Focal-plane electric field sensing with pupil-plane holograms

Emiel H. Por, Christoph U. Keller

Leiden Observatory, Leiden University, P.O. Box 9513, 2300 RA Leiden, The Netherlands

ABSTRACT

The direct detection and spectral characterization of exoplanets requires a coronagraph to suppress the diffracted star light. Amplitude and phase aberrations in the optical train fill the dark zone of the coronagraph with quasi-static speckles that limit the achievable contrast. Focal-plane electric field sensing, such as phase diversity introduced by a deformable mirror (DM), is a powerful tool to minimize this residual star light. The residual electric field can be estimated by sequentially applying phase probes on the DM to inject star light with a well-known amplitude and phase into the dark zone and anazlying the resulting intensity images. The DM can then be used to add light with the same amplitude but opposite phase to destructively interfere with this residual star light.

Using a static phase-only pupil-plane element we create holographic copies of the point spread function (PSF), each superimposed with a certain pupil-plane phase probe. We therefore obtain all intensity images simultaneously while still retaining a central, unaltered science PSF. The electric field sensing method only makes use of the holographic copies, allowing for correction of the residual electric field while retaining the central PSF for uninterrupted science data collection. In this paper we demonstrate the feasibility of this method with numerical simulations.

Keywords: exoplanets, electric field sensing, phase diversity, electric field conjugation, computer-generated holograms, focal-plane wavefront sensing, adaptive optics

1. INTRODUCTION

Since an Earth-like planet is about 10^{-10} fainter than its host star, there are many technical challenges that need to be overcome to achieve the required, high contrast. Coronagraphs are used to suppress the star light in a part of the focal plane, commonly called the dark zone. Many coronagraph types have been designed,[1] but they all suffer, to a varying extent, from aberrations in the optical path; in practice, their performance is limited by these aberrations and not by the coronagraph design. Among these aberrations, telescope vibrations are a major concern. Therefore, coronagraph designs that operate mainly in a pupil plane, such as shaped pupils,[2] phase-induced amplitude apodization[3] and apodizing phase plate coronagraphs,[4] have been predicted to achieve the highest contrast, as pupil plane coronagraphs are largely insensitive to tip-tilt aberrations.

The residual light in the dark zone of the coronagraph can be attributed to different types of speckles. Atmospheric speckles are generated by turbulence in the atmosphere and typically have a relatively short lifetime (~ 20 ms). They are minimized by an adaptive optics (AO) system, which measures the wavefront and corrects aberrations at > 500 Hz. This correction leaves residual atmospheric speckles inside of the control radius of the AO system. These speckles have a lower amplitude, but even shorter lifetimes (< 2 ms), as they are the high-frequency leftovers of the AO system and the atmosphere. These short lifetimes mean that they average out fairly quickly, leaving an incoherent speckle cloud in the science images.

The other types of speckles are caused by the telescope and instruments themselves. Misalignment and imperfect optics lead to static speckles. As they are static, they can be calibrated and removed by post-processing techniques. Quasi-static speckles are generated by slight changes in the optical train, such as bending due to a changing gravity vector and temperature variations. They change on relatively long timescales (~ 10 min) and therefore do not average out during observations. We therefore have to measure them and correct them in post-processing or correct them during the observations.

Correspondence: `por@strw.leidenuniv.nl`

In both cases, we need to estimate these residual speckles, both in amplitude and phase. This estimate cannot be made in the pupil-plane due to frequency folding:[5] high spatial frequency phase aberrations can induce close-in speckles, without any low spatial frequency phase aberrations in the pupil-plane. This means that we have to sample higher spatial frequencies than naively predicted, therefore increasing the noise in the recovered phases. On top of this, only a limited number of combinations of high-frequency phase aberrations actually induce speckles in the dark zone. A pupil-plane wavefront sensor cannot distinguish between them and has to measure them all, while estimation in the focal-plane can directly measure the speckles, and therefore is more photon-efficient. Furthermore, measuring speckles in the focal-plane avoids any non-common-path aberrations that might bias our sensing results.

The most common, and most successful way of focal-plane electric field sensing is DM diversity.[6] This method puts an accurately known phase probe on a deformable mirror in the pupil plane. This phase probe adds light with a well-known amplitude and phase to the dark zone of the coronagraph. This light interferes with the residual light, and the resulting interference pattern is measured. This process is repeated with several different phase probes to provide the phase diversity needed to estimate the residual electric field.

Instead of sequentially acquiring these probe images, we follow the ideas of Keller[7] and obtain them using holographic spots in the focal plane, generated by a phase plate in the pupil plane. Each of the holographic spots contains a different phase probe in the pupil plane and can therefore be used to estimate the electric field. The central spot contains the original point spread function (PSF) and can therefore be used for science, for example for a fiber pickoff in the dark zone. As the estimation algorithm only uses the holographic spots, we can measure and minimize the residual electric field without directly imaging the science PSF.

In this paper, we discuss a focal-plane electric field sensing method that works in parallel with the actual observations. We begin with an overview of the conventional estimation and correction algorithms in Section 2. In Section 3 we explain the design of the holographic phase plate. We discuss the open and closed-loop performance and calibration procedures in Section 4. The influence of noise sources and their impact on the sensitivity is presented in Section 5. Our conclusions are presented in Section 6.

2. ELECTRIC FIELD ESTIMATION AND CORRECTION

In this section, we briefly explain DM diversity in its conventional form and extend the method to large probe amplitudes where non-linear effects of the probes become important. We also discuss our implementation of electric field conjugation.

2.1 Pairwise electric field estimation

The propagation of the electric field at the DM surface to the focal plane is described with a transformation C. As light propagation and apodization is linear in the electric field, C is a linear transformation. The nominal electric field, without any aberrations at the DM surface, is $A = A(\mathbf{x})$. For simplicity, we assume that all aberrations in the optical train can be represented as aberrations in the DM plane, and denote the aberrated electric field there as $E_{\text{aber}}(\mathbf{x}) = A(\mathbf{x})[1+g(\mathbf{x})]$. Note that this is *not* a linear expansion; it includes all non-linear terms as well. Additionally, using both the real and imaginary components of g, this can model both phase and amplitude aberrations.

Including a phase probe ϕ_k on the deformable mirror, we can write the electric field in the focal-plane as

$$\begin{aligned} E_k &= C\{E_{\text{aber}} \exp[i\phi_k]\} && (1)\\ &= C\{A(1+g)\exp[i\phi_k]\} && (2)\\ &= C\{A(1+g)\} + C\{A(1+g)(\exp[i\phi_k]-1)\} && (3)\\ &= C\{A(1+g)\} + C\{A(\exp[i\phi_k]-1)\} + C\{Ag(\exp[i\phi_k]-1)\}, && (4) \end{aligned}$$

where $\phi_k = \phi_k(\mathbf{x})$ is the k-th phase probe on the DM surface. If the aberrations are relatively small, the last term can be neglected, leading to

$$E_k = C\{A(1+g)\} + C\{A(\exp[i\phi_k]-1)\}. \tag{5}$$

Taking the absolute value squared yields the focal-plane intensity

$$\begin{aligned} I_k &= |E_k|^2 && (6)\\ &= |C\{A(1+g)\}|^2 + |C\{A(\exp[i\phi_k]-1)\}|^2 + 2\Re[C\{A(1+g)\}C^*\{A(\exp[i\phi_k]-1)\}], && (7)\end{aligned}$$

where $(\cdot)^*$ denotes the complex conjugate. Taking the intensity difference between two different phase probes ϕ_k and ϕ_ℓ, we obtain

$$I_k - I_\ell = |C\{A(\exp[i\phi_k]-1)\}|^2 - |C\{A(\exp[i\phi_\ell]-1)\}|^2 + 2\Re[C\{A(1+g)\}C^*\{A(\exp[i\phi_k]-\exp[i\phi_\ell])\}]. \quad (8)$$

The first two terms are the incoherent intensity difference between the two probe electric fields. These do not depend on the aberrations present in the system, and therefore can be calculated analytically. The third term modulates the aberrated electric field with the difference of the two probe fields. As we observe only the real part of this interference term, we need at least two image differences for a well-defined inverse.

For small and opposite probes, $\phi_\ell = -\phi_k$ and $|\phi_k| \ll 1$, this expression becomes significantly simpler. The first two terms cancel each other and the third term reduces to the conventional expression

$$I_k - I_{-k} = 4\Re[C\{A(1+g)\}C^*\{iA\phi_k\}], \quad (9)$$

where negative indices indicate the use of the opposite probe phase. While having small and opposite probes leads to symmetric probe electric fields and therefore cancellation of the additive components of the phase probes, there is nothing preventing us from using large probes or intensity differences between different probes. If we have N probes, we should gather $2N$ probe images, and therefore have $2N-1$ independent image differences instead of just N image differences for using only opposite probes.

Note that for large probes there still is a bias present due to non-linear terms. The second-order terms in the squared phase exponential will cancel each other as they have the same sign. The third-order terms, however, will have the opposite sign, thereby creating a bias in the intensity difference. When left uncorrected, this intensity bias will be visible as an electric field bias in the converged image. This electric field will be stable except for the photon noise.

Writing Equation 8 as an inner product of two vectors, and stacking these for N probes, we obtain

$$\begin{pmatrix} I_1 - I_2 \\ \vdots \\ I_{N-1} - I_N \end{pmatrix} = M \begin{pmatrix} \Re[C\{A(1+g)\}] \\ \Im[C\{A(1+g)\}] \end{pmatrix} + I_{\text{bias}}, \quad (10)$$

where

$$M = 2\begin{pmatrix} \Re[C\{iA(\exp[i\phi_1]-\exp[i\phi_2])\}] & \Im[C\{iA(\exp[i\phi_1]-\exp[i\phi_2])\}] \\ & \vdots & \\ \Re[C\{iA(\exp[i\phi_{N-1}]-\exp[i\phi_N])\}] & \Im[C\{iA(\exp[i\phi_{N-1}]-\exp[i\phi_N])\}] \end{pmatrix}; \quad (11)$$

$$I_{\text{bias}} = \begin{pmatrix} |C\{A(\exp[i\phi_1]-1\}|^2 - |C\{A(\exp[i\phi_2]-1)\}|^2 \\ \vdots \\ |C\{A(\exp[i\phi_{N-1}]-1)\}|^2 - |C\{A(\exp[i\phi_N]-1)\}|^2 \end{pmatrix}. \quad (12)$$

With two or more well-chosen probes, this matrix is invertible using the Moore-Penrose pseudo-inverse, allowing us to estimate the electric field using the measured intensity differences. The transformation for all pixels in the dark zone can then be described as a large, blockwise matrix with the matrices M for each pixel on its diagonal.

The probe shapes must be chosen such as to put most light into the dark zone while minimizing non-linear effects. To first order, the probe electric field is given by the Fourier transform of the phase pattern. For rectangular dark zones, we therefore use

$$\phi_p(\mathbf{x}) = \text{sinc}(\mathbf{w}_x\mathbf{x}_x)\text{sinc}(\mathbf{w}_y\mathbf{x}_x)\cos(\mathbf{c}_x\mathbf{x}_x + \theta_x)\cos(\mathbf{c}_y\mathbf{x}_y + \theta_y), \quad (13)$$

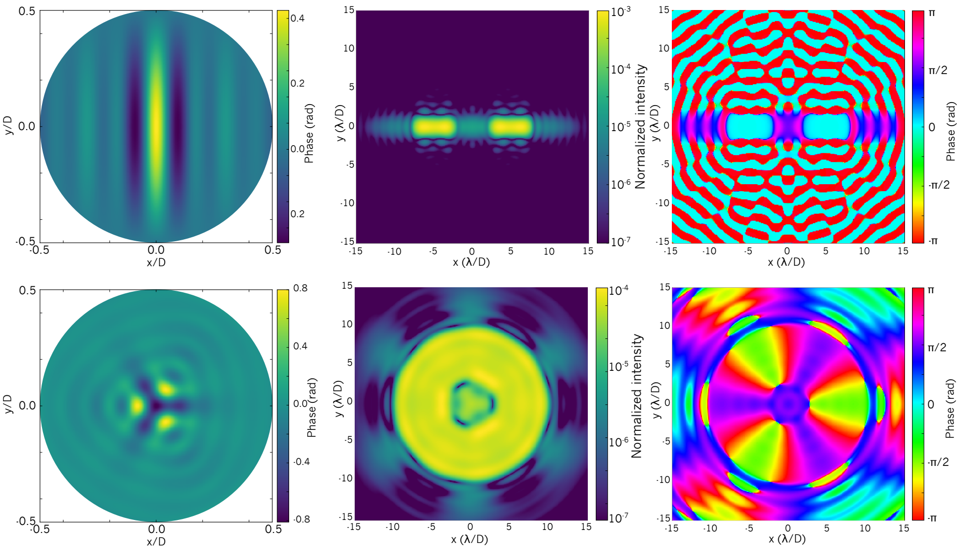

Figure 1. The phase probes (left column) and their corresponding electric fields in intensity (middle column) and phase (right column) for a rectangular (top row) and circular (bottom row) dark-zone. The non-linear effects manifest themselves as broadening of the probe electric fields by frequency folding.

where $\mathbf{w}$ is the size of the dark zone, $\mathbf{c}$ is its center, and θ is an angle that defines the phase of the probe electric field. The probe is oversized to provide sufficient probe strength at the edges of the dark zone.

For dark zones that go all the way around the star the situation becomes more difficult. Because the probe electric field is Hermitian point-symmetric with respect to the star, we cannot choose a single phase for the probe electric field. We instead vary this phase with position angle and restrict the probe electric field to an annulus around the star. The phase probe is then generated by taking the inverse Fourier transform of this field. The other phase probes can be produced by rotating the first probe. This Hermitian symmetry also means that we need at least three probes instead of only two to create the electric field diversity: while using just two probes, there are always certain position angles where the phase of the probe electric field is identical for both probes. The third probe is then used to avoid this degeneracy.

Examples for both types of phase probes can be found in Figure 1, including their corresponding probe electric field in intensity and phase. At the used probe amplitudes, we can clearly see the non-linear effects of the phase probes as frequency folding: the autocorrelation of the dark zone can be seen superimposed in the intensity images.

2.2 Electric field conjugation

Given an estimate for the electric field in the dark zone, we can change the deformable mirror to cancel out this unwanted light. The most straightforward method is electric field conjugation,[8] which finds a phase deformation that diffracts light with the opposite sign from the central Airy core towards the dark zone. It linearizes the phase around the nominal deformable mirror setting, therefore making inversion of the otherwise non-linear problem possible.

The focal-plane electric field after adding a bias phase $\delta\phi_{\mathrm{DM}}$ to the deformable mirror is given by

$$\begin{aligned} E_{\mathrm{fp}} &= C\{A(1+g)\exp[i\delta\phi_{DM}]\} && (14)\\ &\approx C\{A(1+g+i\delta\phi_{\mathrm{DM}}\}, && (15)\\ &= E_{\mathrm{aber}} + C\{iA\delta\phi_{\mathrm{DM}}\}. && (16)\end{aligned}$$

where we have linearized the deformable mirror bias around the nominal value, and again assumed small aberrations and neglected the crossterm of aberrations and DM bias. Combining the real and imaginary parts of the aberrated electric field into a purely real vector to guarantee real actuator values, we can define

$$G = \begin{pmatrix} \Re[iC\text{diag}\{A\}] \\ \Im[iC\text{diag}\{A\}] \end{pmatrix}, \tag{17}$$

where diag$(\cdot)$ is the matrix with the vector $(\cdot)$ on the diagonal. Setting $E_{\text{fp}} = 0$ we obtain

$$\begin{pmatrix} \Re[E_{\text{aber}}] \\ \Im[E_{\text{aber}}] \end{pmatrix} = -G\delta\phi_{\text{DM}}. \tag{18}$$

As the transformation G is linear, it can be inverted, given enough degrees of freedom (i.e. DM actuators) on the deformable mirror. This leads to

$$\delta\phi_{\text{DM}} = -G^{+} \begin{pmatrix} \Re[E_{\text{aber}}] \\ \Im[E_{\text{aber}}] \end{pmatrix}, \tag{19}$$

where G^{+} is the Moore-Penrose pseudo-inverse of the matrix G. Note that $A = A_0 \exp[i\phi_{\text{DM}}]$ includes the current position of the deformable mirror ϕ_{DM}. As each iteration ϕ_{DM} is different, the linear transformation G needs to be recalculated. In our simulations, we calculate G once, and use it for each iteration, meaning that the total DM stroke after convergence has to fall within the linear range (±1 rad). Outside of this range, the algorithm will diverge. Recalculating G for each iteration lifts this assumption, but still requires the DM stroke per iteration to be within the linear range, which can be achieved by reducing the gain of the closed-loop system.

3. PUPIL-PLANE PHASE-ONLY HOLOGRAMS

3.1 Modulated cosine phase gratings

The holographic copies are generated by a single, phase-only optical element located in the pupil plane. Each focal-plane spot is generated with a modulated cosine ripple.[9] All these individual holograms are coherently added, yielding the complex transmittance

$$t(\mathbf{x}) = \exp\left[i \sum_k m_k \cos(a_k \phi_k(\mathbf{x}) + 2\pi i \mathbf{p}_k \cdot \mathbf{x}) \right], \tag{20}$$

where m_k, a_k and $\mathbf{p}_k$ are the the modulation depth, the probe strength and the focal-plane position of the k-th probe image, respectively. Using the well-known Jacobi-Anger expansion

$$\exp[iz\cos\theta] = J_0(z) + 2\sum_{n=1}^{\infty} i^n J_n(z)\cos(n\theta), \tag{21}$$

we end up, neglecting higher-order terms, with the more useful expression

$$t(\mathbf{x}) = \prod_k \left\{ J_0(m_k) + \sum_{n=1}^{\infty} \left[i^n J_n(m_k) \left(\exp[ina_k\phi_k(\mathbf{x}) + 2\pi i \mathbf{x}\cdot\mathbf{p}_k] + \exp[-ina_k\phi_k(\mathbf{x}) - 2\pi i \mathbf{x}\cdot\mathbf{p}_k] \right) \right] + \cdots \right\}. \tag{22}$$

This allows us to more easily write the focal-plane electric field $U_{\text{fp}}(\mathbf{k}) = \mathcal{F}\{U_{\text{pp}}(\mathbf{x})t(\mathbf{x})\}$ for some pupil-plane electric field $U_{\text{pp}}(\mathbf{x})$ transmitted through the optical element as

$$U_{\text{fp}}(\mathbf{k}) = U_{\text{zero}}(\mathbf{k}) + U_{\text{linear}}(\mathbf{k}) + U_{\text{cross}}(\mathbf{k}) + \cdots, \tag{23}$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform, and

$$U_{\text{zero}}(\mathbf{k}) = \mathcal{F}\{U_{\text{pp}}(\mathbf{x})\} * \delta(\mathbf{k}) \prod_k J_0(m_k); \tag{24a}$$

$$U_{\text{linear}}(\mathbf{k}) = i \sum_k b_k \Big\{ \mathcal{F}\left\{U_{\text{pp}} \exp[i a_k \phi_k]\right\} * \delta(\mathbf{k} - 2\pi \mathbf{p}_k) \\ + \mathcal{F}\left\{U_{\text{pp}} \exp[-i a_k \phi_k]\right\} * \delta(\mathbf{k} + 2\pi \mathbf{p}_k) \Big\}; \tag{24b}$$

$$U_{\text{cross}}(\mathbf{k}) = -\frac{1}{2} \sum_{\substack{k,\ell \\ k \neq \ell}} c_{k\ell} \Big\{ \mathcal{F}\left\{U_{\text{pp}} \exp[i(a_k \phi_k + a_\ell \phi_\ell)]\right\} * \delta(\mathbf{k} - 2\pi \mathbf{p}_k - 2\pi \mathbf{p}_\ell) \\ + \mathcal{F}\left\{U_{\text{pp}} \exp[i(a_k \phi_k - a_\ell \phi_\ell)]\right\} * \delta(\mathbf{k} - 2\pi \mathbf{p}_k + 2\pi \mathbf{p}_\ell) \\ + \mathcal{F}\left\{U_{\text{pp}} \exp[i(-a_k \phi_k + a_\ell \phi_\ell)]\right\} * \delta(\mathbf{k} + 2\pi \mathbf{p}_k - 2\pi \mathbf{p}_\ell) \\ + \mathcal{F}\left\{U_{\text{pp}} \exp[i(-a_k \phi_k - a_\ell \phi_\ell)]\right\} * \delta(\mathbf{k} + 2\pi \mathbf{p}_k + 2\pi \mathbf{p}_\ell) \Big\}, \tag{24c}$$

where $*$ is the convolution operation, b_k and $c_{k\ell}$ are the strengths of the first order and cross terms, respectively, given by

$$b_k = J_1(m_k) \prod_{\ell \neq k} J_0(m_\ell); \tag{25a}$$

$$c_{k\ell} = J_1(m_k) J_1(m_\ell) \prod_{n \neq k,\ell} J_0(m_n). \tag{25b}$$

We can clearly see that $U_{\text{zero}}(\mathbf{k})$ is the normal, zero order and is completely unaltered, except for a lower intensity. The linear $U_{\text{linear}}(\mathbf{k})$ terms create symmetrical pairs of spots at different positions in the focal-plane for each phase probe. The two spots are copies of the central spot, but are biased with the positive and negative phase probe $\phi_k(\mathbf{x})$ respectively. Therefore, they are perfectly suited for pairwise electric field sensing.

The cross terms $U_{\text{cross}}(\mathbf{k})$ also produce symmetric spots in the focal plane, now with the sum and difference of each pair of probes. They generally have much smaller amplitudes (quadratic instead of linear in modulation amplitude), meaning increased photon noise compared to the first-order images, and we therefore exclude them from the pairwise electric field estimation, even though their phase probes are well defined.

3.2 Crosstalk and geometry

As the PSF copies are not spatially confined, some light from every copy will end up in the dark zone of the zero-order term, therefore degrading the contrast of the science image. As this extra light is not contained in the linear terms, they will not be measured by the copies themselves. We can correct for this diffracted light using a small-phase aberration containing only low spatial frequencies. This aberration makes the dark zone dark and the low spatial frequencies ensure that the copies are still perfect and not perturbed by the zero-order term.

High spatial frequency speckles that lie on top of the dark zones of the copies, are not as easily corrected: their phase and amplitude are not known beforehand. All speckles from copies will still degrade the contrast. As the power spectrum of speckles decreases for higher spatial frequencies, placing all copies far away from the zero-order term may be sufficient for reducing this effect.

Even more troublesome are the speckles from the zero order term that end up in the dark zones of the PSF copies corresponding to the linear terms. As these linear copies are much dimmer than the zero-order PSF, even a 10^{-7} speckle from the central PSF will be seen as a 10^{-5} speckle in a first-order PSF copy. The naive solution would be to increase the distances between all spots. However when we observe with a finite spectral bandwidth, the spots will smear radially. The radial smearing increases linearly with distance from the center, thereby requiring a smaller bandwidth as we place the copies further out.

A field stop allows us to separate the light for all copies. This ensures that the light from each copy is confined to a certain area. The placement of all copies, that is the geometry of the hologram, can then be chosen such that all terms are separated on the detector. There is one important exception to this rule: if cross terms and higher-order terms perfectly coincide with a linear term, then we can still use the resulting image. In this case

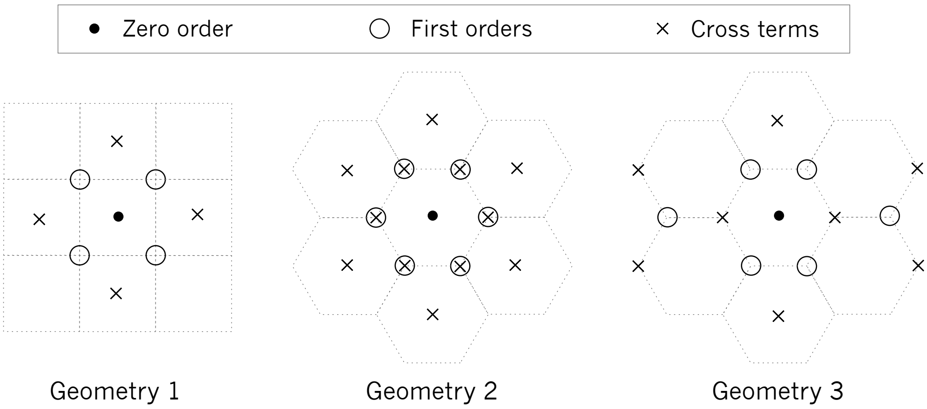

Figure 2. A few example geometries along with cross term and second-order spot positions. For two probes the square geometry 1 is the simplest and provides clean first orders. A detector image for this geometry is shown in Figure 3. For three probes, the compact hexagonal geometry 2 has the disadvantage that cross terms are imaged on top of first-order terms. The more extended geometry 3 avoids this problem.

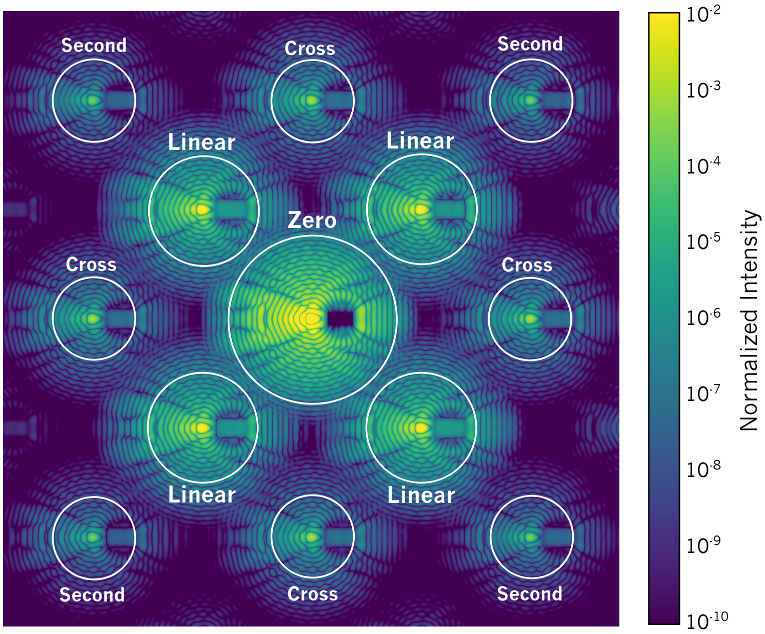

Figure 3. A detector image for the square geometry as indicated in Figure 2. The zero-order and linear terms, cross terms and second-order spots are indicated. Crosstalk between copies is minimized by using a focal-plane mask between the coronagraph and holographic phase plates.

the probe electric fields add up to an effective probe field, and the residual speckles are imaged at the same position so that there is no confusion about speckle location.

In the case of only two probes, we can use a simple square geometry. This ensures that cross terms and higher-order terms end up farther away from the center and will never overlap. For more than two probes, multiple geometries are possible. For three probes, two main classes of geometry can be found. In Figure 2 these three configurations are shown, along with the positions of cross terms and second-order terms. The detector image for the square geometry is shown in Figure 3 where all terms are indicated. For more than three probes, even more geometries are possible.

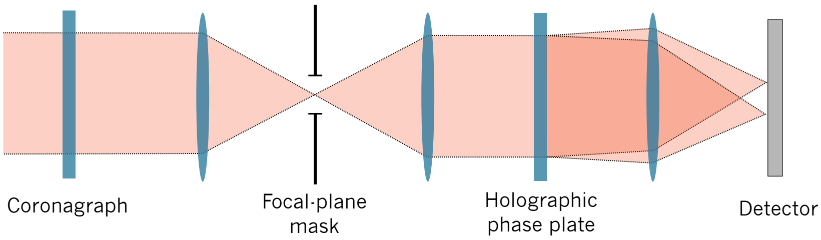

Figure 4. A schematic view of the simulated optical setup. We could avoid the use of two phase plates by multiplexing an apodizing phase plate coronagraph and the holographic phase plate. The coronagraph then depends on the focal-plane mask, although for large masks and shallow coronagraph designs, this dependency is small. In these cases, the aberrated electric field is often smaller than the static speckles and can be reduced in the same way.

3.3 Comparison with other holographic phase gratings

The coronagraphic modal wavefront sensor[10] uses a binary phase grating, which results from two phase gratings that are blazed in opposite direction and creates copies with opposite phase modulations. This increases the diffraction efficiency of the linear terms and simultaneously decreases those of the zeroth order and cross-terms. The zeroth-order term can then be recovered by decreasing the modulation amplitude.

The binary phase grating also introduces a scattered background throughout the focal-plane and therefore reduces the normalized intensity of our coronagraph. This can be corrected using a small phase aberration in the pupil plane, given by Electric Field Conjugation or a Gerchberg-Saxton-like algorithm. This phase aberration must only consist of low spatial frequencies so that it has no influence on the quality of the copies. As the scattered background floods the whole focal plane, it is also visible in the dark zones of the linear copies. This can be corrected by changing the electric field in the copies, as calculated analytically from the holographic phase plate, for the pairwise estimation, instead of the specified phase probes.

In addition, the binary nature means that implementations with pixelated phase-plate holograms cannot represent grating modulations of less than 1 pixel. This means that the phase probe that is actually in the copies might be slightly different from the intended value, especially for small probe amplitudes, as are often used. Again, this can be corrected by using the analytically calculated electric field in the copies for the pairwise estimation.

Although these deficiencies can be corrected, they increase the complexity in generating the holographic phase gratings. For simplicity, we restrict ourselves to using cosine phase gratings.

4. RESULTS

4.1 Open-loop performance

We simulate the optical setup shown in Figure 4. To avoid contamination of the copies by the central spot and each other, we use a focal-plane mask between the coronagraph phase plate and the holographic phase plate. The two phase plates can be combined by simply adding the two phase patterns. However, this means that some light will leak from low to high spatial frequencies due to high spatial frequencies and frequency folding in the APP coronagraph pattern. This can be avoided by designing the APP with the field of view mask in mind, but this adds an additional dependency to the design.

Our APP coronagraph is designed to have a rectangular dark zone from 3.5 to 8.5 λ/D in the horizontal direction and -1 to 1 λ/D in the vertical direction. In this dark zone the contrast is 10^{-10} in the absence of aberrations. We use an unobstructed aperture to simplify the coronagraph design. Even though we can design APPs with a dark zone closer to the star and for apertures with central obscuration, spiders and/or segmentation, this configuration was deemed sufficient for initial testing. The focal-plane mask cuts of any light farther than 16 λ/D away from the star and was smoothed with a two-dimensional Kaiser windowing function with a width of 6 λ/D to avoid ringing effects caused by a hard cutoff in this mask. We oversized the pupil to accommodate any residual ringing of this mask.

In the following we only consider phase and amplitude aberrations in the pupil plane. Although this assumption might not be representative of all aberrations in actual optical systems,[11] it is often used to simulate the effectiveness of speckle reduction techniques. We assume a power-law index of $n_p = n_a = -2$ for phase and amplitude power spectral densities and scale to the required phase or amplitude variances.

We simulate photon noise and detector noise sources including read noise, flat-field noise and dark-current noise. For each copy that we want to use for estimation, we calculate the centroid of its Airy core to correct for tip-tilt errors. Then the dark zone and its surrounding pixels are interpolated using Fourier interpolation to the correct sub-pixel sampling of the dark zone. This step is used to unify the sampling for the dark zone for each of the spots. These interpolated intensities are then used for estimation of the central electric field.

To show the open-loop performance, we show the response to right-singular vectors of the linearized propagation transformation. In Figure 5 we show the modes in the pupil and focal planes for our dark zone. As expected the singular values drop to zero very quickly, indicating that only a limited number of modes is necessary to correct aberrations.

In Figure 6 we show the response of all modes to a single input mode. We can see that most modes have small or no non-linear behaviour. Crosstalk is also minimal, except for a few mid-order modes. These modes still put enough light into the dark zone to trigger frequency folding inside it, leading to non-linear crosstalk. Generally this doesn't matter that much as the power in these modes is substantially lower compared to the lower-order modes.

Figure 7 shows the Pearson product-moment correlation coefficients between measured and actual electric field, both decomposed into the electric field basis obtained from the SVD of the linear propagation transformation. Performing a Monte-Carlo simulation for $N = 2000$ random realizations of the pupil-plane aberrations shows that higher-order modes don't correlate well in the $\sigma = 0.5$rad case, which means that reconstruction of these modes is not possible. Reducing the static aberrations in the system leads to an increase in the number of reconstructible modes. Numerical noise can still be seen in off-diagonal entries.

These plots show that open-loop reconstruction, even for relatively high aberration amplitudes, is possible. If the photon-noise on the static aberrations is limiting our images, then we could do better in closed-loop operation. However, in the case that photon-noise from residual atmospheric speckles is the main noise source, open-loop, post-facto correction should in principle have identical performance to closed-loop operation.

4.2 Closed-loop performance

For closed-loop operation we minimize the light induced by aberrations with a suitable change of the deformable mirror using electric field conjugation. Here we describe our calibration and simulation procedures for the closed-loop simulations. These are an extension of the description of the open-loop configuration.

4.2.1 Calibration

As both the electric field sensing transformation and the electric field conjugation transformation are linear, their combination is also linear. This means that we can use the linear control framework. An interaction matrix can be constructed from theory, by simply multiplying the full phase diversity matrix with the linearized propagation matrix from the electric field conjugation.

Alternatively, the interaction matrix can be found empirically. One image taken without any changes to the deformable mirror serves as our reference image, followed by a sequence of images with modes from a complete mode basis applied to the deformable mirror. The intensities in the dark zones of the copies are extracted from each image, and the reference intensities are subtracted from the sequence. The intensity differences between the copies are then calculated, and the resulting vectors serve as the columns of the interaction matrix, barring a projection matrix onto the used mode basis. If we want to keep the electric field static, we can use the reference intensity differences as a bias correction for the electric field sensing.

The choice of mode basis for this calibration is crucial. A simple actuator poke will not change the intensities much, requiring long exposure times. A better mode basis can be found by taking the right-singular vectors of the theoretical interaction matrix. Their singular values tell us how effective this mode is at putting light into

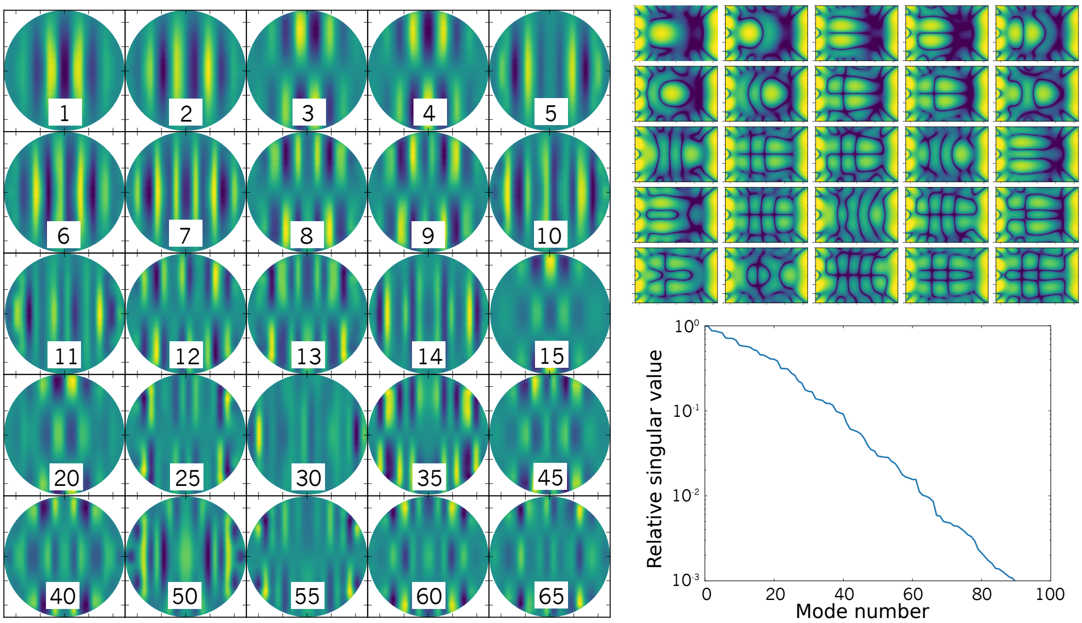

Figure 5. The principal pupil-plane modes and their corresponding focal-plane responses and singular values for an unobstructed aperture and a rectangular dark-zone from 3.5 to $8.5\lambda/D$ in x, and -1 to $1\lambda/D$ in y.

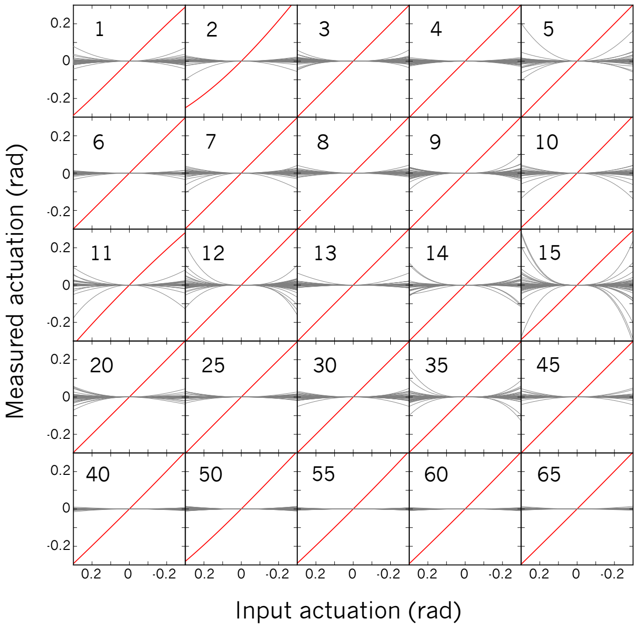

Figure 6. The response of the electric field measured by the holograms while varying the indicated mode in the input. Most modes are reasonably well behaved. Cross talk increases towards higher actuations as expected.

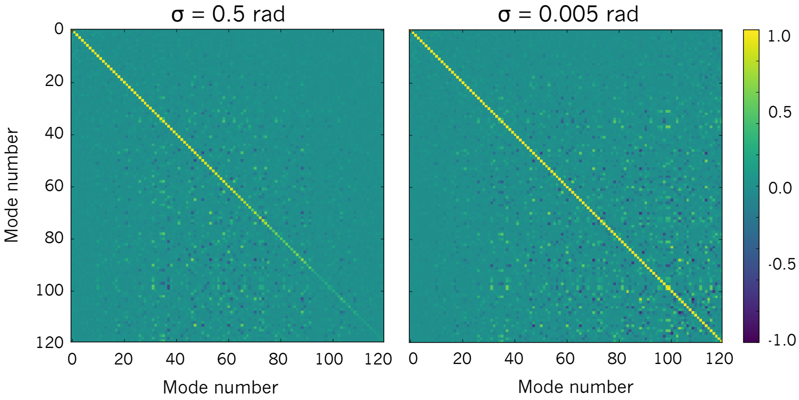

Figure 7. The Pearson product-moment correlation coefficients between the electric fields as measured by the holograms and the actual electric field for $N = 2000$ random realizations of the static aberrations indicated in the text. The left figure has aberrations with a standard deviation of $\sigma = 0.5$rad and the right figure $\sigma = 0.005$rad.

the dark zone. We can therefore increase the stroke of that mode on the deformable mirror during calibration to decrease the required exposure time, while still making sure that we are still within the linear regime.

The interaction matrix can then be inverted to obtain the control or reconstructor matrix. The exact method used for inversion influences the closed-loop performance of the speckle suppression method. A full Moore-Penrose pseudo-inverse will degrade performance by also correcting badly-sensed modes. A modal cutoff, which sets the gain of badly-sensed modes (modes with a low singular value) to zero, can be performed to reduce this effect. Any aberrations in those modes will not be corrected. Alternatively we can use Tikhonov regularization, which simply reduces the gain on badly-sensed modes, instead of totally eliminating them from the correction. In both cases, we have to tune one parameter (number of corrected modes or regularization parameter) to achieve optimal performance. Although better control algorithms can be used, we restrict ourselves to these two for simplicity.

4.2.2 Simulation

The atmosphere is simulated by a phase screen with a Kolmogorov power spectrum. The adaptive optics system is simulated as a transfer function on this power spectrum. The resulting power spectrum is

$$F(u) \propto H(u - u_o)u^{-11/3}, \tag{26}$$

where $H(u)$ is the Heaviside function and u_o is the outer spatial frequency of the adaptive optics control area. Correction was done on a 20×20 deformable mirror. Influence functions were assumed to be Gaussian with a standard deviation of half the interactuator spacing.

For long integration times we need to average our intensity images over many random phase screens. In the limit of infinite integration times we can ignore the speckle pinning term, yielding

$$I_{\text{long}} = S|E|^2 + I_{\text{atmos}} \max[|E|^2], \tag{27}$$

where E is the electric field without atmospheric effects, S is the Strehl ratio of a PSF after the adaptive optics system, and I_{atmos} is the incoherent atmospheric halo after the adaptive optics system. This incoherent halo is given by simulation of many phase screens for an unaberrated PSF and only needs to be calculated once. This approach allows us to perform many simulations averaged over atmospheric speckles without needing an order of magnitude more computing time.

Performing closed-loop simulations, we observe convergence to photon-limited measured electric fields, as can be seen in Figure 8. The choice of control algorithm changes the attainable contrast, as is shown in Figure 9. In the rest of this paper we will use the modal cutoff inversion to set a baseline. Other, more complicated control algorithms, for example Kalman filtering based controllers[12] or predictive control, will improve performance beyond this baseline.

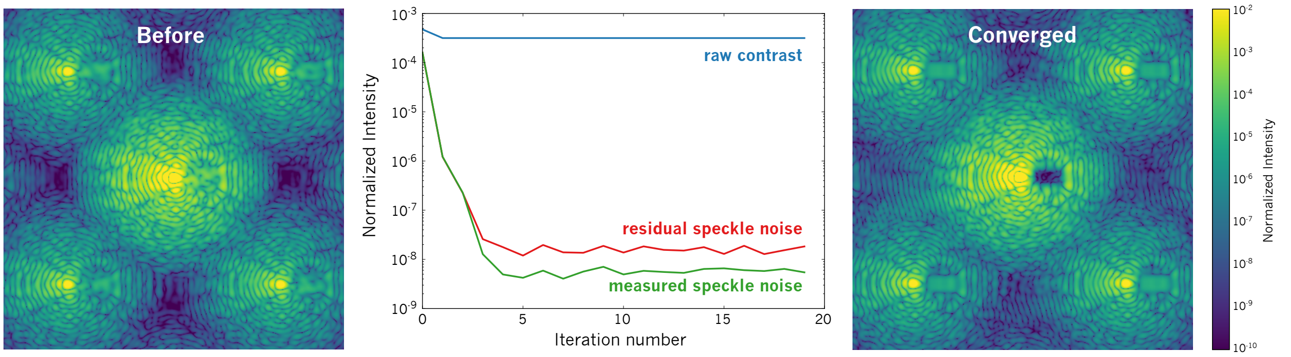

Figure 8. The median intensity in the dark zone of the coronagraph as a function of iteration number for simulated closed-loop operation of the proposed system. A static speckle field was assumed. The total number of detected photons was 10^{11} per iteration, which were used to correct the principal 50 modes. The raw contrast is limited by the normalized intensity of the residual adaptive-optics speckles. The stability on the static speckles is better and is limited by photon noise in the electric field estimation. In the images on the sides, the atmospheric contribution was removed to show the quasi-static speckle reduction.

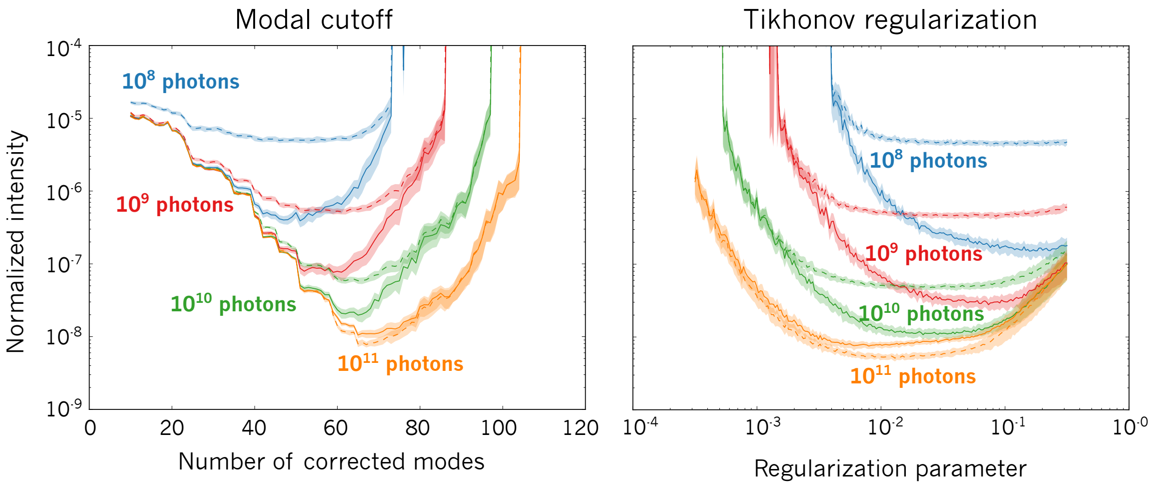

Figure 9. The median normalized intensity inside the dark-zone with closed-loop correction as a function of the number of corrected modes for the modal cutoff controller (left) and the regularization parameter of the Tikhonov regularization (right). Tikhonov regularization performs slightly better and is more stable w.r.t. its tunable parameter, as expected. More complicated controllers will achieve better performance than those simulated here. Solid lines indicate the normalized intensity of the actual electric field in the dark zone, while dashed lines indicate the normalized intensity of the electric field in the dark zone as measured by the holographic phase diversity. The latter in closed-loop operation should be higher than the former due to the averaging effect of the control system. However, a bias between measured and actual electric field will result in convergence of the actual electric field to this bias, therefore limiting its normalized intensity. In this case the measured normalized intensity can be lower than the actual one.

5. ERROR ANALYSIS

By adding noise sources to the closed-loop simulations, we can analyze the noise sensitivity of the speckle suppression algorithm. In this section we describe the most dominant noise components: photon shot noise, residual atmospheric speckle noise, flat-field noise, changing speckle field with time and broadband light effects, from both a theoretical and numerical perspective. Noise sources in phase diversity have been covered before.[13] Instead we will focus on the noise sources where they are relevant or differ in our method.

5.1 Photon shot noise

As photon shot noise is the fundamental limit, we should always strive for photon-noise-limited performance. In that case the variance of the intensity differences is

$$\sigma^2_{I_k - I_\ell} = I_k + I_\ell, \tag{28}$$

where I_k and I_ℓ are the number of photons in each pixel for the respective probe images. For $I_k, I_\ell \gg 1$, we can assume a normal distribution, and these variances can be transformed into covariances of the electric field components as

$$C_E = M^+ \mathrm{diag}[\sigma^2_{I_\mathrm{diff}}](M^+)^T, \tag{29}$$

where M is the DM diversity matrix defined in Equation 10, and $\mathbf{I}_\mathrm{diff}$ is the vector on the left-hand side of the same equation. A ten-fold increase in the number of incoming photons therefore leads to $\sqrt{10}$-fold decrease in the standard deviation of the electric field components, which is a ten-fold decrease in the standard deviation of the speckle intensities in closed-loop operation.

This is exactly what we see in the simulations in Figure 9, at least for the measured electric fields. The actual electric fields still have a bias compared to the measured electric fields and therefore their absolute intensity is higher, although this intensity is more stable compared to lower photon fluxes.

5.2 Fast changing atmospheric speckles

Good performance on ground-based telescopes requires resilience against the fast residual atmospheric speckles remaining after the AO correction. We can describe the interaction of the residual atmospheric speckles with the residual electric field as a speckle-pinning effect.[14] The electric field of the probe images in the focal-plane can be decomposed into three components as

$$E_\mathrm{fp} = E_\mathrm{atmos} + E_\mathrm{probe} + E_\mathrm{resid}, \tag{30}$$

of which we measure the intensity

$$\begin{aligned} I &= |E_\mathrm{fp}|^2 & (31)\\ &= |E_\mathrm{atmos}|^2 + |E_\mathrm{probe}|^2 + |E_\mathrm{aber}|^2 & (32)\\ &\quad + 2\Re[E_\mathrm{atmos}]\Re[E_\mathrm{probe}] + 2\Im[E_\mathrm{atmos}]\Im[E_\mathrm{probe}] &\\ &\quad + 2\Re[E_\mathrm{atmos}]\Re[E_\mathrm{aber}] + 2\Im[E_\mathrm{atmos}]\Im[E_\mathrm{aber}] &\\ &\quad + 2\Re[E_\mathrm{probe}]\Re[E_\mathrm{aber}] + 2\Im[E_\mathrm{probe}]\Im[E_\mathrm{aber}]. & \end{aligned}$$

Every pair of electric fields generates a speckle-pinning term. When integrating on the detector, we measure the time-averaged intensity $\langle I\rangle_t$ during that time. For long integration times we assume $\langle\Re[E_\mathrm{atmos}]\rangle_t = \langle\Im[E_\mathrm{atmos}]\rangle_t = 0$, and we obtain

$$\begin{aligned} \langle I\rangle_t = &\langle|E_\mathrm{atmos}|^2\rangle_t + |E_\mathrm{probe}|^2 + \langle|E_\mathrm{aber}|^2\rangle_t \\ &+ 2\Re[E_\mathrm{probe}]\langle\Re[E_\mathrm{aber}]\rangle_t + 2\Im[E_\mathrm{probe}]\langle\Im[E_\mathrm{aber}]\rangle_t. \end{aligned} \tag{33}$$

We therefore obtain the same result as before when we neglected atmospheric aberrations, except for an incoherent atmospheric background that does not depend on the probe electric field. It is therefore subtracted out in the intensity difference. However, for finite integration times, $\langle \Re[E_{\mathrm{atmos}}] \rangle_t, \langle \Im[E_{\mathrm{atmos}}] \rangle \neq 0$, and we make an error in the estimated aberrated electric field of

$$\sigma_{|E_{\mathrm{aber}}|} = \langle |E_{\mathrm{atmos}}| \rangle_t \sqrt{\frac{t_0}{t_{\mathrm{int}}}}, \tag{34}$$

where t_{int} is the integration time, and t_0 the speckle coherence time of the atmospheric speckles. Within the control area of the AO system, t_0 is given by the speed and gain of the AO system; outside the control area, it is given by the atmospheric coherence time.

This residual atmospheric electric field gives a lower limit on the performance of any electric field sensing method on ground-based telescopes. Without any other noise sources in the simulations, we are indeed limited by these residuals. When introducing photon noise, the incoherent contribution of the residual atmospheric speckles increases the photon shot noise, effectively reducing the number of photons available for estimation.

We can therefore increase the probe amplitudes to increase the response of the electric field estimator until the probes have intensities similar to the residual atmospheric speckles. This, however, increases the non-linear effects of the probes and therefore necessitates the use of the intensity biases explained in Section 2.

5.3 Flat field

Flat-fielding errors lead to an additional bias in the intensities. A zero electric field will be measured as

$$I_{\mathrm{diff,meas}} = f_k I_k - f_\ell I_\ell \neq I_k - I_\ell, \tag{35}$$

where f_k is the flat-field of probe image k. Therefore, the variance of the measured intensity difference is

$$\mathrm{Var}[I_{\mathrm{diff,meas}}] = 2\mathrm{Var}[f_k] I_k^2, \tag{36}$$

where we assumed that the flat-fielding error and the intrinsic intensities were the same. This intensity difference bias results in an electric field bias via the M^+ matrix defined in Equation 10. This means that the electric field bias scales linearly with the intensity in the dark-zone and the flat-fielding error.

This can also be seen in our simulations. In Figure 10 the achieved contrast as a function of flat-fielding error is shown for several photon fluxes. Worse AO performance leads to a larger influence of flat-field errors. If the tip-tilt variance is kept stable, then the flat-fielding error creates a bias in the electric field, which is stable up to photon-noise-limited performance. When the PSF starts to drift however, this bias changes per iteration, resulting in large variances in the closed-loop contrast. The latter effect can also be seen as a solution to the problem. Drift scanning modulates tip-tilt and takes many images during one iteration. These images are then shifted and added to form the image used for correction. This effectively averages the flat field over many pixels, decreasing its standard deviation by a factor $\sqrt{N_{\mathrm{pix}}}$.

5.4 Changing speckles

Under realistic circumstances, the beam may reflect off rotating surfaces, leading to rotating speckle fields. Figure 11 shows the result of a simulation of this scenario. We can clearly see that the closed-loop performance correlates with the rotation angle since the last iteration, indicating that the majority of the residual speckles in closed-loop operation were differential speckles compared to the previous iteration. Performing simulations with changing quasi-static speckles showed similar effects.

A predictive algorithm can significantly improve this performance by predicting the movement or drift of the changing speckle background and correcting for the time lag in our correction algorithm. However, even this simple algorithm improves on the open-loop contrast except at the points of highest rotation speed. Note that these simulations heavily depend on the speed of convergence of the algorithm. In reality, model errors in the correction algorithm limit convergence speed to > 20 iterations instead of ~ 3 iterations in these simulations. This makes predictive algorithms even more important in realistic circumstances.

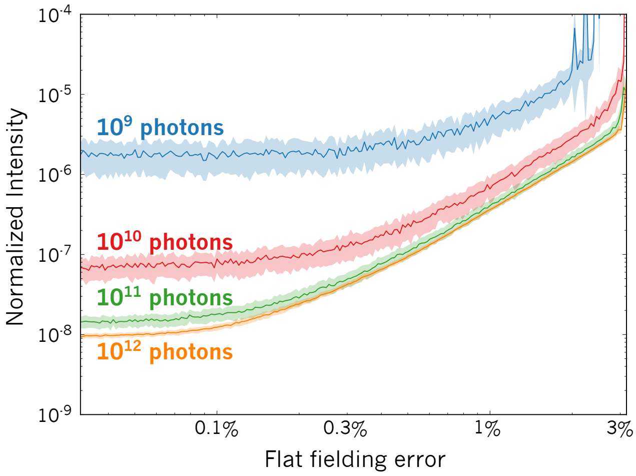

Figure 10. The achieved contrast as a function of flat-fielding error for different photon fluxes. For high photon fluxes, we are limited by the flat-fielding bias at large flat fielding errors. At small flat-fielding errors, we are limited by photon noise in the electric field estimation. An incoherent background from residual adaptive optics speckles of 10^{-3} to $10^{-3.5}$ depending on the position in the dark zone was adopted.

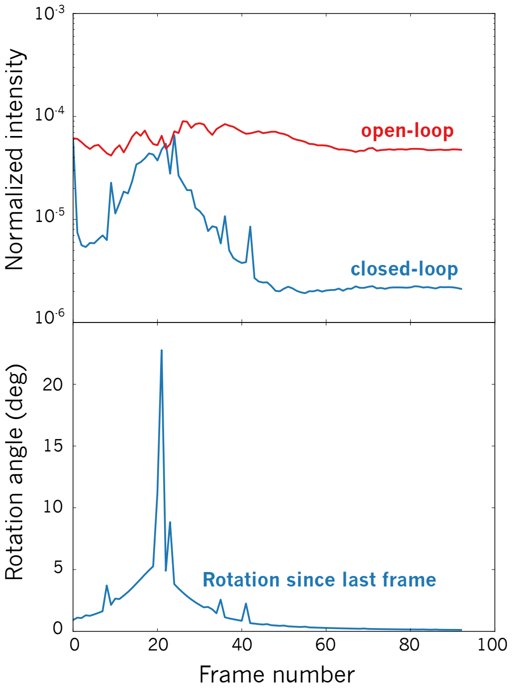

Figure 11. The achieved contrast as a function of frame number for a rotating speckle background. The dark zone in these simulations was larger than in the other simulations, which explains the worse contrast at very slow rotation rates. The closed-loop performance is highly correlated with the rotation speed, indicating that residual speckles are dominated by differential speckles between iterations. Spikes in the differential rotation angle indicate missing frames and therefore more time between corrections and higher differential speckles in that frame.

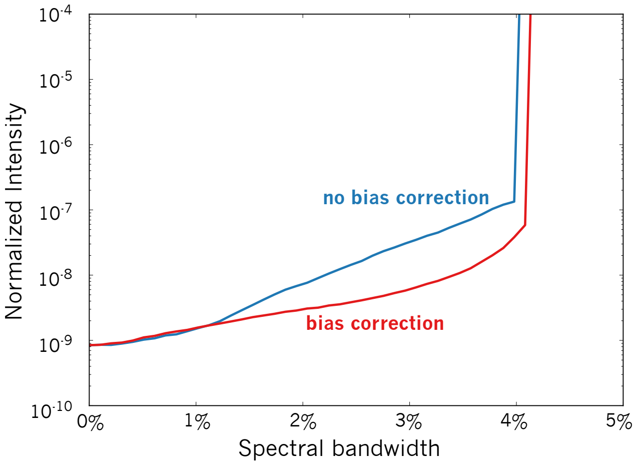

Figure 12. Increasing the spectral bandwidth decreases the achieved contrast. Applying a correction in the bias intensities improves the contrast, indicating that a significant part of the uncorrected contrast was due to an offset in the electric field. Both uncorrected and corrected simulations diverge when the spectral bandwidth reaches $\sim 4\%$, which corresponds to a $\sim 1.2\lambda/D$ shift of the speckles at the positions of the linear terms.

5.5 Spectral bandwidth

When using light with a non-zero spectral bandwidth, the PSF starts to smear in the radial direction; each of the PSF copies will smear in a different direction and with a different strength. If a speckle is smeared by more than $1\lambda/D$, then it starts to overlap with other speckles and the measurements will be incorrect. As our copies are much farther away from the star than the dark zone, we expect to be limited by a much smaller bandwidth than conventional DM-diversity based algorithms.

Our simulations only contain the wavelength scaling of the PSF. The origin of other chromatic aberrations are not well known,[15] and therefore hard to simulate. Broadband wavefront control tries to correct for these aberrations[16] and may indeed be complicated by chromatic electric field measurements.

In Figure 12 the achieved contrast is shown as a function of spectral bandwidth. The contrast steadily increases until divergence at $\sim 4\%$, which corresponds to $\sim 1.2\lambda/D$ at the copy positions. Part of this increase is a change in the bias electric field, which we can correct for by subtracting an intensity bias obtained by simulating the unaberrated system. This decreases the achieved contrast by a factor of ~ 10 at most, but still diverges at the same spectral bandwidth as before.

We can increase the spectral bandwidth by using a chromatic magnification that counteracts the wavelength scaling of the PSF. This puts every spot at the same position in the focal-plane, regardless of wavelength. These kind of optical systems were originally proposed for speckle interferometry,[17] but are suitable for our purpose.

6. CONCLUSIONS

In this paper we studied a new method for focal-plane electric field sensing. It is based on phase diversity, but instead of acquiring the probe images sequentially by changing the DM, a static holographic phase plate in the pupil-plane is used to divert some light from the science beam and use it for phase diversity. The deformable mirror is not used during the electric field estimation, and an unaltered science PSF is still available. This allows for estimation of the residual electric field during the actual observations and for closed-loop correction using electric field conjugation.

We generalized the phase diversity method to large probe amplitudes and explained the design of the holographic phase plate. We used numerical simulations to show open-loop and closed-loop performance with realistic aberrations. These simulations show that this method behaves as expected under photon noise, is resilient to fast atmospheric speckles up to the fundamental limit, and can perform under realistic flat-fielding errors. We have shown that we are able to correct slowly changing speckle backgrounds in closed loop and that the spectral bandwidth is limited to $\sim 2\%$, due to the wavelength scaling of the PSF. The latter can be extended by using a chromatic magnification that counteracts this scaling.

Future work will include implementing this method in the laboratory operating in open-loop and closed-loop. Afterwards, open-loop operation will be attempted on-sky to show resilience against a real fast speckle background.

REFERENCES

[1] Guyon, O., Pluzhnik, E. A., Kuchner, M. J., Collins, B., and Ridgway, S. T., "Theoretical Limits on Extrasolar Terrestrial Planet Detection with Coronagraphs," ApJS **167**, 81–99 (Nov. 2006).

[2] Kasdin, N. J., Vanderbei, R. J., Spergel, D. N., and Littman, M. G., "Extrasolar Planet Finding via Optimal Apodized-Pupil and Shaped-Pupil Coronagraphs," ApJ **582**, 1147–1161 (Jan. 2003).

[3] Guyon, O., "Phase-induced amplitude apodization of telescope pupils for extrasolar terrestrial planet imaging," A&A **404**, 379–387 (June 2003).

[4] Codona, J. L. and Angel, R., "Imaging Extrasolar Planets by Stellar Halo Suppression in Separately Corrected Color Bands," ApJ **604**, L117–L120 (Apr. 2004).

[5] Give'On, A., Kasdin, N. J., Vanderbei, R. J., and Avitzour, Y., "On representing and correcting wavefront errors in high-contrast imaging systems," *Journal of the Optical Society of America A* **23**, 1063–1073 (May 2006).

[6] Give'On, A., Belikov, R., Shaklan, S., and Kasdin, J., "Closed loop, DM diversity-based, wavefront correction algorithm for high contrast imaging systems," *Optics Express* **15**, 12338 (2007).

[7] Keller, C. U., "Novel instrument concepts for characterizing directly-imaged exoplanets," in [*Ground-based and Airborne Instrumentation for Astronomy VI*], Proc. SPIE **9908** (2016).

[8] Give'On, A., "A unified formailism for high contrast imaging correction algorithms," in [*Techniques and Instrumentation for Detection of Exoplanets IV*], Proc. SPIE **7440**, 74400D (Aug. 2009).

[9] Changhai, L., Fengjie, X., Shengyang, H., and Zongfu, J., "Performance analysis of multiplexed phase computer-generated hologram for modal wavefront sensing," Appl. Opt. **50**, 1631 (Apr. 2011).

[10] Wilby, M. J. and Keller, C. U., "The coronagraphic modal wavefront sensor: focal-plane sensing for high-contrast imaging," in [*Adaptive Optics Systems V*], Proc. SPIE **9909** (2016).

[11] Frazin, R. A., "Statistical framework for the utilization of simultaneous pupil plane and focal plane telemetry for exoplanet imaging I Accounting for aberrations in multiple planes," *Journal of the Optical Society of America A* **33**, 712 (Apr. 2016).

[12] Groff, T. D. and Jeremy Kasdin, N., "Kalman filtering techniques for focal plane electric field estimation," *Journal of the Optical Society of America A* **30**, 128 (Jan. 2013).

[13] Groff, T. D., Eldorado Riggs, A. J., Kern, B., and Jeremy Kasdin, N., "Methods and limitations of focal plane sensing, estimation, and control in high-contrast imaging," *Journal of Astronomical Telescopes, Instruments, and Systems* **2**, 011009 (Jan. 2016).

[14] Soummer, R., Ferrari, A., Aime, C., and Jolissaint, L., "Speckle Noise and Dynamic Range in Coronagraphic Images," ApJ **669**, 642–656 (Nov. 2007).

[15] Belikov, R., Give'on, A., Kern, B., Cady, E., Carr, M., Shaklan, S., Balasubramanian, K., White, V., Echternach, P., Dickie, M., Trauger, J., Kuhnert, A., and Kasdin, N. J., "Demonstration of high contrast in 10% broadband light with the shaped pupil coronagraph," in [*Techniques and Instrumentation for Detection of Exoplanets III*], Proc. SPIE **6693**, 66930Y (Sept. 2007).

[16] Groff, T. D., Kasdin, N. J., Carlotti, A., and Riggs, A. J. E., "Broadband focal plane wavefront control of amplitude and phase aberrations," in [*Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*], Proc. SPIE **8442**, 84420C (Sept. 2012).

[17] Wynne, C. G., "Extending the bandwidth of speckle interferometry.," *Optics Communications* **28**, 21–25 (1979).