Validity of the perturbation model for the propagation of MSF structures in 3D

Kevin Liang, G. Forbes, Miguel Alonso

Research Article | Vol. 28, No. 14 / 6 July 2020 / Optics Express 20277
Optics EXPRESS

Validity of the perturbation model for the propagation of MSF structures in 3D

KEVIN LIANG,[1,2,*] G. W. FORBES,[2,3] AND MIGUEL A. ALONSO[1,2,4]

[1]The Institute of Optics, University of Rochester, Rochester, NY 14627, USA
[2]Center for Freeform Optics, University of Rochester, Rochester, NY 14627, USA
[3]Department of Physics and Astronomy, Macquarie University, North Ryde 2109, Sydney, NSW, Australia
[4]Aix Marseille Univ., CNRS, Centrale Marseille, Institut Fresnel, UMR 7249, 13397 Marseille Cedex 20, France
[*]miguel.alonso@fresnel.fr

Abstract: Mid-spatial frequency (MSF) structures on optical surfaces degrade system performance and a perturbation model is typically used to simplify the assessment of their effects. In this simple model, MSF phase structures are dragged along the nominal rays of a system to yield estimates of wavefronts in the exit pupil that may be used for further analysis. However, the validity of the perturbation model remains an open area of study. We extend our previous assessment of the validity of this model [K. Liang, Opt. Express **27**, 3390-3408 (2019)] that was focused on the analysis of single-frequency MSF structures in two dimensions to now include error estimates for broad-spectra MSF structures in three dimensions.

1. Introduction

Mid-spatial frequency (MSF) structures are inevitable in most aspheric and freeform optical systems due to the subaperture tools that are used during the manufacturing process. Their characteristic frequencies lie between those of the common low-order aberrations and high-order scattering, and their detrimental effects on optical performance remain an active area of research. For example, there have been many efforts towards simplifying the tolerancing of optical parts afflicted with MSF [1–6]. To this end, the perturbation model is often used to cut down on the computation time needed to understand the propagation of MSF structures. This model, in which the MSF phase structure (which can vary significantly from part to part) is simply dragged along rays of the nominal system, is often used in order to avoid the need for new ray tracing for each MSF realization [7]. However, the validity of this perturbation model requires further treatment; its analysis in two dimensions was presented in Ref. 8 and those results are extended in Appendix A in a manner that we now generalize to three dimensions.

The mathematical framework used in this manuscript to estimate the error incurred by the perturbation model is based on an asymptotic analysis of the Helmholtz wave equation for the propagation of a monochromatic field in free space. A key step is the placement of the MSF structure at one asymptotic order beneath the nominal wavefront. This framework is similar to that in Ref. 8, but the solution now includes contributions from every asymptotic order (under appropriate approximations). This extension permits the analysis of MSF structures with broad spatial-frequency spectra.

2. Asymptotic propagation estimate based on nominal rays

The propagation of a monochromatic scalar field, $\mathrm{Re}\left[U(\boldsymbol{r})e^{-\mathrm{i}\omega t}\right]$, in a homogeneous medium is governed by the Helmholtz equation

$$\nabla^2 U(\boldsymbol{r}) + k^2 U(\boldsymbol{r}) = 0, \tag{1}$$

#395493 https://doi.org/10.1364/OE.395493
Journal © 2020 Received 23 Apr 2020; revised 29 May 2020; accepted 9 Jun 2020; published 24 Jun 2020

Optics EXPRESS

where $k = \omega/c = 2\pi/\lambda$ is the wavenumber in the medium. The field is taken to be propagating towards larger z and, at some reference plane $z = z_{\mathrm{M}}$, we take the initial value of the field to be given nominally by $U(\boldsymbol{r}_\perp, z_{\mathrm{M}}) = U_0 A(\boldsymbol{r}_\perp)\exp[\mathrm{i}kW(\boldsymbol{r}_\perp)]$, where U_0 is a constant with field units and $\boldsymbol{r}_\perp = (x, y)$ are the transverse coordinates. At $z = z_{\mathrm{M}}$, we also superpose an MSF phase factor of the form $\exp[\mathrm{i}\phi(\boldsymbol{r}_\perp)]$, where $\phi(\boldsymbol{r}_\perp)$ is taken to have zero mean and the magnitude of its variation is less than π. Moreover, given its characterization as an MSF structure, $\phi(\boldsymbol{r}_\perp)$ is assumed to vary more rapidly than either $W(\boldsymbol{r}_\perp)$ or $A(\boldsymbol{r}_\perp)$. The goal now is to derive an estimate, in a manner that is similar to that in Ref. 8, of how this MSF phase structure affects the field under propagation.

We begin by writing $U(\boldsymbol{r}) = U_0\exp[\mathrm{i}k\Phi(\boldsymbol{r})]$, where $\Phi(\boldsymbol{r})$ is a complex quantity that accounts for spatial variations in both the phase and amplitude. With this, Eq. (1) becomes

$$\nabla\Phi\cdot\nabla\Phi - 1 = -\frac{1}{\mathrm{i}k}\nabla^2\Phi, \tag{2}$$

Eq. (2) can be solved upon expressing Φ as an asymptotic series in the parameter $(\mathrm{i}k)^{-1}$:

$$\Phi(\boldsymbol{r}) = \sum_{N=0}^{\infty}\frac{\Phi_N(\boldsymbol{r})}{(\mathrm{i}k)^N}. \tag{3}$$

By using Eq. (3) with Eq. (2) and separating terms of equal powers of k, we arrive at

$$\nabla\Phi_0\cdot\nabla\Phi_0 = 1, \tag{4}$$

$$\nabla\Phi_0\cdot\nabla\Phi_N = -\frac{1}{2}\left(\nabla^2\Phi_{N-1} + \sum_{n=1}^{N-1}\nabla\Phi_n\cdot\nabla\Phi_{N-n}\right), \quad N \geq 1. \tag{5}$$

The initial conditions discussed above can now be stated as $\Phi_0(\boldsymbol{r}_\perp, z_{\mathrm{M}}) = W(\boldsymbol{r}_\perp)$, $\Phi_1(\boldsymbol{r}_\perp, z_{\mathrm{M}}) = \ln[A(\boldsymbol{r}_\perp)] + \mathrm{i}\phi(\boldsymbol{r}_\perp)$, and $\Phi_N(\boldsymbol{r}_\perp, z_{\mathrm{M}}) = 0$ for $N \geq 2$. It is now possible to work to progressively higher orders by integrating in z at each order from these initial conditions.

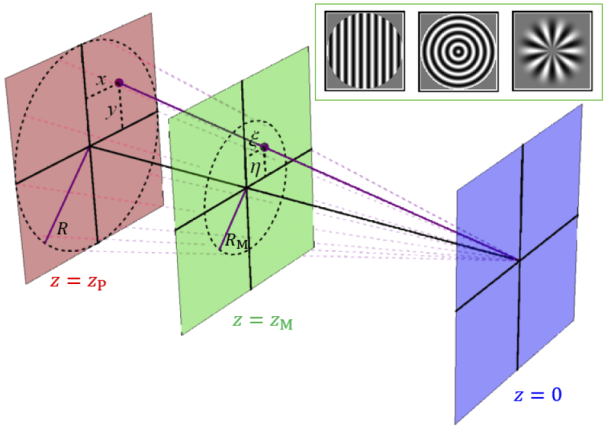

Fig. 1. The image space of an imaging system, where the image plane (blue) is placed at $z = 0$. The exit pupil (red) and the image of the MSF phase structure (green) are located at z_{P} and z_{M}, respectively. Note that $\boldsymbol{\xi}_\perp$ is the location of the intersection of the ray starting at $\boldsymbol{r}_\perp$ in the plane $z = z_{\mathrm{M}}$. The radius of the exit pupil is R and that of the beam footprint at z_{M} is R_{M}. The inset shows, from left to right, examples of MSF structures with what are referred to here as milled, turned, and spoked geometries.

Optics EXPRESS

We begin with Eq. (4), the well-known Hamilton-Jacobi or Eikonal equation, which can be solved in terms of nominal rays by using the following parametrization involving $\boldsymbol{\xi} = (\xi, \eta, s)$:

$$x(\boldsymbol{\xi}) = \xi + s\partial_\xi W(\boldsymbol{\xi}_\perp), \quad y(\boldsymbol{\xi}) = \eta + s\partial_\eta W(\boldsymbol{\xi}_\perp), \quad z(\boldsymbol{\xi}) = z_{\mathrm{M}} + s\chi(\boldsymbol{\xi}_\perp), \tag{6}$$

where $\boldsymbol{\xi}_\perp = (\xi, \eta)$ are the transverse coordinates at the reference plane ($z = z_{\mathrm{M}}$) and s is the arclength along the ray. The direction of each ray is given by the unit vector $[\nabla_{\boldsymbol{\xi}_\perp} W(\boldsymbol{\xi}_\perp), \chi(\boldsymbol{\xi}_\perp)]$, where $\chi(\boldsymbol{\xi}_\perp) \triangleq \sqrt{1 - \left|\nabla_{\boldsymbol{\xi}_\perp} W(\boldsymbol{\xi}_\perp)\right|^2}$ with $\triangleq$ denoting a definition and with $\nabla_{\boldsymbol{\xi}_\perp} \triangleq (\partial_\xi, \partial_\eta)$. Figure 1 shows the relation between $\boldsymbol{r}_\perp$ and $\boldsymbol{\xi}_\perp$ for a nominally converging wavefront. It is shown in Appendix B that the solution to the Eikonal equation is simply

$$\overline{\Phi_0}(\boldsymbol{\xi}) = W(\boldsymbol{\xi}_\perp) + s, \tag{7}$$

where an overline on any function $f(\boldsymbol{r})$ indicates that it is being expressed in terms of the ray parameters $\boldsymbol{\xi}$ by using Eq. (6), i.e., $\overline{f}(\boldsymbol{\xi}) \triangleq f[x(\boldsymbol{\xi}), y(\boldsymbol{\xi}), z(\boldsymbol{\xi})]$. Notice that, as a consequence of placing ϕ one asymptotic order below W, the rays used in this analysis are the nominal rays, which are not affected by the MSF structure.

It is furthermore shown in Appendix B that Eq. (5) can also be solved in terms of the parametrization in Eq. (6). For $N = 1$, the parametrized solution is

$$\overline{\Phi_1}(\boldsymbol{\xi}) = \ln\left[A(\boldsymbol{\xi}_\perp)\sqrt{\frac{\chi(\boldsymbol{\xi}_\perp)}{\Delta(\boldsymbol{\xi})}}\right] + \mathrm{i}\phi(\boldsymbol{\xi}_\perp) + \mathrm{i}\theta_{\mathrm{GM}}, \tag{8}$$

where

$$\Delta(\boldsymbol{\xi}) \triangleq \left|\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\xi}}\right| = \chi + s\left(\chi\nabla^2_{\boldsymbol{\xi}_\perp} W - \nabla_{\boldsymbol{\xi}_\perp}\chi \cdot \nabla_{\boldsymbol{\xi}_\perp} W\right) + s^2 \det(\mathbb{H}_W)/\chi, \tag{9}$$

is the Jacobian determinant of the coordinate transformation given in Eq. (6), $\mathbb{H}_W$ is the Hessian matrix of W, and θ_{GM} is the Gouy-Maslov phase shift. This phase shift is a straightforward extension to three dimensions of that discussed in Appendix C of Ref. 8. As also discussed in Ref. 8, the first term of Eq. (8) accounts for the change in the amplitude due to the bunching or spreading of the nominal rays under propagation; the second term indicates that, asymptotically, the effect of the MSF phase structure can be modeled by simply dragging this phase along the (nominal) rays. This is precisely the perturbation model.

In what follows, the method used for proceeding to larger values of N differs from that presented in Ref. 8. In order to appreciate the three-dimensional results, however, it is helpful to revisit the two-dimensional case and present the mathematical framework upon which the full three-dimensional treatment will follow by analogy, see Appendix A. As a reminder of the derivation in Ref. 8, recall that only the first correction to the perturbation model was analyzed. That is, the series in Eq. (3) is truncated at $N = 2$, hence the field is taken to be approximated by

$$U = U_0 \exp(\mathrm{i}k\Phi) \approx U_0 \exp\left(\mathrm{i}k\Phi_0 + \Phi_1 + \frac{\Phi_2}{\mathrm{i}k}\right). \tag{10}$$

By using the field estimate in Eq. (10), the resulting rules of thumb for the error the perturbation model were ultimately found to be related to the fourth spectral moment of the MSF structure ϕ. This result inspired the development of a new family of rapidly decaying Fourier-like basis functions that yield finite spectral moments [9]. However, an alternative method can be used so that the field estimate contains contributions from every term on the right-hand side of Eq. (3). Although more approximations are necessary for this route, they are consistent with the assumptions regarding ϕ and are detailed in the full derivation shown in Appendix A. As a result, a better error estimate (not limited to $N \leq 2$) for the perturbation model is obtained.

Optics EXPRESS

It is convenient, and necessary with regards to the derivations in Appendices A and B, to now work in image space in cases where, to a good approximation, a wavefront propagating from a point object source converges onto a point on the image plane. As in Ref. 8, we consider for simplicity only the on-axis object point, whose ideal image is located at the origin. It should be noted, however, that the analysis that follows can be used for off-axis object points as well since the choice of the origin was made out of convenience and similar methods can be applied to off-axis field points. Furthermore, we assume that the MSF content on each optical surface is adequately resolved in its corresponding conjugate plane in image space. Under these assumptions, the dominant error of the perturbation model is associated with the process of simply dragging these MSF structures along the nominally converging rays from their conjugate planes to the exit pupil. Henceforth, we will work with a single optical surface (with MSF structures) whose conjugate plane is located at $z = z_{\mathrm{M}}$. Furthermore, the locations of the exit pupil plane and the image plane are taken to be $z = z_{\mathrm{P}}$ and $z = 0$, respectively, see Fig. 1. With this framework, the nominal (converging) wavefront and obliquity factor are given by

$$W(\boldsymbol{\xi}_\perp) = z_{\mathrm{M}}\sqrt{1 + \frac{|\boldsymbol{\xi}_\perp|^2}{z_{\mathrm{M}}^2}} \quad \text{and} \quad \chi(\boldsymbol{\xi}_\perp) = \frac{z_{\mathrm{M}}}{W(\boldsymbol{\xi}_\perp)}. \tag{11}$$

To assess the error incurred by the perturbation model, one must go beyond the $N = 1$ term in Eqs. (3) and (5). It is shown in Appendix B, under the approximations that ϕ is small and that it varies more rapidly than the nominal quantities, that

$$\overline{\Phi}(\boldsymbol{\xi}) \approx \frac{1}{k}\exp\left\{\frac{sz_{\mathrm{M}}\hat{\mathcal{W}}}{2ik\chi^3(\boldsymbol{\xi}_\perp)[s + W(\boldsymbol{\xi}_\perp)]}\right\}\phi(\boldsymbol{\xi}_\perp) + \frac{1}{ik}\ln\left[A(\boldsymbol{\xi}_\perp)\sqrt{\frac{\chi(\boldsymbol{\xi}_\perp)}{\Delta(\boldsymbol{\xi})}}\right] + \overline{\Phi_0}(\boldsymbol{\xi}), \tag{12}$$

where

$$\hat{\mathcal{W}} \triangleq \partial_r^2 + \frac{\chi^2}{r}\partial_r + \frac{\chi^2}{r^2}\partial_\theta^2, \tag{13}$$

is a differential operator, expressed in plane-polar coordinates (r, θ) where $r = \sqrt{\xi^2 + \eta^2}$ and $\theta = \arg(\xi + i\eta)$. This differential operator is discussed further in Appendix C.

The perturbation model is given by

$$\begin{aligned}\overline{U_{\mathrm{P}}}(\boldsymbol{\xi}) &= \bar{U}(\boldsymbol{\xi})\Big|_{\phi=0}\exp[i\phi(\boldsymbol{\xi}_\perp)] \\ &= U_0 A(\boldsymbol{\xi})\sqrt{\frac{\chi(\boldsymbol{\xi}_\perp)}{\Delta(\boldsymbol{\xi})}}\exp\left\{ik[W(\boldsymbol{\xi}_\perp) + s] + \overline{\Omega}(\boldsymbol{\xi})\right\}\exp[i\phi(\boldsymbol{\xi}_\perp)].\end{aligned} \tag{14}$$

We note that the only ϕ-dependent component in $\overline{U_{\mathrm{P}}}$ entered via $\overline{\Phi_1}$. Furthermore, $\overline{\Omega}$ represents contributions from $N \geq 2$ that are independent of ϕ; aside from the stipulation that it is purely imaginary, its explicit form is unimportant here although it is discussed in Appendix B. The (approximate) correction to the perturbation model follows from including the first (ϕ-dependent) term in Eq. (12):

$$\overline{U}(\boldsymbol{\xi}) \approx \overline{U_{\mathrm{P}}}(\boldsymbol{\xi})\exp\left[\mathrm{i}\left(\exp\left\{\frac{sz_{\mathrm{M}}\hat{\mathcal{W}}}{2ik\chi^3(\boldsymbol{\xi}_\perp)[s + W(\boldsymbol{\xi}_\perp)]}\right\} - 1\right)\phi(\boldsymbol{\xi}_\perp)\right]. \tag{15}$$

Note that Eq. (15) represents a more complete expression for the field, beyond the perturbation model since it includes contributions from all N. This is in contrast with the equivalent two-dimensional expression given in Eq. (14) of Ref. 8, which included only the first three terms

Optics EXPRESS

($N \leq 2$) in Eq. (3). It should be noted that the approximations used in Appendix B for the purposes of obtaining Eq. (15) have the consequence of limiting our consideration to MSF structures with small amplitudes ($|\phi|<\pi$), whereas the procedure employed in Ref. 8 explicitly produced a second-order correction term with respect to the amplitude of ϕ. However, our goal is to provide rule-of-thumb error estimates for residual MSF structures from typical processes in freeform manufacturing; these MSF phase structures generally have amplitudes that are a fraction of the wavelength.

3. Simple field error estimates in a homogeneous medium

The root-mean-squared error (RMSE) of the perturbation model, ϵ, can be estimated as a function of propagation distance by integrating over the transverse plane the squared modulus of the difference between U_P and the corrected field estimate in Eq. (15). This is achieved by changing the variable of integration from (x, y) to (ξ, η) by using the differential area transformation

$$\mathrm{d}x\mathrm{d}y = \mathrm{d}\xi\mathrm{d}\eta \left|\frac{\partial \boldsymbol{r}_\perp}{\partial \boldsymbol{\xi}_\perp}\right|, \tag{16}$$

where $\partial \boldsymbol{r}_\perp/\partial \boldsymbol{\xi}_\perp$ is the Jacobian matrix between (x, y) and (ξ, η) after substituting $s = (z_\mathrm{P} - z_\mathrm{M})/\chi(\boldsymbol{\xi}_\perp)$ and holding $(z_\mathrm{P} - z_\mathrm{M})$ constant. The Jacobian determinant is given by

$$\begin{aligned}\left|\frac{\partial \boldsymbol{r}_\perp}{\partial \boldsymbol{\xi}_\perp}\right| &= 1 + (z_\mathrm{P} - z_\mathrm{M})\frac{\chi\nabla^2_{\boldsymbol{\xi}_\perp} W - \nabla_{\boldsymbol{\xi}_\perp}\chi \cdot \nabla_{\boldsymbol{\xi}_\perp} W}{\chi^2} + (z_\mathrm{P} - z_\mathrm{M})^2 \frac{\det(\mathbb{H}_W)}{\chi^4} \\ &= \frac{1}{\chi}\Delta\left[\boldsymbol{\xi}_\perp, \frac{z_\mathrm{P} - z_\mathrm{M}}{\chi}\right].\end{aligned} \tag{17}$$

The squared RMSE is thus

$$\begin{aligned}\epsilon^2(z, z_\mathrm{M}; \phi) &\triangleq \frac{\int_a \left|\overline{U_\mathrm{P}}[\boldsymbol{\xi}_\perp, (z_\mathrm{P} - z_\mathrm{M})/\chi(\boldsymbol{\xi}_\perp)] - \overline{U}[\boldsymbol{\xi}_\perp, (z_\mathrm{P} - z_\mathrm{M})/\chi(\boldsymbol{\xi}_\perp)]\right|^2 \left|\partial \boldsymbol{r}_\perp/\partial \boldsymbol{\chi}_\perp\right| \mathrm{d}\xi\mathrm{d}\eta}{\int_a \left|\overline{U}[\boldsymbol{\xi}_\perp, (z_\mathrm{P} - z_\mathrm{M})/\chi(\boldsymbol{\xi}_\perp)]\right|^2 \left|\partial \boldsymbol{r}_\perp/\partial \boldsymbol{\xi}_\perp\right| \mathrm{d}\xi\mathrm{d}\eta} \\ &\approx \frac{1}{\int_a A^2 \,\mathrm{d}\xi\mathrm{d}\eta}\int_a A^2 \left|1 - \exp\left[\mathrm{i}\exp\left(-\frac{\mathrm{i}r_1^2\hat{\mathcal{W}}}{4\pi\chi^3}\right)\phi - \mathrm{i}\phi\right]\right|^2 \mathrm{d}\xi\mathrm{d}\eta,\end{aligned} \tag{18}$$

where a is the aperture in the initial reference plane and

$$r_1 \triangleq \sqrt{\lambda\left|\frac{(z_\mathrm{P} - z_\mathrm{M})z_\mathrm{M}}{z_\mathrm{P}}\right|}, \tag{19}$$

is the radius of the first Fresnel zone at z_M, as illustrated in Fig. 2(b) of Ref. 8. In the second line of Eq. (18) we used Eqs. (9) and (16) to see that $|\partial \boldsymbol{r}_\perp/\partial \boldsymbol{\xi}_\perp|\left|\overline{U_\mathrm{P}}[\boldsymbol{\xi}_\perp, (z - z_\mathrm{M})/\chi(\boldsymbol{\xi}_\perp)]\right|^2 = U_0^2 A^2(\boldsymbol{\xi}_\perp)$. Note that the integral in the denominator of Eq. (19) is independent of z due to the fact that, when evanescent waves are not included, propagation through lossless media is a unitary operation. Upon defining the weighted average

$$\langle Q\rangle_A \triangleq \frac{\int_\mathrm{a} QA^2 \,\mathrm{d}\xi\mathrm{d}\eta}{\int_\mathrm{a} A^2 \,\mathrm{d}\xi\mathrm{d}\eta}, \tag{20}$$

and expanding the (outer) exponential in Eq. (18), one can write the following simple approximation for ϵ^2:

$$\epsilon^2 \approx 4\left\langle\left|\exp\left(-\frac{\mathrm{i}r_1^2\hat{\mathcal{W}}}{8\pi\chi^3}\right)\sin\left(\frac{r_1^2\hat{\mathcal{W}}}{8\pi\chi^3}\right)\phi\right|^2\right\rangle_A. \tag{21}$$

Optics EXPRESS

As discussed in Sec. 4, there are situations when it is sufficient (and in some ways more insightful) to simply consider the first-order (of the argument within the sine function) approximation of

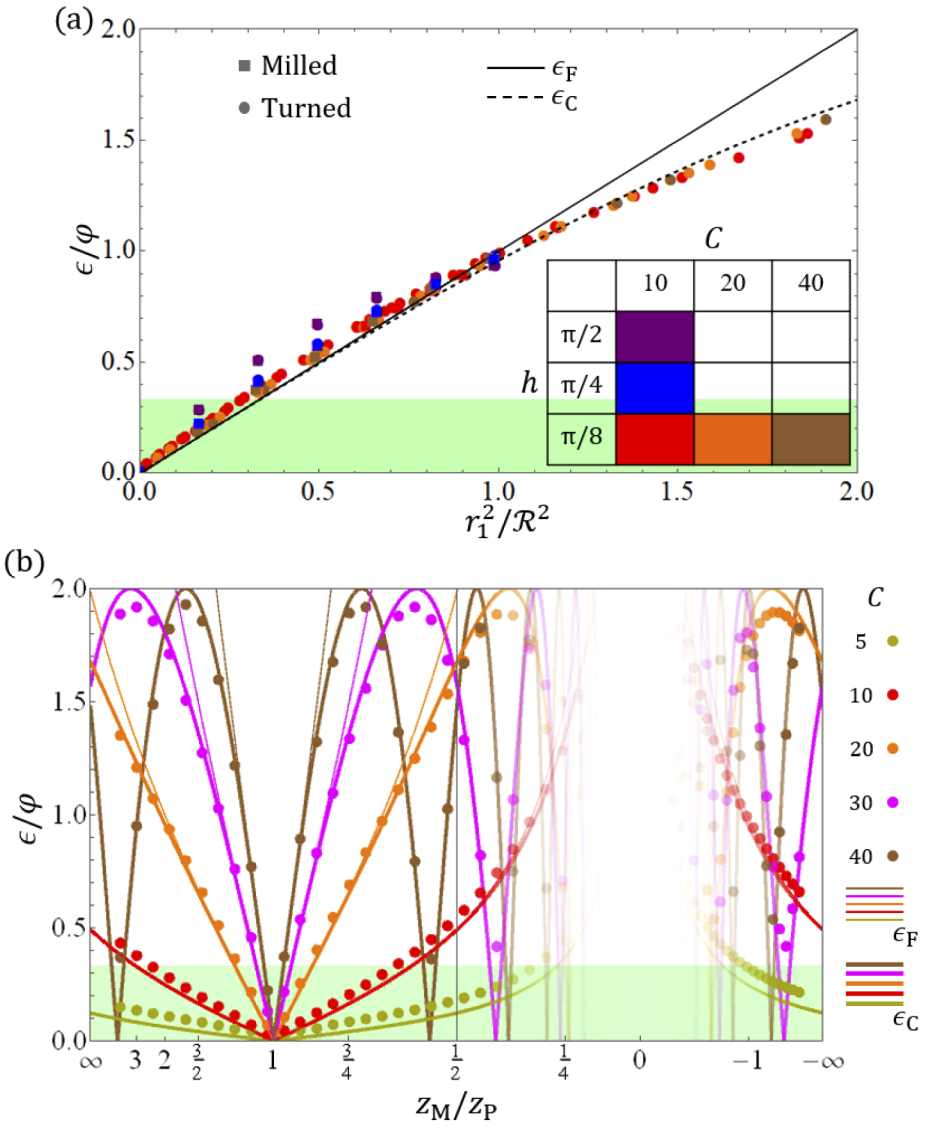

Fig. 2. (a) NRMSE for NA = 0.01 of various values of C and h. The solid black line is given by ϵ_F/φ in Eq. (30) and agrees well with the numerically calculated values for small values of h (circles and squares for the turned and milled cases, respectively). The dashed black line is given by ϵ_C/φ in Eq. (25) and is an even better fit with the numerically calculated values. In these plots, z_M is varied while z_P is fixed; that is, r_1 changes. (b) Numerically calculated NRMSE (for turned MSF structures), with the same NA, for $h = \pi/8$ and various values of C (colored dots), is plotted as a function of z_M/z_P. These values are compared with ϵ_F (thin) and ϵ_C (thick) of Eqs. (23) and (24), respectively. For both (a) and (b), the region of $\epsilon/\varphi<1/3$ is shaded in green as an example of when the perturbation model is valid.

Optics EXPRESS

Eq. (21). That is,

$$\epsilon^2 \approx \frac{r_1^4}{16\pi^2}\left\langle\left(\frac{\hat{\mathcal{W}}\phi}{\chi^3}\right)^2\right\rangle_A. \tag{22}$$

Given that $\hat{\mathcal{W}}$ involves second derivatives, this type of truncation is what led to the aforementioned inspiration to develop the novel bases for the purpose of finding basis sets with finite fourth spectral moments [9]. Upon examination of the complete result in Eq. (21), however, it is clear that such considerations are not necessary.

In Section 4, we use both Eqs. (21) and (22) to obtain rules of thumb regarding the validity of the perturbation model. Although Eqs. (21) and (22) were formally derived for systems with arbitrary NA, the remainder of this work will focus on systems with low to moderate NA. This is because the appropriate analysis for high-NA systems should involve a vector field treatment and the scalar formalism described here is sometimes insufficient. Furthermore, the results for systems with low to moderate NA may be of more interest for manufacturers due to their ease of interpretation and utility. However, for those interested in the behavior of ϵ in the high-NA regime, a discussion of those results from this scalar treatment is included in Appendix C.

4. Rules of thumb for low to moderate NA

In this section, we provide rules of thumb for the error incurred by the perturbation model for an imaging system with low to moderate NA. In this case the expressions for ϵ in Eqs. (21) and (22) can be simplified by using $\chi \approx 1$ and $A \approx 1$. Furthermore, $\hat{\mathcal{W}}$ can be simplified to $\nabla_\perp^2$, the transverse Laplacian operator. With this, Eqs. (22) and (21) become

$$\epsilon^2 \approx \epsilon_{\mathrm{F}}^2 \triangleq \frac{r_1^4}{16\pi^2}\left\langle\left(\nabla_\perp^2\phi\right)^2\right\rangle_1, \tag{23}$$

and

$$\epsilon^2 \approx \epsilon_{\mathrm{C}}^2 \triangleq 4\left\langle\left[\sin\left(\frac{r_1^2\nabla_\perp^2}{8\pi}\right)\phi\right]^2\right\rangle_1, \tag{24}$$

respectively (where the subindices F and C stand for "first term" and "complete", respectively). In the following simulations, with $\lambda = 632$ nm, a low-NA system is used for demonstration. As explained in Ref. 8, the resulting rules of thumb are applicable regardless of whether the MSF is located before or after the aperture stop within the system.

4.1. Rules of thumb for sinusoidal MSF structures in the milled and turned geometries

For MSF structures whose spectra are well-localized, it turns out that $\epsilon \approx \epsilon_{\mathrm{F}}$ in Eq. (23) is sufficient. Note that Eq. (23) can be re-expressed as a normalized RMSE (NRMSE):

$$\frac{\epsilon}{\varphi} \approx \frac{\epsilon_{\mathrm{F}}}{\varphi} = \frac{r_1^2}{\mathcal{R}^2}, \tag{25}$$

where

$$\mathcal{R} \triangleq 2\sqrt{\pi}\left[\frac{\langle\phi^2\rangle_1}{\langle(\nabla_\perp^2\phi)^2\rangle_1}\right]^{1/4}, \tag{26}$$

is a measure of the characteristic feature size of the MSF phase at z_{M} and $\varphi \triangleq \sqrt{\langle\phi^2\rangle_1}$ is the RMS of the MSF structure. The expressions for milled and turned geometries are [in polar coordinates

Optics EXPRESS

(r, θ)] given respectively by

$$\phi_{\mathrm{m}}(r, \theta) = h \cos[2\pi\kappa r \cos(\alpha - \theta) + \beta], \tag{27}$$

$$\phi_{\mathrm{t}}(r, \theta) = h \cos(2\pi\kappa r + \beta). \tag{28}$$

where $2h$ is the PV, κ is the spatial frequency of the MSF structure, β is a phase offset, and the milled groove pattern is perpendicular to the direction that makes an angle α to the x-axis. Both β and α will turn out to be irrelevant for determining error estimates. Note that $\kappa = C/(2R_{\mathrm{M}})$, where, as seen in Fig. 1, $R_{\mathrm{M}} \triangleq R|z_{\mathrm{M}}/z_{\mathrm{P}}|$ is the radius of the beam's circular footprint in the part's conjugate plane and C is the number of cycles across this footprint, where R is the pupil radius. With either Eqs. (27) or (28), Eq. (25) is approximately given by

$$\frac{\epsilon_{\mathrm{F}}}{\varphi} \approx \pi r_1^2(z_{\mathrm{P}}, z_{\mathrm{M}})\kappa^2, \tag{29}$$

for a sufficiently large value of κ. That is, turned and milled MSF structures that are well-approximated with a single frequency obey similar validity measures upon the use of the perturbation model. Note that, in the examples of Eqs. (27) and (28), $\mathcal{R}^2 = 1/(\pi\kappa^2)$. Figure 2(a) shows how the simple estimate of Eq. (30) compares with numerically calculated RMSE for various values of C and h. The angular spectrum approach is used for the numerical simulations.

For completeness and in anticipation of the discussion regarding MSF phases with broad spectra, Fig. 2(b) shows how NRMSE values from a turned MSF surface calculated numerically compare with both ϵ_{F} and ϵ_{C} of Eqs. (23) and (24), respectively, as functions of $z_{\mathrm{M}}/z_{\mathrm{P}}$. Figure 2(b) is an alternative way to view the same data as Fig. 2(a) without having to introduce the notion of r_1, which may obscure the effects of what happens when, for example, the MSF is placed near the exit pupil or the focus. Near $z_{\mathrm{M}}/z_{\mathrm{P}} = 1$, ϵ_{F} is accurate and, as is indicated by the previous discussion regarding Fig. 2(a), the perturbation model is valid there since $\epsilon/\varphi < 1/3$. Furthermore, we point out that it is possible for the perturbation model to be valid [for instance, the case of $C = 5$ in Fig. 2(b)] even for large values of $z_{\mathrm{M}}/z_{\mathrm{P}}$, such as those beyond the image plane. This is in keeping with the fact that ϵ/φ is a closed curve if one were to join the $z_{\mathrm{M}}/z_{\mathrm{P}} = \pm\infty$ edges of Fig. 2(b), as discussed in Ref. 8. For larger values of C, however, the NRMSE begins to oscillate, due to the Talbot effect, for values of $z_{\mathrm{M}}/z_{\mathrm{P}}$ that are sufficiently far from unity (so that $r_1^2/\mathcal{R}^2$ is large) and this behavior is captured only by the complete error estimate. A notable feature of Fig. 2(b) is the region near $z_{\mathrm{M}}/z_{\mathrm{P}} = 0$, where the complete error estimate oscillates rapidly; it is evident that the perturbation model is not valid in this region and this fact is represented by the translucency of the plot. Recall that $\lambda = 632$ nm in these simulations and note that the effect of varying λ on the plots of Fig. 2(b) is to change the value of C to which they correspond, proportionally to $\lambda^{-1/2}$. For example, if λ were increased by a factor of 4, the plot for $C = 20$ would correspond instead to $C = 10$. Figure 2(a), on the other hand, is explicitly independent of λ.

The simple error estimate in Eq. (29) is the analogous rule of thumb, specifically for the milled and turned sinusoidal MSF groove geometries, to the one-dimensional version in Ref. 8. Although it works well for MSF structures in the form of Eqs. (27) and (28), it was observed in Ref. 8 that such an estimate appears to overestimate the error incurred by MSF structures that possess a broad spectrum; this was demonstrated with synthetic MSF structures with spectra that obeyed a power-decay law. For these specific examples, the simple analog of Eq. (24) proved to be sufficient since it accurately predicted the NRMSE behavior in the region $0 < \epsilon/\varphi < 1/3$, which is where the perturbation model was considered acceptable. However, it turns out that the consideration of MSF data requires an extension of the rule of thumb predicted by Eq. (25). For such MSF structures, it is necessary to consider the more complete NRMSE expression of Eq. (24).

Optics EXPRESS

4.2. Rules of thumb for MSF structures with broad spatial spectra

To begin, we present Fig. 3(a) so that it can be used as a reference for further discussion regarding the ineffectiveness of Eq. (25) when applied to MSF structures that are more complicated than those given by Eqs. (27) and (28). There, it is evident that such a simple estimate for ϵ/φ is useful only for small values of $r_1^2/\mathcal{R}^2$. The behavior of the numerically calculated NRMSE departs from the simple estimate very quickly in some examples. Although it is fortunate that Eq. (25) overestimates the true NRMSE, it fails as a rule of thumb for MSF structures with broad spectra (such as those seen in Fig. 3). For instance, someone interested in using the perturbation model with an optical system with MSF structures similar to that color-coded purple in Fig. 3(a), operating at $r_1^2/\mathcal{R}^2 \approx 0.8$ with an NRMSE threshold of 20%, would erroneously believe based on Eq. (25) that this model should not be used. Therefore, there is a need for a more complete estimate that is accurate for MSF structures with broad spectra; this is provided by Eq. (24), which can be rewritten in the Fourier domain, after normalization by φ^2, as

$$\frac{\epsilon_C^2}{\varphi^2} = \frac{4}{\int |\tilde{\phi}(\boldsymbol{\kappa})|^2 \, \mathrm{d}^2\kappa} \int \left| \tilde{\phi}(\boldsymbol{\kappa}) \sin\left(\frac{\pi r_1^2 |\boldsymbol{\kappa}|^2}{2} \right) \right|^2 \mathrm{d}^2\kappa, \tag{30}$$

where the Fourier transform of $\phi(\boldsymbol{\xi}_\perp)$ is taken to be defined by

$$\tilde{\phi}(\boldsymbol{\kappa}) = \int \phi(\boldsymbol{\xi}_\perp) \exp\left(-\mathrm{i}2\pi\boldsymbol{\kappa} \cdot \boldsymbol{\xi}_\perp\right) \mathrm{d}^2\xi_\perp. \tag{31}$$

Eq. (30) shows that the validity of the perturbation model is best understood in the Fourier domain and matches the behavior of the numerically calculated NRMSE values in Fig. 3(a) very well. As done in Fig. 2, an alternative view of Fig. 3(a) is given by Fig. 3(b), which shows the NRMSE as a function of z_M/z_P.

For the numerical simulations performed with the MSF structures shown in Fig. 3, the MSF data were pre-processed to have zero mean and normalized to have RMS of $\pi/8$. As was done with the sinusoidal examples in Sec. 4.1, field propagation was modeled by using the angular spectrum. Specifically, the nominal field is initially generated over an array of points at z_P and propagated to the location z_M, where it is then multiplied by an exponential containing ϕ, the MSF structure. Depending on the value of the ratio $|z_M/z_P|$, the dimensions of the MSF array must be scaled in order to ensure that the converging beam sees the same MSF phase over its diameter. Such a consideration was technically less cumbersome for the analysis in Sec. 4.1 because the MSF structures considered there are periodic and new arrays for ϕ at any z_M can be sampled directly from an analytic sinusoidal function with the appropriate number of cycles. Although the calculated NRMSE does not oscillate quickly near $z_M/z_P = 0$ for MSF phases with broad spectra [compare the complete estimates in Fig. 2(b) and Fig. 3(b)], this region is similarly marked with translucency (to indicate the invalidity of the perturbation model regardless of numerically calculated results) because the simulation procedure described earlier in this process becomes invalid as the MSF phase is placed near the vicinity of the caustic at $z = 0$.

Figure 3 shows that, in the paraxial regime, the more complete estimate in Eq. (30) accurately predicts the NRMSE due to the perturbation model. As mentioned earlier, the simple estimate in Eq. (25) always overestimates the NRMSE predicted by Eq. (30). That is, the perturbation model appears to be more valid for MSF structures with broad spectra, as compared to those that are approximated well with a single frequency. One can understand this by noting that the real MSF data used in the analysis is dominated by low-frequency contributions. As a result, the plots shown in Fig. 3(b) are more akin to the plots in Fig. 2(b) with small values of C (such as $C = 5$ and $C = 10$) that display a less oscillatory behavior. In other words, the calculated NRMSE for MSF data with broad spectra do not display the oscillatory Talbot re-imaging behavior seen in Fig. 2(b) for the larger C values because the MSF data contain multiple frequencies that wash out the Talbot effect (in particular, validity is mainly governed by the low spatial frequencies).

Optics EXPRESS

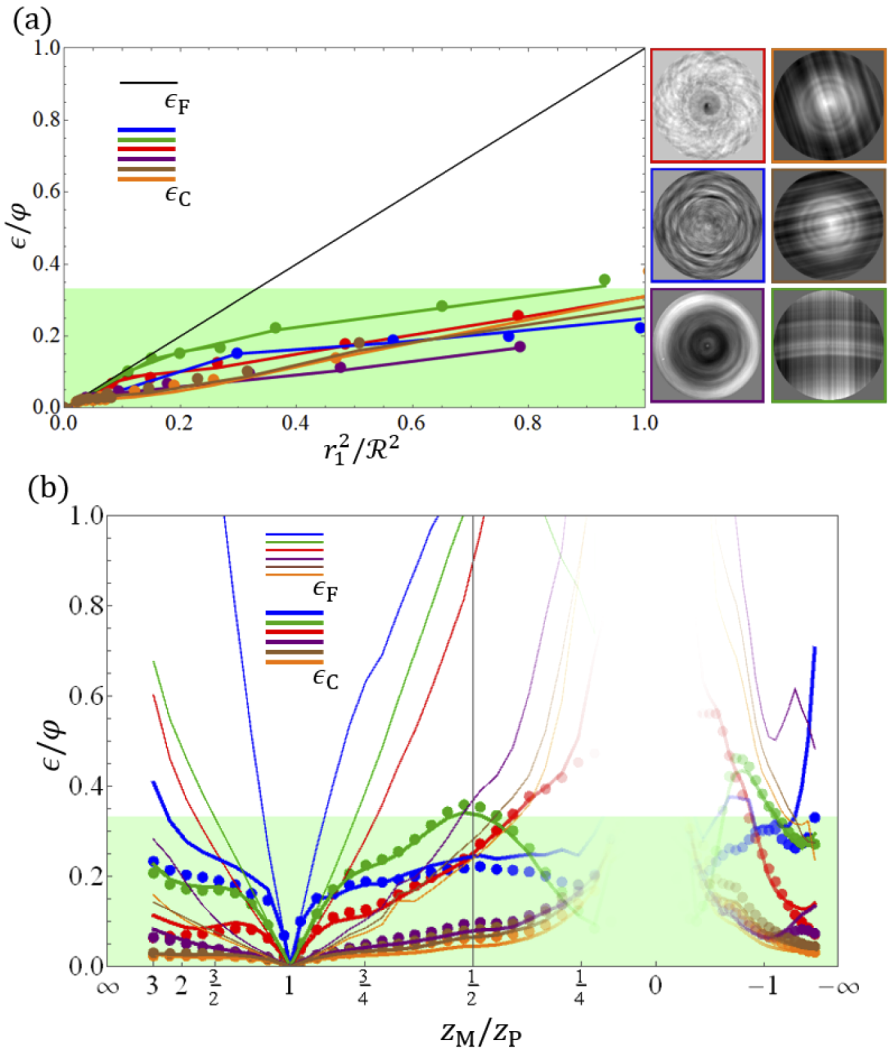

Fig. 3. NRMSE for NA = 0.01 of various examples of real MSF structure data scaled to have RMS values of $\pi/8$ is shown in (a) and (b) as a function of $r_1^2/\mathcal{R}^2$ and z_M/z_P, respectively, and color-matched with the plots. The simple estimate in Eq. (25), shown as a single black curve in (a) and multiple colored thin curves in (b), fails to predict the numerically calculated values of ϵ/φ (dots). The colored thick curves in (a) and (b) are given by the complete error estimate of Eq. (24) [which can also be expressed as Eq. (30)] and match the numerically calculated values well.

A further point should be made regarding the relationship between Eqs. (25) and (30), in which the former is a first-order approximation in $r_1^2|\boldsymbol{\kappa}|^2$ of the latter. Equation (25) is sufficiently accurate for the MSF examples of Eqs. (27) and (28) in Fig. 2(b) within a region that shrinks with larger C (or $\boldsymbol{\kappa}$). However, despite this shrinking region of accuracy, Eq. (25) is sufficiently accurate for $0 \le \epsilon_F/\phi < 1$, which is the region that is relevant for assessing the validity of the perturbation model. The same cannot be said for the MSF examples in Fig. 3(b), whose spectra include both small and large values of $\boldsymbol{\kappa}$. Although the presence of large spatial frequencies consistently explains the small region of accuracy of Eq. (25), it is evident that this simple error estimate fails far before $\epsilon/\varphi \approx 1$.

Optics EXPRESS

5. Concluding remarks

We investigated the validity of the perturbation model in three dimensions by using an asymptotic framework like that in Ref. 8. However, we expanded upon the findings in Ref. 8 not only in the consideration of one extra spatial dimension, but also in the completeness of estimates for the error incurred by the perturbation model. That is, through further consistent approximations within the asymptotic framework, it is possible to solve for the complete correctional term [not just the first correction seen in Eq. (22)], which proves to be necessary when considering realistic MSF structures. This more complete approach gives accurate rules of thumb for the validity of the perturbation model. In particular, for the case of imaging systems with low to moderate NA, a general rule of thumb for the validity of the perturbation model for small-amplitude MSF structures is provided in Eq. (30). This more complete error estimate replaces the one in Ref. 8, which involved a troublesome fourth-order spectral moment of the MSF phase structure. These possibly-divergent moments are evidently replaced by the well-behaved integral of Eq. (30). As observed in Ref. 8, when an optical system has more than a single instance of MSF, the total (mean squared) error incurred upon using the perturbation model is simply given by the sum of the (mean squared) error for each individual MSF structure, provided that the MSF structures are statistically uncorrelated with each other. Furthermore, we mention that our results, which are derived by imaging the MSF structure to its conjugate location in image space and then propagating to the exit pupil, would be the same if we had instead imaged the MSF to the neighborhood of the aperture stop and performed the propagation steps there before imaging to the pupil. This invariance is discussed in Appendix B of Ref. 8.

The general framework in Sec. 2 was specialized to that of imaging systems (in fact, the complete error estimates are accurate only for a spherical nominal wavefront in image space), where only three locations in image space are relevant: the image plane, the exit pupil plane, and the plane that is conjugate to the MSF interface itself. The simplest error estimate, ϵ_{F}, is suitable for imaging systems with low to moderate numerical apertures and MSF structures that are well described with a single spatial frequency with $\epsilon/\varphi<1$. This error estimate is subsumed by the more general one described in Eq. (30), which is nicely represented in the Fourier domain; this estimate was shown to be sufficient in estimating the validity of the perturbation model for MSF structures with broad spectra. The real MSF data used in the analysis contained low spatial frequencies, which dominated the general behavior of the validity of the perturbation model. This, along with the washing-out of the Talbot effect, accounts for why the NRMSE, for many of the real MSF data used in this analysis, is low and not oscillatory when compared with the plots in Fig. 2. Moreover, the simple error estimate of Eq. (25) fails for both the MSF examples in Figs. 2 and 3 outside a domain that shrinks with the presence of higher spatial frequencies. However, the estimate is always sufficiently accurate within the range $\epsilon_{\mathrm{F}}/\varphi<1$ for MSF structures whose spectra are well-localized. For MSF structures with broad spatial spectra, one must instead use the complete error estimate, ϵ_{C}, given by Eq. (30). Even though it was derived in the context of low to moderate NA, we remark that the results in Appendix C lead us to expect that Eq. (30) is, to within a factor of 2 or so, useful for numerical apertures up to 0.8. Therefore, even in the analysis of systems with appreciable NA, it is possible to use Eq. (30) to obtain a rough error estimate of the perturbation model rather than involving the more complicated, and incomplete, discussion regarding systems with high NA in Appendix C.

A. Extended derivation for two spatial dimensions

In this appendix, an alternative approximation for the field, $U(x, z)$, in two spatial dimensions is presented; the definitions of symbols correspond to those given in Appendix A of Ref. 8. One

Optics EXPRESS

begins by considering the general differential equation for $\overline{\Phi_N}$:

$$\partial_s \overline{\Phi_N} = -\frac{1}{2}\overline{\nabla^2 \Phi_{N-1}} - \frac{1}{2}\sum_{n=1}^{N-1} \overline{\nabla \Phi_n \cdot \nabla \Phi_{N-n}}, \tag{32}$$

where the quantities with an overbar are functions of the ray coordinates (ξ, s). In Ref. 8, the equation for $N = 2$, which represents the first correction of using the perturbation model, was solved exactly. It turns out, however, that progress can be made by dropping the last (non-linear) term on the right-hand side of Eq. (32), leaving

$$\partial_s \overline{\Phi_N} \approx -\frac{1}{2}\overline{\nabla^2 \Phi_{N-1}}. \tag{33}$$

This approximation is justified by the fact that ϕ, which appears linearly in $\overline{\Phi}_1$, is small. Equation (33) can be expressed explicitly in terms of derivatives with respect to the ray coordinates as

$$\partial_s \overline{\Phi_N} \approx -\frac{1}{2} J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \overline{\Phi_{N-1}}, \tag{34}$$

where J_{il}^{-1} is the (i, l) element of the $\mathbb{J}^{-1}$ matrix defined in Eq. (37) of Ref. 8. At this point, W is taken to be a converging nominal wavefront, centered at z_{M}:

$$W(\xi) = z_{\mathrm{M}} \sqrt{1 + \frac{\xi^2}{z_{\mathrm{M}}^2}} \quad \text{and} \quad \chi(\xi) = \frac{z_{\mathrm{M}}}{W(\xi)}. \tag{35}$$

This early specification of W is another discrepancy between the method presented here and that in Ref. 8; however, since imaging systems are of interest, where the wavefront converges to a nominal point in image space, this loss of generality is insignificant for our purposes.

An approximation is now made with regards to Eq. (34): the only term retained on the right-hand side is the one that involves ∂_ξ^2. The reason for making this simplification, which is discussed more in Appendix B, is ultimately due to the fact that the derivatives of ϕ (which varies quickly under its characterization as MSF) are large when compared to the other terms. These other terms contain nominal quantities and their transverse derivatives. By doing this, and explicitly using Eq. (35) in the definition of $\mathbb{J}^{-1}$, Eq. (34) is approximated as

$$\partial_s \overline{\Phi_N}(\xi, s) \approx -\frac{1}{2} \frac{z_{\mathrm{M}}^2 + \xi^2}{\chi^2 (s + W)^2} \partial_\xi^2 \overline{\Phi_{N-1}}(\xi, s). \tag{36}$$

We propose a change of variables from s to $\bar{s} \triangleq s/(\chi\Delta)$, where

$$\Delta = \det \mathbb{J} = \frac{s z_{\mathrm{M}}}{\chi^3 (s + W)^2}. \tag{37}$$

Defining $\overline{\Gamma_N}(\xi, \bar{s}) \triangleq \overline{\Phi_N}[\xi, \bar{s}(\xi, s)]$, Eq. (36) becomes

$$\partial_{\bar{s}} \overline{\Gamma_N}(\xi, \bar{s}) \approx -\frac{1}{2} \partial_\xi^2 \overline{\Gamma_{N-1}}(\xi, \bar{s}). \tag{38}$$

Recall that

$$\overline{\Phi_1}(\xi, s) = \mathrm{i}\phi(\xi) + \ln\left[A(\xi) \sqrt{\frac{\chi(\xi)}{\Delta(\xi, s)}} \right], \tag{39}$$

which provides the initial right-hand side for Eq. (38) (for the case $N = 2$). Upon inserting Eq. (39) into Eq. (38) for $N = 2$, it is possible to drop the contribution due to the second term of

Optics EXPRESS

Eq. (39) under the assumption that the nominal field quantities vary significantly more slowly than ϕ. In actuality, this ϕ-independent term is ultimately included in the perturbation model and should be carried along in the following derivation. However, its explicit form (and subsequently derived expressions) is not important and can henceforth be dropped. Equation (38) now leads to

$$\overline{\Gamma_2}(\xi,\bar{s}) \approx -\frac{\mathrm{i}\bar{s}}{2}\partial_\xi^2\phi(\xi). \tag{40}$$

This process can be iterated to give

$$\begin{aligned}\overline{\Gamma_N}(\xi,\bar{s}) &\approx -\frac{\mathrm{i}\bar{s}^{N-1}}{2(N-1)!}\partial_\xi^{2(N-1)}\phi(\xi)\\ \Rightarrow \overline{\Phi_N}(\xi,s) &\approx -\frac{\mathrm{i}}{2(N-1)!}\left[\frac{sz_\mathrm{M}}{\chi^3(s+W)^2}\right]^{N-1}\partial_\xi^{2(N-1)}\phi(\xi).\end{aligned} \tag{41}$$

Note that, Eq. (41) is a further approximation since $\bar{s}$ depends on ξ and strictly cannot be pulled out of the derivatives in ξ. However all the ξ-dependence of $\bar{s}$ is through nominal field quantities and, as was mentioned after Eq. (39), the derivatives of ϕ vary much more quickly. That is, when integrated over the domain of interest, we assume

$$\left|\bar{s}^{N-2}\partial_\xi^{2(N-1)}\phi\right| \gg \left|\partial_\xi^2\bar{s}^{N-2}\partial_\xi^{2(N-2)}\phi\right| \quad \text{and} \quad \left|\bar{s}^{N-2}\partial_\xi^{2(N-1)}\phi\right| \gg \left|\partial_\xi\bar{s}^{N-2}\partial_\xi^{2(N-1)-1}\phi\right|. \tag{42}$$

As a result, in the iterative process of approximately solving for $\overline{\Gamma_N}$, it is possible to pull the factors of $\bar{s}$ out of the derivatives in ξ.

With Eq. (41), it is now possible to calculate

$$\overline{\Phi} = \sum_{N=0}^{\infty}\frac{\overline{\Phi_N}}{(\mathrm{i}k)^N} \approx \frac{1}{k}\exp\left[\frac{sz_\mathrm{M}\partial_\xi^2}{2\mathrm{i}k\chi^3(s+W)}\right]\phi + \ln\left(A\sqrt{\frac{\chi}{\Delta}}\right) + \overline{\Phi_0}. \tag{43}$$

The method presented here differs from that shown in Appendix A of Ref. 8 mainly in two ways. First, the new derivation includes every term in the summation seen in Eq. (43); in Ref. 8, this summation was truncated at $N = 2$. Second, the specification of W to be a converging spherical wavefront was used. These two differences, along with approximations that ϕ is small and varies more quickly than the nominal quantities, W and A, allows for a field estimate that gives rise to a complete error estimate upon using the perturbation model that includes contributions from all N. Although the method in Ref. 8 included higher-order corrections in $\overline{\Phi_2}$ that accounted for larger ϕ, these terms were ultimately discarded in the development of a simple rule of thumb.

B. MSF-independent rays derivation in three spatial dimensions

Equation (4) can be solved with the method of characteristics, which leads to solutions given in the parametrization of Eq. (6). When making this change of variables, it is convenient to use the transpose of the Jacobian matrix

$$\mathbb{J} \triangleq \frac{\partial\mathbf{r}}{\partial\boldsymbol{\xi}} = \begin{bmatrix}\partial_\xi x & \partial_\xi y & \partial_\xi z\\ \partial_\eta x & \partial_\eta y & \partial_\eta z\\ \partial_s x & \partial_s y & \partial_s z\end{bmatrix} = \begin{bmatrix}1+s\partial_\xi^2 W(\xi,\eta) & s\partial_\xi\partial_\eta W(\xi,\eta) & s\partial_\xi\chi(\xi,\eta)\\ s\partial_\xi\partial_\eta W(\xi,\eta) & 1+s\partial_\eta^2 W(\xi,\eta) & s\partial_\eta\chi(\xi,\eta)\\ \partial_\xi W(\xi,\eta) & \partial_\eta W(\xi,\eta) & \chi(\xi,\eta)\end{bmatrix}. \tag{44}$$

With this, the derivatives in Cartesian coordinates $\mathbf{r} = (x, y, z)$ can be written in terms of derivatives in the ray parameters $\boldsymbol{\xi} = (\xi, \eta, s)$ according to the chain rule,

$$\nabla F[x(\boldsymbol{\xi}), y(\boldsymbol{\xi}), z(\boldsymbol{\xi})] = \overline{\nabla F}(\boldsymbol{\xi}) = \mathbb{J}^{-1}\cdot\nabla_{\boldsymbol{\xi}}\overline{F}(\boldsymbol{\xi}), \tag{45}$$

where $\nabla_{\boldsymbol{\xi}} \triangleq (\partial_\xi, \partial_\eta, \partial_s)$.

Optics EXPRESS

To begin, we show that Eq. (7) is the solution to the Eikonal equation in Eq. (4). By using Eq. (45), and with some simplification, we find that

$$\overline{\nabla\Phi_0} = (0\ 0\ 1)\cdot\mathbb{J} = \left(\partial_\xi W, \partial_\eta W, \chi\right). \tag{46}$$

It is simple to see that the dot product of Eq. (46) with itself gives unity, therefore satisfying Eq. (4). Furthermore, by using both Eqs. (45) and (46), we observe that

$$\overline{\nabla\Phi_0}\cdot\overline{\nabla\Phi_N} = (0\ 0\ 1)\cdot\nabla_{\boldsymbol{\xi}}\overline{\Phi_N} = \partial_s\overline{\Phi_N}. \tag{47}$$

Equation (47) allows us to rewrite the left-hand side of Eq. (5) as a derivative in s. The Eikonal can then be expressed as

$$\overline{\Phi_0}(\boldsymbol{\xi}) = W(\boldsymbol{\xi}_\perp) + s. \tag{48}$$

For $N = 1$, Eq. (5) reduces to

$$\nabla\Phi_0\cdot\nabla\Phi_1 = -\frac{1}{2}\nabla^2\Phi_0. \tag{49}$$

By parameterizing in terms of $\boldsymbol{\xi}$ and using Eq. (47), the left-hand side of Eq. (49) simply becomes $\partial_s\overline{\Phi_1}$. For the right-hand side, the parametrization leads to

$$\overline{\nabla^2\Phi_0} = \left(\mathbb{J}^{-1}\cdot\nabla_{\boldsymbol{\xi}}\right)\cdot\overline{\nabla\Phi_0} = \left(\mathbb{J}^{-1}\cdot\nabla_{\boldsymbol{\xi}}\right)\cdot[(0\ 0\ 1)\cdot\mathbb{J}] = \mathrm{Tr}\left(\mathbb{J}^{-1}\cdot\partial_s\mathbb{J}\right) = \partial_s\ln(\Delta), \tag{50}$$

where $\Delta = \det(\mathbb{J})$, as given by Eq. (9). Equation (49) therefore becomes

$$\partial_s\overline{\Phi_1} = -\frac{1}{2}\partial_s\ln(\Delta), \tag{51}$$

which has a simple solution that satisfies the initial conditions of $\overline{\Phi_1}(\xi,\eta,0) = \ln[A(\xi,\eta)] + \mathrm{i}\phi(\xi,\eta)$:

$$\overline{\Phi_1}(\xi,\eta,s) = \ln\left[A(\xi,\eta)\sqrt{\frac{\chi(\xi,\eta)}{\Delta(\xi,\eta,s)}}\right] + \mathrm{i}\phi(\xi,\eta), \tag{52}$$

where the s-dependence is fully encapsulated in Δ. Note that Eq. (52) has a form that is similar to that of the corresponding quantity in the two-dimensional analysis and the logarithmic portion is an amplitude factor that accounts for the bunching of the rays under propagation. This factor diverges at the caustics of these nominal rays.

It is useful to restate the remaining equations for $N \geq 2$. Once again, the left-hand side of Eq. (5) can be simplified to $\partial_s\overline{\Phi_N}$. That is,

$$\partial_s\overline{\Phi_N} = -\frac{1}{2}\overline{\nabla^2\Phi_{N-1}} - \frac{1}{2}\sum_{n=1}^{N-1}\overline{\nabla\Phi_n\cdot\nabla\Phi_{N-n}}, \tag{53}$$

Although progress can be made with Eq. (53), particularly with the case of $N = 2$ (as discussed separately later), as it is presented, we seek an expression for general N. In order to proceed in this direction, several approximations are made; to begin, we neglect the second term on the right-hand side of Eq. (53), which is a nonlinear term for preceding values of Φ_n. This can be justified by the smallness of ϕ. What remains is

$$\begin{aligned}\partial_s\overline{\Phi_N} &\approx -\frac{1}{2}J_{il}^{-1}\partial_l\left(J_{ij}^{-1}\partial_j\overline{\Phi_{N-1}}\right)\\ &= -\frac{1}{2}J_{il}^{-1}J_{ij}^{-1}\partial_l\partial_j\overline{\Phi_{N-1}} - \frac{1}{2}J_{il}^{-1}(\partial_l J_{ij}^{-1})\partial_j\overline{\Phi_{N-1}},\end{aligned} \tag{54}$$

where we use the Einstein implicit summation convention. We now consider only MSF structures ϕ for which there is an appreciable number of cycles across the aperture. This means the

Optics EXPRESS

derivatives of ϕ, which vary quickly, are much greater than those of the nominal quantities. Since the elements of $\mathbb{J}$ include only nominal quantities, the second term of Eq. (54) can be neglected when compared with the first. What remains is then

$$\partial_s \overline{\Phi_N} \approx -\frac{1}{2} J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \overline{\Phi_{N-1}}. \tag{55}$$

Equation (55) is a simplified recursive differential equation for $\overline{\Phi_N}$. At this point, we assume the following form of a converging spherical wavefront for W:

$$W(\xi,\eta) = z_{\mathrm{M}} \sqrt{1 + \frac{\xi^2+\eta^2}{z_{\mathrm{M}}^2}} \quad \text{and} \quad \chi(\xi,\eta) = \frac{z_{\mathrm{M}}}{W(\xi,\eta)}. \tag{56}$$

To see how Eq. (55) can be solved, it is helpful to look at the case of $N = 2$ by itself:

$$\begin{aligned} \partial_s \overline{\Phi_2} &\approx -\frac{1}{2} J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \left[\mathrm{i}\phi + \ln\left(A\sqrt{\frac{\chi}{\Delta}} \right) \right] \\ &= -\mathrm{i}\frac{1}{2} J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \phi - \frac{1}{2} J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \ln\left(A\sqrt{\frac{\chi}{\Delta}} \right), \end{aligned} \tag{57}$$

The first term in Eq. (57) involves the differentiation of only ϕ. Since ϕ is independent of s, the indices l and j there effectively only run through the values of 1 and 2. The s-dependence of the upper-left 2×2 submatrix of $\mathbb{J}^{\mathrm{T}}\mathbb{J}$ can be factored out and Eq. (57) becomes

$$\partial_s \overline{\Phi_2} \approx -\mathrm{i}\frac{1}{2} \frac{\Gamma_{lj}}{(s+W)^2} \partial_l \partial_j \phi - \frac{1}{2} J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \ln\left(A\sqrt{\frac{\chi}{\Delta}} \right), \tag{58}$$

where Γ is an s-independent matrix given by

$$\Gamma = \chi^{-2} \begin{bmatrix} z_{\mathrm{M}}^2 + \xi^2 & \xi\eta \\ \xi\eta & z_{\mathrm{M}}^2 + \eta^2 \end{bmatrix}. \tag{59}$$

Equation (58) can be directly integrated to give

$$\begin{aligned} \overline{\Phi_2} &\approx -\mathrm{i}\frac{1}{2} \Gamma_{lj} \partial_l \partial_j \phi \int_0^s \frac{\mathrm{d}s'}{(s'+W)^2} - \frac{1}{2} \int_0^s J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \ln\left(A\sqrt{\frac{\chi}{\Delta}} \right) \mathrm{d}s' \\ &= -\mathrm{i}\frac{1}{2} \frac{s}{W(s+W)} \Gamma_{lj} \partial_l \partial_j \phi - \frac{1}{2} \int_0^s J_{il}^{-1} J_{ij}^{-1} \partial_l \partial_j \ln\left(A\sqrt{\frac{\chi}{\Delta}} \right) \mathrm{d}s'. \end{aligned} \tag{60}$$

Note that the second term in Eq. (60) involves derivatives of nominal quantities and will henceforth be ignored. It also is independent of ϕ and would ultimately be included in the perturbation model if explicitly included anyway; that is, this term and its subsequent ϕ-independent contributions, for larger N, contribute to $\overline{\Omega}$ in Eq. (14). However, its explicit form is not important and can henceforth be dropped in the following derivation [to be re-included in the expression for the perturbation model in Eq. (14)].

Having obtained $\overline{\Phi_2}$, it is now possible to consider the equation for $\overline{\Phi_3}$:

$$\partial_s \overline{\Phi_3} \approx -\frac{1}{2} J_{ik}^{-1} J_{im}^{-1} \partial_k \partial_m \overline{\Phi_2} = \mathrm{i}\left(-\frac{1}{2}\right)^2 J_{ik}^{-1} J_{im}^{-1} \partial_k \partial_m \left(\frac{s}{W(s+W)} \Gamma_{lj} \partial_l \partial_j \phi \right). \tag{61}$$

Once again, the derivative of nominal quantities are small compared to those of ϕ so we can pull $s/[W(s+W)]$ out of the derivatives in Eq. (61). Doing this once again leaves only s-independent

Optics EXPRESS

quantities within the derivatives and so only the upper-left 2×2 submatrix of $\mathbb{J}^{\mathrm{T}}\mathbb{J}$ is relevant. With this, we arrive at

$$\partial_s \overline{\Phi_3} \approx \mathrm{i}\left(-\frac{1}{2}\right)^2 \frac{1}{(s+W)^2}\frac{s}{W(s+W)}\Gamma_{km}\partial_k\partial_m\left(\Gamma_{lj}\partial_l\partial_j\phi\right), \tag{62}$$

which can now be directly integrated to give

$$\begin{aligned}\overline{\Phi_3} &\approx \mathrm{i}\left(-\frac{1}{2}\right)^2 \Gamma_{km}\partial_k\partial_m\left(\Gamma_{lj}\partial_l\partial_j\phi\right)\int_0^s \frac{s'}{W(s'+W)}\frac{\mathrm{d}s'}{(s'+W)^2}. \\ &= \mathrm{i}\left(-\frac{1}{2}\right)^2 \left\{\frac{1}{2}\left[\frac{s}{W(s+W)}\right]^2\right\}\Gamma_{km}\partial_k\partial_m\left(\Gamma_{lj}\partial_l\partial_j\phi\right).\end{aligned} \tag{63}$$

The processes between Eqs. (61) and (63) can be iterated to give [using an approximation akin to that leading to Eq. (41) in Appendix A]

$$\overline{\Phi_N} \approx \frac{\mathrm{i}}{(N-1)!}\left[-\frac{s}{2W(s+W)}\right]^{N-1}\left\{\mathrm{Tr}[\Gamma\cdot(\nabla\otimes\nabla)]\right\}^{N-1}\phi, \tag{64}$$

where Tr represents a matrix trace. Note that

$$\mathrm{Tr}[\Gamma\cdot(\nabla\otimes\nabla)] = \chi^{-2}W^2\hat{\mathcal{W}}, \tag{65}$$

where $\hat{\mathcal{W}}$ is a differential operator defined as

$$\hat{\mathcal{W}} \triangleq \partial_r^2 + \frac{\chi^2}{r}\partial_r + \frac{\chi^2}{r^2}\partial_\theta^2, \tag{66}$$

for the plane polar coordinates $r = \sqrt{\xi^2+\eta^2}$ and $\theta = \arg(\xi + \mathrm{i}\eta)$. With Eq. (65), we can rewrite Eq. (64) as

$$\overline{\Phi_N} \approx \frac{\mathrm{i}}{(N-1)!}\left[-\frac{sz_{\mathrm{M}}}{2\chi^3(s+W)}\right]^{N-1}\hat{\mathcal{W}}^{N-1}\phi. \tag{67}$$

With Eq. (67), it is now possible to find $\overline{\Phi}$ through an infinite sum.

$$\overline{\Phi} = \sum_{N=0}^{\infty}\frac{\overline{\Phi_N}}{(\mathrm{i}k)^N} \approx \frac{1}{k}\exp\left[\frac{sz_{\mathrm{M}}\hat{\mathcal{W}}}{2\mathrm{i}k\chi^3(s+W)}\right]\phi + \ln\left(A\sqrt{\frac{\chi}{\Delta}}\right) + \overline{\Phi_0}. \tag{68}$$

C. Discussion of the high NA regime

In this section, we discuss the NRMSE expressions given by Eqs. (21) and (22) for systems with high NA. In particular, we examine how going into the high NA regime complicates the simple rules of thumb seen in Sec. 4. The discussion in this section highlights a non-trivial aspect of the generalization to three dimensions; the rule-of-thumb expressions in Eqs. (21) and (22). We restate here that the following results may be inappropriate for a rigorous treatment of high NA systems, where polarization effects can no longer be ignored for field calculations.

As observed earlier, the main distinction between the low/moderate NA and high NA error estimates is comprised of the appearance of the obliquity factor χ and the differential operator $\hat{\mathcal{W}}$ (in place of $\nabla_\perp^2$) for the latter. For convenience, we re-state the approximate NRMSE formulas,

Optics EXPRESS

taken from Eqs. (22) and (23), as

$$\frac{\epsilon_{\rm lm}}{\varphi} \approx \frac{r_1^2}{4\pi}\sqrt{\frac{\langle(\nabla_\perp^2\phi)^2\rangle_1}{\langle\phi^2\rangle_1}} \quad \text{and} \quad \frac{\epsilon_{\rm h}}{\varphi} \approx \frac{r_1^2}{4\pi}\sqrt{\frac{\langle(\chi^{-3}\hat{\mathcal{W}}\phi)^2\rangle_1}{\langle\phi^2\rangle_1}}, \tag{69}$$

for the low/moderate NA and high NA regimes, respectively. In order to quantify their differences, we chose the nominal amplitude to be unity ($A = 1$) and consider the following ratio

$$Q \triangleq \frac{\epsilon_{\rm h}}{\epsilon_{\rm lm}}. \tag{70}$$

From Sec 4, it was evident that the first-order NRMSE expressions in Eq. (69) are accurate for MSF structures that are well represented by a single frequency (such as those with milled and turned geometries; it turns out to be satisfactory for spoked geometry as well, as shown in Fig. 4). Figure 4 shows Q for the case of a turned, milled, and spoked surface with 10 cycles across the aperture; the behavior of Q is fairly insensitive to the value of C. Note that, although we do not show it explicitly in Sec. 4 and Fig. 2, the first-order NRMSE expression in Eq. (25) works well in the domain of ϵ/φ<0.6 for the spoked structure (shown in green) in Fig. 4. Furthermore, Fig. 4 shows examples of Q for other synthetic MSF structures, which are simple combinations of the milled, turned, and spoked geometries. It should be re-emphasized for these MSF structures, however, that Q strictly only informs on the first-order RMS difference between $\epsilon_{\rm lm}$ and $\epsilon_{\rm h}$. However, it turns out that these additions of elementary (milled, turned, and spoked) MSF geometries lead to examples of ϕ for which the first-order NRMSE approximation in Eq. (25) is accurate for a large range of ϵ/φ.

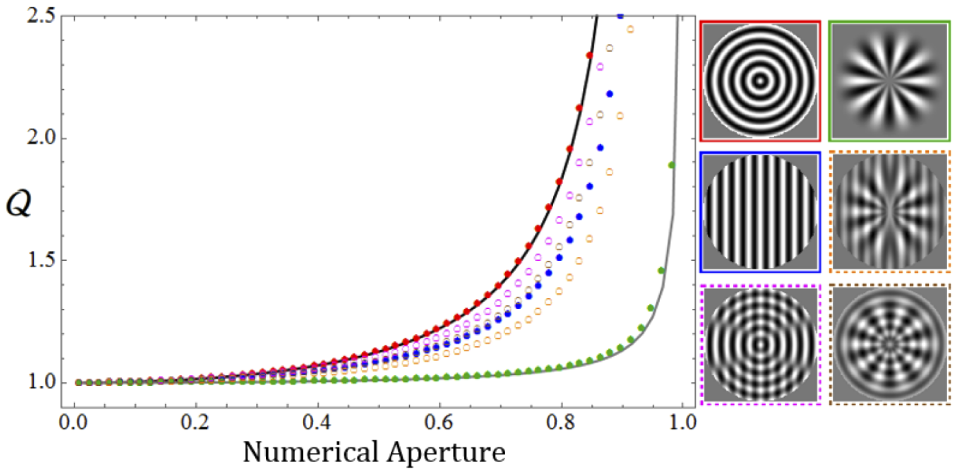

Fig. 4. Plot of Q, which measures the ratio between the first-order expressions of $\epsilon_{\rm lm}$ and $\epsilon_{\rm h}$ various (color-coded) MSF structures ϕ. The elementary geometries of turned, milled, and spoked MSF structures are indicated by solid dots and the remaining, more complicated, examples correspond to hollow dots. The black and gray curves are $Q_{\rm I}$ and $Q_{\rm II}$, which are given by Eq. (72); they represent the bounds of Q for most examples of ϕ.

To make a connection with the rules of thumb found in Ref. 8 regarding the high NA regime, it is prudent to ask which part of $\epsilon_{\rm h}/\varphi$ is responsible for most the behavior of Q as the numerical aperture of the system is increased. From Eq. (69), it is clear that the two possible sources are the factor of χ^{-6} (in the angled brackets of the numerator) and the differential operator $\hat{\mathcal{W}}$. For MSF structures ϕ with more than a few cycles across the aperture, we can further approximate $\epsilon_{\rm h}/\varphi$ by writing the average of a product [χ^{-6} and $(\nabla_\perp^2\phi)^2$] as a product of averages. This approximation

Optics EXPRESS

is justified because χ^{-6} varies much more slowly across the aperture, even for large numerical apertures, when compared with ϕ. Furthermore, as can be seen from Eq. (13), the differential operator $\hat{\mathcal{W}}$ can be approximated by $\nabla_\perp^2$ so long as the dominant variation of ϕ is in r rather than θ. However, if the variation in θ dominates, then $\hat{\mathcal{W}} \approx \chi^2 \nabla_\perp^2$. The NRMSE of the perturbation model, for ϕ that vary dominantly in r or θ, is given by

$$\frac{\epsilon_{\mathrm{np,I}}}{\varphi} \approx \frac{r_1^2}{4\pi}\sqrt{\frac{\langle \chi^{-6}\rangle_A \langle (\nabla_\perp^2 \phi)^2\rangle_A}{\langle \phi^2\rangle_A}} \quad \text{and} \quad \frac{\epsilon_{\mathrm{np,II}}}{\varphi} \approx \frac{r_1^2}{4\pi}\sqrt{\frac{\langle \chi^{-2}\rangle_A \langle (\nabla_\perp^2 \phi)^2\rangle_A}{\langle \phi^2\rangle_A}}, \tag{71}$$

respectively. The corresponding approximate forms of Q are then given by

$$Q_{\mathrm{I}} \approx \sqrt{\langle \chi^{-6}\rangle_1} \quad \text{and} \quad Q_{\mathrm{II}} \approx \sqrt{\langle \chi^{-2}\rangle_1}, \tag{72}$$

which are both independent of the choice of ϕ (aside from the inherent assumption regarding its geometry). Figure 4 shows that Q_{I} matches well with the true value predicted by Eq. (70) for the case where ϕ is a turned MSF structure; for the milled case, there is a notable discrepancy. Once again, this is because a milled MSF structure has non-negligible θ-dependent variations. An example with an even larger discrepancy is that of the spoked geometry, where the variation in θ is much more significant that that in r; in this case, Q_{II} is a much more accurate prediction. For (most) other types of MSF structures, the value of Q (which gives the *first-order* approximation of the RMS difference between the low/moderate-NA and high-NA NRMSE rules of thumb) will lie in between the values predicted by the two expressions in Eq. (72) (between the black and gray curves in Fig. 4). There are unique exceptions; for instance, one may consider MSF structures of the form $\phi(r,\theta) = h(r/r_0)^C \cos(C\theta)$, where C is an integer and r_0 is some constant value. For this particular example, $\nabla_\perp^2 \phi = 0$ but $\hat{\mathcal{W}} \neq 0$. Therefore, $Q \to \infty$ for any numerical aperture. One should keep in mind, though, that this example is very specific and, although $\hat{\mathcal{W}}\phi \neq 0$, it is very near zero [as can be reasoned from the approximations discussed before Eq. (71)].

As a final comment on the comparison of the expressions in Eq. (69), it should be noted that, although Fig. 4 illustrates an intriguing geometrical dependence of Q, its actual value is very close to unity for systems with a moderately large numerical aperature. For an imaging system with a numerical aperture of 0.6, for example, it can be seen that $Q \lesssim 1.2$. Therefore, for the purposes of obtaining a rule-of-thumb error estimate for using the perturbation model in systems with moderate numerical apertures, it may be sufficient to use $\epsilon_{\mathrm{lm}}/\varphi$ in Eq. (69).

Funding

National Science Foundation (1338877); Excellence Initiative of Aix-Marseille Université-A*MIDEX, a French "Investissements d'Avenir" programme.

Disclosures

The authors declare that there are no conflicts of interest related to this article.

References

1. R. J. Noll, "Effect of mid- and high-spatial frequencies on optical performance," Opt. Eng. **18**(2), 182137 (1979).
2. D. Aikens, J. E. DeGroote, and R. N. Youngworth, "Specification and control of mid-spatial frequency wavefront errors in optical systems," in *Frontiers in Optics 2008/Laser Science XXIV/Plasmonics and Metamaterials/Optical Fabrication and Testing*, OSA Technical Digest (CD) (Optical Society of America, 2008), paper OTuA1.
3. J. M. Tamkin, T. D. Milster, and W. Dallas, "Theory of modulation transfer function artifacts due to mid-spatial-frequency errors and its application to optical tolerancing," Appl. Opt. **49**(25), 4825–4835 (2010).
4. G. W. Forbes, "Never-ending struggles with mid-spatial frequencies," Proc. SPIE **9525**, 95251B (2015).
5. K. Liang and M. A. Alonso, "Understanding the effects of groove structures on the MTF," Opt. Express **25**(16), 18827 (2017).

Optics EXPRESS

6. K. Liang and M. A. Alonso, "Effects on the OTF of MSF structures with random variations," Opt. Express **27**(24), 34665 (2019).
7. R. N. Youngsworth and B. D. Stone, "Simple estimates for the effects of mid-spatial-frequency surface errors on image quality," Appl. Opt. **39**(13), 2198–2209 (2000).
8. K. Liang, G. W. Forbes, and M. A. Alonso, "Validity of the perturbation model for the propagation of MSF structure in 2D," Opt. Express **27**(3), 3390–3408 (2019).
9. K. Liang, G. W. Forbes, and M. A. Alonso, "Rapidly decaying Fourier-like bases," Opt. Express **27**(22), 32263 (2019).