Inverse Scattering via Transmission Matrices: Broadband Illumination and Fast Phase Retrieval Algorithms

Manoj Kumar Sharma*, Christopher A. Metzler*, Sudarshan Nagesh, Richard G. Baraniuk, Oliver Cossairt, Ashok Veeraraghavan

***Abstract*—When a narrowband coherent wavefront passes through or reflects off of a scattering medium, the input and output relationship of the incident field is linear and so can be described by a transmission matrix (TM). If the TM for a given scattering medium is known, one can computationally "invert" the scattering process and image through the medium. In this work, we investigate the effect of broadband illumination, i.e., what happens when the wavefront is only partially coherent? Can one still measure a TM and "invert" the scattering?**

To accomplish this task, we measure TMs using the double phase retrieval technique, a method which uses phase retrieval algorithms to avoid difficult-to-capture interferometric measurements. Generally, using the double phase retrieval method requires performing massive amounts of computation. We alleviate this burden by developing a fast, GPU-accelerated algorithm, prVAMP, which lets us reconstruct $256^2 \times 64^2$ TMs in under five hours.

After reconstructing several TMs using this method, we find that, as expected, reducing the coherence of the illumination significantly restricts our ability to invert the scattering process. Moreover, we find that past a certain bandwidth an incoherent, intensity-based scattering model better describes the scattering process and is easier to invert.

***Index Terms*—Transmission Matrix, Phase Retrieval, Scattering.**

I. Introduction

At first glance, imaging through multiple-scattering media seems like an impossible task. Any light incident on the media will undergo multiple reflections. Thus, if we illuminate an object with coherent light, the resulting wavefront will constructively and destructively interfere with itself, and a speckle pattern will be produced on the far side of the scatterer. This speckle pattern generally bears no resemblance to the original image.

A. Inverse Scattering is the Holy Grail of Imaging

Despite the challenge inherent to imaging through scattering media, it is a crucial problem within the optics community. Imaging through scattering media is fundamental to numerous applications including imaging through biological tissues, long-distance imaging through smog, fog, etc., imaging through the atmosphere, and more. Accordingly, numerous attempts have been made to solve the inverse scattering problem. We list some of them next.

B. Imaging Through Scattering Media Today

Confocal microscopy [1], [2] mitigates the scattering effect by using a pinhole to ignore scattered, out-of-focus light. The performance of this method is limited to few attenuation lengths of scatterings and requires very slow, scanned image acquisitions.

Time and/or coherence-gating techniques [3], [4], [5], [6], [7], [8] rely upon the fact that a scattered photon is unlikely to arrive at the same time as a ballistic photon; a photon that went to the target and back without scattering. As a result, one can use pulsed illumination and temporal/coherence gating to separate the ballistic photons from a region of interest. Gating is well-suited to rapidly changing scatterers, like fog, and as it offers rapid acquisition times and does not require the scattering process to be stable over time. However, because these systems only measure ballistic photons, they require powerful and expensive pulsed lasers to image through thick scatterers.

Multi-slice light propagation [9], [10], [11] is another method to image through scattering media. As the name suggests, the multi-slice light-propagation method models a scattering material as a series of 2D scattering slices between which light propagates. Rather than blindly mapping inputs to outputs, the multi-slice propagation method learns a composition of linear transformations. This method is uniquely suited to perform 3D reconstructions; moreover, it provides information about the actual structure of the scattering material. However, the multi-slice approach does not model reflections within the scatterer, and it is unclear how well this method works with thick scattering materials.

Recently phase retrieval techniques are used for wavefront sensing through thin scattring media[[12]] and to image rough distant objects [13], [14], however, when dealing with thin scattering media, single-shot imaging through scattering media is possible using memory-effect based techniques [15], [16]. These methods assume that the scattering medium preserves the strong angular correlation of the speckle patterns; which implies that, with temporally coherent and spatially incoherent illumination, the scattering process follows a convolution model with a speckle-like point spread function. From there

M. Sharma, C. Metzler, R. Baraniuk, and A. Veeraraghavan are with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, 77005 USA e-mail: imaging.mks@gmail.com.

S. Nagesh and O. Cossairt are with Northwestern University, Evanston, IL, 60208 USA.

*M. Sharma and C. Metzler contributed equally.

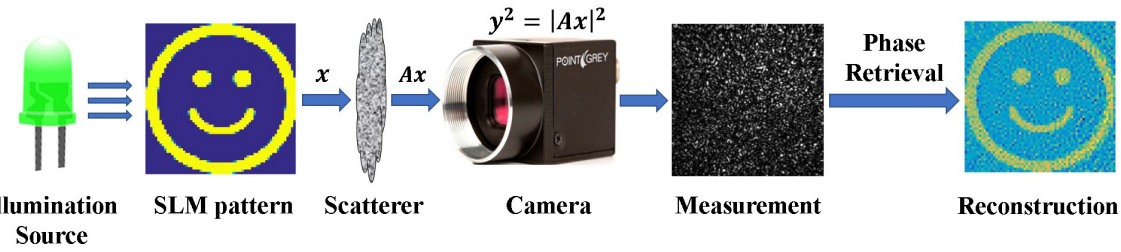

Fig. 1. A sketch of our experimental setup. An illumination source is modulated by an SLM to form a smiley face, the light field scatters and produces a speckle pattern, the intensity of the speckle pattern is measured by a camera, and we use phase retrieval algorithms to reconstruct the smiley face from the speckle pattern.

one can invert the scattering process by estimating the target's auto-correlation function and then performing phase retrieval. This method fails when the scattering medium becomes too thick for the strong memory effect to hold.

C. Transmission Matrices

The gold standard to image through very *thick* and *stable* scattering media is inverse scattering with optical transmission matrices (TMs). Unlike many of the aforementioned methods, inverse scattering with TMs does not require any ballistic photons and the scattering media does not need to exhibit the memory effect. However, TM methods need the scattering media's behavior to be invariant over time; TM techniques are not well-suited to dynamic media like fog. Moreover, TM methods generally assume that you have access to the far side of the scattering media for calibration, which allows you to characterize the relation between the input field, on a spatial light modulator (SLM), and the output field, on a camera.

TM methods are based upon the principle that when dealing with coherent light, the forward scattering process is linear with respect to fields, and so the output field is related to the input field by a matrix A. This matrix is called the TM. If one first measures the TM, one can then "invert" the scattering process and image through [17] or even inside [18] the scattering medium, potentially through the use of phase retrieval algorithms [19]. The imaging process is illustrated in Figure 1.

D. Multi-spectral Transmission Matrices

Traditionally, TMs describe the scattering process when dealing with coherent narrowband light, which lets you ignore the fact that scattering media have spectrally dependent responses [20]. In a recent work [20], the authors demonstrated that characterizing scattering with multi-spectral light is also possible. The key idea is to replace the 2D TM that describes the scattering process at a fixed wavelength, with a 3D tensor, which describes the scattering process at multiple wavelengths. This characterization enables temporal, as well as spatial, focusing through the scattering media [21].

E. Contributions: Measuring the Effects of Temporal Coherence and a Fast Phase Retrieval Algorithm

In this work, we study how effective a single TM is at characterizing and inverting the scattering process associated with a broadband (i.e., partially coherent) illumination source. In essence, we study how well the narrow-band model can approximate a broadband system. We show that as the bandwidth of the illumination increases, the singular values of the measured TM decay towards zero and the speckle encodes less and less information. As the bandwidth increases, we quickly reach a point at which an incoherent scattering model better describes the scattering process and is easier to invert.

To carry out this task efficiently, we learn TMs using the double phase retrieval technique [22], [19]. This method uses phase retrieval algorithms to avoid difficult-to-capture interferometric measurements. Unfortunately, existing double phase retrieval methods require a massive amount of computation. In this work we have developed a fast, GPU-accelerated algorithm which lets us reconstruct 256^2 64^2 TMs in under five hours, as opposed to the tens of thousands that would be required with competing methods. An initial version of this algorithm was presented in [23].

II. Modeling Scattering

In this section, we describe the models underlying TM-based methods.

A. Narrow-band Illumination

When dealing with completely coherent light, the scattering process is linear with respect to *complex-valued* fields. Thus, the complex field z on the far side of the scattering media is related to the field x incident on the scattering media via the *complex-valued monochromatic transmission matrix* A.

That is,

$$\mathsf{z} = \mathsf{A}\mathsf{x} + \ _F; \tag{1}$$

where $_F$ models noise on the field.

Typical cameras capture only intensity information, in which case the measurement process becomes

$$\mathsf{y}^2 = j\mathsf{A}\mathsf{x} + \ _F j^2 + \ _I; \tag{2}$$

where the square is elementwise and $_I$ models noise on the intensity measurement. In this work we ignore $_I$, which we minimize through long exposures and strong illumination, and also take the square root of the intensity measurements. Our simplified measurement model becomes

$$\mathsf{y} = j\mathsf{A}\mathsf{x} + \ j: \tag{3}$$

After one measures the TM, potentially using the double phase retrieval method described in Section III-A, one can use phase retrieval algorithms to reconstruct a signal $\mathbf{x}$ from measurements $\mathbf{y}$. Thus phase retrieval algorithms can be used to see through multiple-scattering media.

B. Incoherent Illumination

At the other extreme, with completely incoherent illumination the scattering process is linear with respect to *real-valued* intensities. That is,

$$\mathbf{y}^2 = \mathbf{A}_{\text{intensity}}\mathbf{x}^2 + \; : \tag{4}$$

This makes reconstructing $\mathbf{x}^2$ from $\mathbf{y}^2$, and thus imaging through scattering media, a simple least squares problem. Unfortunately, the *real-valued intensity transmission matrix* $\mathbf{A}_{\text{intensity}}$ tends to be very poorly conditioned.

C. Multi-spectral Illumination

If one is able to distinguish/measure the various spectral components of the input and output fields, one can form a *multi-spectral transmission matrix/tensor*, $\mathbf{A}(!_k)$ for $k = 1;:::K$, and model scattering as

$$\mathbf{z}(!_k) = \mathbf{A}(!_k)\mathbf{x}(!_k) + \; (!_k); \tag{5}$$

for frequencies $!_1;:::!_K$. Multi-spectral TMs can enable spatio-temporal coherent control of broadband light [20], [21]. However, without access to narrow-band measurements, $\mathbf{A}(!_k)$ remains out of reach.

D. Broadband Illumination

For broadband illumination, where all frequency components are measured at once, the measurement process is best described by

$$\mathbf{y}^2 = \int j\mathbf{z}(!_k)j^2 d!_k = \int j\mathbf{A}(!_k)\mathbf{x}(!_k) + \; (!_k)j^2 d!_k: \tag{6}$$

The intensities measured at the camera consist of the superposition of many different frequency components. However, without applying filters or otherwise capturing spectrally dependent measurements, these components are indistinguishable on the camera.

Thus, while ideally one would learn and invert $\mathbf{A}(!_k)$ across many frequencies, we are motivated to learn a single model for the scattering process and invert this. In effect, we approximate a broadband scattering model (6) using either coherent (2) or incoherent (4) models and study how well one can use said models to image through scattering media. For small illumination bandwidths, where the spectral bandwidth remains within the spectral memory effect [20], the coherent TM model should be quite accurate. However, as we increase the illumination bandwidth the measurements will have contributions from a diverse and uncorrelated set of narrow-band TMs and the model should break down. In contrast, increasing bandwidth should make the incoherent illumination model more accurate.

In the next section, we describe how we measure the coherent (complex-valued field) and incoherent (real-valued intensity) TMs, $\mathbf{A}$ and $\mathbf{A}_{\text{intensity}}$.

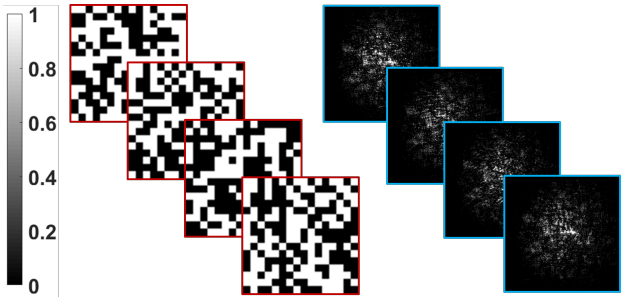

Fig. 2. Binary calibration patterns placed on the SLM and the corresponding measurements recorded by the camera. Random calibration patterns give rise to random uncorrelated speckle patterns.

III. Measuring Field and Intensity Transmission Matrices

TMs are typically measured using holographic interferometry [24]. To learn a TM with interferometry, one illuminates one pixel of the SLM at a time. For each pixel, one uses interferometry to record the complex field incident on the detector. In this way, one is able to directly measure the TM, one column at a time.[1]

While effective, interferometric methods are very sensitive to perturbations; even minute vibrations, such as those caused by an air conditioning unit turning on and off in a lab, can be enough to change the interference pattern and thereby kill this method. This problem becomes particularly pernicious at higher resolutions where the physical stability requirement becomes more and more demanding.

In this work, we avoid the challenge of accurately measuring complex-valued fields by using the double phase retrieval method [22], [19]. The key idea behind double phase retrieval is that if one measures a sufficient number of the intensities of responses to calibration signals, one can use phase retrieval algorithms to learn the TM.

In this section, we explain double phase retrieval and introduce a new, GPU-accelerated algorithm that vastly reduces the computation times associated with this method. We also briefly describe how we measure intensity TMs.

A. Double Phase Retrieval

The double phase retrieval method gets its name from the fact that it requires performing phase retrieval twice; once for calibration, i.e., measuring the TM, and once for imaging.

1) Calibration: One first sends a series of calibration patterns $\mathbf{x}_p \; 2 \; \mathbf{R}^N$ with $p = 1;:::P$, through the scattering media, see Figure 2. For each p, the signal will be transformed by the TM $\mathbf{A} \; 2 \; \mathbf{C}^{M \;\; N}$, to produce measurements $\mathbf{y}_p \; 2 \; \mathbf{R}_+^M$, with

$$\mathbf{y}_p = j\mathbf{A}\mathbf{x}_p + \; _p j;$$

where $_p$ denotes noise.[2] Actual calibration patterns and their corresponding measurements are illustrated in Fig. 2.

[1]This is not the only, nor best, way to measure a TM using interferometry. For instance, [25] more efficiently measures a TM by temporally modulating the phase of an SLM and capturing a video of a dynamic speckle pattern.

[2]Multiple sources of noise, for instance leakage across the SLM, contribute to the noise $_p$. In this work, we assume that $_p$ follows a white circularly-symmetric complex Gaussian distribution.

The sets of calibration, measurement, and noise column vectors are then concatenated with themselves to form $\mathsf{X} = [\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_P] \in \mathbb{R}^{N \times P}$, $\mathsf{Y} = [\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_P] \in \mathbb{R}_+^{M \times P}$, and $\mathsf{E} = [\mathsf{e}_1, \mathsf{e}_2, \ldots, \mathsf{e}_P] \in \mathbb{C}^{M \times P}$.

In this way, the entire calibration process can be described by the equation

$$\mathsf{Y} = |\mathsf{A}\mathsf{X} + \mathsf{E}|.$$

Taking the transpose of the above equality, one obtains

$$\mathsf{Y}^H = |\mathsf{X}^H \mathsf{A}^H + \mathsf{E}^H|.$$

Consider the m^{th} column of Y^H (this consists of the P measurements associated with the m^{th} detector pixel)

$$\mathsf{y}_m^H = |\mathsf{X}^H \mathsf{a}_m^H + \mathsf{e}_m^H|, \tag{7}$$

where a_m^H and e_m^H denote the m^{th} rows of A and E.

Assuming P is sufficiently large (as a rule of thumb $P > 4N$), one can apply phase retrieval algorithms to (7) to recover each row a_m^H; simply treat y_m^H as the measurement and X^H as the measurement matrix. This can be repeated for each of the M rows of A, potentially in parallel, to learn the entire TM.

2) Imaging: After one has an estimate $\hat{\mathsf{A}}$ of the TM, imaging through the scattering medium is straightforward. Given a speckle pattern y, ones needs only apply phase retrieval algorithms to the measurements with $\hat{\mathsf{A}}$ as the known measurement matrix.

B. prVAMP: A Fast and Robust Phase Retrieval Algorithm

The central challenge to using the double phase retrieval method in practice is computation times. Experimental systems often have low signal to noise ratios and when using an amplitude SLM the measurement matrix X^H is nonzero mean.[3] In this context, especially at low sampling rates, few phase retrieval algorithms perform accurate reconstructions. Unfortunately, the algorithm that works by far the best, prSAMP [26], is extremely slow; reconstructing one row of a $256^2 \times 64^2$ TM with prSAMP takes over an hour.

One way to reduce these computation times is to use special block-diagonal measurement matrices, but this comes at a cost in terms of recovery accuracy [27]. We propose a complementary solution.

In this work, we develop a new algorithm, prVAMP, which offers the same accuracy as prSAMP, while running hundreds of times faster. Simulations comparing prVAMP with over a dozen other phase retrieval algorithms are provided in Appendix A. In Appendix B we compare prVAMP and prSAMP with and without the fast, block-diagonal matrices from [27].

prVAMP is a special case of the recently developed Generalized Vector Approximate Message Passing (GVAMP) [28] algorithm. GVAMP is an algorithm for computing approximate minimum means squared error (MMSE) or maximum a posteriori (MAP) solutions to inverse problems involving generalized linear measurements (GLMs); defined to be any measurement of the form

$$\mathsf{y} = Q(\mathsf{z} + \mathsf{w}) \text{ with } \mathsf{z} = \mathsf{A}\mathsf{x}, \tag{8}$$

where A is our measurement matrix (which is a known pattern X^H when calibrating and an estimated TM $\hat{\mathsf{A}}$ when imaging), x is our signal of interest, w is noise, and $Q(\,)$ denotes a simple non-linearity. prVAMP is designed for the special case $Q(\,) = |\,|$.

GVAMP can be broadly understood as an extension to GAMP [29], [30], which infuses the ADMM [31] concept of variable splitting into the algorithm. This modification allows GVAMP to handle non-zero-mean, correlated, and ill-conditioned measurement matrices, which would destabilize the original GAMP algorithm [32], [33].

The pseudo-code for prVAMP is presented in Algorithm 1. prVAMP works by first splitting the vectors x and z into two sets of identical vectors $\mathsf{x_1}$ and $\mathsf{x_2}$ and $\mathsf{z_1}$ and $\mathsf{z_2}$. Iterations of the algorithm then broadly consist of four steps: First, there are two "denoising" steps which impose priors on x and z and ensure they are consistent with the measurements y. Second, there are two linear minimum mean squared error (LMMSE) estimation steps which ensure the estimates of x and z are consistent with each other. Within the algorithm $\hat{\mathsf{x}}$ and $\hat{\mathsf{z}}$ terms are estimates of x and z, r and p terms act as noisy observations of x and z, the α and γ terms track precisions (reciprocals of variances), and the β and δ terms are divergence terms used to estimate these precisions.

We now give explicit forms for the various denoising and estimation functions listed in Algorithm 1. These expressions correspond to the MMSE form of the algorithm under i.i.d. Gaussian priors on the signal and noise. In all these

Algorithm 1 prVAMP

Initialize: r_{10}, p_{10}, γ_{10}, γ_{10}
for k=0, 1, ..., K **do**
  Denoise x:
  $\hat{\mathsf{x}}_{1;k} = g_{x1}(\mathsf{r}_{1;k}; \gamma_{1;k})$,
  $\alpha_{1;k} = h g'_{x1}(\mathsf{r}_{1k}; \gamma_{1;k}) i$
  $\mathsf{r}_{2k} = (\hat{\mathsf{x}}_{1;k} - \alpha_{1;k}\mathsf{r}_{1k})/(1 - \alpha_{1;k})$,
  $\gamma_{2k} = \gamma_{1k}(1 - \alpha_{1;k})/\alpha_{1k}$
  Denoise z:
  $\hat{\mathsf{z}}_{1;k} = g_{z1}(\mathsf{p}_{1k}; \gamma_{1k})$,
  $\gamma_{1k} = h g'_{z1}(\mathsf{p}_{1k}; \gamma_{1k}) i$
  $\mathsf{p}_{2k} = (\hat{\mathsf{z}}_{1k} - \gamma_{1k}\mathsf{p}_{1k})/(1 - \gamma_{1k})$,
  $\gamma_{2k} = \gamma_{1k}(1 - \gamma_{1k})/\gamma_{1k}$
  LMMSE estimation of x:
  $\hat{\mathsf{x}}_{2k} = g_{x2}(\mathsf{r}_{2k}; \mathsf{p}_{2k}; \alpha_{2k}; \gamma_{2k})$,
  $\alpha_{2k} = h g'_{x2}(\mathsf{r}_{2k}\mathsf{p}_{2k}; \alpha_{2k}; \gamma_{2k}) i$
  $\mathsf{r}_{1;k+1} = (\hat{\mathsf{x}}_{2k} - \alpha_{2k}\mathsf{r}_{2k})/(1 - \alpha_{2k})$,
  $\gamma_{1;k+1} = \gamma_{2k}(1 - \alpha_{2k})/\alpha_{2k}$
  LMMSE estimation of z:
  $\hat{\mathsf{z}}_{2k} = g_{z2}(\mathsf{r}_{2k}; \mathsf{p}_{2k}; \delta_{2k}; \gamma_{2k})$,
  $\delta_{2k} = h g'_{z2}(\mathsf{r}_{2k}; \mathsf{p}_{2k}; \delta_{2k}; \gamma_{2k}) i$
  $\mathsf{p}_{1;k+1} = (\hat{\mathsf{z}}_{2k} - \delta_{2k}\mathsf{p}_{2k})/(1 - \delta_{2k})$,
  $\gamma_{1;k+1} = \gamma_{k2}(1 - \delta_{2k})/\delta_{2k}$
Return $\hat{\mathsf{x}}_{1K}$.

[3] In our experiments, the elements of $\mathbf{X}^H$ are i.i.d. Bernoulli distributed with $p = .5$. I.e., each element of $\mathbf{X}$ is 0 or 1, each with probability 0.5.

expressions we have dropped the iteration subscript k for readability.

1) Denoising Steps: On line 5 the algorithm denoises the noisy estimate r_1 of the signal x with the function $g_{x1}(\,)$; that is, it treats r_1 as noisy observation of x with additive white Gaussian noise with variance $1/\tau_1$ and computes the MMSE estimate. When we a priori assume that x is drawn from an i.i.d. zero-mean Gaussian distribution with variance $1/\tau_x$, this function becomes

$$g_{x1}(\mathsf{r}_1;\tau_1) = \mathrm{E}[\mathsf{x}\,|\,\mathsf{r}_1;\tau_1] = \frac{1}{\tau_x+\tau_1}\mathsf{r}_1. \tag{9}$$

With the aforementioned assumptions on the distributions, $g_{x1}(\mathsf{r}_1;\tau_1)$ just scales r_1. An expression for the average partial derivative $\langle g'_{x1}(\,)\rangle$ can be found in [29].

On line 10 the algorithm denoises the noisy estimate p_1 of the signal z with the function $g_{z1}(\,)$; that is, it treats p_1 as noisy observation of z with additive white Gaussian noise with variance $1/\tau_1$ and computes the elementwise MMSE estimate, under a prior determined by the observation y and the distribution of the noise w. If we assume that $\mathsf{y} = |\mathsf{z}+\mathsf{w}|$ with $\mathsf{w} \sim \mathcal{CN}(0;1/\tau_w\mathsf{I})$, where $\mathcal{CN}$ denotes a circular Gaussian distribution (real and imaginary components are i.i.d. with variance $\frac{1}{2}1/\tau_w$), this function, written elementwise, becomes

$$g_{z1;m}(\mathsf{p}_{1:m};\tau_1) = \mathrm{E}[\mathsf{z}_{1:m}\,|\,\mathsf{y}_m;\mathsf{p}_{1:m};\tau_w;\tau_1] = \frac{\mathsf{y}_m^2}{1+\tau_1/\tau_w}R_0\left(\frac{2\mathsf{y}_m|\mathsf{p}_{1;m}|}{1/\tau_w+1/\tau_1}\right)\frac{\mathsf{p}_{1;m}}{|\mathsf{p}_{1;m}|}; \tag{10}$$

for indices $m = 1;\ldots;M$, where $R_0(\,) \equiv \frac{I_1(\,)}{I_0(\,)}$, where $I_0(\,)$ and $I_1(\,)$ denote zeroth and first order Bessel functions. This function, which is specific to the non-linearity $|\,|$, is what distinguishes prVAMP from other GVAMP algorithms. A derivation of this function and expression for its average partial derivative $\langle g'_{z1}(\,)\rangle$ can be found in [30].

2) LMMSE Steps: On lines 14 and 19 the algorithm performs MMSE estimation of x and z under the pseudo-priors that $\mathsf{r}_2 \sim \mathcal{CN}(\mathsf{x};1/\tau_2)$ and $\mathsf{p}_2 \sim \mathcal{CN}(\mathsf{z};1/\tau_2)$, with $\mathsf{x} = \mathsf{z}$, using the functions $g_{x2}(\,)$ and $g_{z2}(\,)$. That is,

$$g_{x2}(\mathsf{r}_2;\mathsf{p}_2;\tau_2;\tau_2) = \mathrm{E}[\mathsf{x}\,|\,\mathsf{r}_2;\mathsf{p}_2;\tau_2;\tau_2];$$

and

$$g_{z2}(\mathsf{r}_2;\mathsf{p}_2;\tau_2;\tau_2) = \mathrm{E}[\mathsf{z}\,|\,\mathsf{r}_2;\mathsf{p}_2;\tau_2;\tau_2].$$

Because these estimates are linear with respect to r_2 and p_2, these are called the LMMSE steps of the algorithm.

The LMMSE steps can be accelerated dramatically if one first computes a singular value decomposition (SVD) of the measurement matrix $\quad$, with $\quad = \mathsf{USV}^t$; before running the algorithm. With this decomposition in hand, the $g_{x2}(\,)$ and $g_{z2}(\,)$ functions can be written as

$$g_{x2}(\mathsf{r}_2;\mathsf{p}_2;\tau_2;\tau_2) = \mathsf{V}\mathsf{D}_k(\tau_2\mathsf{S}^t\mathsf{U}^t\mathsf{p}_2 + \tau_2\mathsf{V}^t\mathsf{r}_2); \tag{11}$$

where D_k denotes an $n \quad n$ diagonal matrix whose i^{th} diagonal entry is defined as $(\tau_2 s_i^2 + \tau_2)^{-1}$, where s_i is the i^{th} singular value of $\quad$ and

$$g_{z2}(\mathsf{r}_2;\mathsf{p}_2;\tau_2;\tau_2) = \quad g_{x2}(\mathsf{r}_2;\mathsf{p}_2;\tau_2;\tau_2). \tag{12}$$

More information about these functions and expression for their average partial derivatives can be found in [28].

3) Tuning and Initialization: prVAMP is a Bayesian algorithm that requires priors on the distribution of the signal and noise. In this work, when reconstructing TMs, we used the following heuristic to estimate these quantities. We first computed an initial solution $\hat{\mathsf{x}}_{\mathrm{GS}}$, using the Gerchberg-Saxton PR algorithm [34]. Then, treating this solution as the ground truth, we estimated the precisions as $\tau_w = \frac{1}{k\mathsf{y}\quad|\hat{\mathsf{x}}_{\mathrm{GS}}|k^2}$ and $\tau_x = \frac{1}{\mathrm{var}(\hat{\mathsf{x}}_{\mathrm{GS}})}$. We assumed that both x and w were zero-mean. The same procedure was used for both the calibration and imaging steps; although we only imaged binary objects, this fact was not used as a prior within the algorithm.

When reconstructing TMs, we used the following heuristic to initialize prVAMP. Using the solution provided by GS, we set $\mathsf{r}_{10} = \hat{\mathsf{x}}_{\mathrm{GS}}$, $\mathsf{p}_{10} = \quad\hat{\mathsf{x}}_{\mathrm{GS}}$, $\tau_{10} = \frac{1}{\mathrm{var}(\hat{\mathsf{x}}_{\mathrm{GS}})}$, and $\tau_{10} = \frac{1}{\mathrm{var}(\quad\hat{\mathsf{x}}_{\mathrm{GS}})}$.

4) GPU Acceleration: The prVAMP algorithm consists mostly of matrix-vector operations. By replacing these matrix-vector operations with matrix-matrix operations, we can solve thousands of phase retrieval problems in parallel. These matrix-matrix operations can be computed very efficiently using a GPU: Recovering 1000 rows of a $256^2 \quad 64^2$ TM in sequence on an Intel 6800k CPU takes 300 minutes. Recovering these rows in parallel on a Pascal Titan X GPU takes 2 minutes.

To implement prVAMP we modified the original GVAMP Matlab code. Code demonstrating prVAMP on synthetic data is available within the GAMP project [35]. Code demonstrating prVAMP on experimental TM data, using GPU computing, is available on our website [36].

C. Measuring Intensity Transmission Matrices

To estimate the incoherent, real-valued intensity TM associated with a series of measurements $\mathsf{y}_p^2 = \mathsf{A}_{\text{intensity}}\mathsf{x}_p^2 + \quad_p$, we assembled the calibration data and the associated measurements into matrices X^{2H} and Y^{2H} (unlike the coherent model, no elementwise square root was taken). From there we first formed a maximum likelihood (ML) estimate of the intensity TM as

$$\mathsf{A}^H_{\text{intensity, ML}} = (\mathsf{X}^{2H})^{y}\mathsf{Y}^{2H}; \tag{13}$$

where y denotes pseudo-inverse.

Next, we assumed that the elements of A and $\quad$ follow an i.i.d. Gaussian distribution with variances τ_a^2, $\quad^2$ and estimated these variances using our ML solution: $\tau_a^2 = \mathrm{var}(\mathsf{A}_{\text{intensity, ML}})$ and $\quad^2 = \frac{k\mathsf{Y}^2 \quad \mathsf{A}_{\text{intensity, ML}}\mathsf{X}^2k_2^2}{M}$. With these priors, the MAP estimate of A becomes

$$\hat{\mathsf{A}}^H_{\text{intensity, MAP}} = \arg\max_{\mathsf{A}} \frac{k\mathsf{Y}^2 \quad \mathsf{A}\mathsf{X}^2k_F^2}{\tau_w^2} \quad \frac{k\mathsf{A}k_F^2}{\tau_a^2}; \tag{14}$$

$$= (\mathsf{X}\mathsf{X}^H + \frac{\tau_w^2}{\tau_a^2}\mathsf{I})^{-1}\mathsf{X}\mathsf{Y}^H. \tag{15}$$

Because solving linear systems is much easier than solving phase retrieval problems, estimating $\mathsf{A}_{\text{intensity}}$ took a few minutes, as opposed to five hours.

We repeated an analogous process, forming an ML solution and then using this quantity to select the priors for our MAP solution, for imaging.