Supplemental Document

Information transfer as a framework for optimized phase imaging: supplement

STEWART KOPPELL* AND MARK KASEVICH

Physics Department, Stanford University, 382 Via Pueblo Mall, Stanford, California 94305, USA
**Corresponding author: skoppell@stanford.edu*

This supplement published with The Optical Society on 8 April 2021 by The Authors under the terms of the Creative Commons Attribution 4.0 License in the format provided by the authors and unedited. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Supplement DOI: https://doi.org/10.6084/m9.figshare.13910009

Parent Article DOI: https://doi.org/10.1364/OPTICA.412129

Information Transfer as a Framework for Optimized Phase Imaging: Supplemental Document

1. THE PHASE GRATING BASIS

We want a spatial frequency parameterization for $\phi(\vec{r}_k)$. For simplicity we will assume that the set of coordinates $\vec{r}_k$ form a regular, square array (so n is square). Since ϕ is strictly real, we should not use a discrete Fourier basis. Instead, we could use a 2D discrete cosine or sine transform. But these contain 'half frequency' elements. It will be more convenient to parameterize ϕ in a basis where every element has a simple Fourier representation. Inspired by the real discrete Fourier transform[1], we use

$$v^{(a)}(\vec{r}_k) = \begin{cases} \sqrt{\frac{1}{n}} & a = 0 \\ c_a \cos(2\pi \vec{q}_a \cdot \vec{r}_k) & a \in A \\ c_a \sin(2\pi \vec{q}_a \cdot \vec{r}_k) & a \notin A \end{cases} \tag{S1}$$

Let $L = \sqrt{n}$. Each value of a can be uniquely expressed as $a = a_0 + La_1$, with $0 \leq a_0, a_1 \leq L-1$. The spatial frequencies are labeled so that $\vec{q}_a = [q_{x;a_0}, q_{y;a_1}]$.

- L **odd**:

$$c_a = \begin{cases} \sqrt{1/n} & a = 0 \\ \sqrt{2/n} & a > 0 \end{cases} \tag{S2}$$

$$A = \{a | 0 < a_1 \leq (L-1)/2 \quad \text{or} \quad (a_0 \leq (L-1)/2 \text{ and } a_1 = 0)\}$$

- L **even**:

$$c_a = \begin{cases} \sqrt{1/n} & [a_0, a_1] \in \{[0,0], [0, L/2], [L/2, 0], [L/2, L/2]\} \\ \sqrt{2/n} & \text{else} \end{cases} \tag{S3}$$

$$A = \{a | a_0, a_1 \leq L/2 \quad \text{or} \quad 0 < a_1 \leq L/2 - 1\}$$

These conditions are shown schematically in Fig. S1.

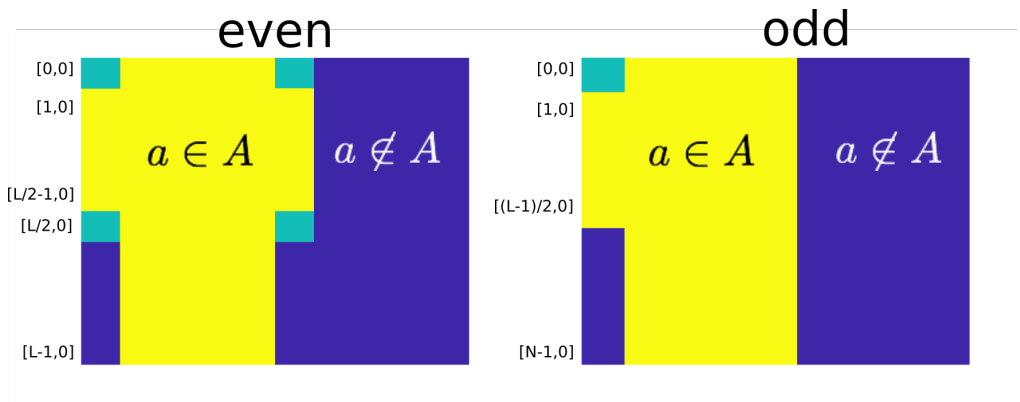

Fig. S1. Schematic of the set A and the values of c_a. The yellow and turquoise regions comprise the set A. $c_a = \sqrt{1/n}$ in the turquoise region and $\sqrt{2/n}$ elsewhere.

2. THE INFORMATION TRANSFER FUNCTION

The definition of contrast most often used to calculate the CTF is the Michelson contrast

$$C^M = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{S4}$$

Suppose the sample is a phase grating with spatial frequency q and amplitude $\delta\theta_q \ll 1$. The CTF can be written

$$\mathcal{C}(q) = C^M / \delta\theta_q \tag{S5}$$

As an example, suppose we use a plane wave probe and the transfer function T applies a phase shift $\mu(q)$ to each Fourier component q. Then $I = |T(\psi)|^2 = |\mathcal{F}\{\mathcal{F}\{\psi\}e^{i\mu}\}|^2$ where $\mathcal{F}$ represents the action of a Fourier lens. Using the WPOA, we find the CTF is

$$\mathcal{C}(q) = 2|\sin(\mu(q) - \mu(q=0)| \tag{S6}$$

In some formulations, the CTF is a signed quantity with negative contrast indicating dark fringes. The CTF is also often normalized so that $|\mathcal{C}(q)| \leq 1$.

Consider a phase object built from a superposition of phase gratings with amplitudes $\Theta = [\theta_1, \theta_2, ..., \theta_n]$. If we perturb one of these amplitudes by a small amount $\delta\theta$, how much contrast will that perturbation generate? The answer using the Michelson contrast depends on the intensities measured by only two detector pixels (at the locations of $I_{\max}$ and $I_{\min}$). Clearly this summary statistic is too coarse-grained to capture the full effect of the perturbation. As an alternative, we can define the CTF using the root-mean square Weber contrast

$$\mathcal{C}(q) = \sqrt{\sum_j \left(\frac{1}{\theta_q} C_j^W(q)\right)^2} \tag{S7}$$

where C^W is the Weber contrast

$$C_j^W(q) = \frac{I_j - I_b; j}{I_b; j} \tag{S8}$$

where $I_b = I|_{\delta\theta_q=0}$. These definitions of the CTF are entirely equivalent in the WPOA. Now consider the square of the Weber CTF for small perturbations $\delta\theta_q \to 0$:

$$\lim_{\delta\theta_q \to 0} \mathcal{C}^2(\theta_q) = \sum_j \left(\frac{1}{\theta_q} \frac{I - I_b}{I_b}\right)^2 \tag{S9}$$

$$= \sum_j \left(\frac{1}{I_b} \frac{\partial}{\partial\theta_q} I_b\right)^2 \tag{S10}$$

$$= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_q} \log(I_b)\right)^2\right] = \mathcal{I}_q \tag{S11}$$

where $\mathbb{E}$ is the expectation value. The final expression is the definition of the FI for parameter θ_q. We can also show this equivalence by calculating the diagonal elements of the FIM for a WPO using the phase grating basis and unitary transfer function T:

$$\mathcal{I}_{a,a}(\Theta = 0, T) = \sum_j \frac{4}{|T(\psi)|_j^2} \Re\left\{\overline{T(\psi)}_j \partial_a T(\psi)_j\right\}^2 \tag{S12}$$

$$= 4/n \sin^2(\mu(\vec{q}_a) - \mu(\vec{q}_0)) \tag{S13}$$

which (apart from the normalization factor $1/n$) is the square of the CTF. However, we cannot interpret the diagonal of the FIM as a transfer function outside of the WPOA, as the off-diagonal elements may be important. Nevertheless, it will be useful to formulate a transfer function based on the FIM to help visualize the properties of a particular measurement. We define the information transfer function (ITF) for a measurement T as the ratio of the maximum variance reduction for parameter θ_a achievable by T (as determined by the van Trees bound) to the maximum variance reduction for parameter θ_a allowed for any measurement (as determined by the GQFI). In general, we write the function as $\text{ITF}(|\vec{q}_a|; \Theta, T)$ assuming Θ is expressed in the phase grating basis and that T and λ respect radial symmetry around $\vec{q} = 0$. Explicitly, the ITF is

$$\mathrm{ITF}(|\vec{q}_a|;\Theta,T) = \frac{\sigma_a^2(\lambda) - (\mathcal{V}^{-1})_{a,a}(\lambda,T)}{\sigma_a^2(\lambda) - (\mathcal{Z}^{-1})_{a,a}(\lambda)} \tag{S14}$$

For large n, $\langle I_a(\Theta,T)\rangle_\lambda \leq \mathcal{J}_a = 4/n \ll 1$ and if $\sigma_a^2(\lambda) \ll n/4$, we can expand $\mathcal{V}^{-1}$ and $\mathcal{Z}^{-1}$ in powers of $\sigma_a^2(\lambda)\langle I\rangle_\lambda$ and $\sigma_a^2(\lambda)J$, respectively:

$$(\mathcal{V}^{-1})_{a,a}(\lambda,T)/\sigma_a^2(\lambda) = 1 - \sigma_a^2(\lambda)\langle \mathcal{I}_a(\Theta,T)\rangle_\lambda + \mathcal{O}\left(\sigma_a^4(\lambda)\langle \mathcal{I}_a(\Theta,T)\rangle_\lambda^2\right) \tag{S15}$$

$$(\mathcal{Z}^{-1})_{a,a}(\lambda)/\sigma_a^2(\lambda) = 1 - \sigma_a^2(\lambda)\mathcal{J}_a + \mathcal{O}\left(\sigma_a^4(\lambda)\mathcal{J}_a^2\right) \tag{S16}$$

and the ITF becomes

$$\mathrm{ITF}(|\vec{q}_a|;\Theta,T) \sim \langle \mathcal{I}_a(\Theta,T)\rangle / \mathcal{J}_a \tag{S17}$$

This approximation is accurate, for example, in the WPOA, in which case the ITF is the square of the CTF as shown above.

3. FISHER INFORMATION FOR WPO IN A STRONG BACKGROUND

The standard formulation of the cost function measures the expected average variance. An optimized measurement will prioritize sensitivity to the parameters with the largest prior variances. Suppose the sample consists of a WPO (the foreground) embedded in a strongly scattering, unknown background. Let parameter vector $\Theta_f = [\theta_{f;0}, \theta_{f;1}, ...]$ with prior distribution $\lambda_f(\Theta_f)$ describe the foreground and $\Theta_b = [\theta_{b;0}, \theta_{b;1}, ...]$ with prior distribution $\lambda_b(\Theta_b)$ describe the background, so the total transmission function is

$$\Phi(\Theta_f)_k \Phi(\Theta_b)_k = \exp\left(i\sum_{a=0}^{n-1}\left(\theta_{f;a} + \theta_{b;a}\right)v_k^{(a)}\right) \tag{S18}$$

We cannot separately measure $\theta_{f;a}$ and $\theta_{b;a}$, but we can adjust the cost function to specifically reward reduction of the foreground variance.

Let λ_{tot} be the prior distribution for $\Theta_{\mathrm{tot}} = [\Theta_f, \Theta_b]$. The covariance matrix for the estimator of the combined parameter vector is constrained by the van Trees bound

$$\left\langle \Sigma_{\hat{\Theta}_{\mathrm{tot}}} \right\rangle_{\lambda_{\mathrm{tot}}} \geq \frac{1}{\mathcal{I}(\lambda_{\mathrm{tot}}) + N\langle \mathcal{I}(\Theta_{\mathrm{tot}})\rangle_{\lambda_{\mathrm{tot}}}} \tag{S19}$$

The cost function

$$\langle C\rangle_{\lambda_{\mathrm{tot}}} = \mathrm{Tr}\left(W_{\mathrm{tot}}\left\langle \Sigma_{\hat{\Theta}_{\mathrm{tot}}}\right\rangle_{\lambda_{\mathrm{tot}}}\right) \tag{S20}$$

is equivalent to the standard cost function when

$$W_{\mathrm{tot}} = \frac{1}{2}\begin{pmatrix} W & 0 \\ 0 & W \end{pmatrix} \tag{S21}$$

but can be specialized to foreground variance reduction using

$$W_{\mathrm{tot}} = \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix} \tag{S22}$$

We can write $\langle \mathcal{I}(\Theta_{\mathrm{tot}})\rangle_{\lambda_{\mathrm{tot}}}$ as a 2×2 block diagonal matrix, where each block is $\langle \mathcal{I}(\Theta)\rangle_{\lambda_{\mathrm{tot}}}$. We will also write $\mathcal{I}(\lambda_{\mathrm{tot}})$ in block form

$$\mathcal{I}(\lambda_{\mathrm{tot}}) = \begin{pmatrix} \mathcal{I}(\lambda_{\mathrm{tot}})_{11} & \mathcal{I}(\lambda_{\mathrm{tot}})_{12} \\ \mathcal{I}(\lambda_{\mathrm{tot}})_{12} & \mathcal{I}(\lambda_{\mathrm{tot}})_{22} \end{pmatrix} \tag{S23}$$

so that the right hand size in Eq. S19 is

$$\begin{pmatrix} \mathcal{I}(\lambda_{\mathrm{tot}})_{11} + N\langle \mathcal{I}(\Theta)\rangle_{\lambda_{\mathrm{tot}}} & \mathcal{I}(\lambda_{\mathrm{tot}})_{12} + N\langle \mathcal{I}(\Theta)\rangle_{\lambda_{\mathrm{tot}}} \\ \mathcal{I}(\lambda_{\mathrm{tot}})_{12} + N\langle \mathcal{I}(\Theta)\rangle_{\lambda_{\mathrm{tot}}} & \mathcal{I}(\lambda_{\mathrm{tot}})_{22} + N\langle \mathcal{I}(\Theta)\rangle_{\lambda_{\mathrm{tot}}} \end{pmatrix}^{-1} \tag{S24}$$

since the weight matrix has three zero quadrants, we need only calculate the upper left quadrant of this matrix inverse to find

$$\left\langle C' \right\rangle_{\lambda_{\rm tot}} \geq \mathrm{Tr}\Big(W\Big(\mathcal{I}(\lambda_{\rm tot})_{11} + N \left\langle \mathcal{I}(\Theta) \right\rangle_{\lambda_{\rm tot}} \tag{S25}$$

$$- \left(\mathcal{I}(\lambda_{\rm tot})_{12} + N \left\langle \mathcal{I}(\Theta) \right\rangle_{\lambda_{\rm tot}} \right) \left(\mathcal{I}(\lambda_{\rm tot})_{22} + N \left\langle \mathcal{I}(\Theta) \right\rangle_{\lambda_{\rm tot}} \right)^{-1} \left(\mathcal{I}(\lambda_{\rm tot})_{12} + N \left\langle I(\Theta) \right\rangle_{\lambda_{\rm tot}} \right) \Big)^{-1} \Big) \tag{S26}$$

If we assume the prior distributions for Θ_f and Θ_b are independent so $\lambda_{\rm tot} = \lambda_f(\Theta_f)\lambda_b(\Theta_b)$, then $\mathcal{I}(\lambda_{\rm tot})_{11} = \mathcal{I}(\lambda_f)$, $\mathcal{I}(\lambda_{\rm tot})_{12} = 0$, $\mathcal{I}(\lambda_{\rm tot})_{22} = \mathcal{I}(\lambda_b)$, and

$$\left\langle C_f \right\rangle_{\lambda_{\rm tot}} \geq \mathrm{Tr}\left(W \frac{1}{\mathcal{I}(\lambda_f) + N \left\langle I(\Theta) \right\rangle_{\lambda_{\rm tot}} - N^2 \left\langle \mathcal{I}(\Theta) \right\rangle_{\lambda_{\rm tot}} \left(\mathcal{I}(\lambda_b) + N \left\langle \mathcal{I}(\Theta) \right\rangle_{\lambda_{\rm tot}} \right)^{-1} \left\langle \mathcal{I}(\Theta) \right\rangle_{\lambda_{\rm tot}}} \right) \tag{S27}$$

As $\mathcal{I}(\lambda_b) \to \infty$ (meaning the background is known and $\lambda_{\rm tot} \to \lambda_f$), this cost function approaches the standard van Trees bound for λ_f. When $\mathcal{I}(\lambda_b)$ is small (i.e. $\mathcal{I}(\lambda_b) \ll N \left\langle \mathcal{I}(\Theta) \right\rangle_{\lambda_{\rm tot}}$) then

$$\left\langle C_f \right\rangle_{\lambda_{\rm tot}} \geq \mathrm{Tr}\left(W \frac{1}{\mathcal{I}(\lambda_f) + \mathcal{I}(\lambda_b) \left(1 - \frac{1}{N} \left\langle I(\Theta) \right\rangle^{-1}_{\lambda_{\rm tot}} \mathcal{I}(\lambda_b) \right)} \right) \tag{S28}$$

As an example, suppose λ_f is independently and identically distributed for each of the parameters so that the prior information matrix is $\mathcal{I}(\lambda_f) = \frac{1}{\sigma^2}\mathbb{I}$. Also suppose $W = \mathbb{I}$ and $\lambda_b(\Theta_b) = \prod_{a=0}^{n-1} \lambda_a(\theta_a)$ where each λ_a is normal with zero mean and variance $\sigma_a^2 \gg N\sigma^2$.

$$\left\langle C' \right\rangle_{\lambda_{\rm tot}} \gtrsim (n-1)\sigma^2 + \frac{\sigma^4}{N} \sum_{a=0}^{n-1} \left(\sigma_a^{-4} \left\langle \mathcal{I}_a(\Theta) \right\rangle_{\lambda_{\rm tot}} - \sigma_a^{-2} \right) \tag{S29}$$

This cost function has a strong preference for measuring parameters with small σ_a, where the foreground is more 'visible' despite the background. For comparison, if we use the weighting in Eq. S21 we get the standard cost

$$\left\langle C \right\rangle_{\lambda_{\rm tot}} \gtrsim \sum_{a=1}^{n-1} \frac{1}{\sigma^2 + \sigma_a^{-2} + N \left\langle \mathcal{I}_a(\Theta) \right\rangle_{\lambda_{\rm tot}}} \tag{S30}$$

which gives priority to increasing $\left\langle \mathcal{I}_a(\Theta) \right\rangle_{\lambda_{\rm tot}}$ for parameters with large σ_a.

4. PROJECTIVE MEASUREMENTS IN THE BAYESIAN REGIME

In multi-phase estimation with $n = 2$ phases, the phase difference $\phi_2 - \phi_1$ can be optimally measured without a reference channel or any prior knowledge of the phases using a 50-50 beam splitter. For $n > 2$ pixels, the phase differences between neighboring channels can be measured using a series of beam splitters and $2n - 1$ detectors. This measurement is impractical for phase imaging, where n is large and space is limited. Instead, we will consider only projective measurements which can be represented by a unitary matrix T of rank n. Here we give an informal argument that projective measurements generally cannot achieve the QFIM in the Bayesian regime.

Using Eq. 5 we can write the FI for parameter θ_a,

$$\mathcal{I}_a(\Theta, T) = 4 \sum_j |T(\partial_a \psi)_j|^2 \sin^2(\gamma_j^{(a)} - \gamma_j) \tag{S31}$$

where $\gamma_j = \arg\{T(\psi)_j\}$ and $\gamma_j^{(a)} = \arg\{T(v^{(a)}\psi)_j\}$. We can easily see that $\mathcal{I}_a(\Theta, T) = \mathcal{J}_a = 4/n$ if and only if $\sin^2(\gamma_j^{(a)} - \gamma_j) = 1$ for all j where $|T(\partial_a\psi)_j|^2 > 0$. This is possible only if $\Lambda = |U\psi|^2$ has no overlap with $\Lambda^{(a)} = |U\partial_a\psi|^2$, and the QFIM can only be achieved if this condition is met for all values of a. Λ can be thought of as the reference component and $\Lambda^{(a)}$ as a signal component. Achieving the QFIM requires the reference to be completely isolated from the signal. It takes $n - 1$ channels to carry information about $n - 1$ independent parameters. This limits Λ to a single channel. In the Bayesian regime, we will not have sufficient prior information to find a measurement basis where Λ occupies a single channel.

5. GENERALIZED ZPC IN THE BAYESIAN REGIME

Here we calculate the FI for GZPC for a particular, simple prior distribution λ. The effect of the ZPC optics is to add a phase shift μ to the zero-frequency component (mean) of the wavefunction exiting the sample. The transfer function can be written

$$T(\psi_j) = \psi_j + (e^{i\mu} - 1)\langle\psi\rangle \tag{S32}$$

We will assume the probe amplitude is uniform and the sample is a pure phase object, so $|\psi_j|^2 = 1/n$. The sample phase is described by n-1 parameters in the vector Θ which weigh phase grating basis elements $v^{(a)}$ (we exclude θ_0, which determines the average phase thickness). If the prior distribution on Θ is $\lambda(\Theta)$, then the expected FI is

$$\left\langle \mathcal{I}_{a,b}(\Theta)\right\rangle_\lambda = \int d^{n-1}\lambda(\Theta)\sum_{j=1}^{n}\frac{1}{I_j}\partial_a I_j \partial_b I_j \tag{S33}$$

$$= \frac{16}{n}\int d^{n-1}\lambda(\Theta)\sum_{j=1}^{n}\frac{\sin^2(\mu/2)\Lambda_0|v_j^{(a)}||v_j^{(b)}|\cos^2(\phi_j-\phi_0-\mu/2)}{1+4\sin^2(\mu/2)\Lambda_0+4\sin(\mu/2)\sqrt{\Lambda_0}\sin(\phi_j-\phi_0-\mu/2)} \tag{S34}$$

where $\Lambda_0 = |\langle\psi\rangle|^2$ and $\phi_0 = \arg\langle\psi\rangle$. Note ϕ_j and ϕ_0 depend on Θ. If we assume that λ is independently and identically distributed for each parameter, then we can also write identical and independent distributions $\lambda(\phi_j) = \lambda(\phi)$ for each ϕ_j. For large n, the distribution for ϕ_0 is narrow (with variance $\sim \frac{1}{n}$) around a mean which we will assume, without loss of generality, is zero. Then

$$\left\langle \mathcal{I}_{a,b}(\Theta)\right\rangle_\lambda = \delta_{a,b}\frac{16}{n}\sin^2(\mu/2)\Lambda_0\int d\phi\lambda(\phi)\frac{\cos^2(\phi-\mu/2)}{1+4\sin^2(\mu/2)\Lambda_0+4\sin(\mu/2)\sqrt{\Lambda_0}\sin(\phi-\mu/2)} \tag{S35}$$

We can now optimize the Zernike phase, μ, for a particular distribution $\lambda(\phi)$. Suppose $\lambda(\phi) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\phi^2/2\sigma^2}$. The ideal choice of μ depends on $\Lambda_0 = e^{-\sigma^2}$:

$$\mu = \begin{cases} \pm\pi/2 & \Lambda_0 \geq \frac{1}{2} \\ \pm 2\arcsin\left(\frac{1}{2\sqrt{\Lambda_0}}\right) & \frac{1}{2} > \Lambda_0 > \frac{1}{4} \end{cases} \tag{S36}$$

The expected FI for $\mu = \pi/2, 2\pi/3$, and π are shown in Fig. S2. The figure also shows phasor diagrams which may provide some intuition for the optimal values of μ.

6. DETAILS OF NUMERICAL CALCULATIONS

.

In order to optimize GCPI, the value of μ and the membership of G must be jointly optimized based on the prior distribution λ. When λ is induced by an expected intensity pattern Λ we sample from λ using the Girchberg-Saxton algorithm with a uniform random initial phase distribution. In order to determine $\mathcal{I}(\lambda)$, we estimate the covariance matrix for λ, Σ_λ, then set $\mathcal{I}(\lambda) = \Sigma_\lambda^{-1}$.

The number of possible sets of G is combiniatorially large. In order to reduce the complexity of optimizing G, we estimate the value V_g of including $g \in G$, then set $G = \{g|V_g \geq V_*\}$ and optimize the threshold value V_*. The estimated value will depend on the cost function. To minimize the weighted average of the expected variance using cost function from Eq. 7, we use

$$V_g = \frac{\Lambda_g}{\sum_a W_{a,a}\Delta_{\max}^{(a)}\left\langle\Lambda_g^{(a)}\right\rangle_\lambda} \tag{S37}$$

where $\Delta_{\max}^{(a)}$ is the maximum variance reduction allowed by the GQFI for θ_a. The denominator is the weighted average of the signal components expected to be carried by eigenvector g, and

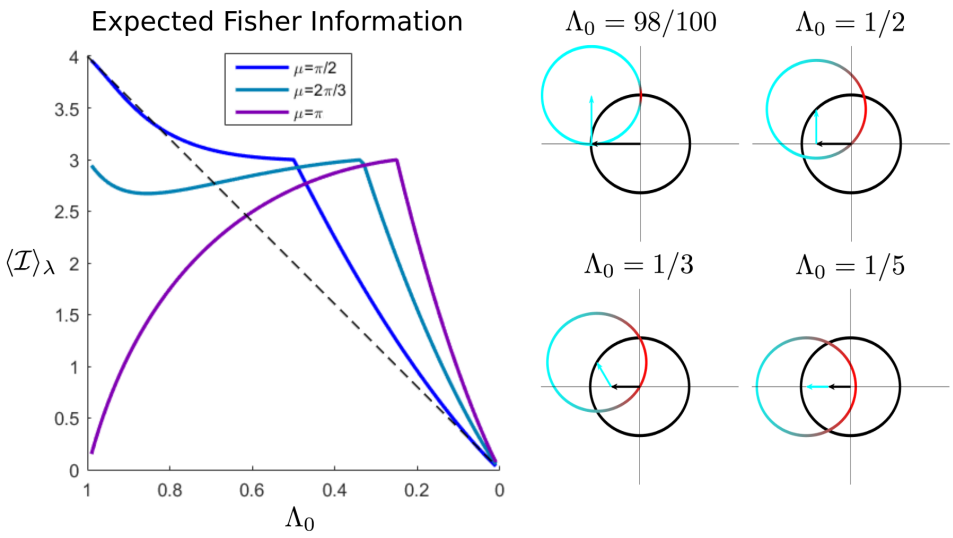

Fig. S2. Left: Expected Fisher Information $\langle \mathcal{I} \rangle_\lambda$ for Zernike phase contrast with a prior λ which is an independent Gaussian distribution with variance σ^2 for each phase. The maximum FI and the ideal Zernike phase shift, μ, depend on the unscattered intensity $\Lambda_0 = e^{-\sigma^2}$. The dashed black line is $4\Lambda_0$, which is a good approximation for $\langle \mathcal{I} \rangle_\lambda$ when $\Lambda_0 > 0.8$. Right: Phasor diagrams which show the the action of the transfer optics on the exit wavefunction (represented by the black unit circle). The cyan circle represents the possible values of the wavefunction at the detector, and the red portion represents the probability distribution of the wavefunction. The black and cyan vectors have length Λ_0 and relative angle μ. For $.25 < \Lambda < .5$, the optimal μ causes the cyan circle to pass through the the origin. For $\Lambda_0 > .5$ and $\Lambda_0 < 0.25$, the optimal values for μ are $\pi/2$ and π, respectively.

estimates the opportunity cost of losing sensitivity to g. When optimizing for foreground variance reduction using the cost function from Eq. 11, we use

$$V_g = \frac{\Lambda_g}{\sum_a W_{a,a} \left(\Delta_{\max}^{(a)}\right)^{-1} \left\langle \Lambda_g^{(a)} \right\rangle_\lambda} \tag{S38}$$

which has higher value when $\Delta_{\max}^{(a)}$, the potential reduction in the background variance, is smaller. In many cases, especially when Λ_g decreases monotonically with g, the same value ranking is obtained simply using $V_g = \Lambda_g$. The optimization proceeds by alternating between minimizing the cost with respect to μ use Matlab's fminbnd (with $\pi/2 < \mu < \pi$), and then minimizing with respect to V_*.

7. THE FI FROM APERTURE LOSSES IS NEGLIGIBLE

In this section we justify neglecting $\mathcal{I}_{\text{Ap}}$, the FI available from a measurement of the total intensity missing at the detector. Rather than presenting a rigorous proof, we give an intuitive argument which we support with numerical calculations.

Consider the FI about θ_a available from measuring the intensities at two locations on the detector, $\mathcal{I}_{1,2;a}$, compared to the FI available from measuring only the sum of the intensities, $\mathcal{I}_{1+2;a}$.

$$\mathcal{I}_{1,2;a} = \frac{(\partial_a I_1)^2}{I_1} + \frac{(\partial_a I_2)^2}{I_2} \geq \frac{(\partial_a I_1 + \partial_a I_2)^2}{I_1 + I_2} = \mathcal{I}_{1+2;a} \tag{S39}$$

The equality occurs when $\partial_a I_1$ and $\partial_a I_2$ are perfectly correlated so that $\langle \partial_a I_1 \partial_a I_2 \rangle_\lambda = \left\langle (\partial_a I_1)^2 \right\rangle_\lambda$. When the two are perfectly anti-correlated, the measurement yields no information. When there is no correlation, $\langle \partial_a I_1 \partial_a I_2 \rangle = 0$ and when $I_1 \approx I_2$, $\mathcal{I}_{1,2;a} = 2\mathcal{I}_{1+2;a}$. This suggests that, in general, the maximum information from a single degree of freedom in the detection scheme is limited to the maximum information available about any single degree of freedom in the sample. Since quantum limit for in-line phase imaging is $\mathcal{I}_a \leq 4/n$, we expect $\mathcal{I}_{\text{Ap}} \leq 4/n$. Indeed, our numerical calculations are consistent with $\text{Tr}\left(\mathcal{I}_{Ap}\right) = \mathcal{O}(1/n)$. Compared to the total information, $\text{Tr}\left(\mathcal{J}\right) = \mathcal{O}(1)$, $\text{Tr}\left(\mathcal{I}_{Ap}\right)$ is negligible for large n.

In figure S3 we show ITFs for a sample known to have a Lorentizan diffraction pattern with unscattered intensity Λ_0. The curve labeled **BF** is the ITF for bright-field imaging and the curve labeled **Ap** is the ITF for a measurement of the total intensity missinig at the detector.

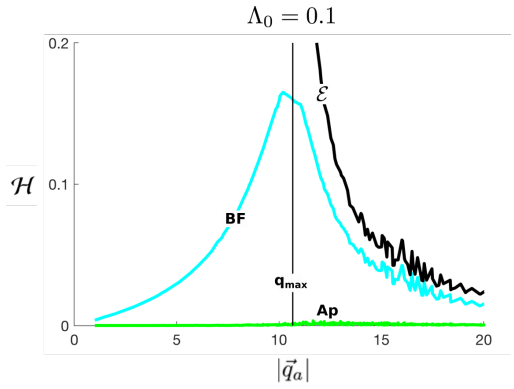

Fig. S3. Information Transfer Function (ITF) for a sample with a prior distribution induced by a known Lorentzian diffraction pattern with $\Lambda_0 = 0.1$ intensity in the zero-order beam. The horizontal axis is the magnitude of the spatial frequency in the sample phase. The black vertical line at $|\vec{q}_a| = q_{\max}$ marks the largest spatial frequency in the exit wavefunction allowed through the Fourier plane aperture. The black envelope labeled $\mathcal{E}$ is the information limit set by the aperture. The other two curves are the ITF for bright field imaging (BF) and for the information gained by measuring the total intensity absorbed by the aperture (Ap).

REFERENCES

1. O. K. Ersoy, "Real discrete fourier transform," IEEE Trans. Acoust. Speech Signal Process. **33**, 880–882 (1985).