# On the computation of the Nijboer–Zernike aberration integrals at arbitrary defocus

A. J. E. M. JANSSEN†, J. J. M. BRAAT‡* and P. DIRKSEN†

†Philips Research Laboratories, 5656 AA Eindhoven, The Netherlands
‡Optics Research Group, Department of Applied Sciences,
Delft University of Technology, 2628 CJ Delft, The Netherlands

**Abstract.** We present a new computation scheme for the integral expressions describing the contributions of single aberrations to the diffraction integral in the context of an extended Nijboer–Zernike approach. Such a scheme, in the form of a power series involving the defocus parameter with coefficients given explicitly in terms of Bessel functions and binomial coefficients, was presented recently by the authors with satisfactory results for small-to-medium-large defocus values. The new scheme amounts to systemizing the procedure proposed by Nijboer in which the appropriate linearization of products of Zernike polynomials is achieved by using certain results of the modern theory of orthogonal polynomials. It can be used to compute point-spread functions of general optical systems in the presence of arbitrary lens transmission and lens aberration functions and the scheme provides accurate data for any, small or large, defocus value and at any spatial point in one and the same format. The cases with high numerical aperture, requiring a vectorial approach, are equally well handled. The resulting infinite series expressions for these point-spread functions, involving products of Bessel functions, can be shown to be practically immune to loss of digits. In this respect, because of its virtually unlimited defocus range, the scheme is particularly valuable in replacing numerical Fourier transform methods when the defocused pupil functions require intolerably high sampling densities.

## 1. Introduction

The computation of strongly defocused amplitude distributions has been considered by several authors. Their effort has been directed towards the stable evaluation of the diffraction integral in the presence of a strongly oscillating defocusing phase factor in the integrand of the basic diffraction integral. We refer to the end of this section for a short survey of numerical and analytic approaches to solve the defocus problem that can be found in the literature [1–5]. In this paper we treat the strong defocus problem in the framework of the recently developed extension of the Nijboer–Zernike approach to the computation of optical point-spread functions of general aberrated optical systems [6–10]. In the extension to the Nijboer–Zernike approach, power series expressions involving the defocus parameter $f$, with coefficients explicitly given in terms of Bessel functions and binomial coefficients, were given for the contribution to the diffraction integral of a single aberration term $\beta_{nm} R_n^m(\rho) \cos m\vartheta$ with $R_n^m(\rho)$ a Zernike polynomial, see

*Author for correspondence.

[11], section 9.2. We follow the developments as given in [7], sections 1–2. Thus, given the Zernike expansion $\sum_{n,m} \beta_{nm} R_n^m(\rho) \cos m\vartheta$ of the pupil function $A \exp[i\Phi]$, with the transmission function $A$ and the aberration phase $\Phi$ assumed to be symmetric in the angular coordinate $\vartheta$, the point-spread function $U$ can be written as

$$U(x, y) = 2 \sum_{n,m} \beta_{nm} i^m V_{nm}(r, f) \cos m\varphi. \tag{1}$$

Here we have used Cartesian coordinates $x, y$ in the image plane that can be transformed to polar coordinates $r, \varphi$ according to $x + iy = r \exp(i\varphi)$. The $V_{nm}$ in (1) are the basic integrals

$$V_{nm}(r, f) = \int_0^1 \exp(if\rho^2) R_n^m(\rho) J_m(2\pi\rho r) \rho \, d\rho, \tag{2}$$

which should be considered for all integer $n, m \geq 0$ with $n - m \geq 0$ and even. For these $V_{nm}$ there holds the power series expansion, see [6, 7],

$$V_{nm}(r, f) = \exp(if) \sum_{l=1}^{\infty} (-2if)^{l-1} \sum_{j=0}^{p} v_{lj} \frac{J_{m+l+2j}(2\pi r)}{l(2\pi r)^l}, \tag{3}$$

with

$$v_{lj} = (-1)^p (m + l + 2j) \left[ \binom{m+j+l-1}{l-1} \binom{j+l-1}{l-1} \binom{l-1}{p-j} \Big/ \binom{q+l+j}{l} \right], \tag{4}$$

for $l = 1, 2, \ldots, j = 0, 1, \ldots, p$ in which

$$p = \tfrac{1}{2}(n - m), \quad q = \tfrac{1}{2}(n + m). \tag{5}$$

This approach of computing point-spread functions has been assessed in [7] from an optical and numerical point of view. As a rule of thumb one should include in the infinite series over $l$ some $3f$ terms to get sufficiently accurate results. The approach has been extended further in [8] so as to cover the cases of high numerical aperture, which requires computation of field components, as well as inclusion of the radiometric effect and the state of polarization. Application of the method in a lithographic context is considered in [9, 10], where the inverse problem of estimating the coefficients $\beta_{nm}$ from measurements of the (intensity) point-spread function in the focal region is solved. In this kind of application the focus variable $f$ is often taken to be complex-valued so that illuminated objects of small but finite size can be accommodated

As said, the series expression in (3) yields accurate results when some $3f$ terms are included in the series over $l$. The evaluation of the required Bessel functions is normally no problem since one can exploit recursion formulas as in [12], expression 9.1.27 on p. 361, when the efficient computation of the Bessel functions is not already available in the software environment of the user. A much more serious problem is posed by the fact that large values of $f$ lead to a considerable loss of digits in the series in (3). Typically, one has terms of the order of magnitude $|f|^l/l!$ in (3) while the $V_{nm}$'s themselves are of the order of unity. Practically, the use of the series in (3) is limited to a range like $|f| \leq 5\pi$, so that an axial range of the order of typically ten focal depths can be handled.

A similar problem was noted in [13] where the attention is limited to radially symmetric aberrations. Here for the computation of the integrals

$$T_{2p,0} = \int_0^1 \exp\left(\mathrm{i}f\rho^2\right)\rho^{2p}\,J_0(2\pi\rho r)\rho\,\mathrm{d}\rho, \tag{6}$$

the $\exp\left(\mathrm{i}f\rho^2\right)$ is expanded in powers of $f$ so that the so-called generalized Jinc functions,

$$\mathrm{Jinc}_n(u) = \frac{1}{u^{2n+2}} \int_0^u v^{2n+1} J_0(v)\,\mathrm{d}v, \quad n = 0, 1, \ldots, \tag{7}$$

appear, for which explicit finite Bessel series are given in [13]. Also in this case one should limit $f$ to a range like $|f| \leq 5\pi$.

In optical problems, axial excursions beyond the described limits are frequently encountered, and we list here some examples:

(1) the self-imaging by periodic structures manifests itself far outside the focal region;
(2) amplitude oscillations close to the geometrical boundary are encountered in the Fresnel diffraction regime;
(3) certain optical problems use the fractional Fourier transform in which light distributions occur that are intrinsically quite remote from the standard Fraunhofer pattern;
(4) in optical recording, the extension to volumetric storage using several recording layers in depth requires propagation of a focused beam through several strongly defocused information layers.

In all these cases a reliable analytical method is required that allows the calculation of strongly defocused, aberrated optical fields. In this paper it is shown how the analysis developed in the framework of our extended Nijboer–Zernike theory can be used to achieve this goal. Below we present the basic features of our approach.

In [14] a method has been analysed to compute Lommel's functions of two variables without loss of digits (the first two Lommel functions can be expressed in terms of $T_{00} = V_{00}$, see (2) and (6)). This method was developed by Nijboer and Zernike [15, 16] and uses Bauer's identity

$$\exp\left(\mathrm{i}f\rho^2\right) = \exp\left(\tfrac{1}{2}\mathrm{i}f\right) \sum_{k=0}^{\infty} (2k+1)\mathrm{i}^k j_k\left(\tfrac{1}{2}f\right) R_{2k}^0(\rho), \tag{8}$$

see [12], formula 10.1.47 on p. 440 (observe that $R_{2k}^0(\rho) = P_k(2\rho^2 - 1)$) or [11], formula (10) on p. 534, with

$$j_k(z) = \left(\frac{\pi}{2z}\right)^{1/2} J_{k+(1/2)}(z), \quad k = 0, 1, \ldots, \tag{9}$$

the spherical Bessel functions of the first kind, see [12], chap. 10. Applying (8) for the computation of $T_{00}$ in (6) we get

$$T_{00} = \exp\left(\tfrac{1}{2}\mathrm{i}f\right) \sum_{k=0}^{\infty} (2k+1)\mathrm{i}^k j_k\left(\tfrac{1}{2}f\right) \int_0^1 R_{2k}^0(\rho)J_0(2\pi\rho r)\rho\,\mathrm{d}\rho. \tag{10}$$

The remaining integrals are computed using the basic result

$$\int_0^1 R_n^m(\rho)J_m(2\pi\rho r)\rho\,\mathrm{d}\rho = (-1)^{(n-m)/2}\frac{J_{n+1}(2\pi r)}{2\pi r} \tag{11}$$

from the 'classical' Nijboer–Zernike theory, see [11], section 9.2, formula (9) on p. 525. Due to certain bounds on the Bessel functions $j_k(\tfrac{1}{2}f)$ and $J_{n+1}(2\pi r)$, it can be shown, see [14], that the resulting method to compute $T_{00}$ does not suffer from loss of digits.

In this paper we extend the method of Nijboer and Zernike to the computation of the $V_{nm}$ in (2). Thus we insert Bauer's formula (8) into the integral at the right-hand side of (2) and interchange integral and summation. Then one is faced with the problem of writing products $R_{2k}^0 R_n^m$ as a linear combination of Zernike polynomials with upper index $m$ so as to be able to apply (11). This problem was attacked by Nijboer in [15] for modest values of $k$, $m$, $n$ by employing recursion formulas for the Zernike polynomials. A systematic procedure for this does not seem to have been devised by Nijboer or thereafter (also see [11], top of p. 535, for this point). By using results of the relatively recent modern theory of orthogonal polynomials, as can be found in [17], sections 6.8 and 7.1, we are able to find finite series expressions, with favourable properties from a computational point of view, for the coefficients needed in the 'linearization' of the products $R_{2k}^0 R_n^m$.

This paper has been organized as follows. In section 2 we present the main result and its derivation, and we comment on the nature and magnitude of the linearization coefficients that yield the main result. In section 3 we show how the main result can be used to evaluate integrals $T_{nm}$ of the type that occurs in (6). In section 4 we briefly comment on how to extend the method to the computation of integrals that occur in the case of high numerical aperture for which a vector formalism as well as inclusion of the radiometric effect and the state of polarization is required. In section 5 we present some examples and results of computations to compare the new scheme with the one based on (3) and the one as follows from the results in [13] for $T_{2p,0}$.

We conclude this section with some comments on different approaches that can be found in the literature. Formally, the phase factor in the diffraction integral is proportional to the projection of the defocus distance onto the direction of the plane wave contribution in the integral. In [1, 2], the evaluation of the diffraction integral is carried out using numerical Fourier transform (FT) techniques, and the effect of the phase factor is mitigated by intentionally introducing a compensating quadratic phase factor that can be incorporated in the FT scheme. An analytic approach to obtain stable expressions for strongly defocused fields is found in [3, 4]. The radiometric effect encountered in the vectorial treatment of the diffraction integral is integrated in the analysis by using a Fourier–Gegenbauer expansion in [3]; along the same lines, the inclusion of circularly symmetric aberrations in the diffraction integral is demonstrated in [4]. In both cases, the field as a function of the distance from the geometrical focus is described in terms of a series containing spherical Bessel functions. A quite different approach to treat the various possible shapes of the radiometric effect function is described in [5]. The electromagnetic field in the aperture is matched by means of an expansion in multipole far-field radiation patterns using spherical harmonics and the field distribution in any defocused position is obtained by applying the known propagation effects to the multipole distributions, involving again well-converging

spherical Bessel functions with the geometrical distance from the focus as argument. In principle, more general aperture functions could be treated by this method although the analysis in [5] is limited to smooth amplitude variations in the aperture due to the considered radiometric functions.

In the light of these earlier approaches, our method has the following salient features. It is analytic in nature and it conforms to the well-established practice of expanding optical aberrations as a series involving Zernike polynomials. As a result, it can be easily inserted into existing computational environments and software packages. Moreover, our Zernike expansion solution covers both amplitude and phase perturbations in the exit pupil of the optical system. Finally, the separation of variables, see (1) and the main result in section 2, equation (26), allows convenient computation of the diffraction pattern in polar coordinates $(r, \phi)$ per defocused plane $(f)$.

## 2. Main result

In this section we present the main result, see subsection 2.2 below, and its proof. For the basic definitions and properties of the Zernike polynomials $R_n^m$ we refer to [11], section 9.2.1 and Appendix VIII. In particular, we have

$$R_{m+2p}^m(\rho) = \rho^m P_p^{(0,m)}(2\rho^2 - 1) \tag{12}$$

with $P_k^{(\alpha,\beta)}$ the Jacobi polynomials in the notation and normalization as occurs in [17, 18].

### 2.1. *Linearization of products of Zernike polynomials*

We present a result on linearization of products $R_{n_1}^{m_1} R_{n_2}^{m_2}$ of Zernike polynomials for which we develop some notation. We let for $m, p, l, s_1, s_2 = 0, 1, \dots$

$$f_{ps}^m = (-1)^{p-s} \frac{2s+1}{p+s+1} \left[ \binom{m+p-s-1}{m-1} \binom{m+p+s}{s} \Big/ \binom{p+s}{s} \right] \quad s = 0, \dots, p, \tag{13}$$

$$g_{ul}^m = \frac{m+2l+1}{m+u+l+1} \left[ \binom{m}{u-l} \binom{u+l}{l} \Big/ \binom{m+l+u}{m+l} \right], \quad u = l, \dots, l+m, \tag{14}$$

$$b_{s_1 s_2 t} = \frac{2s_1 + 2s_2 - 4t + 1}{2s_1 + 2s_2 - 2t + 1} \left( \frac{A_{s_1-t} A_t A_{s_2-t}}{A_{s_1+s_2-t}} \right), \quad t = 0, \dots, \min(s_1, s_2), \tag{15}$$

where $A_k = \binom{2k}{k}$. For $m = 0$ we should read (13) and (14) as

$$f_{ps}^0 = \delta_{ps}, \quad g_{ul}^0 = \delta_{ul}, \tag{16}$$

with $\delta$ being Kronecker's delta.

Now there is the following result. Let $m_1, m_2, p_1, p_2 = 0, 1, \dots$. Then

$$R_{m_1+2p_1}^{m_1} R_{m_2+2p_2}^{m_2} = \sum_l c_l R_{m_1+m_2+2l}^{m_1+m_2}, \tag{17}$$

where the coefficients $c_l$ in (17) are given by

$$c_l = \sum_{s_1,s_2,t} f_{p_1 s_1}^{m_1} f_{p_2 s_2}^{m_2} g_{s_1+s_2-2t,l}^{m_1+m_2}, \tag{18}$$

with the summation range in (18) giving all integer tuples $(s_1, s_2, t)$ with

$$0 \le s_1 \le p_1, \quad 0 \le s_2 \le p_2, \quad 0 \le t \le \min(s_1, s_2),$$
$$l \le s_1 + s_2 - 2t \le l + m_1 + m_2. \tag{19}$$

Furthermore, there holds

$$c_l \ge 0, \quad \sum_l c_l = 1, \tag{20}$$

and $c_l \ne 0$ only for $l$ between $\max(0, p_1 - p_2 - m_2, p_2 - p_1 - m_1)$ and $p_1 + p_2$.

To prove this result we first note that we have from (12)

$$R^{m_1}_{m_1+2p_1}(\rho)R^{m_2}_{m_2+2p_2}(\rho) = \rho^{m_1+m_2} P^{(0,m_1)}_{p_1}(2\rho^2 - 1)P^{(0,m_2)}_{p_2}(2\rho^2 - 1). \tag{21}$$

First we apply [17], Theorem 7.1.2 on p. 358 (with $\alpha = 0$, $\delta = m$; $\beta = 0$; $n = p$, $k = s$) and find after some administration

$$P^{(0,m)}_p = \sum_{s=0}^{p} f^m_{ps} P^{(0,0)}_s. \tag{22}$$

Next we get from [17], Corollary 6.8.3 on p. 320 (with $m = s_1$, $n = s_2$),

$$P^{(0,0)}_{s_1} P^{(0,0)}_{s_2} = \sum_{t=0}^{\min(s_1,s_2)} b_{s_1 s_2 t} P^{(0,0)}_{s_1+s_2-2t}. \tag{23}$$

Finally, we again apply [17], Theorem 7.1.2 on p. 358 (now with $\alpha = 0$, $\delta = 0$; $\beta = m$; $n = u$, $k = l$) and find after some administration

$$P^{(0,0)}_u = \sum_{l=u-m}^{u} g^m_{ul} P^{(0,m)}_l. \tag{24}$$

Then by (12) with $m = m_1 + m_2$ we get (17) with the $c_l$ given in (18) and a summation range for $l$ given in principle by $0, 1, \dots, p_1 + p_2$.

The non-negativity of all $c_l$'s and the fact that the $c_l$'s vanish outside the claimed $l$ range is an immediate consequence of [19], Corollary 5.3. Finally, $\sum_l c_l = 1$ follows from the fact that $R^m_n(\rho = 1) = 1$ for all allowed values of $n$ and $m$.

### 2.1.1. *Notes*

(1) The coefficients $f$, $g$, $b$ as given in (13), (14) and (15) occur as expansion coefficients of certain bounded functions ($R^m_n(\rho)$ has modulus $\le 1$ for all $\rho, 0 \le \rho \le 1$, and all allowed $n$, $m$) with respect to an orthogonal set of functions (with appropriate weight functions). They are therefore quite small, and computation of the $c_l$'s according to (18) and (19) hardly suffers from loss of digits.

(2) It is unlikely that the triple series in (18) can be cast into a reasonably simple closed form. In [20] there is given an explicit expression for the coefficients $d_l$ in the expansion

$$P^{(\alpha,\beta)}_{k_1} P^{(\alpha,\beta)}_{k_2} = \sum_l d_l P^{(\alpha,\beta)}_l. \tag{25}$$

While this may lead somewhat more directly to explicit formulas for the $c_l$'s in (17), it should be said that the expressions in [20] for the $d_l$'s take more than one page of explanation and involve the hypergeometric function $_9F_8$. In fact, we had serious trouble in checking the simple case $\alpha = \beta = 0$, see (15) and (23), from the general result in [20].

### 2.2. *Bessel–Bessel series expression for $V_{nm}$*

We now present the main result of this paper. Let $n$, $m$ be integers $\geq 0$ with $n - m \geq 0$ and even, and let $p = \frac{1}{2}(n - m)$, $q = \frac{1}{2}(n + m)$. Then

$$V_{nm}(r, f) = \exp\left(\tfrac{1}{2}\mathrm{i}f\right) \sum_{k=0}^{\infty} (2k + 1)\mathrm{i}^k j_k\left(\tfrac{1}{2}f\right) \sum_{l=\max(0,k-q,p-k)}^{k+p} (-1)^l w_{kl} \frac{J_{m+2l+1}(2\pi r)}{2\pi r}, \quad (26)$$

where

$$w_{kl} = \sum_{s=0}^{p} \sum_{t=0}^{\min(k,s)} f_{ps}^m b_{kst} g_{k+s-2t,l}^m, \quad (27)$$

with $f$, $b$, $g$ given in (13)–(15). In the special case that $m = 0$ we have that

$$w_{k,k+p-2j} = b_{kpj}, \quad j = 0, 1, \ldots, \min(k, p), \quad (28)$$

while all other $w_{kl}$ vanish. Finally, all $w_{kl} \geq 0$ and $\sum_l w_{kl} = 1$.

To prove this result we start by using Bauer's formula (8) and we obtain

$$V_{nm}(r, f) = \exp\left(\tfrac{1}{2}\mathrm{i}f\right) \sum_{k=0}^{\infty} (2k + 1)\mathrm{i}^k j_k\left(\tfrac{1}{2}f\right) \int_0^1 R_{2k}^0(\rho) R_n^m(\rho) J_m(2\pi\rho r)\rho \, \mathrm{d}\rho. \quad (29)$$

Next we write

$$R_{2k}^0 R_{m+2p}^m = \sum_l w_{kl} R_{m+2l}^m, \quad (30)$$

for which we use the result of subsection 2.1 with $m_1 = 0$, $p_1 = k$, $m_2 = m$, $p_2 = p$, and then (11) and a fair amount of administration yields the result. Here it is useful to note that for the index range in (19) we may restrict to $s_1 = p_1 = k$ because of (16).

We observe that, by analyticity, formula (26) remains valid when $f$ is complex (instead of real) so that (26) can be used in applications such as those described in [9, 10].

We comment on the problem as to how many terms should be included in the series over $k$ in (26) to obtain a desired accuracy. First note that the number of terms in the series over $l$ is bounded by $p + q + 1 = n + 1$ and that $w_{kl} \geq 0$, $\sum_l w_{kl} = 1$. For large $k$ we may consider the case that $m + 2l + 1 = 2k + 1$ as typical since $2k - n + 1 \leq m + 2l + 1 \leq 2k + n + 1$ when $l$ is in the indicated range. Thus we shall bound

$$\left| j_k\left(\tfrac{1}{2}f\right) \right| \left| \frac{2k + 1}{2\pi r} J_{2k+1}(2\pi r) \right|. \quad (31)$$

It follows from [12], 10.1.50 on p. 440 and 9.1.62 on p. 362, and the elementary inequality $\Gamma(k + \frac{3}{2}) \geq (k + \frac{1}{2})^{1/2}\Gamma(k + 1)$ that

$$|j_k(\tfrac{1}{2}f)| \leq \frac{1}{(2k + 1)^{1/2}} \min\left(1, (\pi/2)^{1/2}\frac{|f/4|^k}{k!}\right). \tag{32}$$

Furthermore, by [14], bottom of p. 234 and [12], 9.1.62 on p. 362, we have

$$\left|\frac{2k + 1}{2\pi r}J_{2k+1}(2\pi r)\right| \leq \min\left(1, \frac{(\pi r)^{2k}}{(2k)!}\right). \tag{33}$$

From the bounds (32) and (33), one can easily derive rules of thumb for the truncation of the series over $k$ when a specified accuracy is desired. These rules are somewhat conservative since one can use instead of [12], 9.1.62 on p. 362, the sharper but more complicated inequality 9.1.63 on p. 362. It is useful to note that, due to the occurrence of the *min*-operand in (32) and (33), the product in (31) is small whenever one of the factors is small. When, for instance, $f = 100$ and $k \geq 75$ all quantities in (31) are less than $3 \times 10^{-6}$, no matter how small or large $r$ is. And when $f = 100$ and we include only 50 terms in the series over $k$, we still get reasonably accurate results as long as we restrict to $r$'s with $2\pi r < 70$.

## 3. Evaluation of $T_{nm}$

We shall show now how the integrals

$$T_{nm} = \int_0^1 \rho^n \exp(\mathrm{i}f\rho^2)J_m(2\pi\rho r)\rho \, \mathrm{d}\rho \tag{34}$$

with integer $n, m \geq 0$ such that $n - m \geq 0$ and even can be computed. In [6] the $T_{nm}$ have been studied and a series expansion of the type (3) has been derived for them. Here we want formulas for the $T_{nm}$ of the type as those presented in section 2. Write $n = m + 2k$ with $k = 0, 1, \ldots$. Then we have

$$\rho^{m+2k} = \sum_{p=0}^k h_{kp}^m R_{m+2p}^m(\rho); \quad h_{kp}^m = \frac{m + 2p + 1}{m + p + k + 1}\left[\binom{k}{p}\bigg/\binom{m + k + p}{p}\right], \tag{35}$$

a result that was announced in [6], formula (10), and proved in [8], Appendix A. Accordingly, there holds

$$T_{m+2k,m} = \sum_{p=0}^k h_{kp}^m V_{m+2p,m}, \tag{36}$$

so that the main result of section 2 applies directly. For the special case that $m = 0$, as was considered in [13], a simplification of the eventual computation scheme occurs in accordance with (28).

An alternative method to obtain a stable computation scheme for an integral $T_{m+2p,m}$ with $p = 0, 1, \ldots$, is as follows. We have

$$T_{m+2p,m} = \left(\frac{\mathrm{d}}{\mathrm{i}\,\mathrm{d}f}\right)^p V_{mm}, \tag{37}$$

while

$$V_{mm}(r, f) = \sum_{k=0}^\infty (2k + 1)\mathrm{i}^k \exp\left(\tfrac{1}{2}\mathrm{i}f\right)j_k(\tfrac{1}{2}f) \sum_{l=\max(0,k-m)}^k (-1)^l g_{kl}^m \frac{J_{m+2l+1}(2\pi r)}{2\pi r}. \tag{38}$$

We note here that when $n = m$ certain simplifications occur in (27). We are thus left with the expressions

$$\left(\frac{\mathrm{d}}{\mathrm{i}\,\mathrm{d}f}\right)^p [\exp\left(\tfrac{1}{2}\mathrm{i}f\right)j_k(\tfrac{1}{2}f)]. \tag{39}$$

Using

$$j_k(z) = \frac{1}{2}(-\mathrm{i})^k \int_{-1}^{1} \exp\left(\mathrm{i}zt\right)P_k(t)\,\mathrm{d}t, \tag{40}$$

with $P_k = P_k^{(0,0)}$ the Legendre polynomial of degree $k$, one can write the expression in (39) as a linear combination of $\exp\left(\tfrac{1}{2}\mathrm{i}f\right)j_{k+v}(\tfrac{1}{2}f)$, with $v$ ranging through all integers between $-\min(p, k)$ and $p$ and explicitly given coefficients. The resulting expression for $T_{m,m+2p}$ is of the same complexity as what one gets by using (36).

## 4. Extension to high-numerical aperture systems

In [8], the Nijboer–Zernike approach has been further extended so as to include aberrated optical systems with high numerical aperture using a vectorial diffraction formalism à la Ignatowsky/Richards and Wolf. The basic integrals to be considered can then be expressed in terms of

$$\int_0^1 \rho^{|j|}R_n^m(\rho)\,\exp\left(\mathrm{i}f\rho^2\right)J_{m+j}(2\pi\rho r)\rho\,\mathrm{d}\rho, \tag{41}$$

see [8], formula (17) and the developments in [8], Appendix B. Here $j = -2, -1, 0, 1, 2$. In [8] it has been indicated, through formulas (B30)–(B33), how the products $\rho^{|j|}R_n^m(\rho)$ can be written as a linear combination of at most $|j| + 1$ Zernike polynomials with upper index $m + j$. Then we are in the situation of section 2 and we can apply the main result directly.

## 5. Examples

In this section we compare the methods embodied by formulas (3)–(5) and by the main result in subsection 2.2, respectively, for the computation of some $V_{nm}$'s with respect to the $(r, f)$ range for which they produce accurate results. We do not make a comparison of CPU times here, neither have we made any effort to optimize the computer codes in this respect. Both formulas (3)–(5) and formulas (26)–(27) are converted straightforwardly into computer codes, especially when the available software environment allows accurate computation of high order (spherical) Bessel functions. We shall also briefly compare the methods for computation of $T_{nm}$'s.

### 5.1. *Accuracy comparison for the $V_{nm}$-functions*

The computation of the $V_{nm}$ are in all cases carried out using 15 significant decimal places. We write $V_{nm}^{\mathrm{PBS}}$ (power Bessel series) and $V_{nm}^{\mathrm{BBS}}$ (Bessel–Bessel series) to indicate that $V_{nm}$ has been computed by using formulas (3)–(5) and formulas (26)–(27), respectively.

We first show that $V_{nm}^{\mathrm{PBS}}$ and $V_{nm}^{\mathrm{BBS}}$ produce the same results within the given accuracy when $f$ is small. The role of $r$ is much less critical in this respect due to

the bounds on the Bessel functions as in (33). Thus in figures 1, 2 and 3 we show the real and imaginary parts of $2V_{nm}^{\text{PBS}}$, $2V_{nm}^{\text{BBS}}$ as well as the intensities $|2V_{nm}^{\text{PBS}}|^2$, $|2V_{nm}^{\text{BBS}}|^2$ and their absolute difference $\left||2V_{nm}^{\text{PBS}}|^2 - |2V_{nm}^{\text{BBS}}|^2\right|$ as a function of $v = 2\pi r > 0$ with $f = 0, 2, 4$ for the cases that $(n, m) = (0, 0)$, $(8, 2)$ and $(23, 11)$, respectively. The factor 2 in front of the $V_{nm}$'s has been included to have agreement with formula (1) which shows how the single aberrations $\beta_{nm}R_n^m(\rho)\cos m\theta$ contribute to the point-spread function $U$. We have chosen the number $L_{\max}$ of included terms in the series over $l$ in (3) and over $k$ in (26) to be 150. In view of the estimates (32)–(33) this is quite an overkill for the series in (26). For the series in (3) this is an overkill as well: it follows from the analysis in [7], Appendix B that the $l$th term has a maximum modulus (typically assumed for $2\pi r$ near $m$) of the order $\frac{1}{2}|f|^{l-1}/l!$ or less.

We next consider $V_{00}$ for large values of $f$, and in figure 4 we show the quantity

$$4 \max_{2\pi r \leq 30} \left||V_{00}^{\text{PBS}}(r, f)|^2 - |V_{00}^{\text{BBS}}(r, f)|^2\right| \tag{42}$$

as a function of $f$, $0 \leq f \leq 40$. The terms in the series $V_{00}^{\text{PBS}}$ have largest modulus for $l \approx f$ and $r = 0$ of the order $\exp(f)/2f(2\pi f)^{1/2}$ while $V_{00}$ itself has a modulus of
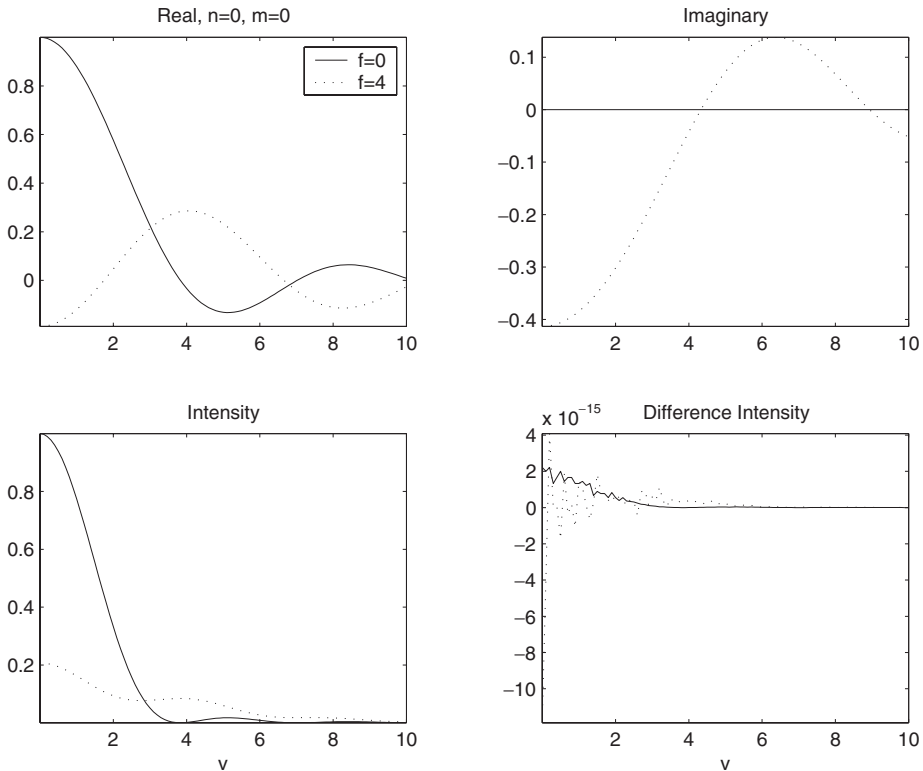


Figure 1. The real and imaginary parts of $2V_{nm}^{\text{PBS}}$ and $2V_{nm}^{\text{BBS}}$ (upper figures) as well as the intensities $|2V_{nm}^{\text{PBS}}|^2$ and $|2V_{nm}^{\text{BBS}}|^2$ (lower left-hand figure) as a function of $v = 2\pi r > 0$ with $f = 0$ (drawn lines) and $f = 4$ (dotted lines). The PBS and BBS curves visually coincide in these figures. In the lower right-hand figure, using a strongly different vertical scale, the absolute intensity difference $\left||2V_{nm}^{\text{PBS}}|^2 - |2V_{nm}^{\text{BBS}}|^2\right|$ has been plotted. In all figures the case $(n, m) = (0, 0)$ has been considered.
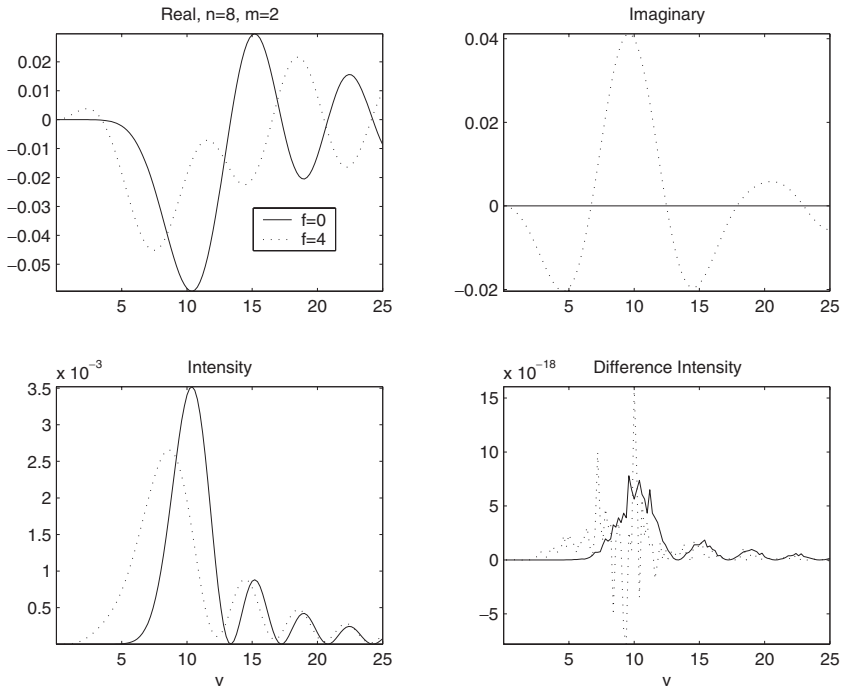
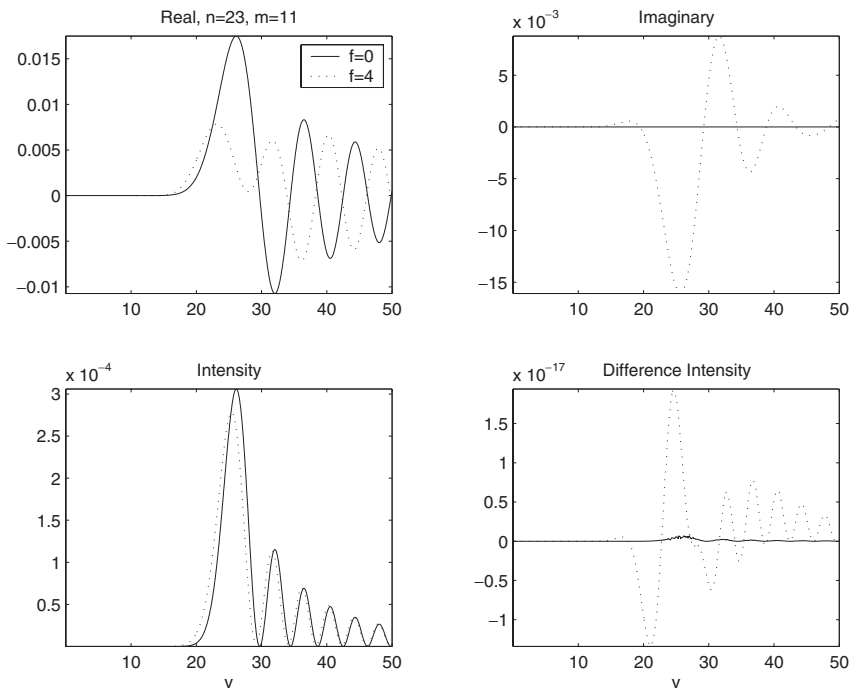Figure 2.    Same legend as in figure 1, case $(n, m) = (8, 2)$.



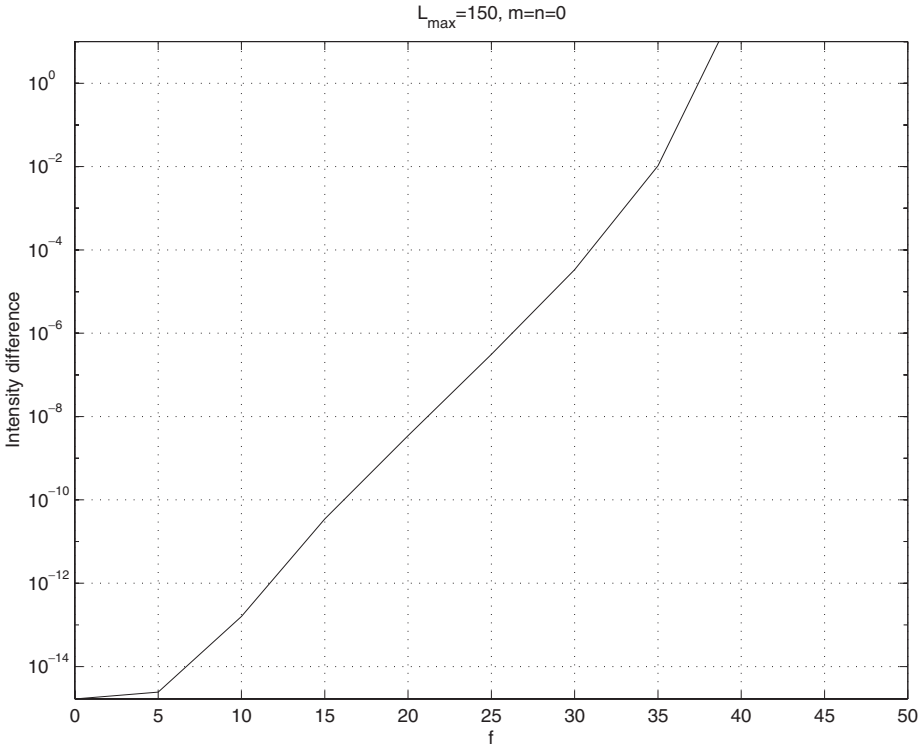Figure 3.    Same legend as in figure 1, case $(n, m) = (23, 11)$.

Figure 4.    The quantity (42) on a $^{10}$log-scale for $0 \leq f \leq 35$.

the order $1/f$. Accordingly, loss of digits in the computation of $V_{00}$ as $V_{00}^{\text{PBS}}$ occurs according to a $f/\ln 10$ law. Figure 4 seems to confirm this statement.

## 5.2. *Strongly defocused fields*

We have carried out a numerical experiment in a lithographic projection system that leads to a large focal shift in the image plane. To this goal, we have introduced a Fresnel zone plate pattern on the mask, comprising 5 zones, and corresponding to an effective focal length of 5 μm at the image side. In the image plane, the intensity distribution is strongly defocused and the sharp focus is only found at a distance 5 μm beyond the paraxial image plane. In figure 5, left, a cross-section of the radially symmetric intensity pattern has been given that contains the $z$ axis. It is clearly seen that the highest intensity is found at a shifted distance $z \approx -5$ μm. In figure 5, right, the axial intensity has been plotted as a function of the axial $z$ coordinate, showing the same focal shift phenomenon due to the presence of the zone plate structure.

We next show in figure 6 the quantity $|2V_{00}(r,f)|^2$ for the (very large) value of $f = 100$, in which we have used $V_{00}^{\text{BBS}}$, as a function of $r$. The estimates in (32)–(33) show that it is sufficient to include some 100 terms in the series over $k$ to guarantee accuracy of the order $10^{-15}$. For large $f$ there is the approximation

$$V_{00}(r,f) \approx \frac{J_0(2\pi r) \exp{(\mathrm{i}f)} - 1}{2\mathrm{i}f}, \tag{43}$$
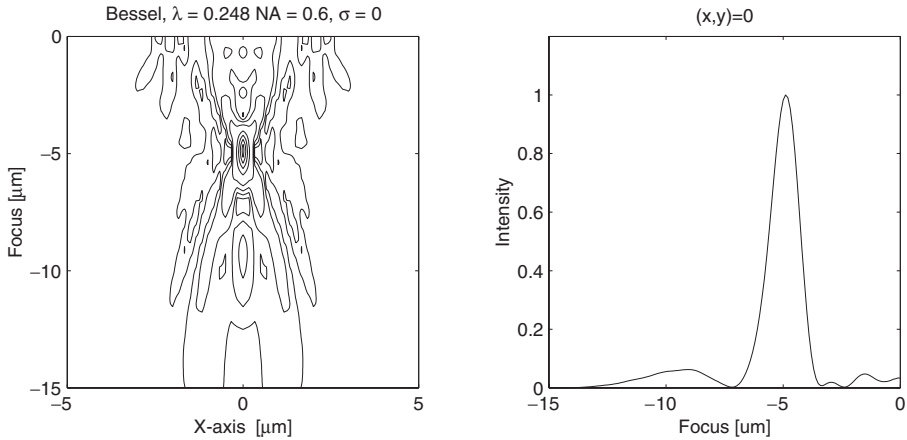
Figure 5. The simulation of a large defocusing by means of a Fresnel zone plate structure in the object plane of a lithographic imaging system. Left-hand panel: an intensity contour plot belonging to a cross-section that contains the $z$ axis. The concentration of intensity in the defocused position due to the action of the zone plate ($z = -5\,\mu m$) is clearly visible. Right-hand panel: a plot of the axial intensity showing again the shifted focal maximum at $z = -5\,\mu m$.

valid for relatively small $r$ and shown to hold by partial integration, while

$$V_{00}(r,f) \approx \frac{1}{2\pi r^{1/2}}\left[\exp(\pi i/4)\int_0^1 \exp(if\rho^2)\exp(-2\pi i\rho r)\rho^{1/2}\,d\rho \right.$$
$$\left. + \exp(-\pi i/4)\int_0^1 \exp(if\rho^2)\exp(2\pi i\rho r)\rho^{1/2}\,d\rho\right], \qquad (44)$$

valid for larger values of $r$ and obtained by using the asymptotics $J_0(x) \approx (2/\pi x)^{1/2}\cos(x-\pi/4)$, $x \to \infty$, in the integral representation of $V_{00}$. From (43)–(44) some of the observations that can be done in figure 6 can be explained. In particular, the low-amplitude wrinkles on the decaying side of $|2V_{00}(r,f)|^2$ around $v = 2\pi r = 200$ can be identified as an interference of the (small) term on the second line of (44) and the (large) term on the first line.

5.3. *Accuracy comparison for the $T_{nm}$ functions*

We conclude by presenting some comparisons of the methods to compute $T_{nm}$. There is the power-Bessel series $T_{nm}^{\mathrm{PBS}}$ of [6],

$$T_{nm}^{\mathrm{PBS}} = \exp(if)\sum_{l=1}^{\infty}(-2if)^{l-1}\sum_{j=0}^{p}t_{lj}\frac{J_{m+l+2j}(2\pi r)}{(2\pi r)^l}, \qquad (45)$$

where $p = \frac{1}{2}(n-m)$, $q = \frac{1}{2}(n+m)$ and

$$t_{lj} = (-1)^j\frac{m+l+2j}{q+1}\left[\binom{p}{j}\binom{m+j+l-1}{l-1}\Big/\binom{q+l+j}{q+1}\right], \quad l,j = 0,1,\ldots. \quad (46)$$
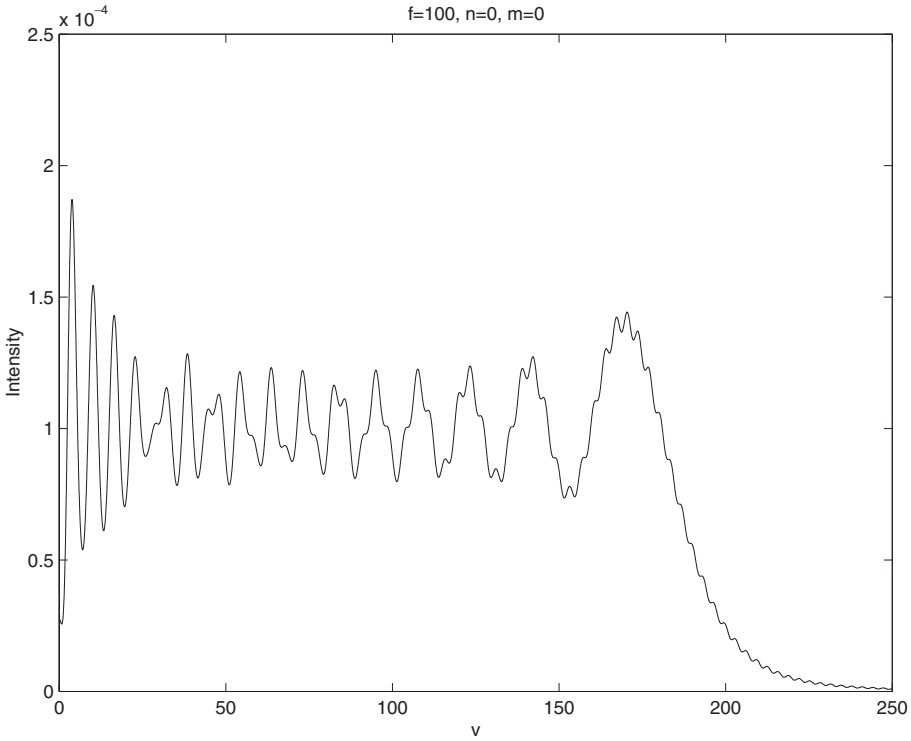
Figure 6.    The quantity $|2V_{00}(r, f = 100)|^2$ as a function of $r$, $0 \le r \le 250$.

Furthermore, there is the Bessel–Bessel series representation $T_{nm}^{\mathrm{BBS}}$ of section 3 via the $V_{nm}^{\mathrm{BBS}}$. Finally, for the special case that $m = 0$, $n = 2p$, there is the representation $T_{2p,0}^{\mathrm{Cao}}$ as given by Cao in [13] via the Jinc functions in (7)

$$T_{2p,0}^{\mathrm{Cao}} = \sum_{l=0}^{\infty} \frac{(\mathrm{i}f)^l}{l!} \, \mathrm{Jinc}_{p+l}(2\pi r) = \sum_{l=0}^{\infty} \frac{(\mathrm{i}f)^l}{l!} \sum_{j=0}^{p+l} (-2)^j \frac{(p+l)!}{(p+l-j)!} \frac{J_{j+1}(2\pi r)}{(2\pi r)^{j+1}}. \qquad (47)$$

The convergence properties of the two series in (45) and (47) are similar to those of $V_{nm}^{\mathrm{PBS}}$; in particular, the maximum modulus of the $l$th term in either series is of the order $|f|^{l-1}/l!$ or less. As to $T_{nm}^{\mathrm{PBS}}$ we refer for this matter to [7], Appendix B. For $T_{2p,0}^{\mathrm{Cao}}$ this is a consequence of the inequality $|\mathrm{Jinc}_p(u)| \le 1/2(p+1)$ (equality if and only if $u = 0$) that was conjectured in [13] and that easily follows from the fact that $|J_0(v)| \le 1$ (equality if and only if $v = 0$).

As an example we consider the case $T_{8,0}$, for which we have the explicit result, see (36),

$$T_{8,0} = \frac{1}{5} V_{0,0} + \frac{2}{5} V_{2,0} + \frac{2}{7} V_{4,0} + \frac{1}{10} V_{6,0} + \frac{1}{70} V_{8,0}. \qquad (48)$$

By producing the analogues of figure 1 with $T_{8,0}^{\mathrm{PBS}}$ and $T_{8,0}^{\mathrm{Cao}}$ instead of $V_{0,0}^{\mathrm{PBS}}$, respectively, and with $T_{8,0}^{\mathrm{BBS}}$ instead of $V_{0,0}^{\mathrm{BBS}}$ as reference, one can conclude that $T_{8,0}^{\mathrm{PBS}}$ and $T_{8,0}^{\mathrm{Cao}}$ are comparably accurate for relatively small values of $f$.
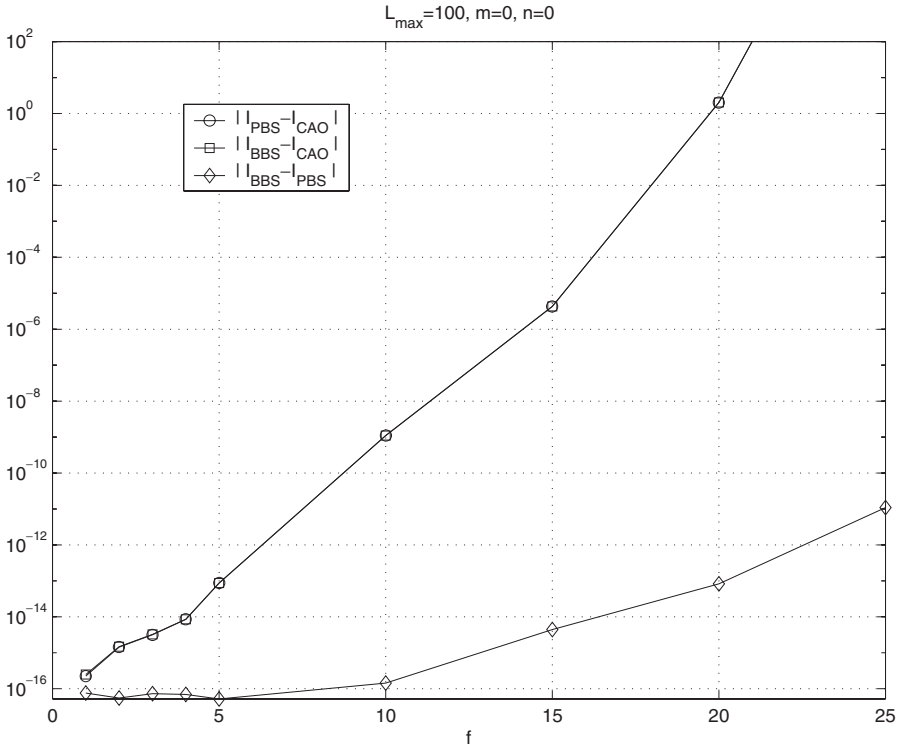
Figure 7. The quantities (49) on a $^{10}$ log scale for $0 \leq f \leq 25$. The curve with the circles and the one with the squares (A and B in the text, respectively) virtually coincide. The lower curve that has been labelled with the rhombs corresponds to case C in the text.

The behaviour of the various calculated $T$ functions for larger values of $f$ is shown in figure 7. We have plotted the quantity

$$4 \max_{2\pi r \leq 30} \left| |T^a_{8,0}(r,f)|^2 - |T^b_{8,0}(r,f)|^2 \right| \tag{49}$$

as a function of $f \leq 25$. Three curves A, B and C have been produced with the following settings:

(i) curve A: a = PBS; b = Cao;
(ii) curve B: a = BBS; b = Cao;
(iii) curve C: a = BBS; b = PBS.

Recalling the results from the preceding sections, we are allowed to take the values calculated according to the Bessel–Bessel method (BBS) as the absolute reference because of the guaranteed convergence of this method. The fact that curves A and B virtually coincide shows that the PBS and BBS methods produce an equal disparity with respect to the Cao method. The third curve, C, shows that the PBS method is inaccurate at a level of $10^{-12}$ with respect to the absolutely accurate BBS method. This permits us to conclude that the highly inaccurate results for large $f$ values of curves A and B can be imputted to the lack of convergence for larger $f$ values of the Cao method.

Although all three methods $T^{\text{Cao}}$, $T^{\text{PBS}}$, $T^{\text{BBS}}$ are easy to implement, it should be said that $T^{\text{Cao}}$ is the simplest in this respect, followed by $T^{\text{PBS}}$. Thus, when a $10^{-10}$ accuracy level for intensity values of $T_{8,0}$ is required, one can safely use $T^{\text{Cao}}$ up to $f = 10$ and $T^{\text{PBS}}$ until $f = 25$ while for values of $f$ in excess of 25, it is imperative to use $T^{\text{BBS}}$.

## 6. Conclusion

A new analytic calculation method has been devised that solves the convergence problem when computing strongly defocused aberrated diffraction patterns in the extended Nijboer–Zernike theory. The resulting analytic expression contains an expansion in terms of well-converging spherical Bessel functions for the axial defocusing parameter and an expansion in Bessel functions of the first kind for the lateral coordinate in the defocused plane. By this separation of the coordinates in our solution, we directly obtain the diffraction pattern in a defocused plane in terms of a set of polar coordinates in that plane. The accuracy of the so-called Bessel–Bessel method is guaranteed to relative intensity values down to machine precision due to the fact that the coefficients of the analytic series expansion are all positive and bounded to the interval $[0, 1]$. The desired accuracy is obtained by using a simple truncation rule for the series expansion.

We have compared the new method with some of the existing ones and concluded that the convergence of former methods (see [6, 7, 13], respectively) is appropriate for a total axial defocusing range of typically ten focal depths. It is shown that the new method is valid for all, including very large, defocusing values so that calculations deep in the Fresnel regime yield reliable and accurate results. The range is only limited by the practical calculation time, not by the convergence of the calculation method. Because of its basic accuracy, the new method can also be used to check numerical methods (e.g. Fourier transform methods) with respect to the required sampling density for achieving a desired precision.

## References

[1] MANSURIPUR, M., 1986, *J. Opt. Soc. Am.* A, **3,** 2086.
[2] MANSURIPUR, M., 1989, *J. Opt. Soc. Am.* A, **6,** 786.
[3] KANT, R., 1993, *J. Mod. Optics*, **40,** 337.
[4] KANT, R., 1993, *J. Mod. Optics*, **40,** 2293.
[5] SHEPPARD, C. J. R., and TÖRÖK, P., 1997, *J. Mod. Optics*, **44,** 803.
[6] JANSSEN, A. J. E. M., 2002, *J. Opt. Soc. Am.* A, **19,** 849.
[7] BRAAT, J. J. M., DIRKSEN, P., and JANSSEN, A. J. E. M., 2002, *J. Opt. Soc. Am.* A, **19,** 858.
[8] BRAAT, J. J. M., DIRKSEN, P., VAN DE NES, A. S., and JANSSEN, A. J. E. M., 2003, *J. Opt. Soc. Am. A*, **20,** 2281.
[9] DIRKSEN, P., BRAAT, J. J. M., JANSSEN, A. J. E. M., and JUFFERMANS, C., 2003, *J. Microlith. Microfab. Microsyst.*, **2,** 61.

[10] DIRKSEN, P., BRAAT, J., JANSSEN, A. J. E. M., JUFFERMANS, C., and LEEUWESTEIN, A., 2003, *Proc. SPIE*, Vol. 5040 (Bellingham, WA: SPIE).

[11] BORN, M., and WOLF, E., 2002, *Principles of Optics* (Cambridge, UK: Cambridge).

[12] ABRAMOWITZ, M., and STEGUN, I. A., 1970, *Handbook of Mathematical Functions* (New York: Dover).

[13] CAO, Q., 2003, *J. Opt. Soc. Am.* A, **20,** 661.

[14] BOERSMA, J., 1962, *Math. Comput.*, **16,** 232.

[15] NIJBOER, B. R. A., 1942, PhD thesis, University of Groningen, The Netherlands.

[16] ZERNIKE, F., and NIJBOER, B. R. A., 1949, *La Théorie des Images Optiques* (Paris: La Revue d'Optique), p. 227.

[17] ANDREWS, G. E., ASKEY, R., and ROY, R., 1999, *Special Functions* (Cambridge, UK: Cambridge).

[18] SZEGÖ, G., 1975, *Orthogonal Polynomials* (Providence: American Mathematics Society).

[19] KOORNWINDER, T., 1978, *J. Lond. Math. Soc.*, **18,** 101.

[20] RAHMAN, M., 1981, *Can. J. Math.*, **33,** 915.