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

**S. van Haver**
svenvanhaver@gmail.com

S[&]T Experts Pool (STEP), P.O. Box 608, 2600 AP Delft, The Netherlands
Optics Research Group, Faculty of Applied Sciences, Technical University Delft, Van der Waalsweg 8, 2628 CH Delft, The Netherlands

**A .J. E. M. Janssen**
a.j.e.m.janssen@tue.nl

Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

The point-spread function (PSF) is used in optics for design and assessment of the imaging capabilities of an optical system. It is therefore of vital importance that this PSF can be calculated fast and accurately. In the past 12 years, the Extended Nijboer-Zernike (ENZ) approach has been developed for the purpose of semi-analytic evaluation of the PSF, for circularly symmetric optical systems, in the focal region. In the earliest ENZ-years, the Debye approximation of the diffraction integral, by which the PSF is given, was considered for the very basic situation of a low-NA optical system and relatively small defocus values, so that a scalar treatment was allowed with a focal factor comprising a quadratic function in the exponential. At present, the ENZ-method allows calculation of the PSF in low- and high-NA cases, in scalar form and for vector fields (including polarization), for large wave-front aberrations, including amplitude non-uniformities, using a quasi-spherical phase focal factor in a virtually unlimited focal range around the focal plane, and no limitations in the off-axis direction. Additionally, the application range of the method has been broadened and generalized to the calculation of aerial images of extended objects by including the finite distance of the object to the entrance pupil. Also imaging into a multi-layer is now possible by accounting for both forward and backward propagation in the layers.
In the advanced ENZ-approach, the generalized, complex-valued pupil function is developed into a series of Zernike circle polynomials, with exponential azimuthal dependence (having cosine/sine azimuthal dependence as special cases). For each Zernike term, the diffraction integral reduces after azimuthal integration to an integral that can be expressed as an infinite double series involving spherical Bessel functions, accounting for the parameters of the optical system and the defocus value, and Jinc functions comprising the radial off-axis value. The contribution of the present paper is the formulation of truncation rules for these double series expressions, with a general rule valid for all circle polynomials at the same time, and a dedicated rule that takes into account the degree and the azimuthal order of the involved circle polynomials to significantly reduce computational cost in specific cases. The truncation rules are based on effective bounds and asymptotics (of the Debye type) for the mentioned spherical Bessel functions and Jinc functions, and show feasibility of computation of practically all diffraction integrals that one encounters in the ENZ-practice. Thus it can be said that the advanced ENZ-theory is more or less completed from the computational point of view by the achievements of the present paper.
[DOI: http://dx.doi.org/10.2971/jeos.2014.14042]

**Keywords:** Diffraction integral, point-spread function (PSF), advanced ENZ-theory, double series, Jinc functions, Debye asymptotics of Bessel functions

## 1 INTRODUCTION AND OVERVIEW

The advanced ENZ-theory of diffraction integrals aims at the computations of the Debye approximation of the Rayleigh integral for the optical point-spread function of radially symmetric optical systems that range from as basic as having low-NA and small defocus values to advanced high-NA systems, with vector fields and polarization, that are meant for imaging of extended objects into a multi-layer structure. Recently, a review of the advanced ENZ-theory has been given in [1]. In this review paper [1], the evolution of the form of the diffraction integral in the process of advancing from basic systems to highly sophisticated systems has been described. Key ENZ-papers in this context are [2]–[8], along with the thesis [9] and the book chapter [10], the latter containing a review of the ENZ-theory until 2008. In [2], the first semi-analytic result for the point-spread function in the focal region has been given for low-NA systems and defocus values that may range to up to 8 focal depths, and in [3] the potential of this semi-analytic result for design and assessment of optical systems has been indicated. The papers [4]–[6] deal with high-NA systems, including vector fields and polarization, and develop the high-NA ENZ-theory both for forward computation [4] and for aberration and birefringence retrieval [5, 6]. In [7], imaging of extended objects located at a finite distance from the entrance pupil of the optical system is considered, which was further extended in [8], to also include image formation inside a multi-layered focal region.

The nature of the semi-analytic result for computing the point-spread function evolved during the process from an infinite series containing products of powers $f^n$ of the defocus value $f$ and Jinc functions, comprising the radial distance $r$ in im-

Received August 13, 2014; revised ms. received September 17, 2014; published September 27, 2014 ISSN 1990-2573

age space, to an infinite double series containing products of spherical Bessel functions (and Hankel functions for high-NA systems) and Jinc functions, see [11]. These latter series have the advantage that the restriction to small-to-moderately-large defocus values can be removed, a point that was already called attention to by Boersma [12] in 1962 for, what we would call now, low-NA, aberration-free optical systems. This approach of Bessel-Bessel series representation of the point-spread function has been worked out in full detail and in all generality in [1], thereby removing clumsy (linearization) procedures in [11] and ad hoc measures in [4]–[8] to treat front factors in the diffraction integrals. The present paper continues and finishes the investigations in [1], in the sense that we develop efficient rules for the truncation of the Bessel-Bessel type series, that allow computation of (almost) all diffraction integrals arising in the ENZ-context with arbitrary, guaranteed accuracy in an acceptable time.

As in classical Nijboer-Zernike theory, the generalized pupil-function $P(\rho,\theta)$ is developed into a series of Zernike terms,

$$P(\rho,\theta) = A(\rho,\theta)\exp\left[i\Phi(\rho,\theta)\right] = \sum_{n,m}\beta_n^m Z_n^m(\rho,\theta)\,,$$
$$0\le\rho\le 1,\ 0\le\theta\le 2\pi\,, \tag{1}$$

in which

$$Z_n^m(\rho,\theta) = R_n^{|m|}(\rho)\exp\left[im\theta\right]\,,\quad 0\le\rho\le 1,\ 0\le\theta\le 2\pi\,, \tag{2}$$

are the Zernike circle polynomials, with exponential azimuthal dependence, where $n$ and $m$ are integer such that $n-|m|$ is even and non-negative. The contribution to the Debye diffraction integral for the point-spread functions is then given per term $\beta_n^m Z_n^m$ as

$$\beta_n^m\int_0^1\int_0^{2\pi} G(\rho;f)\exp\left[2\pi i\rho r\cos(\theta-\phi)\right]Z_n^m(\rho,\theta)\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\,, \tag{3}$$

in which $(\nu,\mu)=(\rho\cos\theta,\rho\sin\theta)$ and $(x,y)=(r\cos\phi,r\sin\phi)$ are the Cartesian-polar coordinates in the exit pupil and the focal planes, respectively, and the front factor $G(\rho;f)$ is a radially symmetric function determined by the parameters of the optical system and the defocus value $f$. By integration over the azimuthal variable $\theta$, the expression in Eq. (3) takes the form

$$\beta_n^m 2\pi i^m\int_0^1 G(\rho;f)J_m(2\pi\rho r)R_n^{|m|}(\rho)\rho\,\mathrm{d}\rho\,. \tag{4}$$

The front factor $G(\rho;f)$ in the remaining integral $I_n^m$ in Eq. (4) is the product of a radially symmetric algebraic factor $a(\rho)$ and a focal factor $f(\rho)$ (comprising the defocus value $f$). This front factor is developed in [1] in a systematic way as a series involving radially symmetric Zernike terms from the Zernike expansions of the algebraic factor $a(\rho)$ and the focal factor $f(\rho)$ using to Clebsch-Gordan coefficients related quantities to linearize products of Zernike terms. Then the remaining integral in Eq. (4) can be expressed as a doubly infinite series

$$I = I_n^m = \sum_{h,t} A_{2t,n,h}^{0mm}(-1)^{\frac{h-m}{2}}\,c_t\,\frac{J_{h+1}(2\pi r)}{2\pi r}\,. \tag{5}$$

In Eq. (5), $m$ 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. (5) 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. (5) 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 $\mathrm{Jinc}_h$ 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. (5) all terms $h,t$ with $0\le h\le H$, $0\le t\le T$ 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 finite magnification. Explicitly, $I$ assumes the form

$$I = I_{VM} = I_{n,VM}^m = \int_0^1 a(\rho)\,f(\rho)\,p(\rho)\,b(\rho)\,\rho\,d\rho\,, \tag{6}$$

where

$$a(\rho) = \frac{(1-s_0^2\rho^2)^{1/2}+(1-s_{0,M}^2\rho^2)^{1/2}}{(1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{3/4}}\,, \tag{7}$$

$$f(\rho) = \exp\left[\frac{if}{u_0}\left(1-\sqrt{1-s_0^2\rho^2}\right)\right]\,, \tag{8}$$

$$p(\rho) = R_n^{|m|}(\rho)\,, \qquad b(\rho) = J_m(2\pi r\rho)\,, \tag{9}$$

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 Section 2 we consider rules for the truncation of the double series in Eq. (5) 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 Section 3 we consider the truncation issues associated with the computation of the structural quantities. In Section 4 the whole computation scheme using the general truncation rule is summarized. In Section 5 we illustrate the performance of the truncation rules by plotting actually achieved accuracy and computation times against required accuracy. In Section 6 we present our conclusions.

An extended version of the present paper is available online in the form of the arXiv publication [13]. It contains a detailed summary of the computation scheme and both versions of the truncation rules, see [13], Sec. 4. Next, in [13], Sec. 5, an extended performance evaluation, comprising 10 pages of plots, of the truncation rules is given. Furthermore, the Appendices A-E in [13] contain all the mathematical details, omitted in the present paper, concerning bounding the Jinc functions and structural quantities and computation and asymptotic behavior of the latter.

## 2 TRUNCATION RULES FOR THE DOUBLE SERIES FOR $I_{VM}$

### 2.1 Double series for $I_{VM}$ and truncation strategy

We have

$$I_{VM} = \sum_{h,t} A^{0mm}_{2t,n,h} (-1)^{\frac{h-m}{2}}\, c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \tag{10}$$

as in Eq. (5), where $c_t$ are the Zernike coefficients of the front factor $a(\rho)\,f(\rho)$, with $a(\rho)$ and $f(\rho)$ as in Eqs. (7)–(8) so that

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

$$= \sum_{t=0}^{\infty} c_t\, R^0_{2t}(\rho)\,. \tag{11}$$

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

$$\sum_h A^{0mm}_{2t,n,h} = 1 = \sum_t \frac{2t+1}{h+1}\, A^{0mm}_{2t,n,h}\,. \tag{12}$$

In Subsection 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 $\varepsilon$ that are relevant in the optical practice, the double series in Eq. (10) 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. (10) by

$$\sum_{h+1\le H, t\le T} A^{0mm}_{2t,n,h} (-1)^{\frac{h-m}{2}}\, c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \tag{13}$$

is safely bounded by

$$\max_{(h,2t)\in S_n^m\,;\, h+1>H \text{ or } t>T} \left| c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \right|, \tag{14}$$

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

In the general truncation rule we devise, the dependence on $n$ and $m$ of the supporting set $S_n^m$ is totally ignored and the functions bounding $\mathrm{Jinc}_{h+1}$ and $c_t$ are replaced by simple functions allowing convenient determination of set points $H$ and $T$ for which

$$\max_{h+1>H \text{ or } t>T} \left| c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \right| \tag{15}$$

is below a specified $\varepsilon > 0$.

In the dedicated rule we devise, we use a more careful approximation of the bounding functions, and we include explicitly the supporting set $S_n^m$. It thus appears that an inspection of the product of the approximated bounding functions along the boundary $\partial\, S_n^m$ of the supporting set in the $(h, 2t)$-plane produces numbers $H = H_n^m$ and $T = T_n^m$ such that the quantity in Eq. (14) is below a specified $\varepsilon > 0$.

### 2.2 Bounding Jinc functions and structural quantities

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

$$\varphi(x\,;\,c) = \begin{cases} 0 & ,\quad 0 \le x \le c\,, \\ x\,\mathrm{arccosh}(x/c) - c\,\sqrt{(x/c)^2 - 1} & ,\quad x \ge c\,, \end{cases} \tag{16}$$

where $\mathrm{arccosh}(y) = \ln(y + \sqrt{y^2-1})$. In [13], Appendix B, the following is shown. Let $r > 0$, and set

$$R = \max\Big(\frac{1}{2\pi}\,, r\Big)\,. \tag{17}$$

Then

$$\left|\frac{J_{h+1}(2\pi r)}{2\pi r}\right| \le \frac{1}{2\pi^2\, R\,\sqrt{R}}\, \exp(-\varphi(h+1\,;\,2\pi R))\,. \tag{18}$$

The bound in Eq. (18) is valid for all $h \ge 0$, except for a small range of $h$'s near $2\pi r$ with $r \to \infty$. In fact, Eq. (18) is valid for

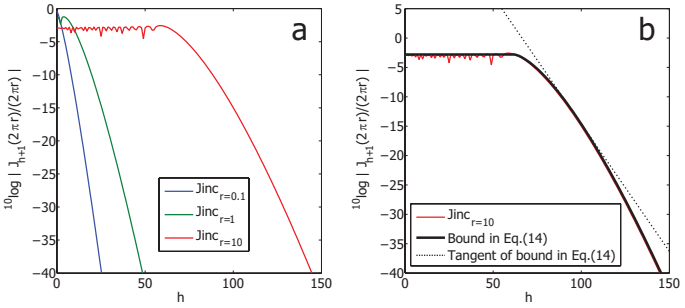


FIG. 1 (a) Plot of $\log_{10}|J_{h+1}(2\pi r)/\,2\pi r|$ as a function of $h = 0, 1, \cdots, 150$ for the case $r = 0.1$ (blue), 1 (green), 10 (red). (b) Plot of $\log_{10}|J_{h+1}(2\pi r)/2\pi r|$ as a function of $h = 0, 1, \cdots, 150$ case $r = 10$ (red), together with the $\log_{10}$ of the bound at the right-hand side of Eq. (18) (solid black) and the tangent line (dashed) corresponding to the right-hand side of Eq. (24).

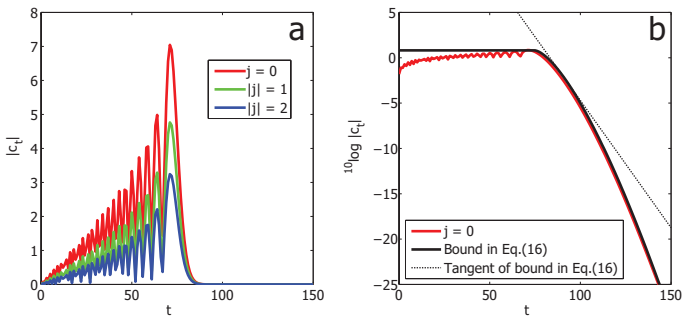


FIG. 2 (a) Plot of $\log_{10}|c_t|$ as a function of $t = 0, 1, \cdots, 150$, for $f = 150$, $s_0 = 0.95$, $s_{0,M} = 0.50$, where $c_t$ are the Zernike coefficients of the front factors that occur in accordance with [1], Eq. (30) for $|j| = 0$ (red), 1 (green), 2 (blue) and of which $c_t$ in Eq. (11) 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. (20) (solid black) and the tangent line (dashed) corresponding to the right-hand side of Eq. (25).

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 1(a), 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 1(b) 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. (18). The (asymptotic) maximum of $\log_{10}|J_{h+1}(2\pi r)/2\pi r|$ can be found from [13], 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 Figure 1(b) the linear function

$$\begin{aligned}&\log_{10}[\exp\{-(h+1-2\pi R\sinh(1))\}/(2\pi^2 R\sqrt{R})]\\ =&28.8387 - 0.4343h\end{aligned}$$

which is a tangent line of the bounding function, see Subsection 2.3.

For the structural quantities $c_t$ a similar result holds. In [13], Appendix C the following is shown. let $f$ be a real number, and set

$$g = \max(1, |f|)\,. \tag{19}$$

Then

$$|c_t| \leq 4w_0\, a_0\, \exp(-\varphi(t\,;\, g/2) + \varphi(t\,;\, g/2v_0))\,, \tag{20}$$

where

$$a_0 = 2\int_0^1 a(\rho)\,\sqrt{1 - s_0^2\rho^2}\,\rho\,d\rho \tag{21}$$

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

$$w_0 = \frac{1}{1+\sqrt{1-s_0^2}}\,, \qquad v_0 = \frac{1-\sqrt{1-s_0^2}}{1+\sqrt{1-s_0^2}}\,. \tag{22}$$

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 Eq. (22) by $s_{0,M}$ and change the right-hand side of Eq. (20) accordingly. The value of $a_0$ is in almost all cases well approximated by

$$A(\tfrac{1}{2}\sqrt{2}) \;\text{ or }\; \tfrac{1}{6}A(0) + \tfrac{2}{3}A(\tfrac{1}{2}\sqrt{2}) + \tfrac{1}{6}A(1) \tag{23}$$

(midpoint rule or Simpson rule for integration over $x = \rho^2$). The bound in Eq. (20) is shown in [13], Appendix C using a somewhat heuristic approach so as to arrive at manageable expressions. As with the bound in Eq. (18) there are small exceptional ranges of $t$ near $\frac{1}{2}g$ and $g \to \infty$, where Eq. (20) holds safe for a factor that grows to infinity very slowly as $g \to \infty$.

In Figure 2(a), 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 2(b) 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. (20). 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 2(b) 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 Subsection 2.4.

In Figure 3, we show the graph of $v_0$, as given in Eq. (22), 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.

## 2.3 General truncation rule

In [13], 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

$$\varphi(h+1\,;\,2\pi R) \geq \max(0, h+1-2\pi R\sinh(1))\,, \tag{24}$$

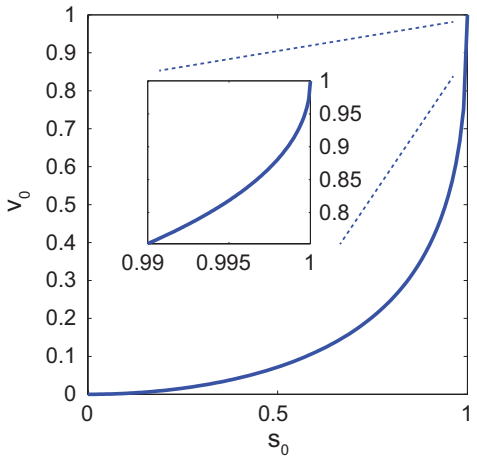


FIG. 3 Graph of $v_0 = (1-\sqrt{1-s_0^2})/(1+\sqrt{1-s_0^2})$ as a function of $s_0$, $0 \le s_0 \le 1$.

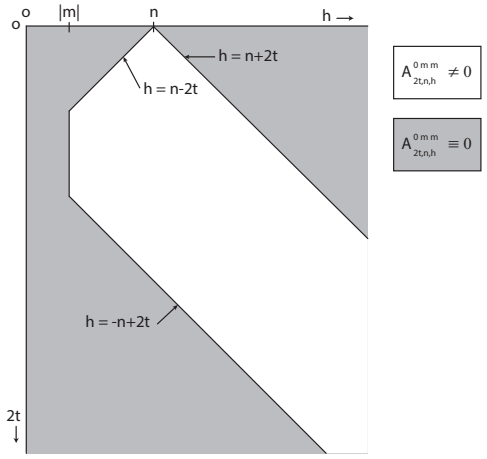


FIG. 4 For given integers $n$ and $m$ with $n-|m|$ even and non-negative, the unshaded set $h \ge |m|$, $|h-n| \le 2t \le h+n$ contains all points $(h,2t)$ with non-negative integer $h$ and $t$ such that $A^{0mm}_{2t,n,h} \neq 0$.

and

$$\varphi(t\,;\,g/2) - \varphi(t\,;\,g/2v_0) \ge \max(0, \gamma t - \tfrac{1}{2}\,g\sinh(\gamma))\,, \quad (25)$$

where

$$\gamma = \min(1, \ln(1/v_0))\,, \quad (26)$$

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

$$B = \max\Big(0, \ln\big(\frac{2w_0a_0}{\pi^2\,\varepsilon\,R\,\sqrt{R}}\big)\Big)\,. \quad (27)$$

Then the quantity in Eq. (15) is less than $\varepsilon$ when

$$\begin{aligned} T &= T^{\mathrm{gen}} = \frac{1}{\gamma}\,B + \tfrac{1}{2}\,g\,\frac{\sinh(\gamma)}{\gamma}\,, \\ H &= H^{\mathrm{gen}} = B + 2\pi R\sinh(1)\,. \end{aligned} \quad (28)$$

See [13], Appendix D for a proof.

By observing that we can write $T$ and $H$ in Eq. (28) as

$$\begin{aligned} T &= \tfrac{1}{2}\,g + \frac{1}{\gamma}\,B + \tfrac{1}{2}\,g\,\frac{\sinh(\gamma)-\gamma}{\gamma}\,, \\ H &= 2\pi R + B + 2\pi R(\sinh(1)-1)\,, \end{aligned} \quad (29)$$

where for $0 < \gamma \le 1$

$$0 < \frac{\sinh(\gamma)-\gamma}{\gamma} \le \sinh(1) - 1 = 0.1752\,, \quad (30)$$

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. (10). We also use better approximations for the functions $\varphi(h+1\,;\,2\pi R)$ and $\varphi(t\,;\,g/2) - \varphi(t\,;\,g/2v_0)$ than those on the left-hand sides of Eqs. (24–25). Thus we consider

$$F(h,t) = \varphi(h+1\,;\,2\pi R) + \varphi(t\,;\,g/2, g/2v_0)\,, \quad (31)$$

where

$$\varphi(t\,;\,g/2, g/2v_0) = \begin{cases} \varphi(t\,;\,g/2) & , \quad 0 \le t \le \tfrac{1}{2}\,g\cosh(\gamma_0)\,, \\ \gamma_0 t - \tfrac{1}{2}\,g\sinh(\gamma_0) & , \quad t \ge \tfrac{1}{2}\,g\cosh(\gamma_0)\,, \end{cases} \quad (32)$$

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 itself is convex in $t \le g/2$ but concave in $t \ge g/2v_0$, from below. The function $\varphi(h+1\,;\,2\pi R)$ is convex in $h \ge 0$. See [13], Appendix A.

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

$$h \ge |m|\,, \qquad |h-n| \le 2t \le h+n\,. \quad (33)$$

We consider the function $F(h,t)$ of Eq. (31) along $\partial\,S_n^m$ with continuous $t \ge 0$, $h \ge 0$. We have that $F(h,t)$ is non-negative and increasing and convex in both $h$ and $t$, and

$$\Big|c_t\,\frac{J_{h+1}(2\pi r)}{2\pi r}\Big| \le \frac{2w_0a_0}{\pi^2\,R\,\sqrt{R}}\,\exp(-F(h,t))\,. \quad (34)$$

We let $B$ as in Subsection 2.3, and we let

$$\begin{aligned} M = \min\{&F(h,t)\,|\,(h,2t) \in \partial\,S_n^m, \\ &h+1 \le H^{\mathrm{gen}},\ t \le T^{\mathrm{gen}}\} \end{aligned} \quad (35)$$

with $H^{\mathrm{gen}}$ and $T^{\mathrm{gen}}$ from Subsec 2.3. From the monotonicity and convexity properties of $F$, we then get, see [13], Appendix D,

– when $M > B$, we have that

$$\max_{(h,2t)\in S_n^m}\Big|c_t\,\frac{J_{h+1}(2\pi r)}{2\pi r}\Big| < \varepsilon\,, \quad (36)$$

– 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 $(h, 2t) \in S_n^m$

$$h \geq \max(h_1, h_2) \text{ or } t \geq \max(t_1, t_2) \Rightarrow F(h,t) \geq B\,. \quad (37)$$

The dedicated truncation rule becomes then as follows. Determine $M$ in Eq. (35). 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 $h+1 \leq H^{\text{gen}}$ & $t \leq T^{\text{gen}}$, for the two points $(h_1, 2t_1)$ and $(h_2, 2t_2)$ satisfying Eq. (37), 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. (14) is less than $\varepsilon$.

By the monotonicity and convexity properties of $F$, the minimum $M$ of $F$ along $\partial S_n^m$ is assumed on the edge $h = n - 2t$. 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 $\hat{H} \geq H^{\text{gen}}$ and $\hat{T}$ is the smallest integer with $\hat{T} \geq T^{\text{gen}}$. 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. (7–8). We make a slight variation of the approach in [1], Sec. 4 and 8, in that we write

$$a(\rho)\, \sqrt{1 - s_0^2 \rho^2} = \sum_{l=0}^{\infty} a_l\, R_{2l}^0(\rho)\,, \quad (38)$$

$$f(\rho) / \sqrt{1 - s_0^2 \rho^2} = \sum_{k=0}^{\infty} b_k\, R_{2k}^0(\rho)\,, \quad (39)$$

and we use linearization coefficients $A$ to write

$$a(\rho)\, f(\rho) = \sum_{t=0}^{\infty} c_t\, R_{2t}^0(\rho)\,, \quad (40)$$

where

$$c_t = \sum_{l,k=0}^{\infty} A_{2l,2k,2t}^{000}\, a_l\, b_k\,. \quad (41)$$

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.

$$b_k = \frac{1}{i u_0}\, \exp\left[i f / u_0\right] (2k+1)\, j_k(f/2)\, h_k^{(2)}(f/2v_0)\,. \quad (42)$$

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

$$j_k(z) = \sqrt{\frac{\pi}{2z}}\, J_{k+1/2}(z)\,, \quad (43)$$

$$\begin{aligned} h_k(z) &= j_k(z) - i\, y_k(z) \\ &= \sqrt{\frac{\pi}{2z}}\, \left(J_{k+1/2}(z) - i\, Y_{k+1/2}(z)\right) \\ &= \sqrt{\frac{\pi}{2z}}\, H_{k+1/2}^{(2)}(z)\,, \end{aligned} \quad (44)$$

with $J_\nu$, $Y_\nu$ and $H_\nu^{(2)}$ the Bessel function of first, second and third kind (Hankel function) and of order $\nu$, see [14], Ch. 10. The quantities $b_k$ can be computed, via Eqs. (43–44) using MatLab routines, efficiently at any desired accuracy.

As to the coefficients $a_l$, we first write, see Eq. (7),

$$\begin{aligned} a(\rho)\, \sqrt{1 - s_0^2 \rho^2} \;=\; & (1 - s_0^2 \rho^2)^{3/4}\, (1 - s_{0,M}^2 \rho^2)^{-3/4} \\ & + (1 - s_0^2 \rho^2)^{1/4}\, (1 - s_{0,M}^2 \rho^2)^{-1/4}\,. \end{aligned} \quad (45)$$

Next, either term on the right-hand side of Eq. (45) is developed into a power series

$$a_{\alpha\beta}(\rho) = (1 - s_\alpha^2 \rho^2)^\alpha\, (1 - s_\beta^2 \rho^2)^\beta = \sum_{N=0}^{\infty} r_N \rho^{2N}\,, \quad (46)$$

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

$$a_{l,\alpha\beta} = \sum_{N=l}^{\infty} b_N(l)\, r_N\,, \qquad l = 0, 1, \ldots\,, \quad (47)$$

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 $\varepsilon$ in the truncation analysis of Section 2 that involves the products of $c_t$'s and Jinc functions $J_{h+1}(2\pi r)/2\pi r$ as in Eqs. (14)–(15). Now $|J_{h+1}(z)/z| \leq 1/2$ for $z \geq 0$. Hence, when $c_t$ is computed with absolute accuracy $\varepsilon$, and the truncation rules of Subsections 2.3–2.4 are used with $\varepsilon/2$ instead of $\varepsilon$, a final absolute accuracy better than $\varepsilon$ results.

Next, given integers $L, K > 0$, the absolute error due to approximating $c_t$ of Eq. (41) by

$$c_{t,LK} = \sum_{l=0}^{L} \sum_{k=0}^{K} A_{2l,2k,2t}^{000}\, a_l\, b_k \quad (48)$$

is, as in Eqs. (13–14), safely bounded by

$$\max_{l>L \text{ or } k>K} |a_l b_k|\,. \quad (49)$$

Now there are the bounds

$$|a_l| \leq \tfrac{16}{3}\,, \quad |b_k| \leq 4\,, \qquad l,k = 0,1,\ldots\,. \tag{50}$$

The second bound in Eq. (50) follows from [13], Appendix C, Eq. (C18), while the first bound is obtained by considering in [13], 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. (49) is less than $\varepsilon$ when $L$ and $K$ are such that

$$l > L \Rightarrow |a_l| < \tfrac{1}{4}\,\varepsilon \quad \& \quad k > K \Rightarrow |b_k| < \tfrac{3}{16}\,\varepsilon\,. \tag{51}$$

As to the second condition in Eq. 51, we have according to [13], Appendix C

$$|b_k| \leq 4\,\exp(-\varphi(k\,;\,g/2) + \varphi(k\,;\,g/2v_0))\,, \tag{52}$$

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

$$k > \frac{1}{\gamma}\,\max\Big(0, \ln\Big(\frac{64}{3\varepsilon}\Big)\Big) + \tfrac{1}{2}\,g\,\frac{\sinh(\gamma)}{\gamma}\,, \tag{53}$$

with $\gamma$ as in Eq. (26). The quantities $b_k$ are computed using Eq. (42), involving the spherical Bessel and Hankel functions $j_k$ and $h_k^{(2)}$ that can be computed using Matlab routines. From [13], Appendix C we have that

$$\begin{aligned} &|j_k(f/2)| \leq \tfrac{2}{g}\,, \\ &|h_k(f/2v_0)| \leq \frac{2^{7/4}v_0}{g}\,\exp(\varphi(k;g/2v_0))\,, \end{aligned} \tag{54}$$

where the first inequality holds for all $f$ and the second inequality only holds when $|f/v_0| \geq 1$. In the case that $|f/v_0| < 1$, the $b_k$ of Eq. (42) is best evaluated using the power series representations of $j_k$ and $h_k^{(2)}$ that follow from [14], 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

$$\frac{3\varepsilon}{32}\cdot\frac{u_0\exp(-\varphi(K;g/2v_0))}{2^{7/4}(2K+1)v_0} \quad \text{and} \quad \frac{3\varepsilon}{32}\cdot\frac{u_0}{2(2K+1)}\,, \tag{55}$$

respectively.

As to the first condition in Eq. (51), we consider the decomposition of $a(\rho)\,\sqrt{1-s_0^2\rho^2}$ in terms $a_{\alpha\beta}(\rho)$ as in Eq. (46) with $\alpha+\beta = 0$ and Zernike coefficients $a_{l,\alpha\beta}$ as in Eq. (47). In [13], Appendix E the following is shown. Let $\delta = |\alpha| = |\beta|$, and let $S = \max(s_\alpha, s_\beta)$. Denoting "the $R_{2l}^0$-coefficient of $A(\rho)$" by $Z\,C_l[A(\rho)]$, we have

$$|a_{l,\alpha\beta}| \leq Z\,C_l[(1-S^2\rho^2)^{-\delta}] \sim \frac{E\,V^l}{(l+1)^{-\delta+1/2}}\,, \tag{56}$$

where

$$E = \frac{2\sqrt{\pi}}{\Gamma(\delta)}\,\frac{(1-S^2)^{-\frac{1}{2}\delta+\frac{1}{4}}}{1+\sqrt{1-S^2}}\,, \qquad V = \frac{1-\sqrt{1-S^2}}{1+\sqrt{1-S^2}}\,. \tag{57}$$

Furthermore, the right-hand side of Eq. (56) is less than $\eta := \varepsilon/8$ when

$$l \geq \frac{\ln(E\eta^{-1}) - (-\delta+1/2)\ln(1+\ln(E\eta^{-1})/\ln(1/V))}{\ln(1/V)}. \tag{58}$$

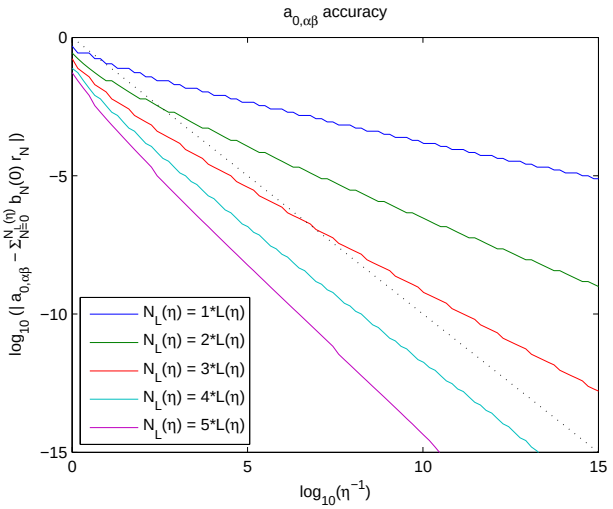


FIG. 5 Plot of $\log_{10}|a_{0,\alpha\beta} - \sum_{N=0}^{N_L(\eta)} b_N(0)r_N|$ as a function of $\log_{10}\eta^{-1} \in [0,15]$, for the case that $a_{0,\alpha\beta}$ is the $R_0^0$-coefficient of $a_{\alpha\beta}(\rho) = (1-s_0^2\rho^2)^\alpha(1-s_{0,M}^2\rho^2)^\beta$ with $\alpha = -\beta = 3/4$ 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. (58). The black (dotted) curve indicates those positions at which accuracy is equal to $\eta$.

Therefore, the first condition in Eq. (51) is satisfied when $L$ is the maximum of the two numbers that occur at the right-hand side of Eq. (58) 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. (47). It is shown in [13], Appendix E that for a given $\varepsilon > 0$ and an integer $L > 0$ such that the right-hand side of Eq. (56) $< \frac{1}{8}\,\varepsilon$ when $l > L$, we have that all numbers $a_{l,\alpha\beta}$, $l = 0,1,\ldots,L$, are computed with absolute accuracy $\varepsilon/16$ when the infinite series in Eq. (47) is truncated at $N_L = 2L/\sqrt{1-S^2}$.

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 $\eta$ 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. (58).

To summarize, for $\varepsilon \in (0,1)$ we replace $c_t$ by $c_{t,LK}$ given in Eq. (48) in which

- $L$ and $K$ are given by the right-hand sides of Eq. (58) and Eq. (51), respectively,
- $b_k$ is as in Eq. (42) for which $j_k(f/2)$ and $h_k^{(2)}(f/2v_0)$ are computed with absolute accuracy as specified in Eq. (55),
- $a_l = a_{3/4,-3/4,l} + a_{1/4,-1/4,l}$ and the two $a_{\alpha,\beta,l}$ are computed by summing the series in Eq. (47) 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 $a_{\alpha,\beta,l}$.

## 4 SUMMARY OF THE TRUNCATION RULES FOR THE GENERAL CASE

We want to compute

$$I = \int_0^1 a(\rho) f(\rho) R_n^{|m|}(\rho) J_m(2\pi r\rho)\rho \,\mathrm{d}\rho \,, \tag{59}$$

where

$$a(\rho) = \frac{(1-s_0^2\rho^2)^{1/2} + (1-s_{0,M}^2\rho^2)^{1/2}}{(1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{3/4}} \,, \tag{60}$$

$$f(\rho) = \exp\left[\frac{if}{u_0}\,(1-\sqrt{1-s_0^2\rho^2})\right] , \tag{61}$$

with integer $n$ and $m$ such that $n-|m|$ is even and non-negative, and for a given real value of $f$ and a positive number $r$, while $0 < s_0,\ s_{0,M} < 1$ are given and $u_0 = 1-\sqrt{1-s_0^2}$. We have the double series representation

$$I = \sum_{h,t} A_{2t,n,h}^{0mm} (-1)^{\frac{h-m}{2}}\, c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \tag{62}$$

with summation over $h$, $t = 0, 1, \ldots$ where $h$ has same parity as $n$ and $m$, and the $A$-coefficients are as in [1], Sec. 5 and Appendix C. Furthermore, the $c_t$ are the Zernike coefficients of $a(\rho)f(\rho)$ and are given in the form

$$c_t = \sum_{l,k=0}^{\infty} A_{2l,2k,2t}^{000}\, a_l\, b_k \,, \tag{63}$$

where the $a_l$ and $b_k$ are the Zernike coefficients of $a(\rho)\sqrt{1-s_0^2\rho^2}$ and $f(\rho)/\sqrt{1-s_0^2\rho^2}$, respectively, and the $A$-coefficients are again as in [1], Sec. 5 and Appendix C.

Let

$$R = \max\left(r, \tfrac{1}{2\pi}\right) , \quad g = \max\left(1, |f|\right) , \tag{64}$$

and

$$S = \max\left(s_0, s_{0,M}\right) , \quad V = \frac{1-\sqrt{1-S^2}}{1+\sqrt{1-S^2}} \,,$$
$$W = \frac{1}{1+\sqrt{1-S^2}} \,, \tag{65}$$

and let $\varepsilon > 0$.

### 4.1 Truncation double series for I

Let

$$B = \max\left(0, \ln\frac{2\pi W a_0}{\pi^2 \varepsilon R\sqrt{R}}\right) , \tag{66}$$

where $a_0 = 2\int_0^1 a(\rho)\sqrt{1-s_0^2\rho^2}\rho\,\mathrm{d}\rho$ is the $R_0^0$-coefficient of $a(\rho)\sqrt{1-s_0^2\rho^2}$. The truncation error when replacing the double series for $I$ in Eq. (62) by

$$\sum_{h+1\le H,\, t\le T} A_{2t,n,h}^{0mm} (-1)^{\frac{h-m}{2}}\, c_t\, \frac{J_{h+1}(2\pi r)}{2\pi r} \,, \tag{67}$$

is less than $\varepsilon$, simultaneously for all $n$ and $m$, when

$$H = B + 2\pi\, R\, \sinh(1) \,, \qquad T = \frac{1}{\gamma}\, B + \tfrac{1}{2}\, g\, \frac{\sinh(\gamma)}{\gamma} \,, \tag{68}$$

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

### 4.2 Truncation double series for $c_t$

Let

$$E = \frac{2\sqrt{\pi}}{\Gamma(3/4)}\, \frac{(1-S^2)^{-1/8}}{1+\sqrt{1-S^2}} \,, \tag{69}$$

and let

$$L = \frac{\ln(8E/\varepsilon) + \frac{1}{4}\ln(1+\ln(8E/\varepsilon)/\ln(1/V))}{\ln(1/V)} \,, \tag{70}$$

$$K = \frac{1}{\gamma}\, \max\Big(0, \ln\frac{64}{3\varepsilon}\Big) + \tfrac{1}{2}\, g\, \frac{\sinh(\gamma)}{\gamma} \,, \tag{71}$$

with $V$ and $\gamma$ as above. The truncation error when replacing the double series for $c_t$ in Eq. (63) by

$$\sum_{l=0}^{L} \sum_{k=0}^{K} A_{2l,2k,2t}^{000}\, a_l\, b_k \tag{72}$$

is less than $\varepsilon$, simultaneously for all $t \le T$.

### 4.3 Truncation issue in computing $a_l$

The $a_l$ are computed as follows. Write

$$\begin{aligned} a(\rho)(1-s_0^2\rho^2)^{1/2} =& (1-s_0^2\rho^2)^{3/4}\,(1-s_{0,M}^2\rho^2)^{-3/4} \\ &+ (1-s_0^2\rho^2)^{1/4}\,(1-s_{0,M}^2\rho^2)^{-1/4} \,. \end{aligned} \tag{73}$$

The two 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. (73) have Zernike coefficients

$$a_{l,\alpha\beta} = \sum_{N=l}^{\infty} b_N(l)\, r_{N,\alpha\beta} \,, \tag{74}$$

with $b_N(l) = \frac{2l+1}{l+1}\,\binom{N}{l}/\binom{N+l+1}{N}$ and where the $r_{N,\alpha\beta}$ are computed recursively according to

$$\begin{aligned} r_{-1} =& 0 \,, \quad r_0 = 1 \,; \\ r_{N+1} =& \frac{1}{N+1}\,[((N-\alpha)\,s_\alpha^2 + (N-\beta)\,s_\beta^2)\,r_N \\ &- (N-1-\alpha-\beta)\,s_\alpha^2\,s_\beta^2\,r_{N-1}] \end{aligned} \tag{75}$$

for $N = 0,\ 1,\ \cdots$. Let

$$N_L = \frac{2L}{\sqrt{1-S^2}}\,, \tag{76}$$

with $L$ given by Eq. (70) and $S$ given by Eq. (65). The truncation error when replacing the series for $a_{l,\alpha\beta}$ in Eq. (74) by

$$\sum_{N=l}^{N_L} b_N(l)\, r_{N,\alpha\beta} \tag{77}$$

is less than $\varepsilon/16$ for all $l = 0,\ 1,\ \cdots,\ L$.

### 4.4 Computation of $b_k$

The $b_k$ are given in terms of spherical Bessel functions $j_k$ and Hankel functions $h_k^{(2)}$ as

$$b_k = \frac{1}{iu_0}\,\exp\left[if/u_0\right](2k+1)\,f\,j_k(f/2)\,h_k^{(2)}(f/2v_0)\,, \tag{78}$$

with
$u_0 = 1-\sqrt{1-s_0^2}$ and $v_0 = (1-\sqrt{1-s_0^2})/(1+\sqrt{1-s_0^2})$, and can be computed using Matlab codes to double precision accuracy.

### 4.5 Computation of Jinc-functions

The Jinc-functions are given in terms of Bessel-functions of the first kind as $J_{h+1}(2\pi r)/(2\pi r)$, and can be computed using Matlab codes to double precision accuracy.

### 4.6 Overall accuracy after assembling

When the $a_l$, $l = 0,\ 1,\ \cdots,\ L$, are computed as in Subsection 4.3 and the $b_k$, $k = 0,\ 1,\ \cdots,\ K$, are computed with absolute accuracy $3\varepsilon/16$, the $c_t$, $t = 0,\ 1,\ \cdots,\ T$, computed as in Subsection 4.2, have absolute accuracy $7\varepsilon/4$. Next, when the Jinc-functions $J_{h+1}(2\pi r)/(2\pi r)$ are computed with absolute accuracy $\varepsilon/(4Wa_0)$ and the truncation rule of Subsection 4.1 is used, the quantity $I$ in Eq. (59) is computed with accuracy $23\varepsilon/8$ for all $n$ and $m$. This overall accuracy number $23\varepsilon/8$ emerges from a worst-case scenario in which all intermediate errors are assumed to contribute maximally to the total error, but actual accuracy will almost always be far better.

## 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 Subsection 2.3 and the dedicated truncation rule of Subsection 2.4 for approximation of the diffraction integral $I$ in Eqs. (5)–(6) as a function of $\varepsilon \in (0,1)$ for a variety of radial values $r$, 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 $\varepsilon$. 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 $\varepsilon/16Wa_0$, 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 Figures 6–8, we show achieved accuracy (a) and computation time (b) against requested accuracy $\varepsilon$ in the range $10^{-15} - 10^{0}$, using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines), for a variety of different parameter combinations. In each figure, going from top to bottom, one or two parameters are varied, while the other parameters are kept constant, in order to illustrate the impact on the accuracy and computation time. In Figure 6, this is done for the magnitudes of the focal and radial parameters $f$ and $r$, in Figure 7, for the numerical aperture value $s_0$, and, finally, in Figure 8, for the degree $n$ and azimuthal order $m$ of the radial part of the Zernike polynomial.

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 $2\pi r$. Also see [13], Sec. 5 for an extended set of illustrations along these lines of the performance of the truncation rules.

## 6 CONCLUSION

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 from the computational point of view.

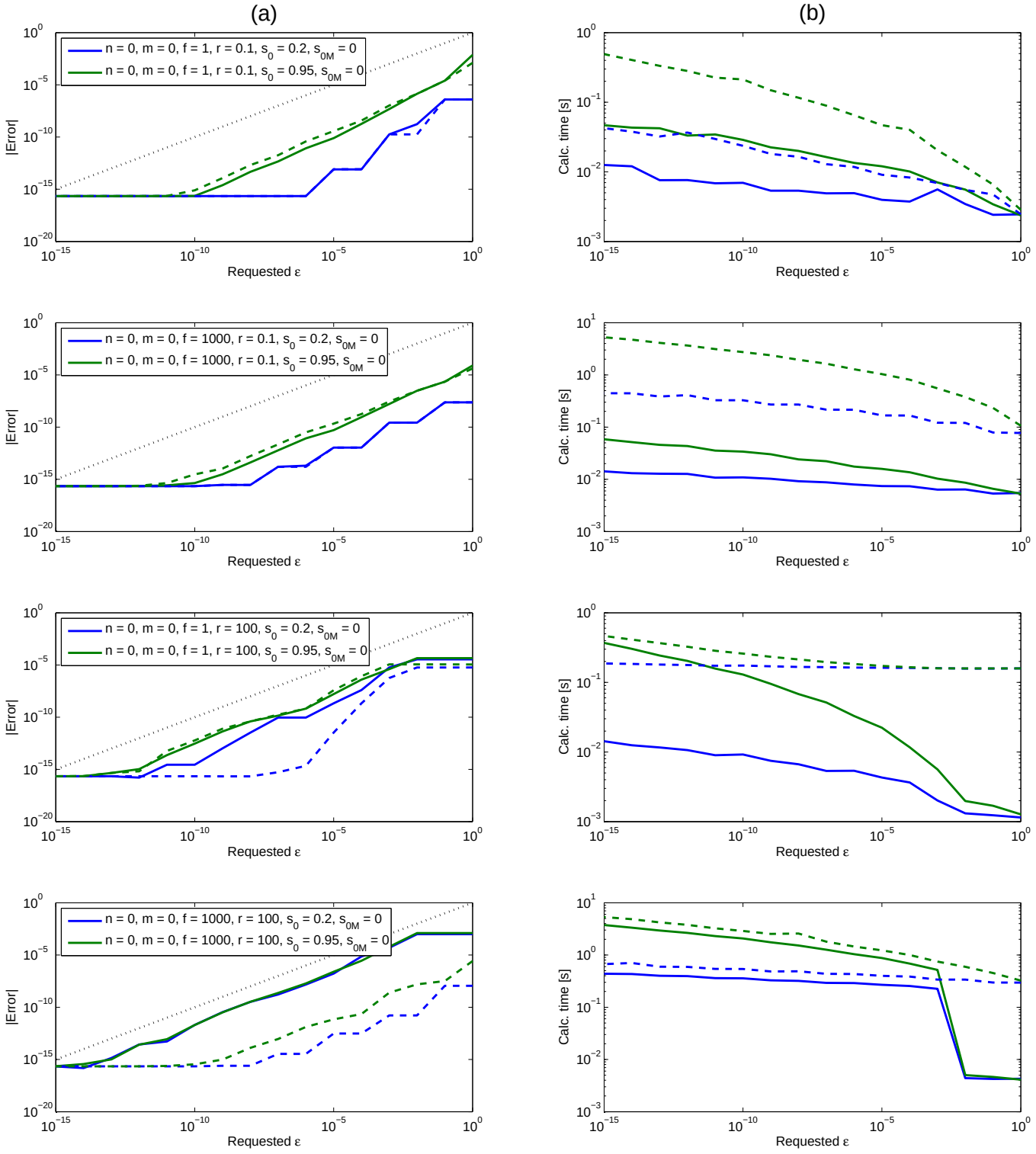


FIG. 6 Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy $\varepsilon$ using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the focal and radial variables $f$ and $r$ from top to bottom according to $(f,r) = (1{,}0.1)$, $(1000{,}0.1)$, $(1{,}100)$, $(1000{,}100)$. Setting of aperture variables: $s_0 = 0.2$, $s_{0,M} = 0$ (blue) and $s_0 = 0.95$, $s_{0,M} = 0$ (green), setting of the degree and azimuthal order of the radial polynomial: $(n,m) = (0{,}0)$.

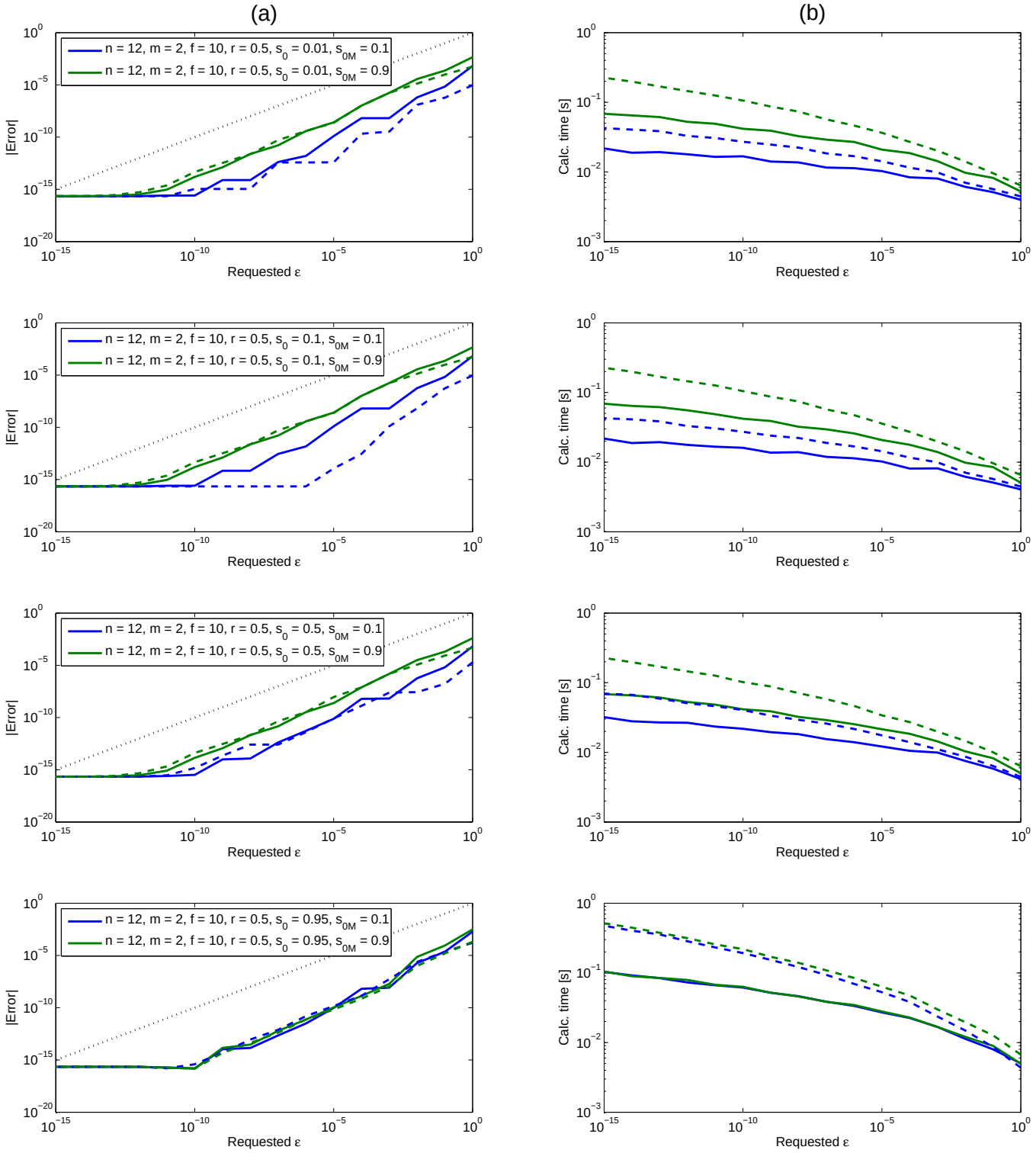


FIG. 7 Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy $\varepsilon$ using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the aperture variable $s_0$ from top to bottom according to $s_0 = 0.01,\ 0.1,\ 0.5,\ 0.95$. Setting of degree and azimuthal order of the radial polynomial: $(n,m) = (12,2)$, setting of aperture variable in object space: $s_{0,M} = 0.1$ (blue) and 0.9 (green), setting of focal and radial variable: $f = 10,\ r = 0.5$.

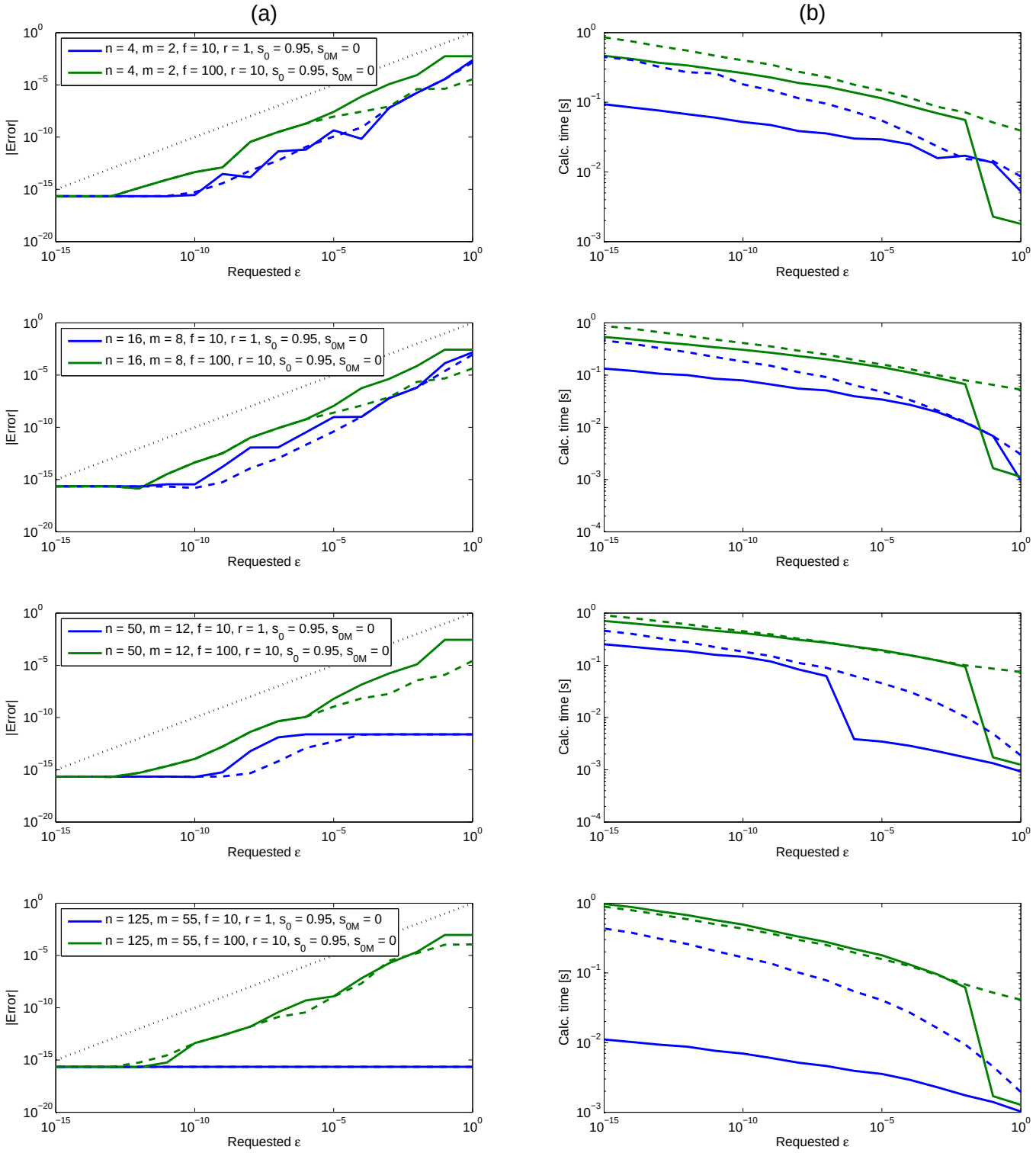


FIG. 8 Absolute accuracy (a) and computation time (b) as a function of requested absolute accuracy $\varepsilon$ using the general truncation rule (dashed lines) and the dedicated truncation rule (solid lines) when varying the degree $n$ and azimuthal order $m$ of the radial polynomial from top to bottom according to $(n,m) = (4,2)$, $(16,8)$, $(50,12)$, $(125,55)$. Setting of aperture variables: $s_0 = 0.95$, $s_{0,M} = 0$, setting of focal and radial variable: $f = 10$, $r = 1$ (blue) and $f = 100$, $r = 10$ (green).

## 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] A. J. E. M. Janssen, "Extended Nijboer-Zernike approach for the computation of optical point-spread functions," J. Opt. Soc. Am. **A19**, 849–857 (2002).

[3] J. J. M. Braat, P. Dirksen, and A. J. E. M. Janssen, "Assessment of an extended Nijboer-Zernike approach for the computation of optical point-spread functions," J. Opt. Soc. Am. **A19**, 858–870 (2002).

[4] J. J. M. Braat, P. Dirksen, A. J. E. M. Janssen, and A. S. van de Nes, "Extended Nijboer-Zernike representation of the vector field in the focal region of an aberrated high-aperture optical system," J. Opt. Soc. Am. **A20**, 2281–2292 (2003).

[5] J. J. M. Braat, P. Dirksen, A. J. E. M. Janssen, S. van Haver, and A. S. van de Nes, "Extended Nijboer-Zernike approach to aberration and birefringence retrieval in a high-numerical-aperture optical system," J. Opt. Soc. Am. **A22**, 2635–2650 (2005).

[6] S. van Haver, J. J. M. Braat, P. Dirksen, and A. J. E. M. Janssen, "High-NA aberration retrieval with the Extended Nijboer-Zernike vector diffraction theory," J. Europ. Opt. Soc. Rap. Public. **1**, 06004 (2006).

[7] S. van Haver, J. J. M. Braat, A. J. E. M. Janssen, O. T. A. Janssen, and S. F. Pereira, "Vectorial aerial-image computations of three-dimensional objects based on the extended Nijboer-Zernike theory," J. Opt. Soc. Am. **A26**, 1221–1234 (2009).

[8] J. J. M. Braat, S. van Haver, A. J. E. M. Janssen, and S. F. Pereira, "Image formation in a multilayer using the extended Nijboer-Zernike theory," J. Europ. Opt. Soc. Rap. Public. **4**, 09048 (2009).

[9] S. van Haver, *The Extended Nijboer-Zernike Diffraction Theory and its Applications* (Ph.D. thesis, Delft University of Technology, 2010).

[10] J. J. M. Braat, S. van Haver, A. J. E. M. Janssen, and P. Dirksen, "Assessment of optical systems by means of point-spread functions," Prog. Optics **51**, 349–468 (2008).

[11] A. J. E. M. Janssen, J. J. M. Braat, and P. Dirksen, "On the computation of the Nijboer-Zernike aberration integrals at arbitrary defocus," J. Mod. Opt. **51**, 687–703 (2004).

[12] J. Boersma, "On the computation of Lommel's functions of two variables," Math. Comput. **16**, 232–238 (1962).

[13] S. van Haver, and A. J. E. M. Janssen, "Truncation strategy for the series expressions in the advanced ENZ-theory of diffraction integrals," arXiv: **1407.6589v1**, (2014).

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