical Bessel and Hankel functions j_k and $h_k^{(2)}$ that can be computed using Matlab routines. From Appendix C we have that

$$|j_k(f/2)| \le \tfrac{2}{g}\;, \quad |h_k(f/2v_0)| \le \frac{2^{7/4}v_0}{g}\,\exp(\varphi(k; g/2v_0))\,, \tag{50}$$

where the first inequality holds for all f and the second inequality only holds when $|f/v_0| \ge 1$. In the case that $|f/v_0| < 1$, the b_k of Eq. (38) is best evaluated using the power series representations of j_k and $h_k^{92)}$ that follow from [2], 10.53. Thus it follows that b_k is computed with absolute accuracy $3\varepsilon/16$ for $k = 0,\ 1,\ \cdots,\ K$ when $j_k(f/2)$ and $h_k^{(2)}(f/2v_0)$ are computed with absolute accuracy

$$\frac{3\varepsilon}{32}\cdot\frac{u_0\exp(-\varphi(K; g/2v_0))}{2^{7/4}(2K+1)v_0} \quad \text{and} \quad \frac{3\varepsilon}{32}\cdot\frac{u_0}{2(2K+1)} \tag{51}$$

respectively.

As to the first condition in Eq. (47), we consider the decomposition of $a(\rho)\,\sqrt{1 - s_0^2\rho^2}$ in terms $a_{\alpha\beta}(\rho)$ as in Eq. (42) with $\alpha + \beta = 0$ and Zernike coefficients $a_{l,\alpha\beta}$ as in Eq. (43). In Appendix E the following is shown. Let

$\delta = |\alpha| = |\beta|$, and let $S = \max(s_\alpha, s_\beta)$. Denoting "the R^0_{2l}-coefficient of $A(\rho)$" by $Z\,C_l[A(\rho)]$, we have

$$|a_{l,\alpha\beta}| \le Z\,C_l[(1 - S^2\rho^2)^{-\delta}] \sim \frac{E\,V^l}{(l+1)^{-\delta+1/2}}\ , \tag{52}$$

where

$$E = \frac{2\sqrt{\pi}}{\Gamma(\delta)}\ \frac{(1-S^2)^{-\frac{1}{2}\delta+\frac{1}{4}}}{1+\sqrt{1-S^2}}\ , \qquad V = \frac{1-\sqrt{1-S^2}}{1+\sqrt{1-S^2}}\ . \tag{53}$$

Furthermore, the right-hand side of Eq. (52) is less than $\eta := \varepsilon/8$ when

$$l \ge \frac{\ln(E\eta^{-1}) - (-\delta+1/2)\ln(1+\ln(E\eta^{-1})/\ln(1/V))}{\ln(1/V)}\ . \tag{54}$$

Therefore, the first condition in Eq. (47) is satisfied when L is the maximum of the two numbers that occur at the right-hand side of Eq. (54) for the choices $\delta = 3/4, 1/4$ (where evidently $\delta = 3/4$ yields the largest value of the two).

We finally address the issue of truncating the series in Eq. (43). It is shown in Appendix E that for a given $\varepsilon > 0$ and an integer $L > 0$ such that all $|a_{l,\alpha\beta}| < \frac{1}{8}\,\varepsilon$ when $l > L$, we have that all numbers $a_{l,\alpha\beta}$, $l = 0, 1, ..., L$, are computed with absolute accuracy $\varepsilon/16$ when the infinite series in Eq. (43) is truncated at $N = 2L/\sqrt{1-S^2}$.

In Figure 5, we show $\log_{10}|a_{0,\alpha\beta} - \sum_{N=0}^{N_L(\eta)} b_N(0) r_N|$ as a function of η with $\log_{10}\eta^{-1} \in [0, 15]$, for the case that $a_{0,\alpha\beta}$ is the R^0_0-coefficient of $a_{\alpha\beta}(\rho) = (1 - s_0^2\rho^2)^\alpha(1 - s_{0,M}^2\rho^2)^\beta$ with $\alpha = -\beta = 3/4$ and $s_0 = 0.50$, $s_{0,M} = 0.90$ and upper summation limit $N_L(\eta) = L(\eta),\ 2L(\eta),\ 4L(\eta),\ 5L(\eta)$, respectively, with $L(\eta)$ the right-hand side of Eq. (54).

To summarize, for $\varepsilon \in (0, 1)$ we replace c_t by $c_{t,LK}$ given in Eq. (44) in which

- L and K are given by the right-hand sides of Eq. (54) and Eq. (47), respectively,

- b_k is as in Eq. (38) for which $j_k(f/2)$ and $h_k^{(2)}(f/2v_0)$ are computed with absolute accuracy as specified in Eq. (51),

- $a_l = a_{3/4,-3/4,l} + a_{1/4,-1/4,l}$ and the two $a_{\alpha,\beta,l}$ are computed by summing the series in Eq. (43) until $N = 2L/\sqrt{1-S^2}$ with $S = \max(s_0, s_{0,M})$.

This results into an absolute error in c_t bounded by $\varepsilon + \frac{1}{2}\varepsilon + \frac{1}{4}\varepsilon = \frac{7}{4}\varepsilon$, due to respectively, truncating the double series over l and k, approximating b_k by computing j_k and $h_k^{(2)}$ using the Matlab-code, and approximating a_l by truncating the series for the two $a_{\alpha,\beta,l}$.

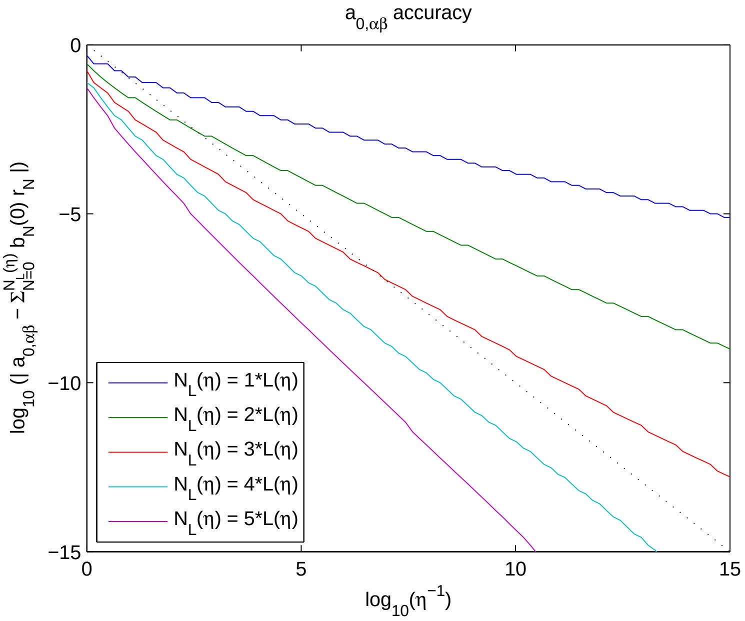

Figure 5: Plot of $\log_{10}|a_{0,\alpha\beta} - \sum_{N=0}^{N_L(\eta)} b_N(0)r_N|$ as a function of $\log_{10}\eta^{-1} \in [0,15]$, for the case that $a_{0,\alpha\beta}$ is the R_0^0-coefficient of $a_{\alpha\beta}(\rho) = (1 - s_0^2\rho^2)^\alpha(1 - s_{0,M}^2\rho^2)^\beta$ with $\alpha = -\beta = 3/4$ and $s_0 = 0.50$, $s_{0,M} = 0.90$. The colored solid lines represent different summation limits $N_L(\eta) = L(\eta)$, $2L(\eta)$, $4L(\eta)$, $5L(\eta)$, respectively, with $L(\eta)$ given by the right-hand side of Eq. (54). The black (dotted) curve indicates those positions at which the observed accuracy of $a_{0,\alpha\beta}$ is equal to η.

4 Summary of the computation scheme and truncation rules

For integer n and m such that $n-|m|$ is even and non-negative, consider

$$I = I^m_{n,VM} = \int\limits_0^1 a(\rho)\,f(\rho)\,p(\rho)\,b(\rho)\,\rho\,d\rho\ , \tag{55}$$

where

$$a(\rho) = \frac{(1-s_0^2\rho^2)^{1/2}+(1-s_{0,M}^2\rho^2)^{1/2}}{(1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{3/4}}\ , \tag{56}$$

$$f(\rho) = \exp\Big[\frac{if}{u_0}\,(1-\sqrt{1-s_0^2\rho^2})\Big]\ , \tag{57}$$

$$p(\rho) = R_n^{|m|}(\rho)\ , \qquad b(\rho) = J_m(2\pi r\rho) \tag{58}$$

with given real f, $r>0$ and $s_0, s_{0,M}\in[0,1)$, and where $u_0 = 1-\sqrt{1-s_0^2}$. There is the double series representation

$$I = \sum_{h,t} A^{0mm}_{2t,n,h}(-1)^{\frac{h-m}{2}}\,c_t\,\frac{J_{h+1}(2\pi r)}{2\pi r} \tag{59}$$

with summation over h, $t = 0,1,...$ and h same parity as n and m. In Eq. (59), we have

$$A^{0mm}_{2t,n,h} = (h+1)\left|\begin{pmatrix} t & \frac{1}{2}n & \frac{1}{2}h \\ 0 & \frac{1}{2}m & -\frac{1}{2}m \end{pmatrix}\right|^2 \tag{60}$$

in terms of the Clebsch-Gordan coefficients in $|\ |^2$ of [2], Chap. 34; the A's are considered in detail in [1], Sec. 5 and Appendix C. Furthermore, the c_t are the Zernike coefficients of the front factor $a(\rho)\,f(\rho)$, so that

$$a(\rho)\,f(\rho) = \sum_{t=0}^{\infty} c_t\,R^0_{2t}(\rho)\ . \tag{61}$$

The c_t have a double series representation

$$c_t = \sum_{l,k=0}^{\infty} A^{000}_{2l,2k,2t}\,a_l\,b_k\ , \tag{62}$$

where the a_l are the Zernike coefficients of $A(\rho) = a(\rho)\sqrt{1-s_0^2\rho^2}$, so that

$$A(\rho) = a(\rho)\sqrt{1-s_0^2\rho^2} = \sum_{l=0}^{\infty} a_l\, R_{2l}^0(\rho)\ , \tag{63}$$

the b_k are the Zernike coefficients of $f(\rho)/\sqrt{1-s_0^2\rho^2}$, so that

$$f(\rho)/\sqrt{1-s_0^2\rho^2} = \sum_{k=0}^{\infty} b_k\, R_{2k}^0(\rho)\ , \tag{64}$$

and the $A_{2l,2k,2t}^{000}$ are related to Clebsch-Gordan coefficients as in Eq. (60). The b_k are given as

$$b_k = \frac{1}{iu_0}\exp\left[if/u_0\right](2k+1)\, f\, j_k(f/2)\, h_k^{(2)}(f/2v_0)\ , \tag{65}$$

with j_k and $h_k^{(2)}$ spherical Bessel and Hankel functions, see [2], Chap. 10, Sec. 10.4.7 and

$$v_0 = \frac{1-\sqrt{1-s_0^2}}{1+\sqrt{1-s_0^2}}\ . \tag{66}$$

The a_l are computed by first writing

$$\begin{aligned} a(\rho)(1-s_0^2\rho^2)^{1/2} \ &= \ (1-s_0^2\rho^2)^{3/4}\,(1-s_{0,M}^2\rho^2)^{-3/4} \\ &\quad + (1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{-1/4}\ , \end{aligned} \tag{67}$$

and then expanding both terms $a_{\alpha\beta}(\rho) = (1-s_\alpha^2\rho^2)^\alpha(1-s_\beta^2\rho^2)^\beta$ at the right-hand side of Eq. (67) into a power series and subsequently into a Zernike series according to

$$a_{\alpha\beta}(\rho) = (1-s_\alpha^2\rho^2)^\alpha\,(1-s_\beta^2\rho^2)^\beta = \sum_{N=0}^{\infty} r_{N,\alpha\beta}\,\rho^{2N} = \sum_{l=0}^{\infty} a_{l,\alpha\beta}\, R_{2l}^0(\rho)\ . \tag{68}$$

The $r_{N,\alpha\beta}$ in Eq. (68) are computed recursively according to

$$\begin{aligned} r_{-1}=0\,,\quad r_0=1\ ;\qquad r_{N+1} \ &= \ \frac{1}{N+1}\left[((N-\alpha)\,s_\alpha^2 + (N-\beta)\,s_\beta^2)\,r_N\right. \\ &\quad \left. -\,(N-1-\alpha-\beta)\,s_\alpha^2\,s_\beta^2\,r_{N-1}\right] \end{aligned} \tag{69}$$

for $N = 0, 1, \ldots$. The $a_{l,\alpha\beta}$ are computed from the $r_{N,\alpha\beta}$ according to

$$a_{l,\alpha\beta} = \sum_{N=l}^{\infty} b_N(l)\, r_{N,\alpha\beta}\ , \qquad l = 0, 1, \ldots\ , \tag{70}$$

where the $b_N(l)$ are given by

$$b_N(l) = \frac{2l+1}{l+1} \binom{N}{l} \Big/ \binom{N+l+1}{N} \ , \tag{71}$$

and can be computed recursively according to [1], Eqs. (42–44).

4.1 Truncating the double series for I

We consider replacing the double series for I in Eq. (59) by

$$\sum_{h+1 \le H,\, t \le T} A_{2t,n,h}^{0mm} (-1)^{\frac{h-m}{2}} \, c_t \, \frac{J_{h+1}(2\pi r)}{2\pi r} \ , \tag{72}$$

where H and T are to be chosen such that the absolute approximation error is less than $\varepsilon \in (0,1)$. Let $R = \max(1/2\pi, r)$, and let $g = \max(1, |f|)$. Furthermore, let

$$B = \max\Big(0, \ln\Big(\frac{2w_0 a_0}{\pi^2 \, \varepsilon \, R \sqrt{R}}\Big)\Big) \ , \tag{73}$$

where $w_0 = (1 + \sqrt{1 - s_0^2})^{-1}$ and a_0 is the R_0^0-coefficient in Eq. (63) so that

$$a_0 = 2 \int_0^1 a(\rho) \sqrt{1 - s_0^2 \rho^2} \, \rho \, d\rho \ . \tag{74}$$

In Eq. (73) and in the definitions of v_0 in Eq. (66) and of w_0 above, we need to replace s_0 by $s_{0,M}$ when $s_{0,M} > s_0$.

4.1.1 General truncation rule

The absolute approximation error is less than ε, simultaneously for all n and m, when

$$H = H^{\text{gen}} = B + 2\pi \, R \sinh(1) \ , \qquad T = T^{\text{gen}} = \frac{1}{\gamma} \, B + \tfrac{1}{2} \, g \, \frac{\sinh(\gamma)}{\gamma} \ , \tag{75}$$

where $\gamma = \min(1, \ln(1/v_0))$.

4.1.2 Dedicated truncation rule

For $c > 0$ and $x \geq 0$, define

$$\varphi(x\,;\,c) = \begin{cases} 0 & , \quad 0 \leq x \leq c \ , \\ x\,\mathrm{arccosh}(x/c) - c\,\sqrt{(x/c)^2 - 1} & , \quad x \geq c \quad , \end{cases} \tag{76}$$

and let for $h \geq 0$ and $t \geq 0$

$$F(h,t) = \varphi(h+1\,;\,2\pi R) + \varphi(t\,;\,g/2, g/2v_0) \ , \tag{77}$$

where for $t \geq 0$

$$\varphi(t\,;\,g/2, g/2v_0) = \begin{cases} \varphi(t\,;\,g/2) & , \quad 0 \leq t \leq \frac{1}{2}\,g\cosh(\gamma_0) \ , \\ \gamma_0 t - \frac{1}{2}\,g\sinh(\gamma_0) & , \quad t \geq \frac{1}{2}\,g\cosh(\gamma_0) \quad , \end{cases} \tag{78}$$

with $\gamma_0 = \ln(1/v_0)$ and v_0 given in Eq. (66).

Let n and m be integers such that $n - |m|$ is even and non-negative. The set S_n^m in the $(h, 2t)$-plane containing all non-zero coefficients $A_{2t,n,h}^{0mm}$ in the double series in Eq. (59) is given by the constraints

$$h \geq |m| \ , \qquad |h-n| \leq 2t \leq h+n \ , \qquad h-n \text{ even} \ . \tag{79}$$

The convex hull of this set S_n^m has a boundary $\partial\,S_n^m$ which is a curve consisting of 4, possibly degenerate, line segments, listed in counterclockwise order as

I. $h = n + 2t$, $t \geq 0$,

II. $h = n - 2t$, $0 \leq t \leq \frac{1}{2}\,(n - |m|)$,

III. $h = |m|$, $\frac{1}{2}\,(n-|m|) \leq t \leq \frac{1}{2}\,(n+|m|)$,

IV. $h = -n + 2t$, $t \geq \frac{1}{2}\,(n + |m|)$.

Let

$$M = \min\left\{F(h,t)\,|\,(h,2t) \in \partial\,S_n^m,\ 0 \leq h \leq H^{\mathrm{gen}},\ 0 \leq t \leq T^{\mathrm{gen}}\right\} \ , \tag{80}$$

with H^{gen} and T^{gen} as in Eq. (75).

The absolute approximation error is less than ε when $H = H_n^m$ and $T = T_n^m$ in Eq. (72) are chosen as follows.

Case $M > B$. Set

$$H = H_n^m = 1 \ , \qquad T = T_n^m = 0 \ . \tag{81}$$

Case $M \leq B$. Follow the boundary curve counterclockwise through points $(h, 2t)$ with integer t and integer h such that $h - n$ is even, starting at the point $(h, 2t)$ on edge I or II with lowest value of h such that $h+1 \geq H^{\text{gen}}$ and ending at the point $(h, 2t)$ on edge II, III or IV with lowest value of t such that $t \geq T^{\text{gen}}$. Let (h_1, t_1) be the first point found in this process for which $F(h_1, t_1) \leq B$, and let (h_2, t_2) be the last point for which $F(h_2, t_2) \leq B$. Set

$$H = H_n^m = h_1 - 1 \ , \qquad T = T_n^m = t_2 \ . \tag{82}$$

4.2 Truncation issues in computing c_t

For $t = 0,\ 1,\ \cdots$ and $0 < \varepsilon < 1$, the quantity

$$c_{t,LK} = \sum_{l=0}^{L} \sum_{k=0}^{K} A^{000}_{2l,2k,2t}\, a_l\, b_k \ . \tag{83}$$

approximates c_t with absolute error less than ε when L and K are such that

$$l > L \Rightarrow |a_l| < \tfrac{1}{4}\,\varepsilon \quad \& \quad k > K \Rightarrow |b_k| < \tfrac{3}{16}\,\varepsilon \ . \tag{84}$$

With $\gamma = \min(1, \ln(1/v_0))$, the second item in Eq. (84) holds when

$$K = \frac{1}{\gamma}\ \max\Big(0, \ln\frac{64}{3\varepsilon}\Big) + \tfrac{1}{2}\, g\, \frac{\sinh(\gamma)}{\gamma} \ . \tag{85}$$

Subsequently, let $S = \max(s_0, s_{0,M})$, and set

$$E = \frac{2\sqrt{\pi}}{\Gamma(3/4)}\ \frac{(1-S^2)^{-1/8}}{1+\sqrt{1-S^2}} \ , \qquad V = \frac{1-\sqrt{1-S^2}}{1+\sqrt{1-S^2}} \ . \tag{86}$$

Then the first item in Eq. (84) is valid when

$$L = \frac{\ln(8E/\varepsilon) + \frac{1}{4}\ln(1 + \ln(8E/\varepsilon)/\ln(1/V))}{\ln(1/V)} \ . \tag{87}$$

Furthermore, when the a_l and b_k required in Eq. (83) are available with absolute accuracy $\frac{1}{4}\,\varepsilon$ and $\frac{3}{16}\,\varepsilon$, respectively, while the K and L of Eqs. (85, 87) are used in Eq. (83), all c_t are approximated with absolute accuracy 2ε.

As to the availability of a_l and b_k for $l = 0, ..., L$ and $k = 0, ..., K$ with a required accuracy we give the following comments. The a_l have the form

$$a_l = a_{l,3/4,-3/4} + a_{l,1/4,-1/4} \ , \tag{88}$$

and either term at the right-hand side of Eq. (88) is computed using the infinite series expression in Eq. (70). When this infinite series is truncated at $N = 2L/\sqrt{1-S^2}$, with $S = \max(s_0, s_{0,M})$, the absolute error is for all $l = 0, ..., L$ and either term at the right-hand side of Eq. (88) less than $\varepsilon/16$, and then all a_l, $l = 0, ..., L$, are computed with absolute error less than $\varepsilon/8$. Finally, the b_k are given by Eq. (65) in terms of spherical Bessel and Hankel functions, and can therefore be computed to any desired accuracy using MatLab routines (employing the expressions for spherical Bessel and Hankel functions in terms of ordinary Bessel and Hankel functions, see [2], Sec. 10.47). When this is done with absolute accuracy $\frac{3}{32}2^{-7/4}\varepsilon u_0 \exp\left(-\varphi(K; g/2v_0)\right)/(2K+1)v_0$ and $3\varepsilon u_0/64(2K+1)$ for j_k and $h_k^{(2)}$, respectively, the b_k are computed for $k = 0,\ 1,\ \cdots,\ K$ with absolute accuracy $3\varepsilon/16$. Using these approximations of a_l and b_k in Eq. (83) with K and L as in Eqs. (85, 87) yields an approximation of c_t with absolute error less than $\frac{7}{4}\varepsilon$.

4.3 Accuracy of assembled scheme

Let $\varepsilon > 0$, and use either one of the truncation rules in Subsec. 4.1. Furthermore, compute c_t as in Subsec. 4.2 with absolute accuracy $\frac{7}{4}\varepsilon$. Finally, compute the Bessel function $J_{h+1}(2\pi r)$ with absolute accuracy $2\pi r\varepsilon/4w_0a_0$, with w_0 and a_0 given in Subsec. 4.1, using Matlab-codes. Then the quantity I in Eq. (59) is approximated with an absolute error that can be bounded by $\varepsilon + \frac{1}{2}\frac{7}{4}\varepsilon + \varepsilon = \frac{23}{8}\varepsilon$, due to, respectively, truncation of the double series in Eq. (59), approximating c_t as in Subsec. 4.2, and approximating the Jinc function $J_{h+1}(2\pi r)/2\pi r$ by computing J_{h+1} using the Matlab-code.

5 Illustration of the truncation rules

In this section, we show the absolute truncation error and the computation time, using the general truncation rule of Subsec. 2.3 and the dedicated truncation rule of Subsec. 2.4 for approximation of the diffraction integral I in Eqs. (1-2) as a function of $\varepsilon \in (0, 1)$ for a variety of radial values r, maximum defocus values f, numerical aperture values s_0 and $s_{0,M}$, and Zernike circle polynomial degrees and orders n and m. The truncation rules are used with $\varepsilon/2$ instead of ε. The structural quantities c_t and Jinc functions $J_{h+1}(2\pi r)/2\pi r$ are computed with absolute accuracies $\varepsilon/2$ and $\varepsilon/16w_0a_0$, respectively, so that the absolute error due to using these computed quantities is bounded by $\varepsilon/2$ for all n and m simultaneously. The total absolute error using the truncated series with the computed quantities is then expected to be less than $\frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon$.

In all figures, we show achieved accuracy (a) and computation time (b) against requested accuracy ε in the range $10^{-15} - 10^{0}$, using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines). The graphs result from specification of

A. the values of the aperture parameters s_0, $s_{0,M}$,

B1. the value of the focal parameter f,

B2. the value of the radial parameter r,

C. the degree n and order m of the radial polynomial R_n^m.

In the presented figures, the item(s) in 3 of the groups **A**, **B1**, **B2**, **C** are varied over at most two cases, while the item(s) in the remaining set is varied over several cases. Thus. schematically, we have in Figs. 6-15 the cases as defined in Table 1. In general, it can be said that the requested accuracy is achieved amply: the graphs in (a) stay well below and parallel to the graph $(\varepsilon, \varepsilon)$ (dotted lines). The performance of the dedicated rule in terms of accuracy is most of the time slightly worse but comparable to that of the general rule, while the performance in terms of computation time can be significantly better. The latter situation occurs especially when the degree and order of the radial polynomial are large compared to $f/2$ and $2\pi r$.

Figure	s_0, $s_{0,M}$	f	r	R_n^m
6, 7	fixed	2 cases		varied
8	fixed	varied	fixed	2 cases
9	fixed	fixed	varied	2 cases
10	varied	fixed	fixed	2 cases
11	varied	2 cases	fixed	fixed
12	varied	fixed	2 cases	fixed
13	fixed	varied	2 cases	fixed
14	2 cases	fixed	2 cases	varied
15	2 cases	fixed	fixed	varied

Table 1: Schematic overview indicating the item(s) in the which groups **A**, **B1**, **B2**, **C** are varied in Figs. 6-15.

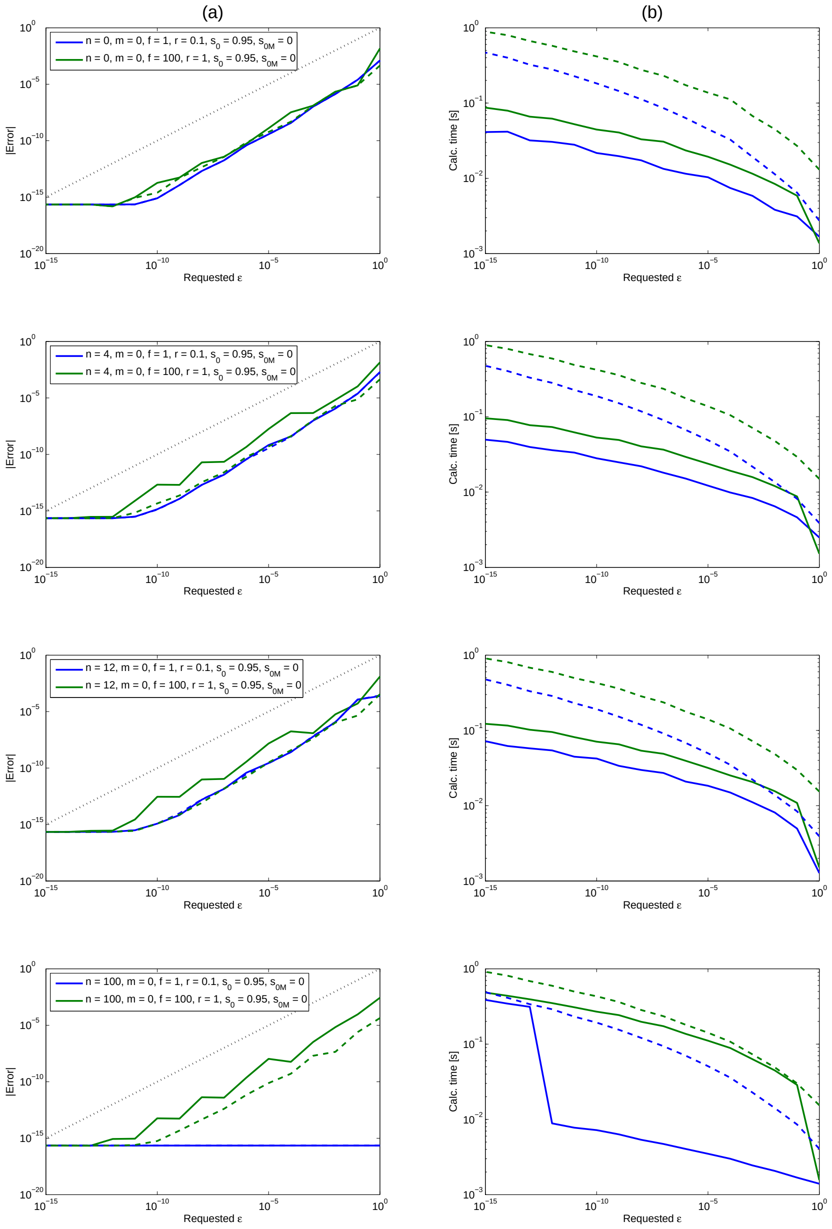

Figure 6: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the degree n and azimuthal order m of the radial polynomial from top to bottom according to $(n, m) = (0, 0),\ (4, 0),\ (12, 0),\ (100, 0)$. Setting of aperture variables: $s_0 = 0.95,\ s_{0,M} = 0$, setting of focal and radial variable: $f = 1,\ r = 0.1$ and $f = 100,\ r = 1$.

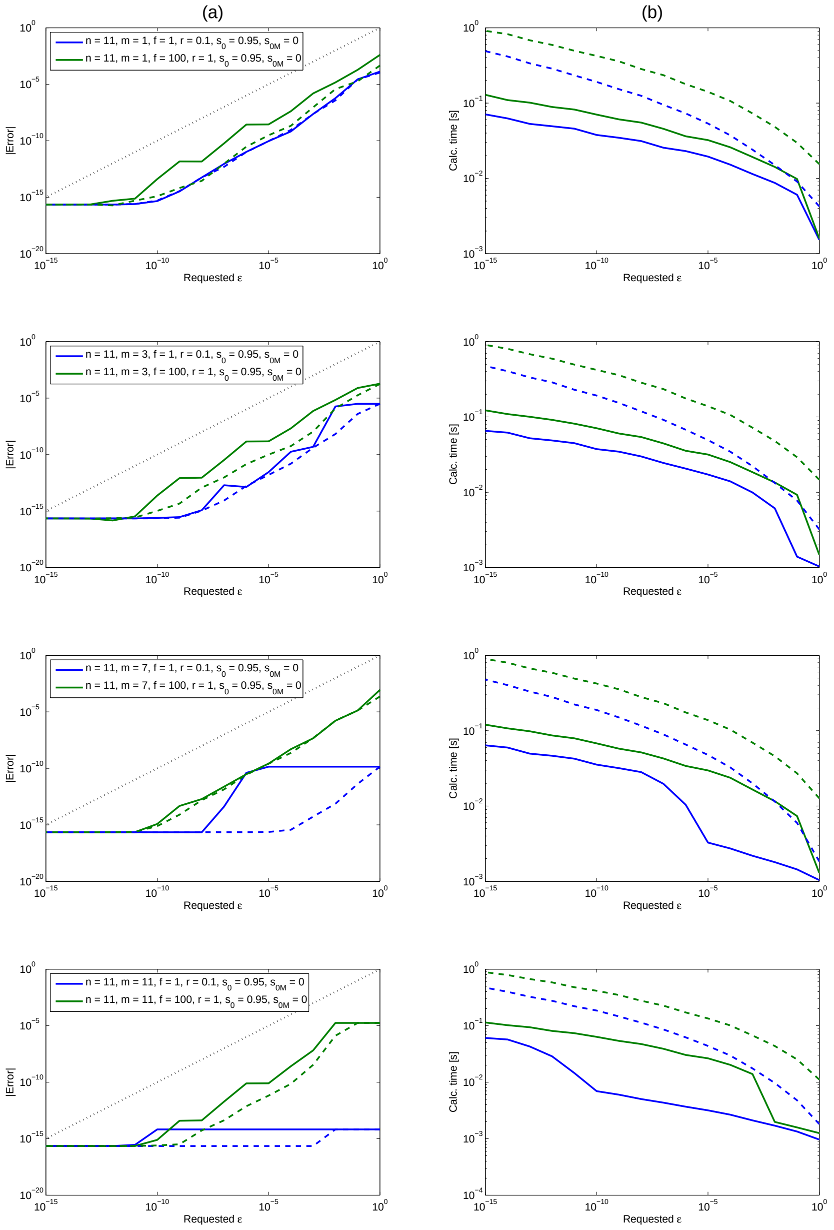

Figure 7: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the degree n and azimuthal order m of the radial polynomial from top to bottom according to $(n, m) = (11, 1),\ (11, 3),\ (11, 7),\ (11, 11)$. Setting of aperture variables: $s_0 = 0.95,\ s_{0,M} = 0$, setting of focal and radial variable: $f = 1,\ r = 0.1$ and $f = 100,\ r = 1$.

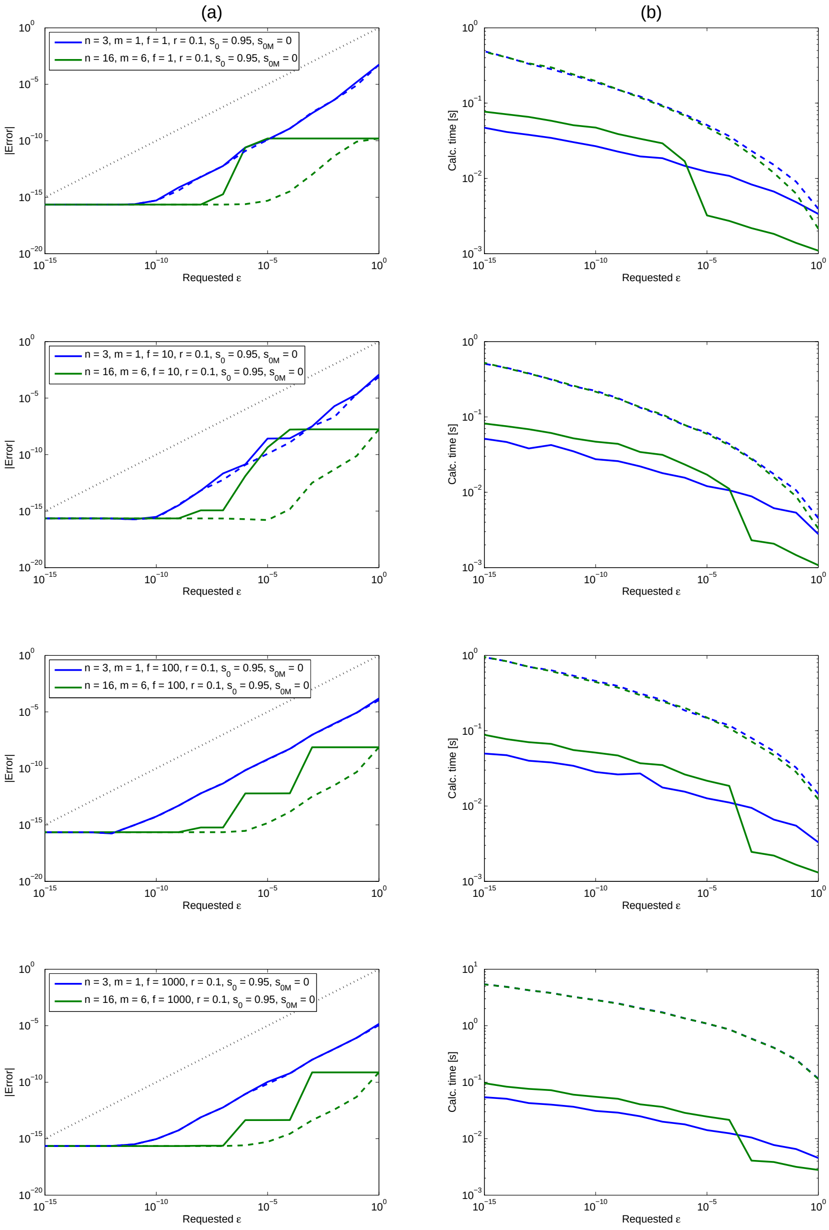

Figure 8: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the focal variable f from top to bottom according to $f = 1,\ 10,\ 100,\ 1000$. Setting of aperture variables: $s_0 = 0.95,\ s_{0,M} = 0$, setting radial variable: $r = 0.1$, setting of the degree and azimuthal order of the radial polynomial: $(n, m) = (3, 1)$ and $(16, 6)$.

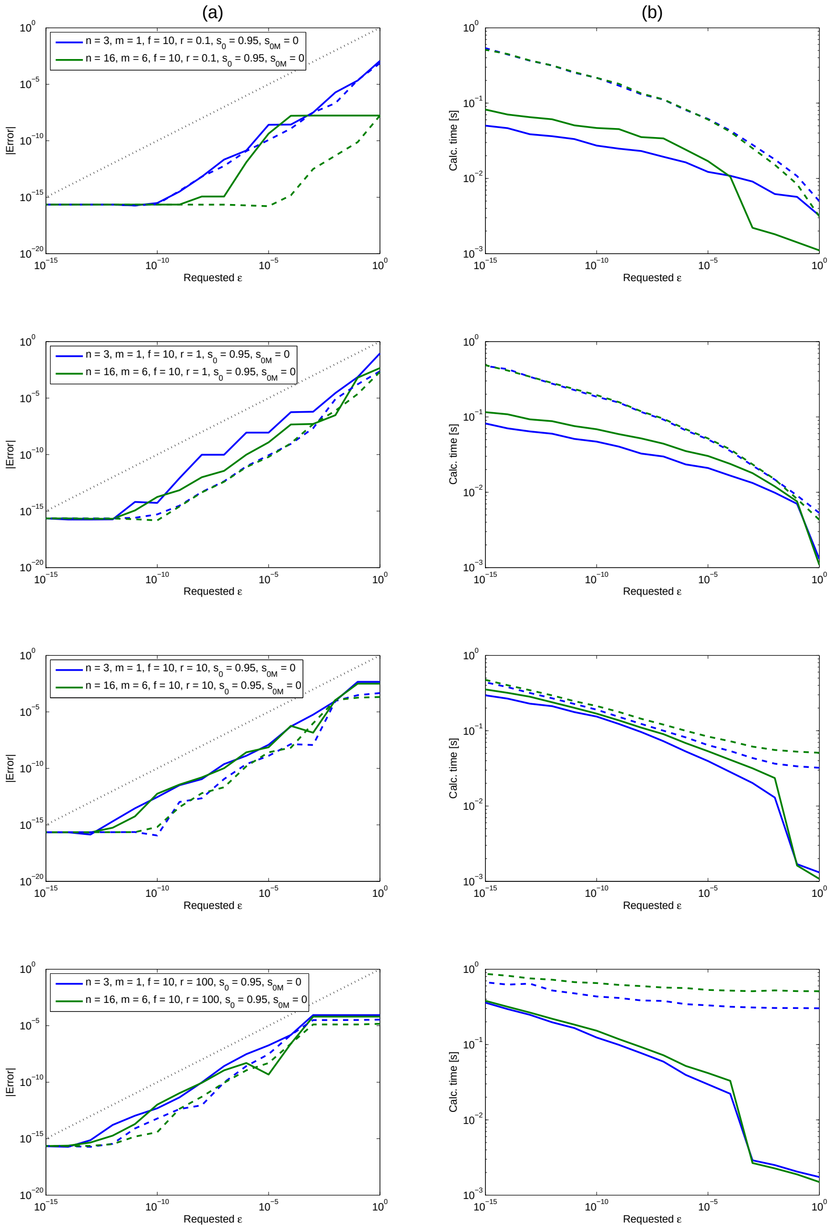

Figure 9: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the radial variable r from top to bottom according to $r = 0.1,\ 1,\ 10,\ 100$. Setting of aperture variables: $s_0 = 0.95,\ s_{0,M} = 0$, setting focal variable: $f = 10$, setting of the degree and azimuthal order of the radial polynomial to $(n, m) = (3, 1)$ and $(16, 6)$.

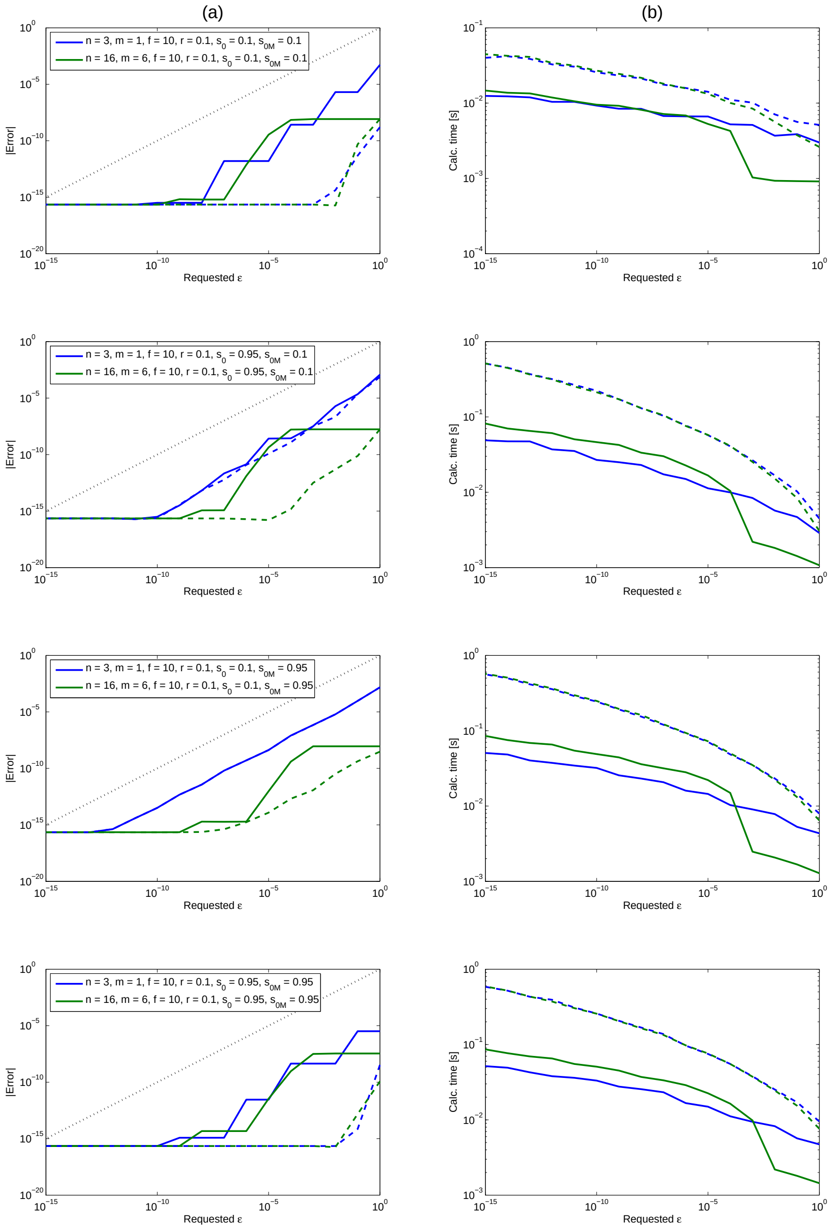

Figure 10: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the aperture variables s_0 and $s_{0,M}$ from top to bottom according to $(s_0, s_{0,M}) = (0.1, 0.1)$, $(0.95, 0.1)$, $(0.1, 0.95)$, $(0.95, 0.95)$. Setting of degree and azimuthal order of the radial polynomial: $(n, m) = (3, 1)$ and $(16, 6)$, setting of focal and radial variable: $f = 10$, $r = 0.1$.

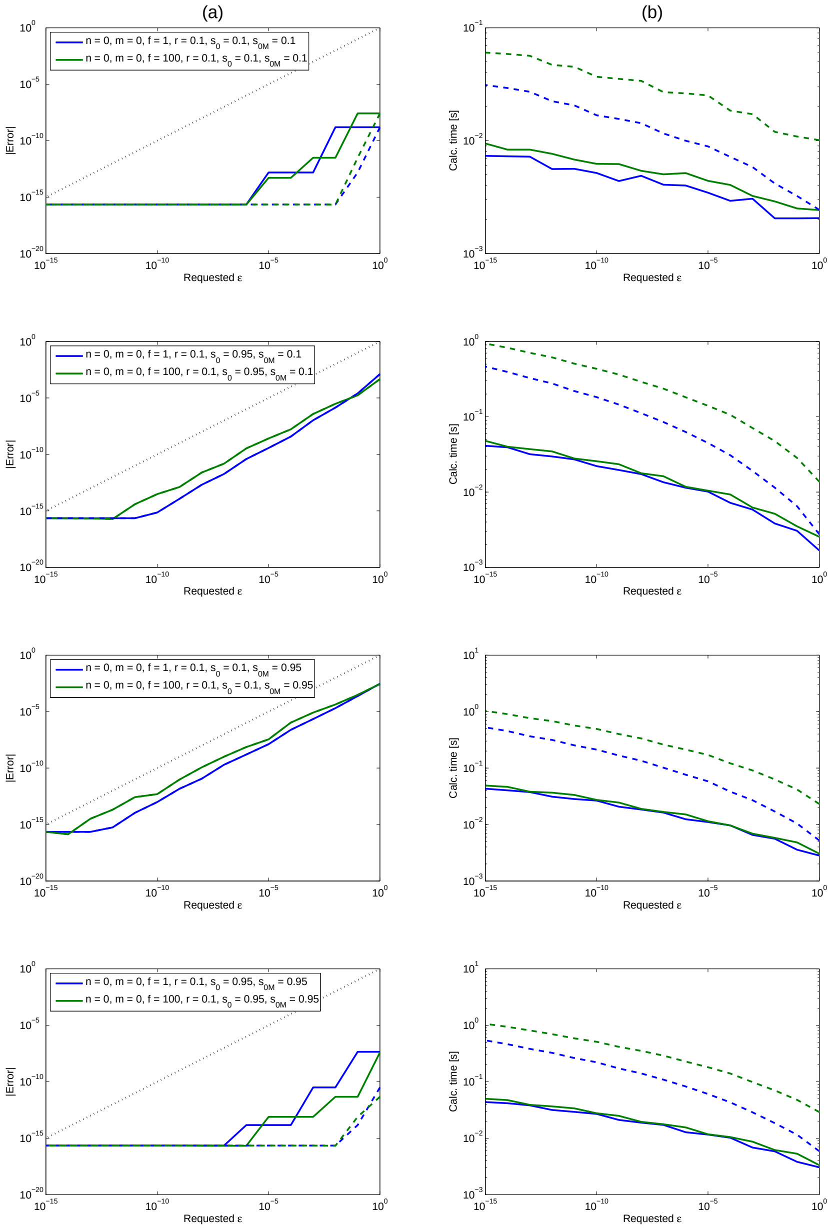

Figure 11: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the aperture variables s_0 and $s_{0,M}$ from top to bottom according to $(s_0, s_{0,M}) = (0.1, 0.1)$, $(0.95, 0.1)$, $(0.1, 0.95)$, $(0.95, 0.95)$. Setting of degree and azimuthal order of the radial polynomial: $(n, m) = (0, 0)$, setting of focal and radial variable: $f = 1$, $r = 0.1$ and $f = 100$, $r = 0.1$.

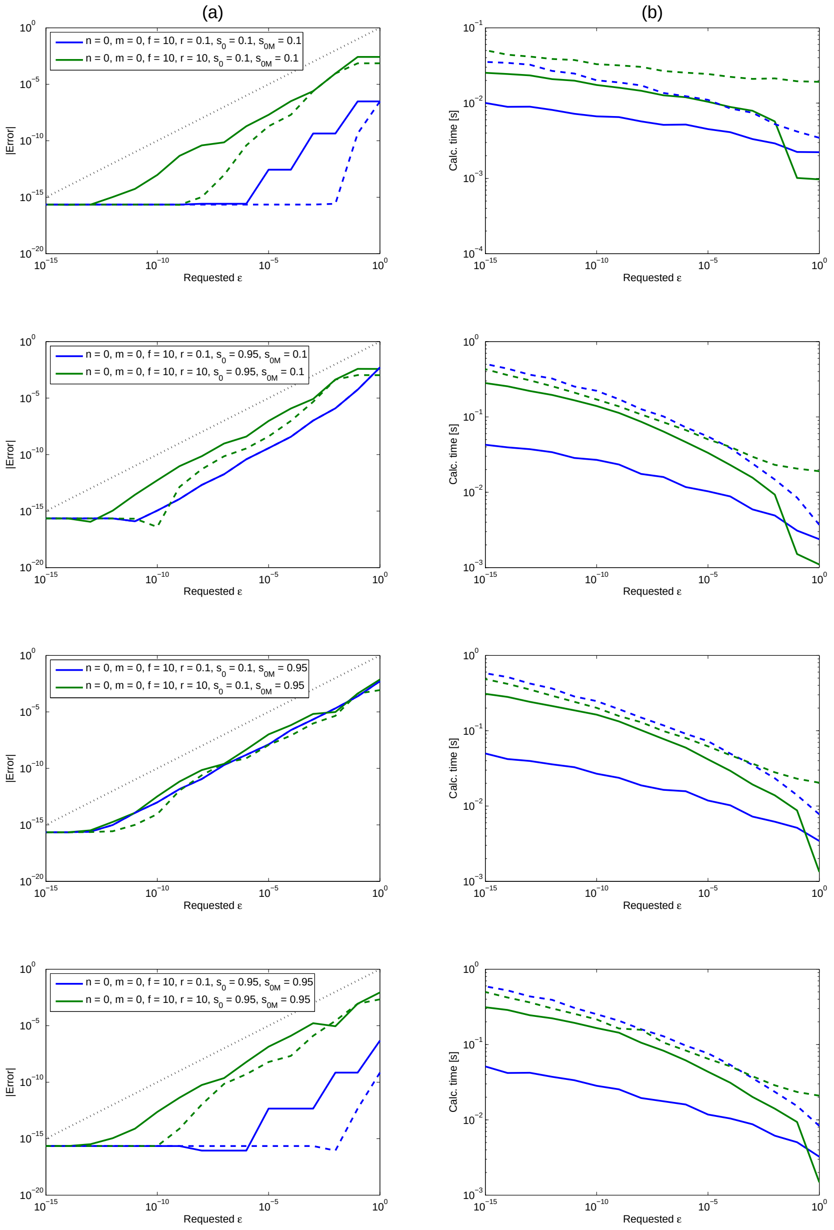

Figure 12: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the aperture variables s_0 and $s_{0,M}$ from top to bottom according to $(s_0, s_{0,M}) = (0.1, 0.1),\ (0.95, 0.1),\ (0.1, 0.95),\ (0.95, 0.95)$. Setting of degree and azimuthal order of the radial polynomial: $(n, m) = (0, 0)$, setting of focal and radial variable: $f = 10,\ r = 0.1$ and $f = 10,\ r = 10$.

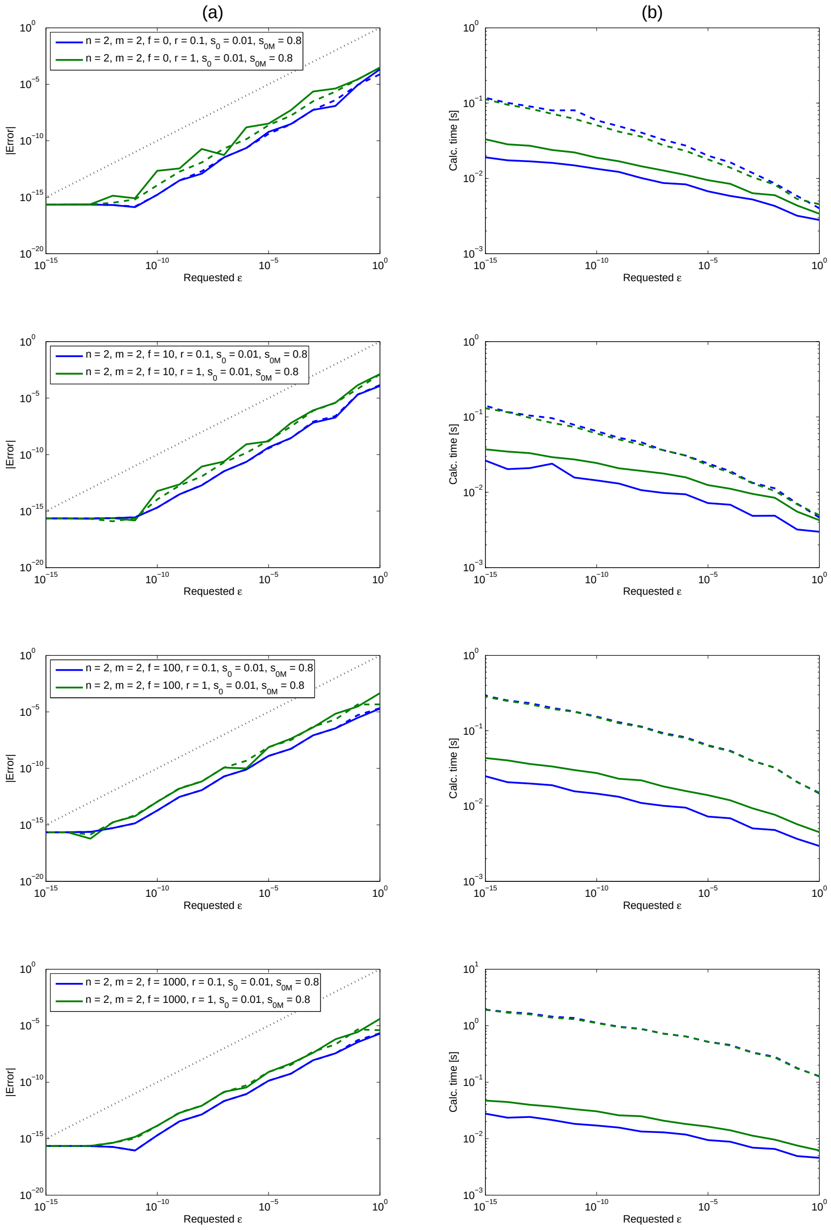

Figure 13: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the focal variable f from top to bottom according to $f = 0,\ 10,\ 100,\ 1000$. Setting of aperture variables: $s_0 = 0.01,\ s_{0,M} = 0.8$, setting radial variable: $r = 0.1$ and $r = 1$, setting the degree and azimuthal order of the radial polynomial: $(n, m) = (2, 2)$.

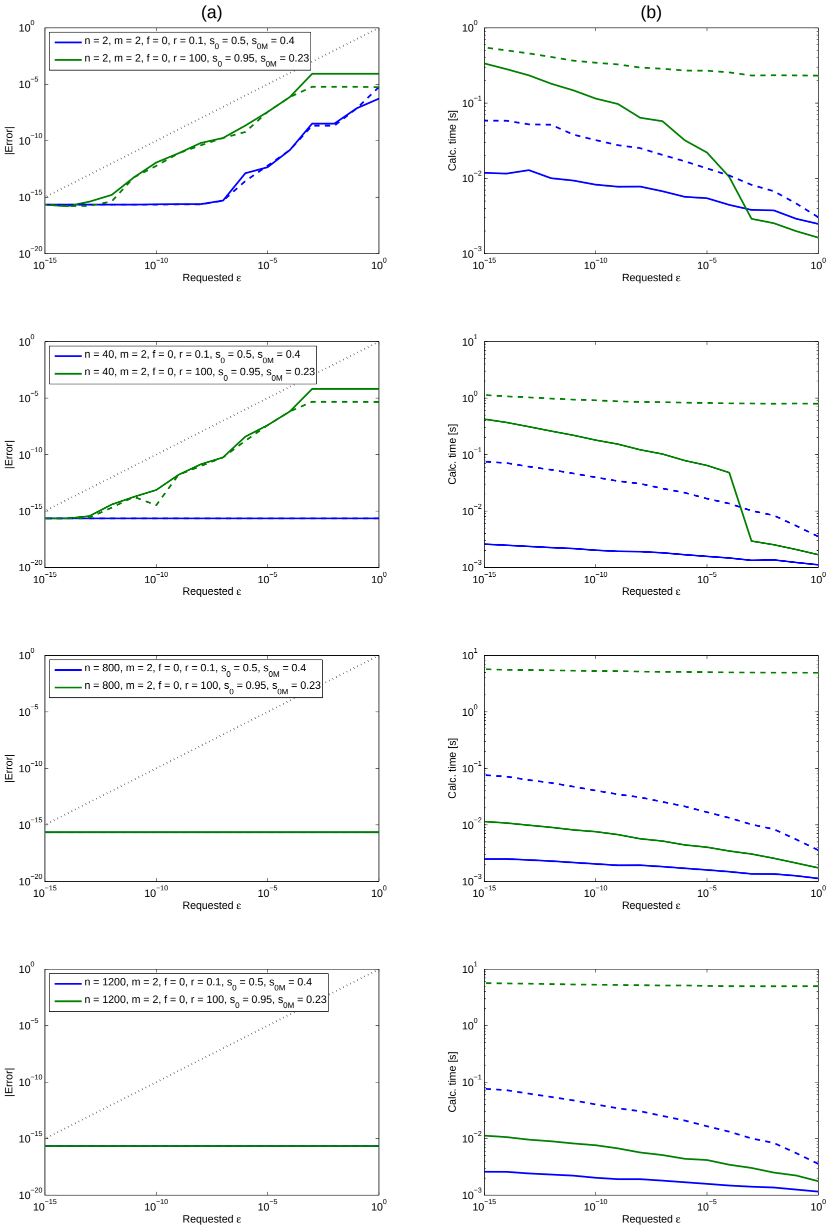

Figure 14: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the degree n and azimuthal order m of the radial polynomial from top to bottom according to $(n, m) = (2, 2),\ (40, 2),\ (800, 2),\ (1200, 2)$. Setting of aperture variables: $s_0 = 0.5,\ s_{0,M} = 0.4$ and $s_0 = 0.95,\ s_{0,M} = 0.23$, setting of focal and radial variable: $f = 0,\ r = 0.1$ and $f = 0,\ r = 100$.

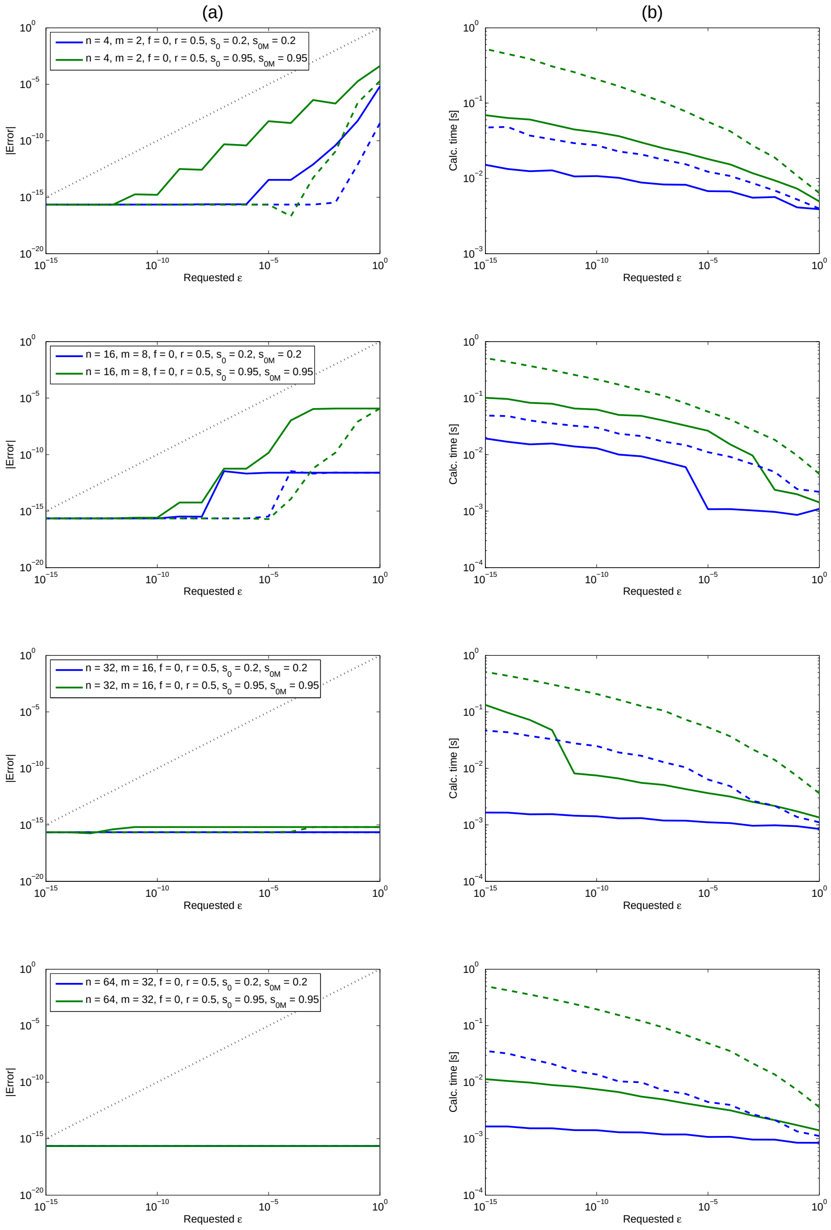

Figure 15: Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy ε using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the degree n and azimuthal order m of the radial polynomial from top to bottom according to $(n, m) = (4, 2),\ (16, 8),\ (32, 16),\ (64, 32)$. Setting of aperture variables: $s_0 = 0.2,\ s_{0,M} = 0.2$ and $s_0 = 0.95,\ s_{0,M} = 0.95$, setting of focal and radial variable: $f = 0,\ r = 0.5$.

6 Conclusions

We have formulated and verified truncation rules for the double series expressions that emerge from the advanced ENZ-theory for the computation of the optical diffraction integrals pertaining to optical systems with high NA, vector fields, polarization, and meant for imaging of extended objects. These rules have been devised for the central case $j = 0$ in the vectorial framework, which can be considered to be representative for all occurring diffraction integrals. Two versions of the truncation rule have been developed. The general rule gives precision to the rule-of-thumb that the required summation range is of the order $2\pi r$ times $\frac{1}{2}|f|$ with r and f the values of the (normalized) radial and the focal parameters in image space, irrespective of the degree and order of the radial polynomial involved in the diffraction integral. In the dedicated rule, we have also accounted for the specific way the radial polynomial influences the actual summation range, leading to performances comparable in terms of accuracy and better in terms of computation time than what is offered by the general truncation rule. A salient feature of the double series that manifest itself through the truncation rules is that the computation times stay well within what can be considered practicable, more or less independently of the values of the aperture parameters and the magnitudes of the focal and radial variable. In the case that circle polynomials of very high degree and/or order are involved in the diffraction integrals, the general truncation rule becomes impracticable, and one has to resort to using the dedicated rule. With this full understanding of the double series with regard to truncation matters, it can be said that the advanced ENZ-theory is more or less completed.

A Results on φ-functions

In this appendix, we present results on the functions

$$\varphi(x\,;\,c) = \begin{cases} 0 & , \quad 0 \le x \le c \ , \\ x\,\mathrm{arccosh}(x/c) - c\,\sqrt{(x/c)^2 - 1} & , \quad x \ge c \quad , \end{cases} \tag{A1}$$

and

$$\psi(x\,;\,c,d) = \varphi(x\,;\,c) - \varphi(x\,;\,d) \ , \qquad x \ge 0 \ , \tag{A2}$$

where $d > c > 0$. In Eq. (A1), we have

$$\mathrm{arccosh}(y) = \ln(y + \sqrt{y^2 - 1}) = \int\limits_1^y \frac{dz}{\sqrt{z^2 - 1}} \ , \qquad y \ge 1 \ , \tag{A3}$$

and this is a non-negative, non-decreasing function of y. Furthermore, with "$'$" denoting differentiation with respect to x,

$$\varphi'(x\,;\,c) = \begin{cases} 0 & , \quad 0 \le x \le c \ , \\ \operatorname{arccosh}(x/c) \ , & \quad x \ge c \qquad , \end{cases} \tag{A4}$$

so that $\varphi(x\,;\,c)$ is continuously differentiable in $x \ge 0$. From Eqs. (A3–A4), it is seen that $\varphi(x\,;\,c)$ is non-negative, non-decreasing and convex in $x \ge 0$, and strictly so in $x > c$. Also, $\varphi(x\,;\,c)$ behaves like $x \ln(2x/ec)$ for large $x > 0$, and grows therefore super-linearly.

We next consider $\psi(x\,;\,c,d)$ in Eq. (A2). From

$$\frac{\partial\varphi}{\partial c}\,(x\,;\,c) = \begin{cases} 0 & , \quad 0 \le x \le c \ , \\ \dfrac{-1}{c}\,\sqrt{x^2-c^2} \ , & \quad x \ge c \qquad , \end{cases} \tag{A5}$$

we have that $\varphi(x\,;\,c)$ is decreasing in $c > 0$ for any x, and so $\psi(x\,;\,c,d)$ is non-negative. Furthermore,

$$\psi'(x\,;\,c,d) = \begin{cases} 0 & , \quad 0 \le x \le c \ , \\ \operatorname{arccosh}(x/c) & , \quad c \le x \le d \ , \\ \operatorname{arccosh}(x/c) - \operatorname{arccosh}(x/d) \ , & \quad x \ge d \qquad , \end{cases} \tag{A6}$$

and this shows that $\psi(x\,;\,c,d)$ is non-decreasing in $x \ge 0$, and strictly so in $x \ge c$. Moreover, we have for $x > d$

$$\psi''(x\,;\,c,d) = \frac{1}{\sqrt{x^2-c^2}} - \frac{1}{\sqrt{x^2-d^2}} < 0 \ , \tag{A7}$$

and so $\psi(x\,;\,c,d)$ is strictly concave in $x > d$, while $\psi(x\,;\,c,d)$ is strictly convex in $x \in (c,d)$. Finally,

$$\psi'(x\,;\,c,d) = \ln\Big(\frac{d}{c}\Big) + \ln\Big(\frac{x+\sqrt{x^2-c^2}}{x+\sqrt{x^2-d^2}}\Big) > \ln\Big(\frac{d}{c}\Big) \ , \qquad x > d \ , \tag{A8}$$

which shows that $\psi'(x\,;\,c,d)$ decreases to $\ln(d/c)$ as $x \to \infty$, and we have directly from Eqs. (A1-A3)

$$\psi(x\,;\,c,d) - x\ln\Big(\frac{d}{c}\Big) = -\,\frac{d^2-c^2}{4x} + O\Big(\frac{1}{x^3}\Big) \ , \tag{A9}$$

for $x > d$, so that $\psi(x\,;\,c,d) - x\ln(d/c)$ increases to 0 as $x \to \infty$.

In the formulation of the general truncation rule, it has been used that one can find piecewise linear functions bounding $\varphi(x\,;\,c)$ and $\psi(x\,;\,c,d)$ from below. Furthermore, in the design of the dedicated truncation rule, it is convenient to have convex functions bounding $\psi(x\,;\,c,d)$ from below (since $\varphi(x\,;\,c)$ is itself convex, such an effort does not have to be made for φ).

By convexity of $\varphi(x\,;\,c)$, the graph of φ lies above any tangent line, and so for any $x_0 > 0$, we have

$$\varphi(x\,;\,c) \geq \varphi(x_0\,;\,c) + (x - x_0)\,\varphi'(x_0\,;\,c)\ , \qquad x \geq 0\ . \tag{A10}$$

For a linear lower bound on $\psi(x\,;\,c,d)$, one must choose $x_0 \in (c,d)$ such that $\psi'(x_0\,;\,c,d) \leq \ln(d/c)$, see Eq. (A8), and then

$$\psi(x\,;\,c,d) \geq \psi(x_0\,;\,c,d) + (x - x_0)\,\psi'(x_0\,;\,c,d)\ , \qquad x \geq 0\ . \tag{A11}$$

Since $x_0 \in (c,d)$ and $\psi(x\,;\,c,d) = \varphi(x\,;\,c)$ for $c \leq x \leq d$, we have from Eqs. (A1, A6) that

$$\begin{aligned} \psi(x\,;\,c,d) &\geq x\,\mathrm{arccosh}(x_0/c) - c\,\sqrt{(x_0/c)^2 - 1} \\ &= \gamma x - c\sinh(\gamma)\ , \qquad x \geq 0\ , \end{aligned} \tag{A12}$$

where we have set $\gamma = \mathrm{arccosh}(x_0/c)$. Choosing the largest possible $x_0 \in (c,d)$, so that

$$\psi'(x_0\,;\,c,d) = \ln(d/c) =: \gamma_0\ , \tag{A13}$$

we have

$$x_0 = c\cosh(\gamma_0)\ , \qquad \psi(x_0\,;\,c,d) = \gamma x_0 - c\sinh(\gamma_0)\ . \tag{A14}$$

Hence, for any $\gamma \in (0, \gamma_0]$, we have

$$\psi(x\,;\,c,d) \geq \gamma x - c\sinh(\gamma)\ , \qquad x \geq 0\ . \tag{A15}$$

Evidently, since $\varphi(x\,;\,c) \geq \psi(x\,;\,c,d)$, the latter bound is also valid for $\varphi(x\,;\,c)$, without a restriction on γ. The choice $\gamma = 1$ leads to

$$\varphi(x\,;\,c) \geq x - c\sinh(1)\ , \qquad x \geq 0\ . \tag{A16}$$

The largest convex functon bounding $\psi(x\,;\,c,d)$ from below is given by

$$\varphi(x\,;\,c,d) = \begin{cases} \varphi(x\,;\,c) & ,\quad 0 \leq x \leq c\cosh(\gamma_0)\ , \\ \gamma_0 x - c\sinh(\gamma_0)\ , & \quad x \geq c\cosh(\gamma_0)\ . \end{cases} \tag{A17}$$

We conclude this appendix by showing 3 inequalities. The first one of these reads

$$\varphi(x\,;\,c)+\tfrac{3}{2}\ln c\geq\varphi(x\,;\,1)\ ,\qquad 0<c\leq 1\ ,\tag{A18}$$

when $x\geq\frac{1}{2}\sqrt{13}$, and is required in Appendix B. We have by Eq. (A5) for $0<c\leq 1\leq x$ that

$$\frac{d}{dc}\,[\varphi(x\,;\,c)+\tfrac{3}{2}\ln c]=\frac{1}{c}\,(\tfrac{3}{2}-\sqrt{x^2-c^2})\ ,\tag{A19}$$

and this is negative for all $c\in(0,1]$ when $\sqrt{x^2-1}\geq 3/2$, i.e., when $x\geq\sqrt{13}$. Since there is equality in Eq. (A18) when $c=1$, we get the result.

Next, we show that for $\alpha>0$ and $x\geq c\geq\alpha\geq 0$

$$\varphi(x+\alpha\,;\,c)-\varphi(x\,;\,c)-\alpha\ln\Bigl(\frac{x+\alpha}{c}\Bigr)\geq 0\ .\tag{A20}$$

This is required in Appendix C with $\alpha=1/2$ and $c\geq 1/2$.

To show Eq. (A20), we let $b=\alpha/c$, and we observe from Eq. (A4) that Eq. (A20) holds for $x\geq c$ if and only if

$$\Phi(w\,;\,b):=\int\limits_{w}^{w+b}\operatorname{arccosh}(v)\,dv-b\ln(w+b)\geq 0\tag{A21}$$

holds for $w:=x/c\geq 1$. Now $\Phi(w\,;\,b=0)=0$, and

$$\frac{\partial\Phi}{\partial b}\,(w\,;\,b)=\ln\Bigl[1+\Bigl(1-\frac{1}{(w+b)^2}\Bigr)^{1/2}\Bigr]-\frac{b}{w+b}\tag{A22}$$

increases in $w\geq 1$ for fixed $b\geq 0$. Hence, when $b_0>0$ is such that

$$\frac{\partial\Phi}{\partial b}\,(1\,;\,b)\geq 0\ ,\qquad 0\leq b\leq b_0\ ,\tag{A23}$$

we have that

$$\frac{\partial\Phi}{\partial b}\,(w\,;\,b)\geq 0\ ,\qquad 0\leq b\leq b_0\,,\quad w\geq 1\ ,\tag{A24}$$

and so, from $\Phi(w\,;\,b=0)=0$, that $\Phi(w\,;\,b)\geq 0$ for $w\geq 1$ and $0\leq b\leq b_0$. Now with $z=\frac{1}{1+b}\in(0,1]$,

$$\frac{\partial\Phi}{\partial b}\,(1\,;\,b)=\ln(1+(1-z^2)^{1/2})-1+z\tag{A25}$$

is a concave function of $z \in (0,1]$, since

$$\frac{d}{dz}\,[\ln(1+(1-z^2)^{1/2})] = \frac{-z}{1-z^2+(1-z^2)^{1/2}} \tag{A26}$$

decreases from 0 at $z=0$ to $-\infty$ at $z=1$. Furthermore, the right-hand side of Eq. (A25) vanishes at $z=1$, has the value $\ln 2 - 1 < 0$ at $z=0$, and the value $\ln(1+\frac{1}{2}\sqrt{3}) - \frac{1}{2} = 0.12\ldots > 0$ at $z = 1/2$. Therefore, the right-hand side of Eq. (A25) is non-negative for $\frac{1}{2} \le z \le 1$. Hence, with $b = \alpha/c \in [0,1]$, so that $z = (1+b)^{-1} \in [\frac{1}{2}\,,1]$, we have that Eqs. (A23–A24) hold with $b_0 = 1$. It follows that Eq. (A21) holds for $0 \le b \le 1 \le w$, as required.

An inequality converse to Eq. (A20) reads

$$\varphi(x+\alpha\,;\,c) - \varphi(x\,;\,c) - \alpha\ln\Big(\frac{x+\alpha}{c}\Big) \le \alpha\ln 2 \tag{A27}$$

when $\alpha > 0$ and $x \ge 0$, $x+\alpha \ge c \ge 0$, and follows easily from Eqs. (A3-A4).

In Appendix C, the inequality in Eq. (A20) is required for all $x \ge 0$. We shall comment on this below.

We next show that for $x \ge d \ge c \ge \alpha \ge 0$

$$\begin{aligned}\Big[\varphi(x+\alpha\,;\,c) - \alpha\ln\Big(\frac{x+\alpha}{c}\Big)\Big] - \Big[\varphi(x+\alpha\,;\,d) - \alpha\ln\Big(\frac{x+\alpha}{d}\Big)\Big]&\\ \ge\ \varphi(x\,;\,c) - \varphi(x\,;\,d)\ .&\end{aligned} \tag{A28}$$

This is required in Appendix C with $\alpha = 1/2$ and $c \ge 1/2$. For $x \ge d$, we have by Eqs. (A3–A4)

$$\begin{aligned}&\Big[\varphi(x+\alpha\,;\,c) - \varphi(x\,;\,c) - \alpha\ln\Big(\frac{x+\alpha}{c}\Big)\Big]\\ &\qquad -\ \Big[\varphi(x+\alpha\,;\,d) - \varphi(x\,;\,d) - \alpha\ln\Big(\frac{x+\alpha}{d}\Big)\Big]\\ &= \int_x^{x+\alpha} \Big(\operatorname{arccosh}\Big(\frac{y}{c}\Big) - \operatorname{arccosh}\Big(\frac{y}{d}\Big)\Big)\,dy + \alpha\ln\Big(\frac{c}{d}\Big)\\ &= \int_x^{x+\alpha} \ln\Big(\frac{y+\sqrt{y^2-c^2}}{y+\sqrt{y^2-d^2}}\Big)\,dy \ge 0\ ,\end{aligned} \tag{A29}$$

and this is the required inequality.

The inequalities in Eqs. (A20, A28) are required in Appendix C for all $x \geq 0$. Since $\varphi(x+\alpha\,;\,c)$ and $\varphi(x\,;\,c)$ vanish when $x+\alpha \leq c$, we have that Eq. (A20) holds for all $x \geq 0$, except perhaps when $c-\alpha \leq x \leq c$. In this latter case, we have that $\varphi(x\,;\,c) = 0$, and therefore the left-hand side of Eq. (A20) can be written as

$$\varphi(x+\alpha\,;\,c) - \alpha \ln\Bigl(\frac{x+\alpha}{c}\Bigr) = \alpha\,[c'(v \operatorname{arccosh} v - \sqrt{v^2-1}) - \ln v]\ , \tag{A30}$$

where we have set $c' = c/\alpha$ and $v = (x+\alpha)/c \in [1, 1+1/c']$. Now the minimum of

$$c'(v \operatorname{arccosh} v - \sqrt{v^2-1}) - \ln v \tag{A31}$$

is assumed at v such that $v \operatorname{arccosh} v = 1/c'$ (this v is indeed in $[1, 1+1/c']$), and this minimum increases in c'. For the case that $c' = c/\alpha = 1$, we find numerically the minimum value -0.109709667. Hence, for the case that $\alpha = 1/2$, as considered in Appendix C, we are dealing with a minimum value of the whole left-hand side of Eq. (A20) of the order -0.05. This can safely be ignored, and so we declare Eq. (A20) to be valid for all $x \geq 0$.

A similar situation arises for the inequality in Eq. (A28) whose validity is ensured for $x \geq d$, $x \leq c-\alpha$ and $c \leq x \leq d-\alpha$ (in the latter case, the second term in [ ] in Eq. (A29) is non-positive, while the first term in [ ] is non-negative by Eq. (A20)). So we only need to consider $c-\alpha \leq x \leq c$ and $d-\alpha \leq x \leq d$ (these two x-intervals overlap when $d-\alpha \leq c$). The minimum value of the first term in [ ] in Eq. (A29) has been bounded from below by -0.109709667α. The second term can be written on $d-\alpha \leq x \leq d$ as

$$\alpha\,[d'(v \operatorname{arccosh} v - \sqrt{v^2-1}) - \ln v] \tag{A32}$$

with $d' = d/\alpha \geq 1$ and $v = (x+\alpha)/d \in [1, 1+1/d']$. The function

$$f(v) = d'(v \operatorname{arccosh} v - \sqrt{v^2-1}) - \ln v\ , \qquad v \geq 1\ , \tag{A33}$$

is convex, and so its maximum over $[1, 1+1/d']$ occurs at $v=1$, with value $f(1) = 0$, or at $v = 1+1/d'$, with value

$$\frac{1}{u}\,[(1+u)\operatorname{arccosh}(1+u) - \sqrt{(1+u)^2-1}] - \ln(1+u)\ , \tag{A34}$$

where $u = 1/d' \in [0,1]$. An elementary analysis of the function in Eq. (A34) shows that it is maximal at $u = 0.191487884$, with maximal value 0.2486813544. Hence, for the case $\alpha = 1/2$, as considered in Appendix C, we are dealing with a maximum value of the second term in [ ] in Eq. (A29) that can be bounded by $1/8$. This can be safely ignored, and we thus declare Eq. (A28) to be valid for all $x \geq 0$ and $d \geq c \geq \alpha$.

B Bounding Jinc functions

In this appendix, we bound and estimate Jinc functions $J_{h+1}(2\pi r)/2\pi r$ for $h = 0, 1, ...$ and $r > 0$.

We first consider the case that $h+1 < 2\pi r$. Let $\beta \in (0, \pi/2)$ be fixed, and let $\xi = \nu(\tan\beta - \beta) - \frac{1}{4}\pi$. With $\sec\beta = 1/\cos\beta > 1$, the first term of Debye's asymptotic result [2], 10.19.6, p. 231 as $\nu \to \infty$ yields the approximation

$$J_\nu(\nu\sec\beta) + i\,Y_\nu(\nu\sec\beta) \approx \Big(\frac{2}{\pi\nu\tan\beta}\Big)^{1/2} e^{i\xi} \ , \tag{B1}$$

where J_ν and Y_ν are the Bessel functions of first and second kind, respectively, and of order ν. With $\nu = h+1$ and β such that $h+1 = 2\pi r\cos\beta$, we have

$$\Big(\frac{2}{\pi\nu\tan\beta}\Big)^{1/2} = \frac{1}{\pi\sqrt{r}}\Big(1 - \Big(\frac{h+1}{2\pi r}\Big)^2\Big)^{-1/4} \ . \tag{B2}$$

The factor $(1 - ((h+1)/2\pi r)^2)^{-1/4}$ is close to 1 on a large part of the range $0 \le h+1 < 2\pi r$, and we shall replace it by 1 (this issue is further addressed below). We thus estimate

$$\Big|\frac{J_{h+1}(2\pi r)}{2\pi r}\Big| \le \frac{1}{2\pi^2\, r\sqrt{r}} \ , \qquad 0 \le h+1 < 2\pi r \ . \tag{B3}$$

We next consider the case that $h+1 > 2\pi r$. With $\operatorname{sech}\alpha = 1/\cosh\alpha < 1$, the first term of Debye's asymptotic result [2], 10.19.3, p. 231 as $\nu \to \infty$ yields the approximation

$$J_\nu(\nu\operatorname{sech}\alpha) \approx \frac{\exp(\nu(\tanh\alpha - \alpha))}{(2\pi\nu\tanh\alpha)^{1/2}} \ . \tag{B4}$$

With $\nu = h+1$ and α such that $h+1 = 2\pi r\operatorname{sech}\alpha$, we have

$$\Big(\frac{1}{2\pi\nu\tanh\alpha}\Big)^{1/2} = \frac{1}{2\pi\sqrt{r}}\Big(\Big(\frac{h+1}{2\pi r}\Big)^2 - 1\Big)^{-1/4} \ . \tag{B5}$$

We replace the factor $(((h+1)/2\pi r)^2 - 1)^{-1/4}$ at the right-hand side of Eq. (B5) by 1 as before, and we observe that

$$\begin{aligned} \nu(\tanh\alpha - \alpha) &= 2\pi r\Big(\Big(\frac{h+1}{2\pi r}\Big)^2 - 1\Big)^{1/2} - (h+1)\operatorname{arccosh}\Big(\frac{h+1}{2\pi r}\Big) \\ &= -\varphi(h+1\,;\,2\pi r) \ , \end{aligned} \tag{B6}$$

with φ as in Appendix A.

We thus get on the whole range $h \geq 0$ the estimate

$$\Big|\frac{J_{h+1}(2\pi r)}{2\pi r}\Big| \leq \frac{1}{2\pi^2\, r\,\sqrt{r}}\, \exp(-\varphi(h+1\,;\,2\pi r))\ . \tag{B7}$$

In deriving the bound in Eq. (B7), we have set the $(\ )^{-1/4}$-factors in Eq. (B2, B5) equal to 1. We shall now assess the amount by which the bounding function in Eq. (B7) is off by this simplification. At the point $h+1=2\pi r$ we have $\varphi(h+1\,;\,2\pi r)=0$, and we are thus comparing the bound $(2/\pi\nu)^{1/2}$ for $J_\nu(\nu)$ by its actual value when $\nu=h+1=2\pi r\to\infty$. In [2], 10.14.2, p. 227, there is the bound, for $0<x<\nu$,

$$0<J_\nu(x)<J_\nu(\nu)=\frac{2^{1/3}}{3^{2/3}\,\Gamma(2/3)\,\nu^{1/3}}=0.4473\nu^{-1/3}\ . \tag{B8}$$

The asymptotic value of the maximum of $|J_\nu(x)|$ over all $x>0$ is $\approx 0.6748\nu^{-1/3}$ (assumed near $x=\nu+(\nu/2)^{1/3}$), and this has to be compared with $(2/\pi\nu)^{1/2}$. The ratio of the asymptotic maximum value and $(2/\pi\nu)^{1/2}$ is $\approx\ 0.8457\nu^{1/6}$. The quantity $0.8457\nu^{1/6}$ equals 1, 2 and 4 for $\nu=2.73$, 175 and 11194, respectively.

The bound in Eq. (B7) is somewhat awkward to use when r is close to 0. With $R=\max(1/2\pi,r)$, we have

$$\Big|\frac{J_{h+1}(2\pi r)}{2\pi r}\Big| \leq \frac{1}{2\pi^2\, R\,\sqrt{R}}\, \exp(-\varphi(h+1\,;\,2\pi R))\ . \tag{B9}$$

Indeed, when $r\geq 1/2\pi$, the two right-hand sides of Eqs. (B7, B9) are equal. When $0<r<1/2\pi$ and $h\geq 1$, the right-hand side of Eq. (B7) is less than the right-hand side of Eq. (B9) which follows from Eq. (A18) with $0<c=r/R<1$ and $x=h+1\geq 2>\frac{1}{2}\sqrt{13}$. The case $h=0$ needs separate consideration. The inequality to be proved is then

$$\Big|\frac{J_1(x)}{x}\Big| \leq \frac{1}{y}\sqrt{\frac{2}{\pi y}} \tag{B10}$$

when $x>0$ and $y=\max(1,x)$. When $0<x\leq 1$, we have $y=1$ and the right-hand side of Eq. (B10) equals $\sqrt{2/\pi}$ which exceeds the maximum value 1/2 of $|J_1(x)/x|$. When $x\geq 1$, the inequality to be shown reads $x\,J_1^2(x)\leq 2/\pi$. It follows from [3], §13.74 that $x(J_1^2(x)+Y_1^2(x))$ decreases to $2/\pi$ when $x\to\infty$. The maximum value of $x\,J_1^2(x)$ is just slightly larger than $2/\pi$ (0.6652 near $x=2.00$, compared to $2/\pi=0.6366$). We shall ignore this minor excess.

C Bounding structural quantities

In this appendix, we bound and estimate the structural quantities c_t required in Eq. (1). At this point, we are interested in a manageable bound that can be used to formulate transparent truncation rules. To achieve this, we argue somewhat heuristically. We make the observation that the algebraic factor $a(\rho)$ is composed from functions $(1-s^2\rho^2)^\delta$ with $|\delta| \leq 3/4$. Any such function can be written as

$$\begin{aligned}(1-s^2\rho^2)^\delta &= \exp(2\delta \ln(1-(1-(1-s^2\rho^2)^{1/2}))) \\ &\approx \exp(-2\delta(1-(1-s^2\rho^2)^{1/2}))\ , \end{aligned} \tag{C1}$$

where the latter function has the appearance of a focal factor with imaginary value of the normalized focal parameter f/u_0 of order unity. Moving a factor $\sqrt{1-s_0^2\rho^2}$ from the focal factor to the algebraic factor, see Eqs. (34–35), we are led to estimate the Zernike coefficients c_t of $a(\rho)\, f(\rho)$ by those of

$$\frac{a_0}{\sqrt{1-s_0^2\rho^2}}\, \exp\Big(\frac{ig}{u_0}\,(1-\sqrt{1-s_0^2\rho^2})\Big)\ , \tag{C2}$$

where $g = \max(1, |f|)$ and a_0 is the R_0^0-coefficient of $a(\rho)\,\sqrt{1-s_0^2\rho^2}$ as in Eq. (17). Using the explicit form of the Zernike coefficients $b_t(g)$ of the modified focal factor, see Eqs. (4, 35, 38), we thus postulate for c_t the bound

$$a_0\,|b_t(g)| = a_0\,\frac{2t+1}{u_0}\,g\,|j_t(g/2)|\,|h_t^{(2)}(g/2v_0)|\ . \tag{C3}$$

Here it has been assumed that $s_0 \leq s_{0,M}$. In the case that $s_{0,M} > s_0$, we should replace in the above all s_0 by $s_{0,M}$.

We next estimate j_t and $h_t^{(2)}$ using Debye's asymptotic results. We have from Eq. (39) and Appendix B

$$|j_t(g/2)| = \sqrt{\frac{\pi}{g}}\,|J_{t+1/2}(g/2)| \leq \frac{2}{g}\ , \qquad 0 \leq t+1/2 \leq g/2\ , \tag{C4}$$

where we have replaced a factor $(1-((2t+1)/g)^2)^{-1/4}$ by 1. Similarly, we have from Eq. (40) and Appendix B

$$|h_t^{(2)}(g/2v_0)| \leq \frac{2v_0}{g}\ , \qquad 0 \leq t+1/2 \leq g/2v_0\ , \tag{C5}$$

where we have replaced a factor $(1-((2t+1)\,v_0/g)^2)^{-1/4}$ by 1. Hence,

$$|b_t(g)| \leq 4\,\frac{v_0}{u_0}\,\frac{2t+1}{g} \leq 4\,\frac{v_0}{u_0}\ , \qquad 0 \leq t+1/2 \leq g/2\ . \tag{C6}$$

On the range $t+1/2 \geq g/2$, we need to be more careful since the factor $(2t+1)\,g$ at the right-hand side of Eq. (C3) can become arbitrarily large. We estimate now, in accordance with the equality in Eq. (C4) and Eqs. (B4–B5) with $h+1 = t+1/2$ and $2\pi r = g/2$

$$\begin{aligned} &|j_t(g/2)| \\ \leq\ & \sqrt{\frac{\pi}{g}}\,\frac{1}{2\pi\,\sqrt{g/4\pi}}\,\Big(\Big(\frac{t+1/2}{g/2}\Big)^2 - 1\Big)^{-1/4} \\ &\cdot\ \exp\Big(-\Big((t+1/2)\,\mathrm{arccosh}\Big(\frac{t+1/2}{g/2}\Big)\Big) - \frac{g}{2}\,\Big(\Big(\frac{t+1/2}{g/2}\Big)^2 - 1\Big)^{1/2}\Big) \\ =\ & \frac{1}{g}\,\Big(\Big(\frac{t+1/2}{g/2}\Big)^2 - 1\Big)^{-1/4}\,\exp(-\varphi(t+1/2\,;\,g/2))\ . \end{aligned} \tag{C7}$$

On the range $(t+1/2) \leq \sqrt{2}\,(g/2)$ we replace the factor $(((t+1/2)/(g/2))^2 - 1)^{-1/4}$ by 1 at the expense of an error whose impact has been assessed in Appendix B, see around Eq. (B8). For $(t+1/2) \geq \sqrt{2}\,(g/2)$, we have

$$\Big(\Big(\frac{t+1/2}{g/2}\Big)^2 - 1\Big)^{-1/4} \leq 2^{1/4}\Big(\frac{g/2}{t+1/2}\Big)^{1/2}\ . \tag{C8}$$

Hence, we estimate

$$|j_t(g/2)| \leq \frac{2^{1/4}}{g}\,\Big(\frac{g/2}{t+1/2}\Big)^{1/2}\,\exp(-\varphi(t+1/2\,;\,g/2))\ , \qquad t+1/2 \geq g/2\ . \tag{C9}$$

Combining this with the estimate in Eq. (C5), we arrive at

$$|b_t(g)| \leq 2^{5/4}\,\frac{v_0}{u_0}\,\Big(\frac{t+1/2}{g/2}\Big)^{1/2}\,\exp(-\varphi(t+1/2\,;\,g/2))\ , \qquad g/2 \leq t+1/2 \leq g/2v_0\ . \tag{C10}$$

We proceed in a similar way on the range $t+1/2 \geq g/2v_0$ for $h_t^{(2)}(g/2v_0)$, using Debye's asymptotic result, [2], 10.19.3, p. 231

$$Y_\nu(\nu\,\mathrm{sech}\,\alpha) \approx \frac{\exp(\nu(\alpha - \tanh\alpha))}{(\frac{1}{2}\,\pi\nu\tanh\alpha)^{1/2}} \tag{C11}$$

with $\nu = t+1/2$ and $\nu\,\mathrm{sech}\,\alpha = g/2v_0$. The right-hand side of Eq. (C11) equals

$$\Big(\frac{4v_0}{\pi g}\Big)^{1/2}\,\Big(\Big(\frac{t+1/2}{g/2v_0}\Big)^2 - 1\Big)^{-1/4}\,\exp(\varphi(t+1/2\,;\,g/2v_0))\ . \tag{C12}$$

Then from Eq. (40) and ignoring the relatively small quantity $J_{t+1/2}(g/2v_0)$, we estimate

$$\begin{aligned}|h_t^{(2)}(g/2v_0)| &\approx \frac{2v_0}{g}\,\Big(\Big(\frac{t+1/2}{g/2v_0}\Big)^2-1\Big)^{-1/4}\,\exp(\varphi(t+1/2\,;\,g/2v_0))\\ &\le \frac{2^{5/4}v_0}{g}\,\Big(\frac{g/2v_0}{t+1/2}\Big)^{1/2}\,\exp(\varphi(t+1/2\,;\,g/2v_0))\ ,\\ &\qquad\qquad t+1/2\ge g/2v_0\ ,\end{aligned}\tag{C13}$$

where the factor $(((t+1/2)/(g/2v_0))^2-1)^{-1/4}$ has been dealt with in the same way as with the corresponding factor in Eq. (C7).

Combining Eqs. (C9, C13), we get the estimate

$$\begin{aligned}|b_t(g)| &\le 2^{3/2}\,\frac{v_0}{u_0}\,\Big(\frac{t+1/2}{g/2}\Big)^{1/2}\,\exp(-\varphi(t+1/2\,;\,g/2))\\ &\quad\cdot\Big(\frac{g/2v_0}{t+1/2}\Big)^{1/2}\,\exp(\varphi(t+1/2\,;\,g/2v_0))\ ,\qquad t+1/2\ge g/2v_0\ .\end{aligned}\tag{C14}$$

We have established now the estimates in Eqs. (C6, C10, C14) on $|b_t(g)|$ on the ranges $0\le t+1/2\le g/2$, $g/2\le t+1/2\le g/2v_0$ and $t+1/2\ge g/2v_0$, respectively. According to Appendix A, Eqs. (A20, A28), extended to all $x\ge 0$ at the expense of a negligible error when $\alpha=1/2$, see end of Appendix A, we thus have

$$|b_t(g)|\le 4\,\frac{v_0}{u_0}\ ,\qquad 0\le t+1/2\le g/2\ ,\tag{C15}$$

$$|b_t(g)|\le 2^{5/4}\,\frac{v_0}{u_0}\,\exp(-\varphi(t\,;\,g/2))\ ,\qquad g/2\le t+1/2\le g/2v_0\ ,\tag{C16}$$

$$|b_t(g)|\le 2^{3/2}\,\frac{v_0}{u_0}\,\exp(-\varphi(t\,;\,g/2)+\varphi(t\,;\,g/2v_0))\ ,\quad t+1/2\ge g/2v_0\ .\tag{C17}$$

Since $\varphi(t\,;\,g/2)=0$ for $t\le g/2$ and $\varphi(t\,;\,g/2v_0)=0$ for $t\le g/2v_0$, the three estimates in Eqs. (C15–C17) can be combined into a single one, viz.

$$|b_t(g)|\le 4\,\frac{v_0}{u_0}\,\exp(-\varphi(t\,;\,g/2)+\varphi(t\,;\,g/2v_0))\ ,\qquad t\ge 0\ .\tag{C18}$$

Using this in Eq. (C3), we see that $|c_t|$ is estimated by

$$4a_0\,w_0\,\exp(-\varphi(t\,;\,g/2)+\varphi(t\,;\,g/2v_0))\ ,\qquad t\ge 0\ ,\tag{C19}$$

where

$$w_0 = \frac{v_0}{u_0} = \frac{1}{1+\sqrt{1-s_0^2}} \ . \tag{C20}$$

The validity of Eq. (C19) as a bound for $|c_t|$ should be subjected to the same side comment as validity of Eq. (B7) for the Jinc function $J_{h+1}(2\pi r)/2\pi r$. There are now two relatively small regions, around $t+1/2 = g/2$ and around $t+1/2 = g/2v_0$, where the bound in Eq. (C19) is too low by a factor that increases very slowly as $g \to \infty$. Fortunately, we consider values of $s_0 \leq 0.99$, which implies that $v_0 \leq 0.75$, so that the exceptional regions do not overlap as $g \to \infty$.

For the sake of computation of the quantities b_k in Eq. (38), involving the products of spherical Bessel and Hankel functions, with a specified accuracy, we note the bounds for $k \geq 0$

$$|j_k(g/2)| \leq \frac{2}{g} \ , \ \ |h_k(g/2v_0)| \leq \frac{2^{7/4}v_0}{g} \exp\left(\varphi(k; g/2v_0)\right) \ . \tag{C21}$$

The first bound follows from Eqs. (C4), (C9) and (A20) with $\alpha = 1/2$ and $c = g/2$, and the second bound follows from Eqs. (C5), (C13) and (A27) with $\alpha = 1/2$ and $c = g/2v_0$. Since $|j_k(f/2)| \leq 1$, we may replace the argument $g/2$ in the first inequality in Eq. (C21) by $f/2$. In the second inequality, we can replace the argument $g/2v_0$ by $f/2v_0$ only when $|f/v_0| \geq 1$.

D Proof of validity of truncation rules

In this appendix, we give the proofs for the results in Subsecs. 2.3–2.4 on truncation rules. We first show that the quantity in Eq. (11) is less than $\varepsilon \in (0,1)$ when H and T are chosen according to Eq. (24) with B given in Eq. (23).

From Appendix A, we have for $d \geq c > 0$ that

$$\varphi(x\,;\,c) \geq \varphi(x\,;\,c) - \varphi(x\,;\,d) \geq \gamma x - c\sinh(\gamma) \ , \qquad x \geq 0 \ , \tag{D1}$$

where $\gamma \leq \ln(d/c)$. Taking $x = h+1$, $c = 2\pi R$, $d = ec$ (so that $\gamma \leq 1$), we get by taking $\gamma = 1$

$$\varphi(h+1\,;\,2\pi R) \geq h+1 - 2\pi R\sinh(1) \ . \tag{D2}$$

The right-hand side of Eq. (D2) exceeds B of Eq. (23) when $h+1 \geq H$, and then from Eq. (14)

$$\left|\frac{J_{h+1}(2\pi r)}{2\pi r}\right| \leq \frac{\varepsilon}{4w_0 a_0} \ , \qquad h+1 \geq H \ . \tag{D3}$$

Next, for $x = t$, $c = g/2$, $d = g/2v_0$ in Eq. (D1) (so that $\gamma \leq \ln(1/v_0)$), we get by taking $\gamma = \min(1, \ln(1/v_0))$

$$\varphi(t\,;\, g/2) - \varphi(t\,;\, g/2v_0) \geq \gamma t - \tfrac{1}{2}\, g \sinh(\gamma)\ . \tag{D4}$$

The right-hand side of Eq. (D4) exceeds B of Eq. (23) when $t \geq T$, and then from Eq. (16)

$$|c_t| \leq 2\varepsilon\, \pi^2\, R\sqrt{R}\ , \qquad t \geq T\ . \tag{D5}$$

Since for all $h \geq 0$, $t \geq 0$ by Eqs. (14, 16)

$$\Big|\frac{J_{h+1}(2\pi r)}{2\pi r}\Big| \leq \frac{1}{2\pi^2\, R\sqrt{R}}\ , \qquad |c_t| \leq 4w_0 a_0\ , \tag{D6}$$

we find that

$$|c_t|\, \Big|\frac{J_{h+1}(2\pi r)}{2\pi r}\Big| < \varepsilon \tag{D7}$$

when $h + 1 \geq H$ and/or $t \geq T$. This means that the quantity in Eq. (11) is less than ε.

As to the dedicated truncation rule, we use continuity, monotonicity and convexity of $F(h, t)$ as a function of both h and t, see Eqs. (27–28). It thus follows easily that the right-hand side of Eq. (30) is less than ε when $h+1 > H$ or $t > T$ when H and T are chosen as $H = H_n^m = \max(h_1, h_2)+1$, $T = T_n^m = \max(t_1, t_2)$ (for the case that M in Eq. (31) $\leq$ B; otherwise we simply have $H = 1$, $T = 0$). Here the points (h_1, t_1) and (h_2, t_2) are found as the first and the last point $(h, 2t)$ on ∂S_n^m with $F(h, t) > B$ when inspecting the 4 line segments of the boundary ∂S_n^m in counterclockwise manner through integer h and t with h same parity as n. This means that with this choice of H and T the quantity in Eq. (10) is less than ε.

It also follows that $F(h, t)$ increases along both edge I and edge IV in Fig. 1 when $(h, 2t) \to \infty$. Also $F(h, t)$ increases along edge III when t increases and h is kept fixed at $|m|$. Therefore, the minimum M in Eq. (31) is to be found on edge II. On this edge II, it follows from convexity of F that the minimum is attained on a set of points $(n - 2t, 2t)$ with t in a closed interval contained in $[0, \frac{1}{2}\,(n - |m|)]$ (which reduces to a single point t when F is strictly convex on edge II).

E Asymptotics, bounds and truncation issues for coefficients of algebraic functions

We consider in this appendix the (computation of the) Zernike coefficients of the modified algebraic function

$$A(\rho) = a(\rho)\,\sqrt{1-s_0^2\rho^2} = \sum_{l=0}^{\infty} a_l\, R_{2l}^0(\rho)$$

$$= \ (1-s_0^2\rho^2)^{3/4}\,(1-s_{0,M}^2\rho^2)^{-3/4} + (1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{-1/4}\ , \tag{E1}$$

see Eqs. (3, 41). This A is the sum of two functions

$$a_{\alpha\beta}(\rho) = (1-s_\alpha^2\rho^2)^{\alpha}\,(1-s_\beta^2\rho^2)^{\beta} = \sum_{l=0}^{\infty} a_{l,\alpha\beta}\, R_{2l}^0(\rho)\ . \tag{E2}$$

We let for such an $a_{\alpha\beta}$

$$S = \max(s_\alpha, s_\beta)\ , \qquad s = \min(s_\alpha, s_\beta)\ , \tag{E3}$$

$$\Delta = \arg(S)\ , \qquad \delta = \arg(s)\ , \tag{E4}$$

so that

$$a_{\alpha\beta}(\rho) = (1-s^2\rho^2)^{\delta}\,(1-S^2\rho^2)^{\Delta}\ . \tag{E5}$$

Observe that in the cases in Eq. (E1) we have $\Delta+\delta=0$.

We consider the power series coefficients $r_{N,\alpha\beta}$ of $a_{\alpha\beta}(\rho)$, and the computation of the $a_{l,\alpha\beta}$ according to

$$a_{l,\alpha\beta} = \sum_{N=l}^{\infty} b_N(l)\, r_{N,\alpha\beta}\ ; \qquad b_N(l) = \frac{2l+1}{l+1}\binom{N}{l}\Big/\binom{N+l+1}{N}\ . \tag{E6}$$

It will be shown below that for $\delta\in(-1,1)$ and $N=1,2,...$

$$r_{N,\delta,-\delta} = \frac{1}{\pi}\sin(\pi\delta)\int\limits_{1/S^2}^{1/s^2}\Big(\frac{1-s^2x}{S^2x-1}\Big)^{\delta}\,\frac{dx}{x^{N+1}}\ . \tag{E7}$$

Hence, $r_{N,\delta,-\delta}$ has the sign of δ, and it will also be shown that for $\delta\in(0,1)$ and $N=0,1,...$

$$r_{N,\delta,-\delta} \ge |r_{N,-\delta,\delta}|\ . \tag{E8}$$

It follows easily from Eq. (E7) that $r_{N,\delta,-\delta}$ decreases as a function of $s\in(0,S]$ when $\delta>0$. Hence, for $\delta\in(0,1)$,

$$r_{N,\delta,-\delta} \le \lim_{s\downarrow 0} r_{N,\delta,-\delta} = C_{\rho^{2N}}[(1-S^2\rho^2)^{-\delta}]\ . \tag{E9}$$

Since the $b_N(l)$ in Eq. (E6) are all non-negative, it follows from Eqs. (E8, E9) that for $\delta \in (0,1)$

$$|a_{l,-\delta,\delta}| \le a_{l,\delta,-\delta} \le ZC_l\left[(1-S^2\rho^2)^{-\delta}\right] , \tag{E10}$$

where ZC_l abbreviates "the $l^{\rm th}$ Zernike coefficient of the function in [ ]".

We shall show below that for $\Delta \in (-1,1)$, the asymptotic behavior of the Zernike coefficients of $(1-S^2\rho^2)^\Delta$ is given by

$$ZC_l\left[(1-S^2\rho^2)^\Delta\right] \sim \frac{2\sqrt{\pi}}{\Gamma(-\Delta)}\,\frac{(1-S^2)^{\frac{1}{2}\Delta+\frac{1}{4}}}{1+\sqrt{1-S^2}}\,\frac{V^l}{(l+1)^{\Delta+\frac{1}{2}}} \tag{E11}$$

as $l \to \infty$, where

$$V = \frac{1-\sqrt{1-S^2}}{1+\sqrt{1-S^2}} . \tag{E12}$$

For $\Delta = 0$, we have $\Gamma(-\Delta) = \infty$, and the right-hand side of Eq. (E11) vanishes. For $\Delta = -1/2$, the right-hand side of Eq. (E11) is exactly equal to $ZC_l\left[(1-S^2\rho^2)^{-1/2}\right]$, see [1], Eq. (134), and also for the case that $\Delta = 1/2$, there is good agreement between $ZC_l\left[(1-S^2\rho^2)^{1/2}\right]$, given by [1], Eq. (135), and the right-hand side of Eq. (E11).

The maximum modulus of the right-hand side of Eq. (11) occurs at $l = 0$ and decreases in $l = 0, 1, ...$ unless $\Delta < -1/2$ and S is extremely close to 1. In the relevant case that $\Delta = -3/4$, monotonicity of the modulus is guaranteed as long as $V \le 2^{-1/4}$, i.e., $S \le 2(2^{-1/8}+2^{1/8})^{-1} = 0.9963$.

In Sec. 3, Eqs. (52–54), it is required to find for a given $\eta > 0$, $E > 0$, and $\Delta \in (-1,0)$, $V \in (0,1)$ an $L > 0$ such that

$$l \ge L \Rightarrow \frac{E\,V^l}{(l+1)^{\Delta+1/2}} < \eta . \tag{E13}$$

Under the monotonicity assumption, an approximation of the required L is found by rewriting the equation $E\,V^L(L+1)^{-\Delta-1/2} = \eta$ for L as

$$L = \frac{\ln(E/\eta) - (\Delta+1/2)\ln(L+1)}{\ln(1/V)} , \tag{E14}$$

and to iterate this equation twice, starting with $L = 0$. This yields the quantity at the right-hand side of Eq. (54), with $\Delta = -\delta$.

We next address the truncation issue when computing $a_{l,\alpha\beta}$ according to Eq. (E6). It is sufficient to consider this for the function $(1-S^2\rho^2)^\Delta$ with $\Delta \in (-1,0)$, see Eqs. (E9, E10). We have

$$C_{\rho^{2N}}\left[(1-S^2\rho^2)^\Delta\right] = \frac{\Gamma(N-\Delta)\,S^{2N}}{\Gamma(-\Delta)\,\Gamma(N+1)} \sim \frac{S^{2N}}{\Gamma(-\Delta)\,N^{\Delta+1}} . \tag{E15}$$

Thus, the terms in the series in Eq. (E6) are approximated as

$$\frac{(2l+1)\Gamma^2(N+1)}{\Gamma(N+1+l)\Gamma(N+1-l)}\frac{S^{2N}}{\Gamma(-\Delta)N^{\Delta+1}(N+l+1)}\ . \tag{E16}$$

For a given N, the maximum of

$$\frac{(2l+1)\Gamma^2(N+1)}{\Gamma(N+1+l)\Gamma(N+1-l)}\ ,\quad l=0,\ 1,\ \cdots,\ N\ , \tag{E17}$$

is approximately $\sqrt{2N/e}$ and occurs at l near $\sqrt{N/2}$. Thus the truncation errors $\sum_{N=N_L}^{\infty} b_n(l) r_{N,\alpha\beta}$ for the series in Eq. (E6) are all bounded by

$$\sqrt{\frac{2}{e}}\frac{1}{\Gamma(-\Delta)}\sum_{N=N_L}^{\infty}\frac{S^{2N}}{N^{\Delta+3/2}}\ . \tag{E18}$$

Now, by partial integration and $\Delta+3/2>0$,

$$\begin{aligned}\sum_{N=N_L}^{\infty}\frac{S^{2N}}{N^{\Delta+3/2}} &\approx \int_{N_L}^{\infty}\frac{e^{-x\ln(S^{-2})}}{x^{\Delta+3/2}}\mathrm{d}x < \frac{e^{-N_L\ln(S^{-2})}}{N_L^{\Delta+3/2}\ln(S^{-2})}\\ &= \frac{S^{2N_L}}{N_L^{\Delta+3/2}\ln S^{-2}} < \frac{S^{2N_L}}{N_L^{\Delta+3/2}(1-S^2)}\ ,\end{aligned} \tag{E19}$$

and so the quantity in Eq. (E18) is realistically bound by

$$\sqrt{\frac{2}{e}}\frac{S^{2N_L}}{\Gamma(-\Delta)N_L^{\Delta+3/2}(1-S^2)}\ . \tag{E20}$$

We recall that $a_{l,\alpha\beta}$ are required for all $l\le L$, where L satisfies $V^L=\frac{\eta}{E}(L+1)^{\Delta+1/2}$, see Eq. (E13), with

$$E=\frac{2\sqrt{\pi}}{\Gamma(-\Delta)}\frac{(1-S^2)^{\frac{1}{2}\Delta+\frac{1}{4}}}{1+\sqrt{1-S^2}}\ . \tag{E21}$$

We now propose to take $N_L=2L/\sqrt{1-S^2}$. Then

$$S^{2N_L}=\exp\left(\frac{2L\ln(S^2)}{\sqrt{1-S^2}}\right)<V^L=\frac{\eta}{E}(L+1)^{\Delta+1/2}\ , \tag{E22}$$

where the inequality in Eq. (E22) follows from

$$\frac{2}{y}\ln(1-y^2)<\ln\left(\frac{1-y}{1+y}\right)\ ,\quad 0<y<1\ , \tag{E23}$$

with $y = \sqrt{1-S^2}$. Thus, all truncation errors are bounded by

$$\sqrt{\frac{2}{e}}\frac{(L+1)^{\Delta+1/2}\eta}{E\Gamma(-\Delta)N_L^{\Delta+3/2}(1-S^2)} \approx \frac{\eta\sqrt{2/e}}{2\sqrt{\pi}\,2^{\Delta+3/2}}\frac{1+\sqrt{1-S^2}}{L\sqrt{1-S^2}} \ , \tag{E24}$$

where we have used the definitions of E and N_L. This quantity (E24) is well below $\eta/2$ for somewhat larger values of L. In fact, from Eq. (E22) and in the relevant case $\Delta = -3/4$ (so that $(L+1)^{\Delta+1/2} \leq 1$)

$$\begin{aligned} \ln\left(\frac{\eta}{E}\right) > \ln V^L &= L\ln\left(\frac{1-\sqrt{1-S^2}}{1+\sqrt{1-S^2}}\right) \\ &= L\ln\left(1-\frac{2\sqrt{1-S^2}}{1+\sqrt{1-S^2}}\right) \approx -\frac{2L\sqrt{1-S^2}}{1+\sqrt{1-S^2}} \ , \end{aligned} \tag{E25}$$

and so the quantity in Eq. (E24) is realistically estimated at ($\Delta = -3/4$)

$$\frac{\eta\sqrt{2/e}}{\sqrt{\pi}\,2^{\Delta+3/2}\ln(E/\eta)} = \frac{0.2878\,\eta}{\ln(E/\eta)} \ . \tag{E26}$$

We still owe the reader a proof of the results in Eq. (E7, 11). As to Eq. (E7), we consider the general case in Eq. (E5). Setting $x = \rho^2$, we have by Cauchy's formula

$$r_{N,\alpha\beta} = C_{x^N}\left[(1-s^2x)^\delta(1-S^2x)^\Delta\right] = \frac{1}{2\pi i}\oint\frac{(1-s^2z)^\delta(1-S^2z)^\Delta}{z^{N+1}}\,dz \ , \tag{E27}$$

with integration contour a circle of radius $< 1/S^2$ in positive sense. We choose principal values of the roots $(1-s^2z)^{\delta,\Delta}$, and we deform the contour so that the positive real axis from the first branch point $z = 1/S^2$ onwards, passing along the second branch point $z = 1/s^2$, to $z = \infty$ is enclosed. When $N = 1, 2, ...$ and $\delta, \Delta > -1$, $\delta + \Delta < 1$, this can be done without problems. Since

$$(1-s^2(x\pm io))^\delta = (s^2x-1)^\delta\, e^{\mp\pi i\delta} \ , \quad x > 1/s^2 \ , \tag{E28}$$

$$(1-S^2(x\pm io))^\Delta = (S^2x-1)^\Delta\, e^{\mp\pi i\Delta} \ , \quad x > 1/S^2 \ , \tag{E29}$$

it follows that

$$\begin{aligned} r_{N,\alpha\beta} &= \frac{1}{2\pi i}\int\limits_{1/S^2}^{1/s^2}(1-s^2x)^\delta\,(S^2x-1)^\Delta\,(e^{-\pi i\Delta}-e^{\pi i\Delta})\,\frac{dx}{x^{N+1}} \\ &\quad+\frac{1}{2\pi i}\int\limits_{1/s^2}^{\infty}(s^2x-1)^\delta\,(S^2x-1)^\Delta\,(e^{-\pi i(\delta+\Delta)}-e^{\pi i(\delta+\Delta)})\,\frac{dx}{x^{N+1}} \\ &= \frac{-\sin\pi\Delta}{\pi}\int\limits_{1/S^2}^{1/s^2}\frac{(1-s^2x)^\delta\,(S^2x-1)^\Delta}{x^{N+1}}\,dx \\ &\quad-\frac{\sin(\delta+\Delta)}{\pi}\int\limits_{1/s^2}^{\infty}\frac{(s^2x-1)^\delta\,(S^2x-1)^\Delta}{x^{N+1}}\,dx\ . \end{aligned} \tag{E30}$$

When $\delta+\Delta=0$, the second integral in Eq. (E30) is canceled, and we get Eq. (E7).

We now show Eq. (E8). We have for $\delta\in(0,1)$

$$r_{0,\delta,-\delta}=r_{0,-\delta,\delta}=1\ , \qquad r_{1,\delta,-\delta}=-r_{1,-\delta,\delta}=(S^2-s^2)\,\delta \tag{E31}$$

as readily follows from Eq. (E5). From Eq. (E7) we have

$$\begin{aligned} &r_{N,\delta,-\delta}+r_{N,-\delta,\delta} \\ &= \frac{\sin\pi\delta}{\pi}\int\limits_{1/S^2}^{1/s^2}\Big[\Big(\frac{1-s^2x}{S^2x-1}\Big)^\delta-\Big(\frac{1-s^2x}{S^2x-1}\Big)^{-\delta}\Big]\frac{dx}{x^{N+1}}\ , \end{aligned} \tag{E32}$$

and this vanishes when $N=1$. The function $g(x)$ in [ ] in the integral in Eq. (E32) decreases in $x\in[1/S^2,1/s^2]$ since $\delta>0$, and has there a single zero, at $x=2/(s^2+S^2)=:x_0$. Then for $N>1$, we have

$$\int\limits_{1/S^2}^{1/s^2}\frac{g(x)}{x^{N+1}}\,dx=\int\limits_{1/S^2}^{1/s^2}\frac{g(x)}{x^2}\Big(\frac{1}{x^{N-1}}-\frac{1}{x_0^{N-1}}\Big)\,dx\ , \tag{E33}$$

and this is positive since the integrand of the second integral is positive for all $x\neq x_0$. Since $r_{N,\delta,-\delta}$ is positive and $r_{N,-\delta,\delta}$ is negative, see Eq. (E7), we get Eq. (E8).

We finally show the asymptotic result in Eq. (E11). We have from $R_{2l}^0(\rho) = P_l(2\rho^2-1)$, where P_l is the Legendre polynomial of degree l, the substitutions

$$x = 2\rho^2 - 1 \in [-1,1]\ , \qquad a = 1 - \tfrac{1}{2}\,S^2\ , \quad b = \tfrac{1}{2}\,S^2\ , \tag{E34}$$

Rodriguez' formula

$$P_l(x) = \frac{(-1)^l}{2^l\,l!}\Big(\frac{d}{dx}\Big)^l\,[(1-x^2)^l]\ , \tag{E35}$$

and l partial integrations, that

$$\begin{aligned} ZC_l\,[(1-S^2\rho^2)^\Delta] &= 2(2l+1)\int\limits_0^1 (1-S^2\rho^2)^\Delta\,R_{2l}^0(\rho)\,\rho\,d\rho \\ &= \frac{(l+1/2)\,\Gamma(l-\Delta)}{l!\,\Gamma(-\Delta)}\int\limits_{-1}^1 (a-bx)^\Delta\Big(\frac{b}{2}\,\frac{1-x^2}{a-bx}\Big)^l\,dx\ . \end{aligned} \tag{E36}$$

The remaining integral in Eq. (E36) can be approximated by using Laplace's method. The stationary point of the integrand is found by setting $((1-x^2)/(a-bx))' = 0$, and this yields $x = V$ when we restore the parameter S, see Eqs. (E34, 12). We have furthermore

$$a - bx|_{x=V} = \sqrt{1-S^2}\ , \qquad \frac{b}{2}\,\frac{1-x^2}{a-bx}\Big|_{x=V} = V\ , \tag{E37}$$

and

$$\Big(\ln\Big(\frac{1-x^2}{a-bx}\Big)\Big)''\Big|_{x=V} = -\,\frac{(1+\sqrt{1-S^2})^2}{2\,\sqrt{1-S^2}}\ . \tag{E38}$$

This then yields

$$\begin{aligned} ZC_l\,[(1-S^2\rho^2)^\Delta] &\approx \frac{(l+1/2)\,\Gamma(l-\Delta)}{\Gamma(l+1)\,\Gamma(-\Delta)}\,(\sqrt{1-S^2})^\Delta\,V^L \\ &\quad\cdot\int\limits_{-\infty}^{\infty}\exp\Big(-l\,\frac{(1+\sqrt{1-S^2})^2}{4\,\sqrt{1-S^2}}\,(x-V)^2\Big)\,dx \\ &= \frac{2\sqrt{\pi}}{\Gamma(-\Delta)}\,\frac{(1-S^2)^{\frac{1}{2}\Delta+\frac{1}{4}}}{1+\sqrt{1-S^2}}\,\frac{(l+1/2)\,\Gamma(l-\Delta)}{\Gamma(l+1)\,l^{1/2}}\,V^l \\ &\approx \frac{2\sqrt{\pi}}{\Gamma(-\Delta)}\,\frac{(1-S^2)^{\frac{1}{2}\Delta+\frac{1}{4}}}{1+\sqrt{1-S^2}}\,\frac{V^l}{(l+1)^{\Delta+\frac{1}{2}}}\ , \end{aligned} \tag{E39}$$

as required.

References

[1] S. van Haver and A.J.E.M. Janssen, "Advanced analytic treatment and efficient computation of the diffraction integrals in the Extended Nijboer-Zernike theory", J. Europ. Opt. Soc. Rap. Public. **8**, 13044 (2013).

[2] F.W.J. Olver, D.W. Lozier, R.F. Boisvert and C.W. Clark, *NIST Handbook of Mathematical Functions* (Cambridge University Press, Cambridge, United Kingdom, 2010).

[3] G.N. Watson, *Theory of Bessel Functions* (Merchant Books, USA, 2008).