From Fienup's phase retrieval techniques to regularized inversion for in-line holography: tutorial

Fabien Momey, Loïc Denis, Thomas Olivier, Corinne Fournier

Beyond Fienup's phase retrieval techniques: regularized inversion for in-line holography

FABIEN MOMEY,[1,*] LOÏC DENIS,[1], THOMAS OLIVIER,[1] AND CORINNE FOURNIER[1]

[1] *Laboratoire Hubert Curien UMR CNRS 5516, Université Jean Monnet, F-42000 Saint-Étienne, France*
[*] *fabien.momey@univ-st-etienne.fr*

Abstract: This paper includes a tutorial on how to reconstruct in-line holograms using an inverse problems approach, starting with modeling the observations, selecting regularizations and constraints, and ending with the design of a reconstruction algorithm. A special focus is made on the connections between the numerous alternating projection strategies derived from Fienup's phase retrieval technique and the inverse problems framework. In particular, an interpretation of Fienup's algorithm as iterates of a proximal gradient descent for a particular cost function is given. Reconstructions from simulated and experimental holograms of micrometric beads illustrate the theoretical developments. The results show that the transition from alternating projection techniques to the inverse problems formulation is straightforward and advantageous.

1. Introduction

The imaging of samples at macro, micro, and nano scales is important in many fields of research from physics (fluid mechanics, materials) to biology (cells, bacteria, viruses). Imaging remains challenging and advanced techniques are still emerging. These new techniques often exploit physical principles involved in light/matter interactions from a variety of radiation sources (X-rays, electrons, visible or near-visible light, laser). In this context, imaging techniques aim to record the perturbation of the incident electromagnetic wave by the sample of interest. The perturbed wave depends both on the absorption and phase-shift properties of the sample. Thus, being able to precisely measure the shape of this complex wave - the amplitude and phase of the electromagnetic field in space and time - is crucial to fully quantitatively characterize the sample.

Holography, invented by Dennis Gabor in 1948 [1], is a technique to reconstruct the complex wave (amplitude and phase) due to the diffraction of light when a scattering sample is illuminated by a coherent source. This principle gave rise to a wide variety of imaging techniques (off-axis holography, phase-shifting, in-line holography, diffractive tomography, X-ray diffractive microscopy, *etc.*) [2]. Since only intensity measurements of light are available, the problem of retrieving the phase on the sensor has fueled many studies. In contrast to interferometric setups like off-axis or phase-shifting holography, iterative phase retrieval techniques are algorithms that estimate the phase of light in the plane of in-line holograms.

Digital hologram reconstruction is typically performed by backpropagating the hologram from the sensor plane to the object plane. In the absence of phase information with in-line holograms, the reconstruction suffers from artifacts called *the twin-image*. To suppress the twin-image, phase retrieval techniques were introduced in the 1970s and 1980s, and since then, improved algorithms have regularly been described in the literature. Most phase retrieval techniques are still derived from the methods of alternating projections initially proposed by Gerchberg and Saxton [3] and popularized and extended by Fienup [4, 5]. This class of methods is still widely used today [6–13], with improvements to enforce *a priori* knowledge (support of the objects, admissible values domain, sparsity constraints) [14–16].

Inverse problem approaches take a different point of view from phase retrieval techniques: rather than recovering the phase on the sensor plane (which does not completely solve the sample reconstruction problem), they focus on the reconstruction of the complex-valued transmittance in the object plane. Because of measurement noise and hologram truncation at the borders of

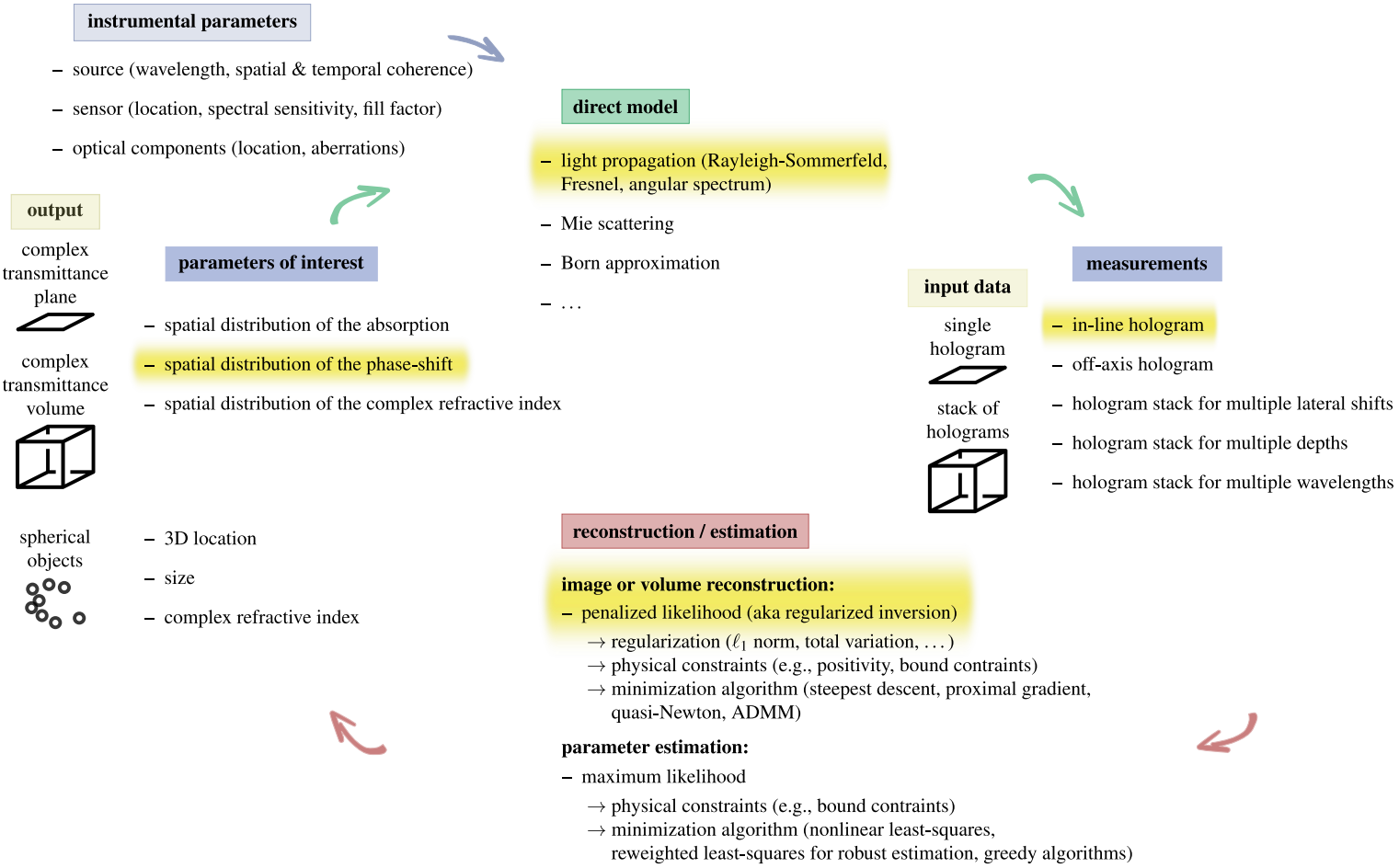


Fig. 1. Hologram reconstruction based on inverse problem approaches: the direct model connects the sample of interest to the measurements (the holograms), inverting this model leads to a reconstruction and/or estimation of the optical parameters of the sample. The case chosen as an illustration in this paper is highlighted in yellow.

the sensor, backpropagating the complex amplitude in the sensor plane does not lead to perfect sample reconstruction, i.e., restoring the sensor phase and reconstructing the complex wave in the sample plane are not strictly equivalent problems.

Figure 1 gives an overview of the different ingredients of the inverse problems methodology in the context of holographic imaging. The "direct model" relates a description of the "parameters of interest" (*i.e.* the sample) and the "measurements" (*i.e.* the hologram), taking into account the "instrumental parameters" (*i.e.* the imaging setup). The sample may be described as a complex-valued transmittance plane [17], a volume of transmittance planes [18], or as a collection of objects [19]. Depending on the application, the sample is absorbing, and/or translucent (it induces a phase-shift due to a refractive index difference with respect to the surrounding medium). It is described by the spatial distribution of the absorption, phase-shift or complex refractive index within a plane or a volume. When individual objects are considered, their 3D location, size, and complex refractive index are the parameters that characterize the objects. Any appropriate optical model can then be used to describe how an incident illumination wave gets diffracted by the sample and propagates to the sensor: diffraction by a plane (Rayleigh-Sommerfeld, Fresnel approximation, angular spectrum) [20], diffraction by a volume (Born approximation, Rytov approximation) [21], diffraction by a sphere (Mie scattering) [22], free-space propagation. Multiple holograms can be recorded to improve information diversity by including lateral or axial shifts of the sensor [16, 23], multiple illumination wavelengths [24], or multiple illumination angles [25, 26]. When a suitable model of the observations has been defined, a "reconstruction" (or "estimation") method can be selected. If the sample is described by a spatial distribution within a plane or a volume, the problem amounts to an "image or volume reconstruction". If a collection of objects is to be identified, the geometrical and optical parameters of each object have to be estimated ("parameter estimation"). "Image or volume reconstruction" problems are generally solved by regularized inversion [27–37], i.e., by solving a minimization problem whose objective function is defined by a data-fitting term that penalizes the deviation of the model from the data. Regularization terms are added to favor reconstructions with desirable properties (e.g., smooth images, sharp edges, zero-valued background). Additional (hard) constraints are also enforced to guarantee that the reconstruction fulfills basic physical properties such as the positivity of the absorption or the sign of the phase-shift induced by the objects (e.g., thin objects with an optical index larger than that of the surrounding medium induce a positive phase-shift). Depending on the type of cost function to minimize (quadratic, non-linear but differentiable (smooth), non-differentiable (non-smooth), non-convex), different strategies can be used. "Parameter estimation" of a collection of objects generally follows a different path because the reduced number of unknowns makes the regularization unnecessary. Rather than considering a fixed discretization of space, the 3D location of each object is generally sought in a continuous domain (*i.e.* subpixel location). Objects are often detected and characterized one after another, and the measurements are described by the superimposition of the diffraction patterns due to each object [19, 38]. Nonlinear least-squares minimization techniques are applied to estimate the geometrical and optical parameters of each object. In the case where the data contains some signatures that can not be fitted by the model (outliers), *e.g.* out-of-field objects, the estimations can be improved by using a robust signal processing approach such as the reweighted least-squares [39].

The hologram formation model depends on "instrumental parameters" such as the wavelength and coherence of the source, the sample-to-sensor distance, or the sensor response (fill-factor and spectral sensitivity). The model of the data not only underlies the reconstruction algorithm, but also provides a way to calibrate the experimental parameters.

The goal of this paper is to illustrate the inverse problems methodology and to draw connections with the popular phase retrieval techniques based on Fienup's approach. We selected a problem that can be addressed using either approach: the reconstruction of an in-line hologram of a non-

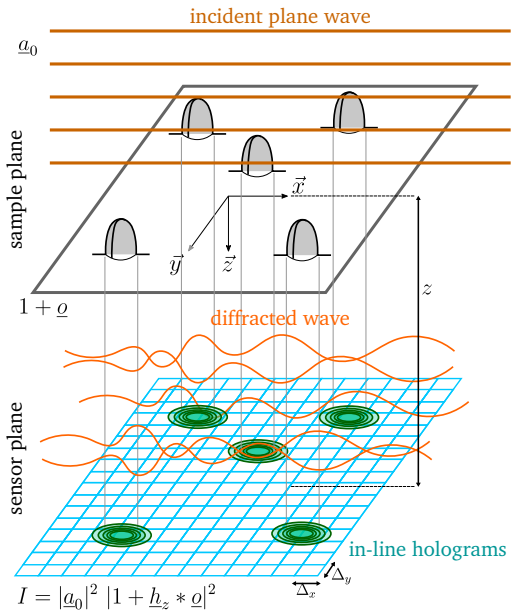


Fig. 2. In-line holography principle: a plane wave illuminates the sample plane with normal incidence. The presence of inhomogeneities in the spatial distribution of the complex-valued refractive index, in the sample plane, distorts the illumination wave. A diffraction pattern is recorded on the sensor plane.

absorbing plane sample that induces spatially varying phase-shifts. We discuss the advantages provided by the inverse problems framework and interpret Fienup's as an instance of a proximal gradient algorithm applied to a specific cost function.

The paper is organized as follows: next, we present the phase retrieval problem in the context of in-line digital holography. We then present the standard alternating projections strategy proposed by Gerchberg and Saxton, Fienup, and subsequent variants. In section 4, we present the inverse problems methodology applied to the reconstruction of in-line holograms of phase-only samples. We validate and compare the algorithms in section 5, first in numerical simulations, then on experimental in-line holograms of microscopic beads. Finally, we discuss the possible extensions of inverse problems approaches for phase retrieval.

2. The inverse problem of image reconstruction in in-line holography

The object of interest, *i.e.* the sample, can be modeled as a two-dimensional (2D) complex transmittance plane $\underline{t}(\boldsymbol{r})$ ($\boldsymbol{r} = (x, y)$ is the vector of 2D spatial coordinates and all complex-valued variables are underlined in this paper). A transmittance $\underline{t}(\boldsymbol{r}) = 1$, $\forall \boldsymbol{r}$, means a fully transparent plane that leaves any illumination wave unperturbed after passing through it. Consequently, the presence of scattering objects in this plane will be revealed by a certain deviation $\underline{o}\,(\boldsymbol{r})$ from a unit transmittance:

$$\underline{t}(\boldsymbol{r}) = 1 + \underline{o}(\boldsymbol{r}) = \rho(\boldsymbol{r}) e^{i\varphi(\boldsymbol{r})} \tag{1}$$

$\underline{o}(\boldsymbol{r})$ is the quantity of interest. It models the distribution of the phase $\varphi(\boldsymbol{r})$ and amplitude $\rho(\boldsymbol{r})$, directly linked to the complex refractive index difference between the observed objects (beads, droplets, cells, *etc.*) and the medium.

Digital in-line holography (DIH) (*cf.* Fig. (2)) consists in illuminating a sample plane with a

coherent (generally plane) wave $\underline{a}_0(\boldsymbol{r})$ of wavelength λ, and in recording on a digital sensor the intensity $I(\boldsymbol{r})$ of the total diffracted wave $\underline{a}_z(\boldsymbol{r})$ at a distance z from the sample plane (the sensor plane being parallel to the sample plane). From the Huygens-Fresnel principle, the physical data acquisition process writes:

$$I(\boldsymbol{r}) = \left| \iint_{\mathbb{R}^2} \underline{h}_z(\boldsymbol{r} - \boldsymbol{r}')\, \underline{a}_0(\boldsymbol{r}')\, \underline{t}(\boldsymbol{r}')\, \mathrm{d}\boldsymbol{r}' \right|^2 = |\underline{h}_z \underset{\boldsymbol{r}}{*} (\underline{a}_0 \cdot \underline{t})\,|^2 \tag{2}$$

where $*$ is a convolution operator and $\underline{h}_z$ is the propagation kernel. Depending on the approximations considered, the Rayleigh-Sommerfeld and the Fresnel kernels are used as $\underline{h}_z$ [20]. In this paper, we use the Fresnel kernel (paraxial approximation), which writes:

$$\underline{h}_z(\boldsymbol{r}) = \frac{1}{i\lambda z} \exp\left(\frac{i\pi}{\lambda z}||\boldsymbol{r}||^2\right) \tag{3}$$

The backpropagation kernels $\underline{h}_{-z}$ can be used to recover the original complex wave from a diffracted (i.e., propagated) wave.

Considering Eq. (1), and under the assumption that the incident wave is plane ($\underline{a}_0(\boldsymbol{r}) = \underline{a}_0$), Eq. (2) becomes:

$$I(\boldsymbol{r}) = |\underline{a}_0|^2\, |1 + \underline{h}_z \underset{\boldsymbol{r}}{*} \underline{o}|^2 = |\underline{a}_0|^2 \underbrace{\left(1 + 2\,\Re\mathrm{e}(\underline{h}_z \underset{\boldsymbol{r}}{*} \underline{o}) + |\underline{h}_z \underset{\boldsymbol{r}}{*} \underline{o}|^2\right)}_{\text{model } m(\underline{o})}. \tag{4}$$

The intensity $I(\boldsymbol{r})$ is recorded by a digital sensor. Therefore, effective intensity measurements correspond to sampled (and noisy) intensity values that can be collected in the form of a column vector $\boldsymbol{d} = \left(d_q\right)^{\mathrm{T}}_{q=1\ldots M}$ with M data measurements. This also leads to considering the targeted physical quantity $\underline{o}$ as a vector $\underline{\boldsymbol{o}} = \left(x_q\right)^{\mathrm{T}}_{q=1\ldots N}$ of N unknowns. In imaging systems, N is often equal to M as we aim to recover an image of the same size as the hologram, *i.e.* the field of view. In the context of inverse approaches, the unknown field of view of $\underline{\boldsymbol{o}}$ can be extended, leading to $N > M$ (*cf.* Sec. 6). We note $\mathbb{D}$ and $\mathbb{O}$ respectively the feasible domains of $\boldsymbol{d}$ and $\underline{\boldsymbol{o}}$, *i.e.* the domains of valid values for each quantity. From Eq. (4), $\mathbb{D} \subset \mathbb{R}^M$ and $\mathbb{O} \subset \mathbb{C}^N$. As discussed in Sec. 1, in order to ensure physically relevant properties of the object of interest, the feasible domain $\mathbb{O}$ can be restricted, for example to impose a positive absorption or the sign of the phase shift. Such properties are applied in the regularized inverse approach developed in our case study in Sec. 4.2.

The algebraic process that "transforms" $\underline{\boldsymbol{o}}$ in $\boldsymbol{d}$ is modeled by the mapping $\boldsymbol{m}$ which goes from $\mathbb{D}$ to $\mathbb{O}$, and is a discretized version of the physical - analytic - formulation m in Eq. (4). Finally, we get the following algebraic formulation of the image formation model:

$$\boldsymbol{d} = c\, \boldsymbol{m}(\underline{\boldsymbol{o}}) + \boldsymbol{\eta} = c\, |\mathbf{1} + \underline{\mathbf{H}}_z\, \underline{\boldsymbol{o}}|^2 + \boldsymbol{\eta} \tag{5}$$

where the square modulus operator $|\boldsymbol{x}|^2$ of a vector $\boldsymbol{x}$ is applied component-wise, leading to an M-dimensional vector, and $\mathbf{1}$ is a vector of 1 of size M. $\boldsymbol{\eta} = \left(\eta_q\right)^{\mathrm{T}}_{q=1\ldots M}$ corresponds to a vector of M noise values to model electronic noise in the measurement system, and mathematical approximations involved in the model, *etc.* $\underline{\mathbf{H}}_z$ is the complex-valued convolution operator that calculates the discrete propagation of the object $\underline{\boldsymbol{o}}$: $\underline{\mathbf{H}}_z$ is a $M \times N$ matrix such that $[\underline{\mathbf{H}}_z\, \underline{\boldsymbol{o}}]_q = [\underline{h}_z * \underline{o}]_q$, for all pixels q. The Hermitian transpose of $\underline{\mathbf{H}}_z$, written $\underline{\mathbf{H}}_z^\dagger$, corresponds to the backpropagation operator $\underline{\mathbf{H}}_{-z}$. Because of sampling and field truncation at the border of the images, the backpropagation does not exactly inverts the propagation, yet the approximation $\underline{\mathbf{H}}_z^\dagger \underline{\mathbf{H}}_z \approx \mathbf{I}$ is often used to get an intuitive understanding of holographic reconstruction. Finally,

c is a scalar scaling factor that accounts for the intensity of the incident wave $|\underline{a}_0|^2$ as well as the detector gain and quantum efficiency.

Retrieving the unknown image of the sample $\underline{\boldsymbol{o}}$, noted $\underline{\boldsymbol{o}}^*$, from the observed hologram $\boldsymbol{d}$, based on the hologram formation model in Eq. (5), constitutes the inverse problem of holographic reconstruction.

3. Fienup's alternating projections strategies

The standard alternating projection strategy, still widely used to date, was first proposed by Fienup in 1978 [4] as the Error-Reduction (ER) algorithm. It is an upgraded variant of the initial method proposed by Gerchberg and Saxton in 1972 [3] from two-plane intensity measurements.

In [5], Fienup has proposed several variants of the ER algorithm, such as the Basic Input-Output (BIO) or the Hybrid Input-Output (HIO), involving a parameter β in the "projection on data" step that relaxes the strict projection on the feasible domain. These variants are known to be more efficient that the classical ER strategy (HIO is considered as the most efficient). In this paper, we focus on the classical approach ER, which we find the most intuitive way to understand the alternating projections strategy. We show in Sec. 4.3.3 that this algorithm can be reformulated to fit the inverse problems framework. Thus, our goal is not an exhaustive comparison of inverse approaches with all variants of the alternating projections strategy, but rather to give an "inverse problems" interpretation of Fienup's strategy and show that inverse problems offer a framework that as accessible, yet more flexible and powerful than alternating projections.

Algorithm (1) summarizes the Fienup ER strategy. Details of implementation are given in Appendix A, in Algo. (3).

Fienup's ER algorithm only requires one intensity measurement image $\boldsymbol{d}$ at the sensor plane at distance z. This measurement image has to be normalized so that the background equals 1, to ensure that the retrieved transmittance plane $\underline{\boldsymbol{o}}$ is normalized the same way ($\underline{\boldsymbol{o}} = \mathbf{1}$ stands for a fully transparent plane, see Sec. 2), and that appropriate physical constraints can be applied. The normalized data image, noted $\bar{\boldsymbol{d}}$, can be obtained for example by dividing $\boldsymbol{d}$ by its mean or its median (almost valid in case of a low-density distribution of objects).

The principle is to iteratively alternate the following steps starting from an estimate of the object $\underline{\boldsymbol{o}}^{(i)}$ at the sample plane to get a new estimate $\underline{\boldsymbol{o}}^{(i+1)}$:

- The estimate $\underline{\boldsymbol{o}}^{(i)}$ is propagated to the sensor plane to get a simulated complex wave $\underline{\boldsymbol{a}}_z^{(i+1/2)}$ corresponding to the total diffracted wave in this plane.
- At the sensor plane, a "projection on data" step forces the amplitude of $\underline{\boldsymbol{a}}_z^{(i+1/2)}$ to match the square root of the normalized measurement vector $\bar{\boldsymbol{d}}$ (projection on a non-convex set), to get a modified wave $\underline{\boldsymbol{a}}_z^{(i+1)}$.
- The modified wave $\underline{\boldsymbol{a}}_z^{(i+1)}$ is backpropagated to the object plane to get a new estimate $\underline{\boldsymbol{o}}^{(i+1/2)}$.
- At the sample plane, $\mathfrak{Re}(\underline{\boldsymbol{o}}^{(i+1/2)})$ and $\mathfrak{Im}(\underline{\boldsymbol{o}}^{(i+1/2)})$ are forced to be positive or negative quantities (depending on the hypothesis concerning the object of interest), and/or forced to be zero outside a restricted support. This step, that leads to the new estimate $\underline{\boldsymbol{o}}^{(i+1)}$, can be summed up as a projection step $\mathcal{P}_{\mathbb{O}}(\underline{\boldsymbol{o}}^{(i+1/2)})$ on the feasible domain $\mathbb{O} \subset \mathbb{C}^N$ (usually a convex set).

This approach is based on the approximation $\underline{\mathbf{H}}_z^{\dagger}\underline{\mathbf{H}}_z \approx \mathbf{I}$ that makes it possible to go back and forth between the sample plane and the sensor plane. Since $\underline{\mathbf{H}}_z^{\dagger}$ is not the exact inverse of $\underline{\mathbf{H}}_z$, there are strong limitations related to sampling and truncation effects (e.g., objects on the border of the field-of-view are inevitably distorted). If the algorithm is stopped after several iterations

and the sensor plane constraint applied, the data are augmented by the phase information (i.e., the lost phase on the sensor plane is retrieved by the algorithm). Thanks to this restored phase, the twin-image phenomenon typical of in-line holography is strongly reduced.

The great popularity of this class of algorithms is due to the fact that the strategy is very intuitive as well as being easy to implement. Likewise, these algorithms have been proven to converge to a local minimum [40–42]. Such approaches have been mainly exploited in the fields of X-ray coherent diffraction imaging [7, 43–46], digital holographic microscopy [29, 47, 48].

As a result, research in this field is still active and new approaches are often proposed. In the past 30 decades, it has benefited from many practical improvements [6, 10, 11, 13, 49] and theoretical studies [9, 41, 50]. Recent approaches include new constraint enforcement strategies using sparsity constraints in the object or data spaces [51, 52] or in the wavelets domain [16, 53] .

Algorithm 1: Fienup's Error-Reduction (ER) algorithm [4]

Input:

$\bar{\boldsymbol{d}}$; {*normalized intensity measurements (background = 1) at sensor plane at distance z*}

$\underline{\mathbf{H}}_z, \underline{\mathbf{H}}_{-z}$; {*Propagation and backpropagation operators*}

$\underline{\boldsymbol{o}}^{(0)}$; {*initial estimate of* $\underline{\boldsymbol{o}}$*:* e.g. *the backpropagated measured amplitude at sensor plane* $\underline{\mathbf{H}}_{-z}\left(\sqrt{\bar{\boldsymbol{d}}} - \mathbf{1}\right)$ *or simply random values*}

Output:

$\underline{\boldsymbol{o}}^{*}$; {*estimated complex transmittance at sample plane*}

begin

1 $i \leftarrow 0$;

repeat

2 $\underline{\boldsymbol{a}}_z^{(i+1/2)} \leftarrow \mathbf{1} + \underline{\mathbf{H}}_z\, \underline{\boldsymbol{o}}^{(i)}$; {*Step 1: propagation to the sensor plane*}

3 $\underline{\boldsymbol{a}}_z^{(i+1)} \leftarrow \sqrt{\bar{\boldsymbol{d}}} \odot \left(\underline{\boldsymbol{a}}_z^{(i+1/2)} / |\underline{\boldsymbol{a}}_z^{(i+1/2)}|\right)$; {*Step 2: enforce the measured amplitude at sensor plane*}

{*N.B.: the product* $\odot$*, the division* $/$*, and the modulus* $|\cdot|$ *are applied pixelwise*}

4 $\underline{\boldsymbol{o}}^{(i+1/2)} \leftarrow \underline{\mathbf{H}}_{-z}\,(\underline{\boldsymbol{a}}_z^{(i+1)} - \mathbf{1})$; {*Step 3: backpropagation to the sample plane*}

5 $\underline{\boldsymbol{o}}^{(i+1)} \leftarrow \mathcal{P}_{\mathbb{O}}\left(\underline{\boldsymbol{o}}^{(i+1/2)}\right)$; {*Step 4: projection on the domain* $\mathbb{O}$}

6 $i \leftarrow i + 1$;

until *Maximum number of iterations reached*;

7 $\underline{\boldsymbol{o}}^{*} \leftarrow \underline{\boldsymbol{o}}^{(i)}$;

end

4. Inverse problems methodology for the reconstruction of in-line holograms: a tutorial

4.1. A linearized image formation model

The first step of the inverse problems methodology is to derive a suitable image formation model, by analyzing the object and data characteristics. Here we describe a linearized model based on two hypotheses: (i) the sample is a distribution of purely and weakly dephasing objects ($\rho(\boldsymbol{r}) = 1$, see Eq. (1)) such that:

$$\underline{t}(\boldsymbol{r}) = 1 + \underline{o}(\boldsymbol{r}) = e^{i\varphi(\boldsymbol{r})} \approx 1 + i\varphi(\boldsymbol{r}) = 1 + io(\boldsymbol{r}) \; ; \tag{6}$$

Thus, we define a new unknown image $o(\boldsymbol{r})$ which is real and stands for the phase image $\varphi(\boldsymbol{r})$.

(ii) we neglect the 2^{nd} order term in Eq. (4), leading to the following approximated real and linear image formation model:

$$\tilde{m}\,(o) = 1 - 2\,\Im\mathrm{m}(\underline{h}_z) * o \tag{7}$$

The new discrete image formation model writes:

$$\boldsymbol{d} = c\;\tilde{\boldsymbol{m}}(\boldsymbol{o}) + \boldsymbol{\eta} \quad \text{with} \quad \tilde{\boldsymbol{m}}(\boldsymbol{o}) = (\mathbf{1} + \mathbf{G}_z\boldsymbol{o}) \tag{8}$$

where $\mathbf{G}_z = -2\,\Im\mathrm{m}(\underline{\mathbf{H}}_z)$ is a linear propagation operator. It is a real-valued convolution operator: $\mathbf{G}_z$ is still a $M \times N$ matrix such that $[\mathbf{G}_z\;\boldsymbol{o}]_q = [a_z]_q = [-2\,\Im\mathrm{m}(\underline{h}_z) * o]_q$, for all pixels q. We note the convolution kernel $g_z = -2\,\Im\mathrm{m}(\underline{h}_z)$. As it is real and even, the operator $\mathbf{G}_z$ is symmetric. Therefore $\mathbf{G}_z^{\mathrm{T}} = \mathbf{G}_z$. The noise $\boldsymbol{\eta}$ now also includes modeling errors due to the above mentioned approximations. Model approximations are very often considered because of their computational interest in leading to tractable mathematical and numerical resolution strategies. **The inverse problem is now fully real-valued (model $\tilde{\boldsymbol{m}}$ and unknown $\boldsymbol{o}$), and constitutes a simple case-study to pedagogically illustrate the basic ingredients and tools required to build a relevant inverse problems approach and draw a parallel with Fienup's alternating projections approaches.** It is of course also possible to consider non-linear direct models and both the attenuation and phase-shift induced by the sample. An example of such an approach can be found in [17].

4.2. Building a regularized inverse approach

First, we recall the definition of the l_2 and l_1 norms applied to a vector $\boldsymbol{x}$: $||\boldsymbol{x}||_2 = \sqrt{\sum_q x_q^2}$ and $||\boldsymbol{x}||_1 = \sum_q |x_q|$. The notation $||\boldsymbol{x}||_{\mathbf{W}}$, where $\mathbf{W}$ is a positive definite matrix, corresponds to: $||\boldsymbol{x}||_{\mathbf{W}}^2 = \boldsymbol{x}^{\mathrm{T}}\mathbf{W}\boldsymbol{x}$.

4.2.1. Data-fidelity

An inverse approach consists in retrieving an optimal solution $\boldsymbol{o}^*$ to Eq. (8), knowing an approximate numerical model $\tilde{\boldsymbol{m}}(\boldsymbol{o})$ of the data formation process.

To do so, we define a calculable metric $\mathcal{J}_{\mathsf{fid}}(\boldsymbol{o}, \boldsymbol{d})$ that evaluates the error between the data $\boldsymbol{d}$ and the model $\tilde{\boldsymbol{m}}(\boldsymbol{o})$ obtained from a guessed $\boldsymbol{o}$. This error term $\mathcal{J}_{\mathsf{fid}}(\boldsymbol{o}, \boldsymbol{d})$, also called the data-fidelity term, acts as a penalization: the larger the error, the farther the guess $\boldsymbol{o}$ is from the solution and the more it has to be corrected to reach a better estimate.

The most popular data-fidelity term is the weighted least squares criterion, which derives from the assumption of Gaussian errors. The criterion corresponds to the $\mathbf{W}$-weighted squared l_2 norm of the differences between $\boldsymbol{d}$ and $\tilde{\boldsymbol{m}}(\boldsymbol{o})$, which writes:

$$\begin{aligned}
\mathcal{J}_{\mathsf{fid}}(c, \boldsymbol{o}, \boldsymbol{d}) &= \|c\;\tilde{\boldsymbol{m}}(\boldsymbol{o}) - \boldsymbol{d}\|_{\mathbf{W}}^2 \\
&= (c\;\tilde{\boldsymbol{m}}(\boldsymbol{o}) - \boldsymbol{d})^{\mathrm{T}}\;\mathbf{W}\;(c\;\tilde{\boldsymbol{m}}(\boldsymbol{o}) - \boldsymbol{d}) \\
&= \sum_q w_q \left(\tilde{m}_q(\boldsymbol{o}) - d_q\right)^2 \qquad \text{(if errors are uncorrelated, } \textit{i.e.}\text{, } \mathbf{W} \text{ is diagonal)} \\
&= \sum_q w_q \left(1 + [g_z * o]_q - d_q\right)^2
\end{aligned} \tag{9}$$

where $\tilde{m}_q(\boldsymbol{o})$ is the q-th pixel of the model $\tilde{\boldsymbol{m}}(\boldsymbol{o})$, and d_q is the q-th pixel of the hologram. w_q (q-th element of the diagonal of matrix $\mathbf{W}$) is a weight associated with pixel q on the sensor. In imaging problems, this allows to avoid unmeasured pixel data y_q in the field of view, by setting to 0 the corresponding weight w_q. As discussed later (see Sec. 6), these unmeasured pixels could

be those outside the sensor field of view, when out-of-field objects have to be taken into account because of their contribution to the recorded hologram [1].

To solve the inverse problem, the objective is to find the solution $\boldsymbol{o}^*$ that minimizes the criterion Eq. (9), thus the error between the data $\boldsymbol{d}$ and the model $\tilde{\boldsymbol{m}}(\boldsymbol{o}^*)$. Such a minimization problem writes:

$$\{\boldsymbol{o}^*, c^*\} = \underset{\boldsymbol{o},c}{\arg\min} \quad \mathcal{J}_{\text{fid}}(c, \boldsymbol{o}, \boldsymbol{d}) \tag{10}$$

and is the standard formulation of an inverse problem. In this formulation of our particular problem, the optimal solution also depends on the scalar scaling factor c. The optimal value for c can be obtained in closed form:

$$c^*(\boldsymbol{o}) = \frac{\tilde{\boldsymbol{m}}(\boldsymbol{o})^{\mathrm{T}}\mathbf{W}\boldsymbol{d}}{\tilde{\boldsymbol{m}}(\boldsymbol{o})^{\mathrm{T}}\mathbf{W}\tilde{\boldsymbol{m}}(\boldsymbol{o})}\,. \tag{11}$$

If matrix $\mathbf{W}$ is diagonal, the expression of c^* takes a simpler form:

$$c^*(\boldsymbol{o}) = \frac{\sum_q w_q \tilde{m}_q(\boldsymbol{o}) d_q}{\sum_q w_q \tilde{m}_q(\boldsymbol{o})^2}\,, \tag{12}$$

We can then solve the following problem with regards to $\boldsymbol{o}$:

$$\boldsymbol{o}^* = \underset{\boldsymbol{o}}{\arg\min} \quad \mathcal{J}_{\text{fid}}(c^*(\boldsymbol{o}), \boldsymbol{o}, \boldsymbol{d}) \tag{13}$$

We discuss how this minimization problem can be addressed in Sec. 4.3.2.

4.2.2. Constraints and regularizations

The minimization problem Eq. (10) is not sufficient to obtain a satisfactory solution because measurements are noisy and there are too many unknowns for a reliable estimation based only on data fitting. As a result, unsatisfactory reconstructions can be found that perfectly match the data, *i.e.* solutions of the problem where the noise is also fitted by the model.

These issues are the characteristics of what is called an ill-posed problem. To overcome these limitations, the resolution of the problem can not be achieved by only considering the data-fidelity term Eq. (10). One has to find a solution $\boldsymbol{o}^*$ that best fits a particular prerequisite on the targeted information: one has to enforce some prior knowledge about the unknown $\boldsymbol{o}$.

Such prior knowledge can be injected in the minimization problem Eq. (10), taking the form of hard and/or soft constraints. The new minimization problem writes:

$$\boldsymbol{o}^* = \underset{\boldsymbol{o}\in\mathbb{O}}{\arg\min} \quad \mathcal{J}_{\text{fid}}(c^*(\boldsymbol{o}), \boldsymbol{o}, \boldsymbol{d}) + \mathcal{J}_{\text{reg}}(\boldsymbol{o}, \boldsymbol{\theta}) \tag{14}$$

where $\boldsymbol{o} \in \mathbb{O}$ and $\mathcal{J}_{\text{reg}}(\boldsymbol{o}, \boldsymbol{\theta})$ correspond to the enforced prior knowledge, that take respectively the form of hard and soft constraints, see for example [55, 56].

$\boldsymbol{o} \in \mathbb{O}$ imposes the feasible domain of values $\mathbb{O}$ for the estimate $\boldsymbol{o}$. In many inverse problems involving real quantities, a relevant constraint is to enforce positivity. This is for example the case in positron emission tomography (PET), where the quantity of interest is a photon count [57].

[1] $\mathbf{W}$ can be viewed from a statistical point of view: if the noise statistics is known *a priori*, $\mathbf{W}$ can be set to the inverse of the noise covariance matrix $\mathbf{C}_{\boldsymbol{\eta}} = \mathbb{E}\left[\boldsymbol{\eta}\,\boldsymbol{\eta}^{\mathrm{T}}\right]$ (where $\mathbb{E}[\cdot]$ stands for the expectation) [54]. In this case, the weighted least squares criterion corresponds to the co-log-likelihood, where the noise follows Gaussian statistics. Under the additional assumption of an uncorrelated noise, $\mathbf{W}$ is a diagonal matrix and the diagonal elements w_q correspond to the inverse of the variance at the pixel y_q (the larger the variance of the noise in a given pixel y_q, the smaller the weight assigned to this pixel in the criterion Eq. (9)). The maximum likelihood estimation theory suggests minimizing the weighted least-squares under a Gaussian assumption [55].

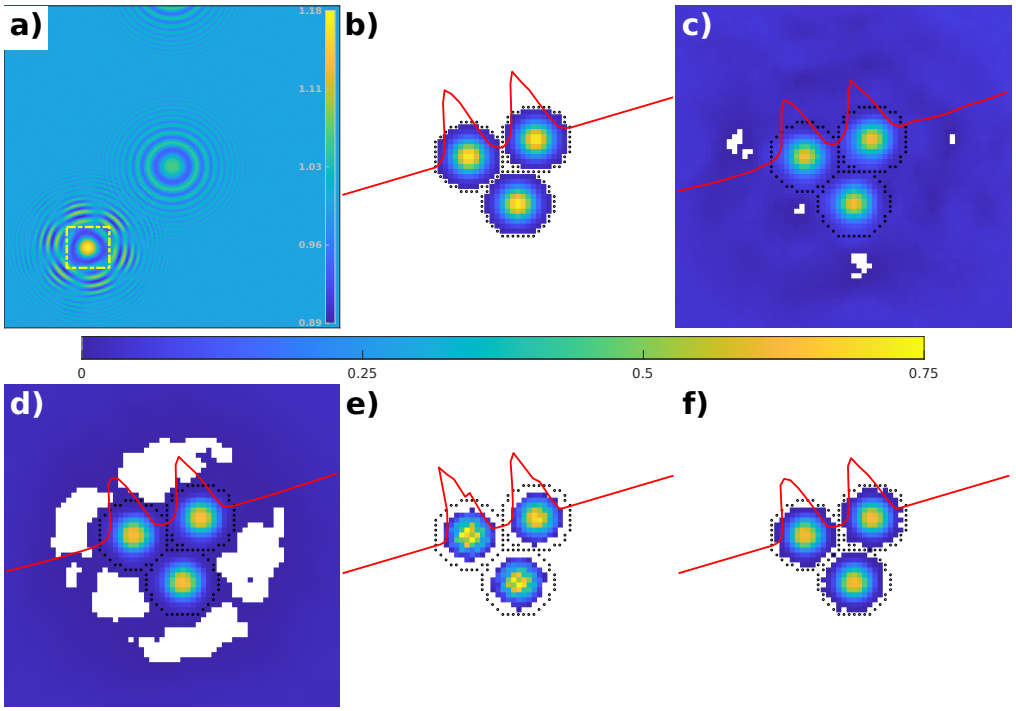


Fig. 3. Illustration of the behavior of several regularization terms on reconstructions (in a region of interest) of a simulated hologram (a) using a regularized inverse problems approach. Details on the method, the simulation and reconstruction parameters are given in Sec. 5, Tab. (1) and Tab. (2). a) Data (in-line hologram). The yellow frame indicates the region of interest that is extracted from the field of view for visualization. b) Ground truth phase. c-f) Reconstructed images using the following regularizations: c) the squared l_2 norm of the gradient $\mathcal{J}_{l_2\nabla}$ (Eq. (15)), d) the edge-preserving smoothness $\mathcal{J}_{\mathrm{TV}_\epsilon}$ (Eq. (16)), e) the l_1 sparsity constraint $\mathcal{J}_{l_1}$ (Eq. (17)), f) a composition of the l_1 sparsity constraint and the edge-preserving smoothness $\mathcal{J}_{\mathrm{reg}} = \mathcal{J}_{l_1} + \mathcal{J}_{\mathrm{TV}_\epsilon}$ (Eq. (18)). Red curves show a line profile passing through two particles of the image.

$\mathcal{J}_{\mathrm{reg}}(\boldsymbol{o}, \boldsymbol{\theta})$ are soft constraints imposed as a combination of regularization terms, parameterized by the set of hyperparameters $\boldsymbol{\theta}$ that balance each term with respect to the data-fidelity term. In general, standard regularizations aim to smooth the estimate $\boldsymbol{o}$ to filter high frequencies due to noise. In the following, we present three popular regularizations in image processing. Figure 3 illustrates their behavior on reconstructions, using a regularized inverse problems approach, of the simulated hologram that will be used in Sec. 5 for further demonstrations. More details on the method, the simulation and reconstruction parameters are also provided in Sec. 5.

A possible strategy is to minimize the squared l_2 norm of the gradient image $\nabla\boldsymbol{o}$:

$$\mathcal{J}_{l_2\nabla}(\boldsymbol{o}, \mu) = \mu \sum_q ||\nabla_q \boldsymbol{o}||_2^2 \tag{15}$$

where the gradient operator ∇_q corresponds to the finite difference operator at pixel q (a 2D vector with the differences in x and in y directions). However, this solution tends to oversmooth the sharp features of the estimate $\boldsymbol{o}$ (*cf.* Fig. (3)(c)) and is consequently not suitable for images with almost sharp edges. To overcome this limitation, more sophisticated regularizations can enforce piecewise smoothness, *i.e.* edge preservation. This is the targeted behavior of the very

popular total variation (TV) [58], which enforces sparsity of the gradient image by minimizing the sum of the l_2 norm of the gradient vector at each pixel. This regularization favors piecewise continuous images, i.e., sharp edges. Because it promotes the appearance of flat areas, which is not desired in cases where smooth variations are expected, it is often replaced by a generalized version, also known as "edge-preserving smoothness" [59]:

$$\mathcal{J}_{\mathrm{TV}_\epsilon}(\boldsymbol{o}, \mu, \epsilon) = \mu \sum_q \sqrt{||\nabla_q \boldsymbol{o}||_2^2 + \epsilon^2} \tag{16}$$

where ϵ, which must be different from zero to ensure the differentiability of the criterion, tunes the regularization behaviour. For values of the gradient norm much higher than ϵ the regularization acts as TV (preservation of sharp edges), while for much lower values, the regularization is almost quadratic (it almost acts as the regularizer of Eq. (15)) and the estimate $\boldsymbol{o}$ will be smoothed (*cf.* Fig. (3)(d)). Thus this criterion is more flexible and can impose more natural constraints than TV.

Direct sparsity of the image can also be enforced. This strategy is very often used to favor a low density distribution of objects in the image. This is implemented by minimizing the separable l_1 norm of the image:

$$\mathcal{J}_{l_1}(\boldsymbol{o}, \mu) = \mu \, \|\boldsymbol{o}\|_1 = \mu \sum_q |o_q| \tag{17}$$

This type of regularization limits the size of the objects support [29] by favoring reconstructions with pixels at zero. Indeed, looking at Fig. (3)(e), we clearly see that the objects of interest constitute the only signal reconstructed, while the remaining background is set to 0. Note that when $\boldsymbol{o}$ is constrained to be nonnegative, the above equation is just μ times the sum of the values in $\boldsymbol{o}$, hence a linear term in $\boldsymbol{o}$ (see Fig. (5)).

These regularizations can be mixed to benefit from each of their advantages (*cf.* Fig. (3)(f)), leading $\mathcal{J}_{\mathrm{reg}}(\boldsymbol{o}, \boldsymbol{\theta})$ in Eq. (14) to be a combination of some of these terms. This increased flexibility comes at the cost of a more difficult tuning of the reconstruction algorithms because of the increase of the number of hyperparameters. For example, in Fig. (3)(f) and reconstruction experiments in Sec. 5, we use a combination of a separable sparsity prior $\mathcal{J}_{l_1}$ (Eq. (17)) and an edge-preserving regularization $\mathcal{J}_{\mathrm{TV}_\epsilon}$ (Eq. (16)), leading to the following expression for the global regularization term $\mathcal{J}_{\mathrm{reg}}(\boldsymbol{o}, \boldsymbol{\theta})$:

$$\mathcal{J}_{\mathrm{reg}}(\boldsymbol{o}, \boldsymbol{\theta}) = \mathcal{J}_{l_1}(\boldsymbol{o}, \mu_{l_1}) + \mathcal{J}_{\mathrm{TV}_\epsilon}(\boldsymbol{o}, \mu_{\mathrm{TV}}, \epsilon_{\mathrm{TV}}) \tag{18}$$

with $\boldsymbol{\theta} = \{\mu_{l_1}, \mu_{\mathrm{TV}}, \epsilon_{\mathrm{TV}}\}$, the set of all tuning hyperparameters.

4.3. Case-study: solving a standard sparsity-based problem

4.3.1. The problem

Following the methodology presented in the previous section, we now focus on the resolution of the following criterion that corresponds to a popular sparsity-based approach:

$$\begin{aligned} \boldsymbol{o}^* &= \underset{\boldsymbol{o} \geq \boldsymbol{0}}{\arg\min} \; \underbrace{\mathcal{J}_{\mathrm{fid}}(c^*(\boldsymbol{o}), \boldsymbol{o}, \boldsymbol{d})}_{\text{smooth part } \mathcal{G}} + \underbrace{\mathcal{J}_{l_1}(\boldsymbol{o}, \mu)}_{\text{non-smooth part } \mathcal{H}} \\ &= \underset{\boldsymbol{o} \geq \boldsymbol{0}}{\arg\min} \; \|c^*(\boldsymbol{o}) \, \tilde{\boldsymbol{m}}(\boldsymbol{o}) - \boldsymbol{d}\|_{\mathbf{W}}^2 + \mu \, \|\boldsymbol{o}\|_1 \end{aligned} \tag{19}$$

The criterion minimizes the sum of a least squares data-fidelity term developed in Eq. (9) and an l_1 regularization (*cf.* Eq. (17)), weighted by the hyperparameter μ, to enforce the sparsity of the solution in the spatial domain. A positivity constraint is imposed on the solution ($\boldsymbol{o} \geq \boldsymbol{0}$).

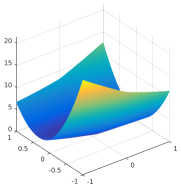
MINIMIZING A NON-SMOOTH CONVEX COST FUNCTION WITH PROXIMAL GRADIENT METHODS ISTA/FISTA

Decomposition of the cost function:

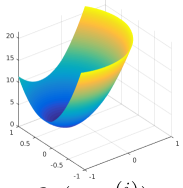
$$\mathcal{J}_{\text{fid}}(\boldsymbol{o}) + \mathcal{J}_{\text{reg}}(\boldsymbol{o}) = \underbrace{\mathcal{G}(\boldsymbol{o})}_{\text{smooth part}} + \underbrace{\mathcal{H}(\boldsymbol{o})}_{\text{non-smooth part}}$$

$\boldsymbol{o}^{(i)}$: current reconstruction → $\boldsymbol{o}^{(i+1)}$: new reconstruction

Since directly minimizing the cost function is too difficult, the ISTA iteration minimizes a simpler problem:

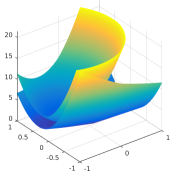
original cost function: 

$\mathcal{G}(\boldsymbol{o}) + \mathcal{H}(\boldsymbol{o})$

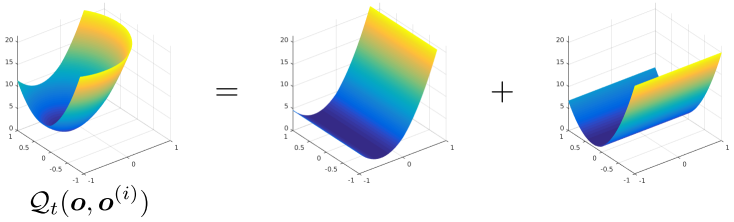
local approximation: 

$\mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)})$

$$\mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)}) := \underbrace{\mathcal{G}(\boldsymbol{o}^{(i)}) + (\boldsymbol{o} - \boldsymbol{o}^{(i)})^{\mathrm{T}} \nabla \mathcal{G}(\boldsymbol{o}^{(i)}) + \frac{1}{2t} \|\boldsymbol{o} - \boldsymbol{o}^{(i)}\|_2^2}_{\text{a separable quadratic model of } \mathcal{G}(\boldsymbol{o})} + \mathcal{H}(\boldsymbol{o})$$

$\mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)})$ is a majorant approximation: 

if the non-smooth part is separable, the approximation is also separable:



$\mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)})$

Minimizing the majorant approximation $\mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)})$ improves the solution:

$$\boldsymbol{o}^{(i+1)} = \arg\min_{\boldsymbol{o}} \mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)}) \quad \Rightarrow$$

$$\mathcal{G}(\boldsymbol{o}^{(i+1)}) + \mathcal{H}(\boldsymbol{o}^{(i+1)}) \leq \mathcal{Q}_t(\boldsymbol{o}^{(i+1)}, \boldsymbol{o}^{(i)}) \leq \mathcal{Q}_t(\boldsymbol{o}^{(i)}, \boldsymbol{o}^{(i)}) = \mathcal{G}(\boldsymbol{o}^{(i)}) + \mathcal{H}(\boldsymbol{o}^{(i)})$$

$\mathcal{Q}_t$ is a majorant; $\boldsymbol{o}^{(i+1)}$ is a minimizer

Fig. 4. The principle that underlies proximal gradient methods is the iterative minimization of a local approximation of the original cost function. If the parameter t is chosen small enough, the approximation is majorant and minimizing the approximation improves the current solution until convergence. When the non-smooth component $\mathcal{H}$ is separable (as is the case of the ℓ_1 norm), minimizing the local approximation is easily done (see Fig. (5)).

MINIMIZING THE MAJORANT APPROXIMATION $\mathcal{Q}_t$: COMPUTATION OF THE PROXIMAL OPERATOR

$\mathcal{Q}_t$ takes the generic form:

$$\mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)}) = \frac{1}{2t}\|\boldsymbol{o} - \boldsymbol{x}\|_2^2 + \mathcal{H}(\boldsymbol{o}) + \text{const}$$
$$\text{with } \boldsymbol{x} = \boldsymbol{o}^{(i)} - t\,\nabla\mathcal{G}(\boldsymbol{o}^{(i)})$$

The mapping $\boldsymbol{x} \mapsto \arg\min_{\boldsymbol{o}} \frac{1}{2t}\|\boldsymbol{o} - \boldsymbol{x}\|_2^2 + \mathcal{H}(\boldsymbol{o})$ is called the *proximal operator* of $\mathcal{H}(\boldsymbol{o})$

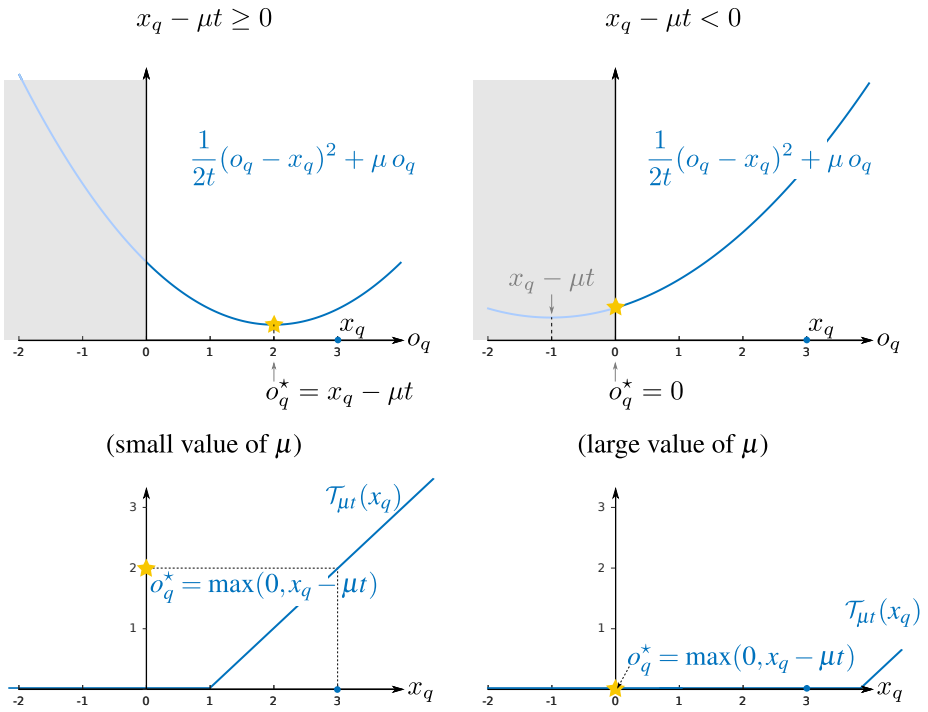
The closed-form expressions of many proximal operators are given in the literature.

Here is a derivation of the proximal operator when $\mathcal{H}(\boldsymbol{o}) = \mu\|\boldsymbol{o}\|_{l_1}$ and $\boldsymbol{o} \geq 0$:

$$\arg\min_{\boldsymbol{o}\geq 0} \frac{1}{2t}\|\boldsymbol{o} - \boldsymbol{x}\|_2^2 + \mu\|\boldsymbol{o}\|_1 \quad \Leftrightarrow \quad \arg\min_{\boldsymbol{o}\geq 0} \sum_q \left[\frac{1}{2t}(o_q - x_q)^2 + \mu\, o_q\right]$$

no absolute value here because $\boldsymbol{o}$ is non-negative

The sum involves terms that depend on a single o_q at a time. It is minimal when each term of the sum is minimal, i.e., when o_q is such that $\frac{1}{2t}(o_q - x_q)^2 + \mu\, o_q$ is minimal while being non-negative. The unconstrainted minimum of this quadratic function is $o_q = x_q - \mu t$. There are two cases:



The optimal value $o_q^\star$ is given by: $o_q^\star = \mathcal{T}_{\mu t}(x_q) = \max(0, x_q - \mu t)$, i.e., by soft-thresholding and clipping negative values to 0.

This leads to the following iteration:

$$\boldsymbol{o}^{(i+1)} = \arg\min_{\boldsymbol{o}} \mathcal{Q}_t(\boldsymbol{o}, \boldsymbol{o}^{(i)}) = \mathcal{T}_{\mu t}\left(\boldsymbol{o}^{(i)} - t\nabla\mathcal{G}(\boldsymbol{o}^{(i)})\right)$$

Fig. 5. Computation of the proximal operator: an illustration with the case of the l_1 norm under positivity constraints. Values that are smaller than μt are mapped to 0 by the proximal operator. Larger values of the regularization weight μ therefore lead to more values being set to 0, i.e., a sparser solution.

The problem Eq. (19) constitutes our case-study from which we must find suitable minimization strategies.

In the following, we introduce a simple yet efficient method to solve this inverse problem, named Iterative Shrinkage-Thresholding Algorithm (ISTA). This method belongs to the class of **proximal gradient methods** [60], whose general principle is depicted in Fig. (4). The key idea is to replace the original minimization problem by a sequence of simpler minimization problems involving a surrogate function that majorizes the original function. The cost function is decomposed into the sum of a smooth (i.e., differentiable) and a non-smooth (i.e., non differentiable) term. The smooth term is replaced by a separable quadratic approximation. Minimizing this approximation corresponds to computing the proximal operator associated to the non-smooth term [60]. In the case of a regularization by an l_1 norm, under a positivity constraint, we recall in Fig. (5) that the proximal operator corresponds to a soft-thresholding operation with a clipping of negative values to 0. In the following, we also establish that there is a strong connection between Fienup's alternating projections method and this strategy applied to a particular formulation of the minimization problem.

4.3.2. Optimization strategy

The application of the proximal gradient framework to the minimization of Eq. (19) leads to the following iteration:

$$\begin{aligned}
\boldsymbol{o}^{(i+1)} &= \mathcal{T}_{\mu t}\left(\boldsymbol{o}^{(i)} - t\ \nabla\mathcal{G}(\boldsymbol{o}^{(i)})\right) \\
&= \mathcal{T}_{\mu t}\left(\boldsymbol{o}^{(i)} - t\ \nabla\mathcal{J}_{\mathsf{fid}}(c^*(\boldsymbol{o}^{(i)}), \boldsymbol{o}^{(i)}, \boldsymbol{d})\right) \\
&= \mathcal{T}_{\mu t}\left(\boldsymbol{o}^{(i)} - 2\ t\ c^*(\boldsymbol{o}^{(i)})\ \mathbf{G}_z^{\mathrm{T}}\mathbf{W}\big(c^*(\boldsymbol{o}^{(i)})\ \tilde{\boldsymbol{m}}(\boldsymbol{o}^{(i)}) - \boldsymbol{d}\big)\right)
\end{aligned} \tag{20}$$

where the soft-thresholding operator $\mathcal{T}_{\mu t}$ is applied to the current reconstruction, improved by a steepest gradient descent step (t is the step length). As proved in Fig. (5), the proximal operator for the l_1 norm under positivity constraint is defined by:

$$\mathcal{T}_\alpha(\boldsymbol{o})_q = \max(0,\ o_q - \alpha)\,. \tag{21}$$

This soft-thresholding sets to 0 all values below a threshold α, which effectively denoises the reconstruction and tends to remove small fluctuations in the background. This projection step enforces sparsity directly in the spatial domain. It is also possible to derive proximal operators that enforce sparsity in "transformed" domains, such as the image gradient space (total-variation regularization, see [61]) or the wavelet domain (see [62]).

To ensure the convergence of the algorithm, the step length t has to be chosen adequately. For our problem, a suitable value depends on the maximum eigen value of $\mathbf{G}_z^{\mathrm{T}}\mathbf{G}_z$ (i.e., the Lipschitz constant of $\mathcal{J}_{\mathsf{fid}}$, see [60,61]). As $\mathbf{G}_z$ is a convolution operator, this is equivalent to the maximum squared modulus of the Fourier transform of the kernel g_z. It is also possible to use back-tracking methods to reduce the step length t if it is found that the surrogate function does not majorize the original cost function for the current value of t [61].

In [61], the authors propose a strategy to accelerate the convergence rate of ISTA, leading to the popular algorithm named Fast Iterative Shrinkage-Thresholding Algorithm (FISTA). We detail the general steps of FISTA in Algorithm (2). Implementation details are given in Appendix A, Algo. (4). In Algo. (2), if the scalar factor s is kept to the value 1, we fall back to the simpler ISTA algorithm.

Algorithm 2: FISTA algorithm [61] for solving problem Eq. (19)

Input:
$\boldsymbol{d}$; {*intensity measurements at sensor plane* z_{det}}
$\boldsymbol{o}^{(0)}$; {*initial estimate of* $\boldsymbol{o}$*:* e.g. $\mathbf{0}$ *or random values*}
μ ;
t ;
Output:
$\boldsymbol{o}^{*}$; {*estimated phase map on the sample plane* z_{obj}}

begin
1 $\quad \boldsymbol{u}^{(0)} \leftarrow \boldsymbol{o}^{(0)}$;
2 $\quad s^{(0)} \leftarrow 1$;
3 $\quad i \leftarrow 0$;
$\quad$ **repeat**
4 $\quad\quad \boldsymbol{o}^{(i+1)} \leftarrow \mathcal{T}_{\mu t}\left(\boldsymbol{u}^{(i)} - 2\,t\,c^{*}(\boldsymbol{u}^{(i)})\,\mathbf{G}_z^{\mathrm{T}}\mathbf{W}\big(c^{*}\,\tilde{\boldsymbol{m}}(\boldsymbol{u}^{(i)}) - \boldsymbol{d}\big)\right)$; {*ISTA iteration (*cf. *Eq. (20))*}
5 $\quad\quad s^{(i+1)} \leftarrow \frac{1}{2}\left(1 + \sqrt{1 + 4\,(s^{(i)})^2}\right)$; {*get a new interpolation coefficient*}
6 $\quad\quad \boldsymbol{u}^{(i+1)} \leftarrow \boldsymbol{o}^{(i)} + \frac{s^{(i)}-1}{s^{(i+1)}}\left(\boldsymbol{o}^{(i+1)} - \boldsymbol{o}^{(i)}\right)$; {*move according to previous iterate*}
7 $\quad\quad i \leftarrow i + 1$;
$\quad$ **until** *convergence*;
8 $\quad \boldsymbol{o}^{*} \leftarrow \boldsymbol{o}^{(i)}$;
end

4.3.3. The alternating projections point of view

We analyze the alternating projections strategy under the ISTA formulation presented in Fig. (4). From steps 1 to 4 in Algo. (1), we derive the equivalent iteration step of Fienup's algorithm as follows:

$$\underline{\boldsymbol{o}}^{(i+1)} = \mathcal{P}_{\mathbb{O}}\left(\underline{\mathbf{H}}_z^{\dagger}\left(\sqrt{\bar{\boldsymbol{d}}} \odot \frac{\underline{\boldsymbol{a}}_z^{(i+1/2)}}{|\underline{\boldsymbol{a}}_z^{(i+1/2)}|} - \mathbf{1}\right)\right) \tag{22}$$

with $\underline{\boldsymbol{a}}_z^{(i+1/2)} = \mathbf{1} + \underline{\mathbf{H}}_z\ \underline{\boldsymbol{o}}^{(i)}$, and given that $\underline{\mathbf{H}}_z^{\dagger} = \underline{\mathbf{H}}_{-z}$. $\odot$ stands for the Hadamard pixelwise product, and the division and modulus operators are also applied pixelwise. Since $\underline{\mathbf{H}}_z^{\dagger}\,\underline{\mathbf{H}}_z \approx \mathbf{I}$, we can rewrite Eq. (22) as follows:

$$\underline{\boldsymbol{o}}^{(i+1)} \approx \mathcal{P}_{\mathbb{O}}\left[\underline{\boldsymbol{o}}^{(i)} - \underline{\mathbf{H}}_z^{\dagger}\left(\frac{\underline{\boldsymbol{a}}_z^{(i+1/2)}}{|\underline{\boldsymbol{a}}_z^{(i+1/2)}|} \odot \left(|\underline{\boldsymbol{a}}_z^{(i+1/2)}| - \sqrt{\bar{\boldsymbol{d}}}\right)\right)\right] \tag{23}$$

The term in the square brackets on which the projector $\mathcal{P}_{\mathbb{O}}$ is applied corresponds to a steepest gradient descent step that decreases the following data-fidelity criterion:

$$\mathcal{J}_{\mathsf{fid}}(\boldsymbol{o}, \bar{\boldsymbol{d}}) = \left\| |\underline{\boldsymbol{a}}_z(\underline{\boldsymbol{o}})| - \sqrt{\bar{\boldsymbol{d}}} \right\|_2^2 = \left\| |\mathbf{1} + \underline{\mathbf{H}}_z\ \underline{\boldsymbol{o}}| - \sqrt{\bar{\boldsymbol{d}}} \right\|_2^2 . \tag{24}$$

Contrary to the previous data-fidelity (weighted least-squares), this term involves the square root of the observations $\bar{\boldsymbol{d}}$ and does not correspond to the co-log-likelihood under a Gaussian assumption. Deriving the gradient term $\nabla\mathcal{J}_{\mathsf{fid}}$ must be done carefully because $\underline{\boldsymbol{o}}$ is complex-valued.

Separating the real and imaginary parts, or alternatively using Wirtinger calculus, leads to the gradient expression in (Eq. (23)), see for example [63].

The projector $\mathcal{P}_{\mathbb{O}}$ is the orthogonal projection onto the convex set $\mathbb{O}$, and corresponds in our case to forcing positive or negative values of the real and/or imaginary parts of $\underline{\boldsymbol{o}}$, and/or to forcing to zero all pixels that are outside the support of the object. $\mathcal{P}_{\mathbb{O}}$ is in fact the proximal operator associated with the indicator function $\iota_{\mathbb{O}}$ of the domain $\mathbb{O}$ [60]:

$$\iota_{\mathbb{O}}(\underline{\boldsymbol{o}}) = \begin{cases} 0 & \text{if } \underline{\boldsymbol{o}} \in \mathbb{O} \\ +\infty & \text{otherwise} \end{cases} \tag{25}$$

Then, applying Fienup's alternating projections strategy given in Eq. (23) is analogous to performing ISTA iterations with a fixed step length $t = \frac{1}{2}$, derived from the following problem:

$$\underline{\boldsymbol{o}}^* = \arg\min_{\underline{\boldsymbol{o}}} \underbrace{\left\| |\mathbf{1} + \underline{\mathbf{H}}_z\, \underline{\boldsymbol{o}}| - \sqrt{\bar{\boldsymbol{d}}} \right\|_2^2}_{\text{smooth part } \mathcal{G}} + \underbrace{\iota_{\mathbb{O}}(\underline{\boldsymbol{o}})}_{\text{non-smooth part } \mathcal{H}} \tag{26}$$

where the smooth part $\mathcal{G}$ correponds to the data-fidelity term $\mathcal{J}_{\text{fid}}$ in Eq. (24) and the non-smooth part $\mathcal{H}$ is the indicator function $\iota_{\mathbb{O}}$ in Eq. (25).

The above analysis is important since it demonstrates that the alternating projections method underlies a particular inverse problems approach, and thus falls within this rigorous framework. The same as with the method developed in Sec. 4.3.2, the accelerated FISTA algorithm can then be applied to this problem, and it is even possible to replace the standard projection $\mathcal{P}_{\mathbb{O}}$ by, for example a soft-thresholding proximal operator $\mathcal{T}_{\mu t}$ to enforce sparsity of $\underline{\boldsymbol{o}}$ (real or/and imaginary part), or other regularization terms (provided that an efficient computation of the associated proximal operator is possible).

This interpretation of the alternating projections strategy in the framework of inverse problems and convex optimization methods has already been noted in previous works. In [5] Fienup already analyzed his variants of the ER algorithm as particular steepest descent strategies. Levi and Stark [50] have established that this method can be analyzed as an nonconvex instance of the *Projection Onto Convex Sets (POCS)* algorithm. This was rigorously demonstrated by Bauschke *et al.* [41] since Fienup's BIO and HIO are shown to be instances of Dykstra and Douglas-Rachford algorithms. In [12], alternating projections strategies are reformulated as proximity operators derived from the maximum likelihood point of view.

Conversely, this analysis makes it possible to interpret the ISTA strategy in Eq. (20) from the alternating projections point of view, since we can deduce that this algorithm performs a succession of a propagation step $\mathbf{G}_z$ of the image $\boldsymbol{o}$ to get a model of the data $\tilde{\boldsymbol{m}}\ (\boldsymbol{o})$, followed by a pseudo backpropagation step $\mathbf{G}_z^{\mathrm{T}}$ of an error term (difference between $\tilde{\boldsymbol{m}}\ (\boldsymbol{o})$ and $\boldsymbol{d}$) and finally a projection step $\mathcal{T}_{\mu t}$ on constraints.

With the inverse problems point of view, Fienup's ER algorithm becomes much more flexible. We can for example avoid the data normalization step $\bar{\boldsymbol{d}}$ and inject the optimal parameter $c^*(\boldsymbol{o})$. Taking the square root of the data is not necessarily a good solution since it does change the noise statistics. Moreover, the image formation model $\boldsymbol{m}$ of our approach directly calculates an intensity image $\tilde{\boldsymbol{m}}\ (\boldsymbol{o})$ that can match the raw intensity measurements $\boldsymbol{d}$, with or without the approximations of Sec. 4.1. Therefore, the approach developed in Sec. 4.3.1 is to be preferred, as it can be considered as a similar but more sophisticated alternating projections strategy.

With the above interpretation, we want to demonstrate that the inverse problems framework, in some specific but adapted cases, can keep the intuitive nature and ease of implementation property that generally constitutes the argument for preferring the use of Fienup's method. This is an original contribution of this paper compared with the previous works.

Table 1. Experimental and simulation parameters of in-line holograms data.

		Simulation	Experiment
illumination wavelength (in nm)	λ	532	
propagation distance (in μm)	z	12.5	7.3
refractive index of the medium	n_0	1.52	
refractive index of the beads	n_{bead}	1.59	1.59 *(commercial data)*
diameter of the beads (in μm)	d_{bead}	1.0	1.0 *(commercial data)*
magnification factor (microscope objective)	mag	56.7	
size of the image (in pixels)	$N = M$	512×512	1920×1080
pixel size (in μm)	$\Delta_x = \Delta_y$	4.4	2.2
noise type	$\boldsymbol{\eta}$	*Gaussian i.i.d.*	-
signal-to-noise ratio	$< \boldsymbol{d} > / \sigma_{\boldsymbol{\eta}}$	200	-

5. Results

In this section, we present the reconstruction results obtained with the proposed inverse problems approach formulated in Eq. (19), using the ISTA iteration given in Eq. (20) and its acceleration with the FISTA algorithm Algo. (2). We compare this method with Fienup's alternating projections approach defined by the iteration Eq. (23) using two versions: (i) the standard way with positivity constraint enforcement, (ii) an upgraded way with a soft-thresholding step (l_1 sparsity constraint). All the reconstruction strategies also enforce a positivity constraint. In the following interpretations of the results, our regularized inversion approach method is referred to as "the ri method", while Fienup's alternating projections approach is referred to as "the Fienup method".

5.1. Data

We reconstruct two data-sets: an experimental in-line hologram of polystyrene beads in immersion oil (*cf.* Fig. (7)(a)), and a simulated in-line hologram of beads in similar conditions (*cf.* Fig. (6)(a)). In this simulation, the transmittance o is modeled as 2D truncated Gaussian footprints (the Gaussians are truncated so that the footprint has a diameter of 6σ, and 0 outside.). The peak value of this footprint is set to a refractive index difference $(n_0 - n_{\text{bead}}) = 0.07$. The in-line hologram is simulated under Fresnel approximation. Table 1 summarizes the experiment and simulation parameters.

Looking at the data in Fig. (6)(a), we can see that a particle outside the field of view is also present. Its diffraction fringes are truncated at the top of the hologram. We intentionally included this out-of-field object to illustrate that the inverse approach framework allows the reconstruction of holograms in a wider-field of view than that seen by the sensor. We discuss this particular point in Sec. 6.

In these conditions, the maximum phase-shift induced by the particles is equal to $\Delta\varphi = 2\pi d_{\text{bead}}(n_0 - n_{\text{bead}})/\lambda \approx 0.83$rad. Then, the hypothesis (weakly dephasing objects) made to derive the model proposed in Sec. 4.1 are not valid. Still, in addition to the fact that the problem is simplified (only a real image has to be estimated), we show that this approximate model leads to good reconstructions.

5.2. Reconstructions

For each experiment, all the reconstruction parameters in Fig. (6) and Fig. (7) are summarized in Tab. (2).

Table 2. Reconstruction parameters for simulations and experiments in Fig. (3), Fig. (6), Fig. (7) and Fig. (8).

		Reconstruction parameters						
Figure	**Method**	**Gradient descent step**	**Positivity**	**Gradient smoothness**	**l_1-sparsity**	**Edge-preserving smoothness**		**max. nb. of iterations**
		t		$\mu_{l_2\nabla}$	μ_{l_1}	μ_{TV}	ϵ_{TV}	
Fig. (3)(b) ; Fig. (6)(b)	Ground truth							
Fig. (6)(d)	Fienup	1.0	*yes*	-	-	-	-	1000
Fig. (6)(e)	Fienup	1.0	*yes*	-	0.01	-	-	1000
Fig. (6)(f)	Fienup	1.0	*yes*	-	0.01	-	-	10
Fig. (3)(c)	ri | VMLMB	-	*yes*	1.0	-	-	-	1000
Fig. (3)(d)	ri | VMLMB	-	*yes*	-	-	0.1	0.01	1000
Fig. (3)(e) ; Fig. (6)(g)	ri | FISTA	0.1	*yes*	-	0.1	-	-	1000
Fig. (3)(f) ; Fig. (6)(h)	ri | FISTA	0.05	*yes*	-	0.1	0.1	0.01	1000
Fig. (6)(i)	ri | FISTA	0.05	*yes*	-	0.1	0.1	0.01	10
Fig. (7)(c)	Fienup	1.0	*yes*	-	-	-	-	100
Fig. (7)(d)	Fienup	1.0	*yes*	-	0.1	-	-	100
Fig. (7)(e)	Fienup	1.0	*yes*	-	0.1	-	-	10
Fig. (7)(f)	ri | FISTA	0.1	*yes*	-	0.5	-	-	100
Fig. (7)(g)	ri | FISTA	0.1	*yes*	-	0.5	0.01	0.01	100
Fig. (7)(h)	ri | FISTA	0.1	*yes*	-	0.5	0.01	0.01	10
Fig. (8)(c)	ri | FISTA	0.01	*yes*	-	0.1	0.001	0.01	100
Fig. (8)(d)	ri | FISTA	0.01	*yes*	-	0.1	0.1	0.01	100

Figure 6 shows the results of the reconstruction of the simulated hologram data, using the two methods presented above, with different parameters. For each reconstructed image in Fig. (6)(d-i), the values of hyperparameters are given in Tab. (2). First, we can observe that both reconstruction methods using a soft-thresholding step give a satisfying estimation of the objects' support. We see that the Fienup method enforcing this sparsity constraint is clearly better than the standard approach with a positivity constraint alone. However, the simulated objects have a spatially extended support, *i.e.* they cannot be considered as point objects. Thus, the sparsity constraint tends to create tiny holes within the objects. The sparsity constraint alone is not well-adapted to this particular hologram reconstruction problem. The RI method can be improved by adding an edge-preserving regularization in the form of the edge-preserving term presented in Eq. (16). The new minimization problem writes:

$$\begin{aligned} \boldsymbol{o}^* &= \underset{\boldsymbol{o}\geq\boldsymbol{0}}{\arg\min} \; \underbrace{\mathcal{J}_{\mathsf{fid}}(c^*(\boldsymbol{o}),\boldsymbol{o},\boldsymbol{d}) + \mathcal{J}_{\mathrm{TV}_\epsilon}(\boldsymbol{o},\mu_{\mathrm{TV}})}_{\text{smooth part } \mathcal{G}} \; + \; \underbrace{\mathcal{J}_{l_1}(\boldsymbol{o},\mu_{l_1})}_{\text{non-smooth part } \mathcal{H}} \\ &= \underset{\boldsymbol{o}\geq\boldsymbol{0}}{\arg\min} \; \|c^*(\boldsymbol{o})\,\tilde{\boldsymbol{m}}(\boldsymbol{o}) - \boldsymbol{d}\|^2_{\mathbf{W}} \; + \; \mu_{\mathrm{TV}} \sum_q \sqrt{||\nabla_q \boldsymbol{o}||_2^2 + \epsilon^2} \; + \; \mu_{l_1} \|\boldsymbol{o}\|_1 \; . \end{aligned} \tag{27}$$

As this additional regularization term is differentiable, it is included in the smooth part of the proximal gradient algorithm (see Fig. (4)). Thus the development of the ISTA iteration yields

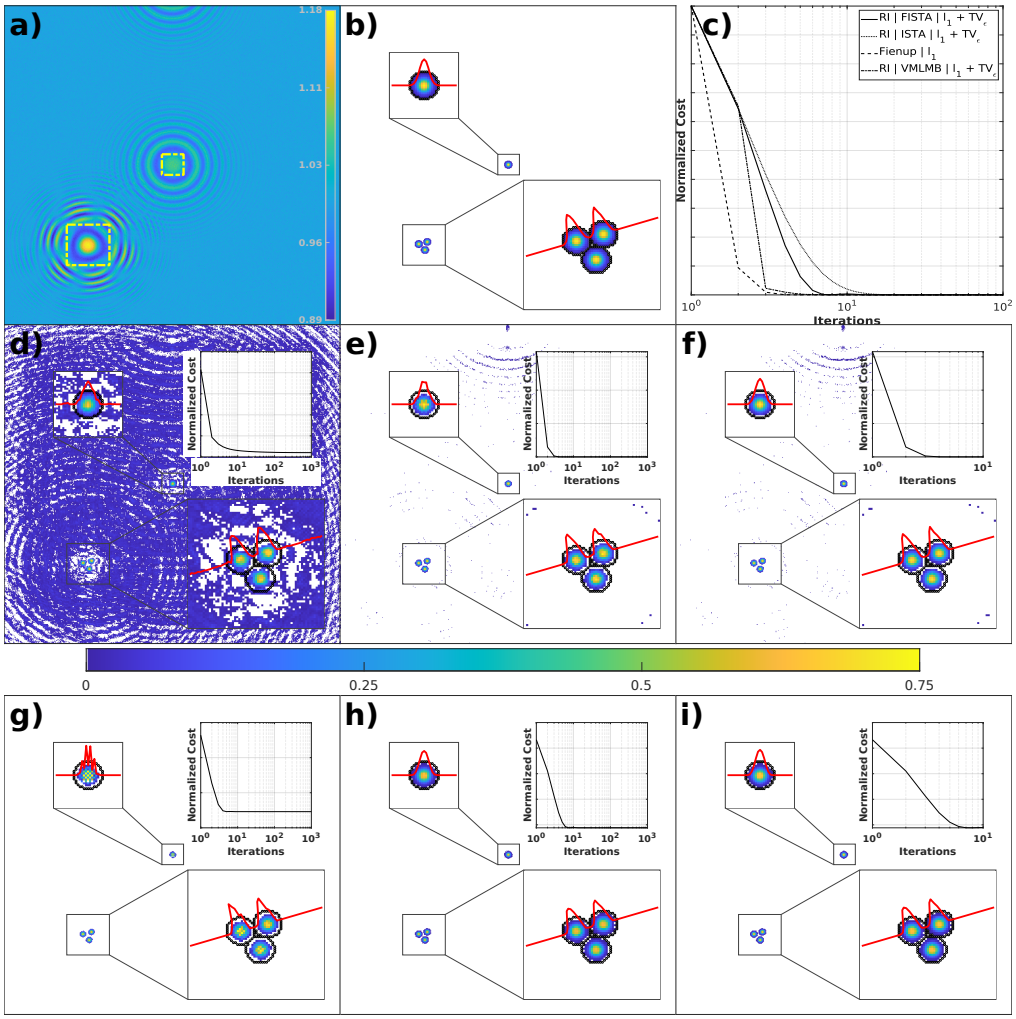


Fig. 6. Phase image reconstructed from simulated data with different methods and regularization terms. a) Data (in-line hologram). The yellow frames indicate the regions of interest that are extracted from the field of view for visualization. b) Ground truth phase. c) Evolution of the normalized cost criterion (minimum and maximum values ranged between 0 and 1) for each reconstruction. d-i) Reconstructed images using: d) Fienup method, e) Fienup method + soft-thresholding, f) Fienup method + soft-thresholding, stopped at 10 iterations, g) ri method using FISTA with soft-thresholding, h) RI method using FISTA with soft-thresholding + edge-preserving, i) RI method using FISTA with soft-thresholding + edge-preserving, stopped at 10 iterations. For each reconstruction, a positivity constraint is imposed to the solution. The values of the hyperparameters are indicated in Tab. (2). Black dots delimit the contour of the ground truth image. Red curves show line profiles passing through particles of the image.

(see Fig. 9):

$$\begin{aligned}\boldsymbol{o}^{(i+1)} &= \mathcal{T}_{\mu_{l_1} t}\left(\boldsymbol{o}^{(i)} - t\left(\nabla\mathcal{J}_{\text{fid}}(c^*(\boldsymbol{o}^{(i)}), \boldsymbol{o}^{(i)}, \boldsymbol{d}) \;+\; \nabla\mathcal{J}_{\text{TV}_\epsilon}(\boldsymbol{o}^{(i)}, \mu_{\text{TV}})\right)\right) \\ &= \mathcal{T}_{\mu_{l_1} t}\left(\boldsymbol{o}^{(i)} - 2\,t\,c^*(\boldsymbol{o}^{(i)})\,\mathbf{G}_z^{\text{T}}\mathbf{W}\left(c^*\,\tilde{\boldsymbol{m}}(\boldsymbol{o}^{(i)}) - \boldsymbol{d}\right) - t\,\mu_{\text{TV}}\sum_q \nabla_q^{\text{T}}\left(\frac{\nabla_q \boldsymbol{o}^{(i)}}{\sqrt{||\nabla_q \boldsymbol{o}^{(i)}||_2^2 + \epsilon^2}}\right)\right).\end{aligned} \tag{28}$$

The reconstruction via this combination of regularizations is shown in Fig. (6)(h), where a clear benefit of considering both a sparsity constraint and an edge-preserving smoothing constraint is observed, as already demonstrated in Fig. (3)(f). Using the inverse problems interpretation of the Fienup method, we could also adapt Fienup's algorithm to include both regularizations.

In Fig. (6)(c), we show the evolution over the iterations of the global (normalized) criterion of the RI and Fienup methods, respectively the criteria Eq. (27) and Eq. (26). We illustrate the convergence of the RI method with several optimization strategies: FISTA, ISTA, and VMLMB [64, 65], a quasi-Newton gradient Variable Metric method with Limited Memory requirements and possibly Bound constraints enforcement on the unknowns. Whatever the method, an empirical convergence of the criteria values is reached in less than ten iterations. This is confirmed in the reconstructions Fig. (6)(f,i) which show the same reconstructions as Fig. (6)(e,h) after only 10 iterations. We see that both the reconstructions (ri and Fienup) are almost the same as the solution obtained after 1000 iterations.

Figure 7 shows reconstructions of the experimental hologram data, using the two methods presented above with different parameters. The field of view has been cropped to 1080×1080 images. For each reconstructed image in Fig. (7)(c-h), the values of the hyperparameters are given in Tab. (2). In Fig. (7)(b), the evolution over the iterations of the global normalized criterion of the methods is shown.

The observations and analysis that can be performed on these reconstructions match those made on the simulated data: the objects of interest are clearly detected. Moreover, the estimation of their shape, diameter (see Fig. (7)(h)) and phase-shift are coherent with the expected values. Indeed, from the experimental parameters in Tab. (1), the maximum phase-shift should be around 0.83rad. Again, we notice that the convergence of the criteria is almost reached in a few tenth of iterations.

6. Discussion

As pointed out in Sec. 5.1, the linear model derived in this paper is only an approximation. Nevertheless, we observe, both in the simulated and experimental results, that the shape of the beads is correctly retrieved and that the phase-shift values estimated remain close to the expected values (the difference is less than one order of magnitude). The values obtained on the experimental holograms even match the expected phase-shift at the center of the beads ($\Delta\varphi \approx 0.83$rad).

This illustrates that the regularizations and constraints introduced to solve the inverse problem balance the modeling errors and lead to a satisfying solution: the reconstruction is robust to measurement and modeling errors.

In Sec. 5.1, we mentioned that an out-of-field particle was considered in the simulation, leading to cropped diffraction fringes in the upper part of the data (*cf.* Fig. (8)(a)). The inverse problems framework makes it possible to account for this data truncation (a problem also frequently faced in X-ray computerized tomography [66]) and reconstruct a wider-field of view. This can be done by two equivalent ways: (i) an extension of the size of both the reconstructed object plane and the measured hologram; (ii) a field extension of only the reconstructed object plane. In the first case, unmeasured pixels (i.e., pixels outside the field-of-view of the sensor) are given a weight $w_q = 0$ in Eq. (9). In the second case, the model is rewritten to include a truncation operator (i.e., a rectangular matrix obtained by chopping off rows corresponding to unmeasured pixels):

$$\tilde{\boldsymbol{m}}(\boldsymbol{o}) = \mathbf{T}\ (\mathbf{1} + \mathbf{G}_z \boldsymbol{o}) \tag{29}$$

Figure 8 shows reconstructions of the simulated hologram with the RI method, where the object field of view is doubled compared to the data field of view. These reconstructions illustrate that the out-of-field particle can be detected. Since a major part of the diffraction pattern of this bead is missing, the reconstruction is not as good as for other beads. If a proper regularization is used,

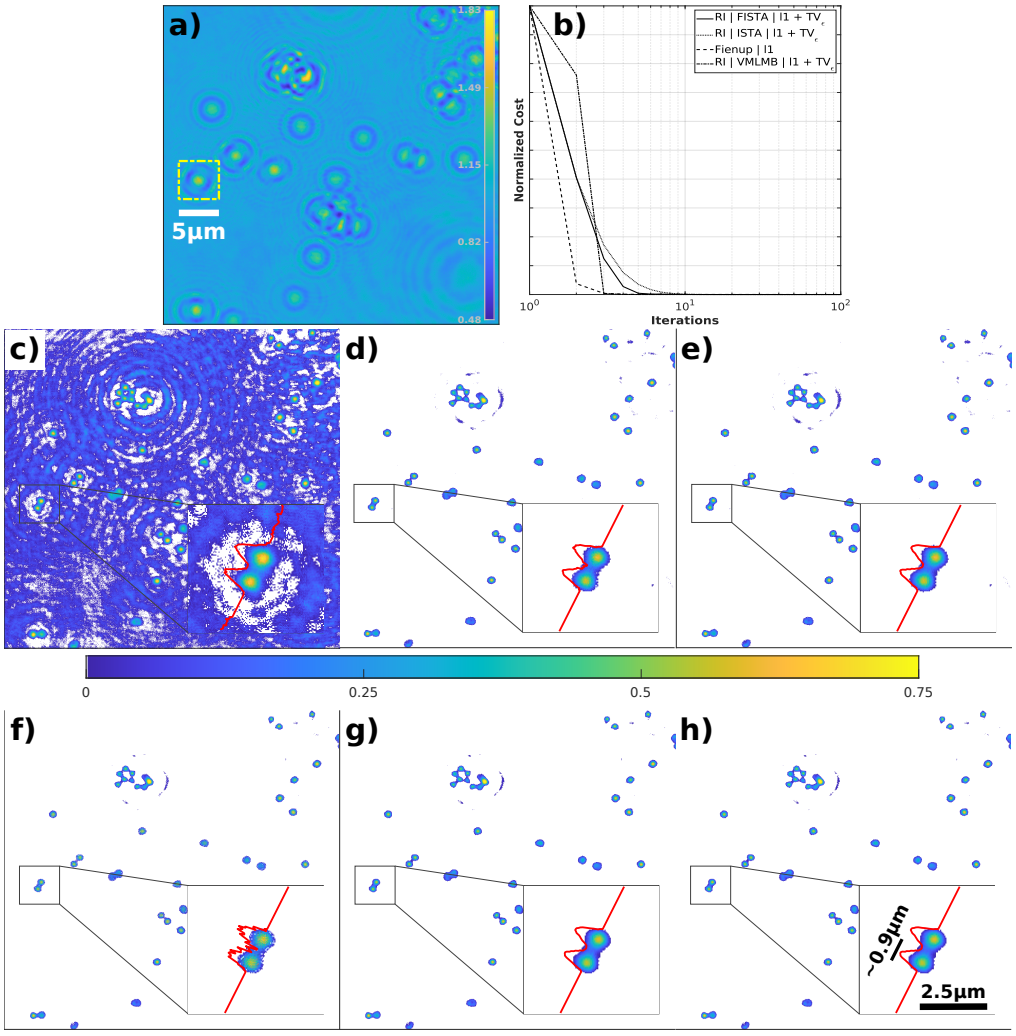


Fig. 7. Phase image reconstructed from experimental data with different methods and regularization terms. a) Data (in-line hologram). The yellow frame indicates the region of interest that is extracted from the field of view for visualization. b) Evolution of the normalized cost criterion (minimum and maximum values ranged between 0 and 1) for each reconstruction. c-h) Reconstructed images using: c) Fienup method, d) Fienup method + soft-thresholding, e) Fienup method + soft-thresholding, stopped at 10 iterations, f) RI method using FISTA with soft-thresholding, g) RI method using FISTA with soft-thresholding + edge-preserving, h) RI method using FISTA with soft-thresholding + edge-preserving, stopped at 10 iterations. For each reconstruction, a positivity constraint is imposed to the solution. The values of the hyperparameters are indicated in Tab. (2). Red curves show a line profile passing through two particles of the image.

the bead can be detected (a "too large" sparsity regularization would tend to suppress it). Here, the signal-to-noise ratio and the approximate direct model lead to an imperfect reconstruction of the morphology of the bead (asymmetrical shape). Other "reconstruction" approaches based on a parametric model of the spherical beads coupled with an adequate inversion algorithm (e.g., continuous matching pursuit algorithm) lead to accurate estimations even out of the field of

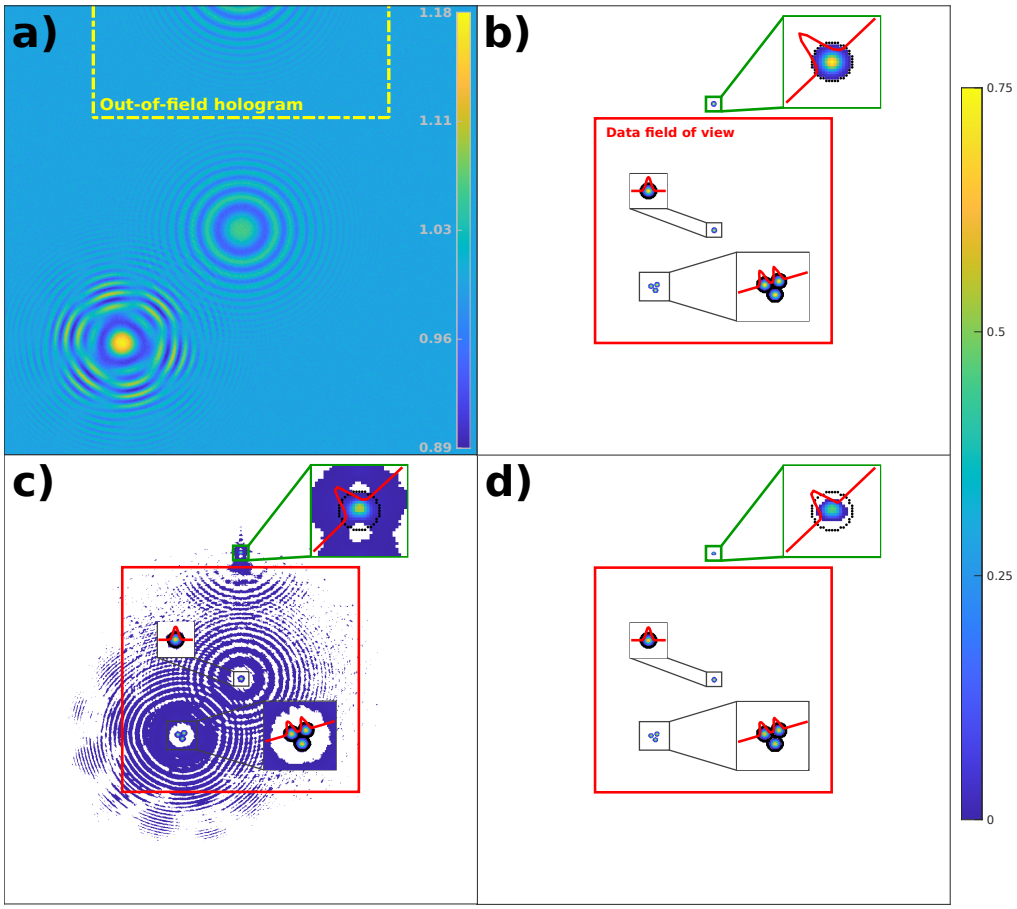


Fig. 8. Reconstruction (detection) of an out-of-field particle using RI method and field of view extension. a) Data in-line hologram highlighting (in yellow) a part of the out-of-field hologram. b) Ground truth image. c,d) Reconstructions with two different sets of regularization weights values (*cf.* Tab. (2)). The red frames show the initial data field of view (pixels seen by the detector). The green frames are zooms on the out-of-field particle.

view [19].

These discussions give an overview of the refinements made possible by a properly built inverse approach to enhance the quality of reconstructions and extract the most relevant information from the data.

7. Conclusion

In this work, we have presented the inverse problems methodology applied to the reconstruction of the phase information from in-line intensity holograms. The main goal was to provide a tutorial for this reconstruction strategy, in comparison with the standard alternating projections method proposed by Fienup. To this end, we have presented the overall methodology for building an inverse approach dedicated to the targeted application, from the modeling of data formation to the choice of suitable constraints and regularizations. Deriving a reconstruction algorithm from a cost function is straightforward, for classical regularization terms, by following a decomposition into smooth and non-smooth components, as summarized in figure 9. We have shown that Fienup's

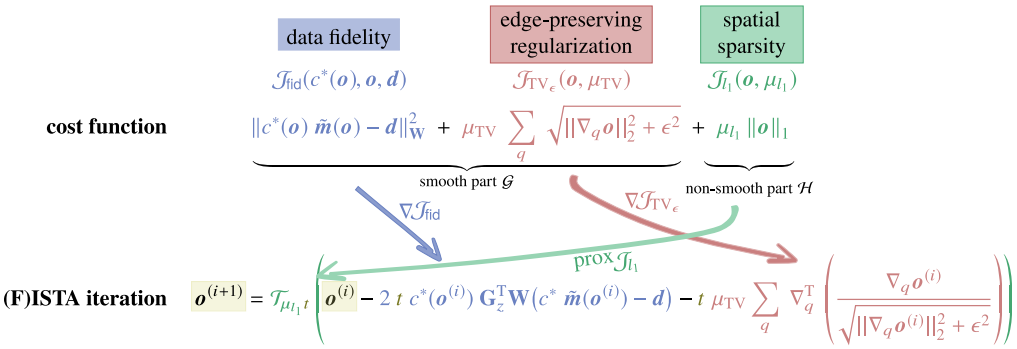


Fig. 9. Derivation of a reconstruction algorithm from a cost function.

method is analogous to a particular formulation of this phase retrieval problem when solved by proximal gradient descent iterations, i.e., the ISTA algorithm. We have shown reconstructions from our proposed RI method compared with standard and upgraded versions of Fienup's method, and for which we have provided algorithmic details of implementation. The analysis of the results has shown that the reconstruction quality can be quite similar when the Fienup method is refined using the inverse problems framework. We hope that we have convinced the reader that developing a reconstruction strategy following the inverse problems methodology can bridge a gap in reachable reconstruction quality, while allowing much more flexibility to extract relevant information from the data, at no cost in terms of implementation efforts and computational burden. We think that this unifying point of view on various approaches to hologram reconstruction will help the cross-fertilization of algorithmic ideas.

A. Appendix: detailed implementations of Algo. (1) and Algo. (2)

In order to facilitate the derivation of the equations and the algorithms, we used the formalism of linear algebra in the main body of the paper. While it simplifies the expressions, it also makes the implementation of the algorithms less straightforward. In this appendix, we rewrite the algorithms to explicitly define which operations are involved (pixel by pixel operation, 2D discrete convolution).

The image of the object plane o (or $\underline{o}$) that has to be reconstructed, as well as intermediate images such as the propagated wave $\underline{a}_z$, the convolution kernels g_z (or $\underline{h}_z$ and $\underline{h}_{-z}$), and others, are implemented as 2D arrays of size $N_x \times N_y$. In the absence of field extension or pixel super-resolution, the number of pixels $N_x \times N_y$ in the object plane is equal to the number of pixels in the measured data. We define each 2D array x as follows: $x \in \mathbb{T}^{N_x \times N_y}$, $\mathbb{T}$ being the real domain $\mathbb{R}$ or the complex domain $\mathbb{C}$. We keep the notation x_q to represent the q-th pixel of image x. A 2D discrete convolution $h * x$ with a kernel h is typically computed in the Fourier space

using fast Fourier transforms (FFTs) and adequate 0-padding to prevent periodization artifacts.

Algorithm 3: Detailed implementation of Algo. (1)

Input:
$d \in \mathbb{R}^{N_x \times N_y}$; {intensity measurements}
$\underline{h}_z \in \mathbb{C}^{N_x \times N_y}$; {propagation kernel $\underline{h}_z$ }
$\underline{h}_{-z} \in \mathbb{C}^{N_x \times N_y}$; {backpropagation kernel $\underline{h}_{-z}$ }
$\Lambda rmin \in \mathbb{R}^{N_x \times N_y}$; {minimum bound constraint for each pixel of $\mathfrak{Re}(\underline{o})$ (feasible domain $\mathbb{O}$)}
$\Lambda rmax \in \mathbb{R}^{N_x \times N_y}$; {maximum bound constraint for each pixel of $\mathfrak{Re}(\underline{o})$ (feasible domain $\mathbb{O}$)}
$\Lambda imin \in \mathbb{R}^{N_x \times N_y}$; {minimum bound constraint for each pixel of $\mathfrak{Im}(\underline{o})$ (feasible domain $\mathbb{O}$)}
$\Lambda imax \in \mathbb{R}^{N_x \times N_y}$; {maximum bound constraint for each pixel of $\mathfrak{Im}(\underline{o})$ (feasible domain $\mathbb{O}$)}
$maxiter$; {maximum number of iterations}
Output:
$\underline{o} \in \mathbb{C}^{N_x \times N_y}$; {unknown deviation from unit transmittance $\underline{t}$ }
Allocations:
$\underline{a} \in \mathbb{C}^{N_x \times N_y}$; {array for storing the simulated diffracted wave a_z }

begin
{Step 0: initializations}
1 $d \leftarrow \text{normalize}(d)$; {normalize the data hologram so that the background equals 1}
{N.B.: for instance divide d by its mean or its median (almost valid in case of a low-density distribution of objects).}
2 $\underline{o} \leftarrow \underline{h}_{-z} * (\sqrt{d} - 1)$; {First guess: for instance direct backpropagation of the data (or initialization with random values).}
{N.B.: the square root $\sqrt{}$ is applied pixelwise.}
for $i \leftarrow 1 : maxiter$ **do**
3 $\underline{a} \leftarrow 1 + \underline{h}_z * \underline{o}$; {Step 1: propagation to the sensor plane}
for $q \leftarrow 1 : N_x \cdot N_y$ **do** {Step 2: enforce the measured amplitude at sensor plane}
if $\underline{a}_q \neq 0$ **then**
4 $\underline{a}_q \leftarrow \sqrt{d_q} \cdot \left(\underline{a}_q / |\underline{a}_q|\right)$;
else
5 $\underline{a}_q \leftarrow 0$;
end
end
6 $\underline{o} \leftarrow \underline{h}_{-z} * (\underline{a} - 1)$; {Step 3: backpropagation to the sample plane}
for $q \leftarrow 1 : N_x \cdot N_y$ **do** {Step 4: projection on the domain $\mathbb{O}$}
7 $\mathfrak{Re}(\underline{o})_q \leftarrow \min\left(\mathfrak{Re}(\underline{o})_q, \Lambda rmax_q\right)$;
8 $\mathfrak{Re}(\underline{o})_q \leftarrow \max\left(\mathfrak{Re}(\underline{o})_q, \Lambda rmin_q\right)$;
9 $\mathfrak{Im}(\underline{o})_q \leftarrow \min\left(\mathfrak{Im}(\underline{o})_q, \Lambda imax_q\right)$;
10 $\mathfrak{Im}(\underline{o})_q \leftarrow \max\left(\mathfrak{Im}(\underline{o})_q, \Lambda imin_q\right)$;
{N.B.: The operators $\mathfrak{Re}()$ and $\mathfrak{Im}()$, giving respectively the real and imaginary parts (arrays) of the array $\underline{o}$, are applied pixelwise.}
end
end
end

Algorithm 4: Detailed implementation of Algo. (2)

Input:
$d \in \mathbb{R}^{N_x \times N_y}$ {intensity measurements}
$g_z \in \mathbb{R}^{N_x \times N_y}$; {propagation kernel g_z }
$w \in \mathbb{R}^{N_x \times N_y}$; {confidence weights applied to each pixel data (these array's values constitutes to the diagonal of the matrix $\mathbf{W}$ in Eq. (9))}
$flag_{pos}$; {flag (***true*** or ***false***) for enforcing a positivity constraint}
μ ; {hyperparameter value for the soft-thresholding (sparsity contraint)}
t ; {steepest gradient descent step length}
Output:
$o \in \mathbb{R}^{N_x \times N_y}$ {unknown deviation from unit transmittance $\underline{t}$ }
Allocations:
$o_{prev} \in \mathbb{R}^{N_x \times N_y}$; {estimate at previous iterate}
$\tilde{m} \in \mathbb{R}^{N_x \times N_y}$; {array for storing the direct model}
$r \in \mathbb{R}^{N_x \times N_y}$; {array for storing the residues (difference between the model and the data)}
$u \in \mathbb{R}^{N_x \times N_y}$; {intermediate array}
$c \in \mathbb{R}$; {intensity hologram scaling factor}
$s \in \mathbb{R}$; {scalar factor for the acceleration step}
$s_{prev} \in \mathbb{R}$; {previous scalar factor for the acceleration step}

begin
{Step 0: initializations}
1 $o_{prev} \leftarrow \text{zeros}(N_x, N_y)$; {First guess: for instance $\text{zeros}(N_x, N_y)$ returns a $N_x \times N_y$ array of zeros.}
2 $u \leftarrow o_{prev}$;
3 $s_{prev} \leftarrow 1$;
repeat
4 $\tilde{m} \leftarrow 1 + [g_z * u]$; {Step 1: calculate the direct model}
5 $c \leftarrow \sum_q w_q \tilde{m}_q d_q \;/\; \sum_q w_q \tilde{m}_q^2$; {Step 2: get optimal current scaling factor}
for $q \leftarrow 1 : \; N_x \cdot N_y$ **do**
6 $r_q \leftarrow w_q (c\, \tilde{m}_q - d_q)$; {Step 3: get the weighted residual pixel values}
end
7 $r \leftarrow g_z * r$; {Step 4: backpropagation of the residues}
{N.B.: the backpropagation kernel is equal to the propagation kernel (*cf.* Sec. 4.1)}
for $q \leftarrow 1 : \; N_x \cdot N_y$ **do**
8 $o_q \leftarrow u_q - 2\, t\, c\, r_q$; {Step 5: steepest gradient descent step}
if $flag_{pos} =$ ***true*** **then** {Step 6: soft-thresholding}
9 $o_q \leftarrow \max\left(0, o_q - \mu t\right)$;
else
10 $o_q \leftarrow \text{sign}(o_q) \max\left(0, |o_q| - \mu t\right)$;
end
end
11 $s \leftarrow \frac{1}{2}\left(1 + \sqrt{1 + 4\,(s_{prev})^2}\right)$; {Step 7: get new scalar factor s for the acceleration step}
12 $u \leftarrow o + \frac{s_{prev} - 1}{s}\left(o - o_{prev}\right)$; {Step 8: new intermediate array u according to previous iterate}
{N.B.: this step is applied pixelwise.}
13 $s_{prev} \leftarrow s$;
14 $o_{prev} \leftarrow o$; {Step 9: update memory of previous values}
until *convergence*;
end

In Algo. (4), if the scalar factor s is kept to the value 1, we fall back to the simpler ISTA algorithm.

Funding

Région Auvergne-Rhône-Alpes. French National Research Agency (ANR) (ANR-11-LABX-0063, ANR-11-IDEX-0007).

Acknowledgements

The authors would like to warmly thank the anonymous Reviewers for their careful reading and their numerous comments and suggestions which helped to significantly improve the paper.

This work has been supported in part by the project DIAGHOLO, funded by "Région Auvergne-Rhône-Alpes". It was also performed within the framework of the LABEX PriMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

The algorithmic tools (optimization strategies, models, regularizations) presented in this work have been implemented within the framework of the Matlab library GlobalBioIm [67,68] (https://biomedical-imaging-group.github.io/GlobalBioIm/index.html).

Disclosures

The authors declare no conflicts of interest.

References

1. D. GABOR, "A New Microscopic Principle," Nature **161**, 777–778 (1948).
2. V. Micó, J. Zheng, J. Garcia, Z. Zalevsky, and P. Gao, "Resolution enhancement in quantitative phase microscopy," Adv. Opt. Photonics **11**, 135–214 (2019).
3. R. Gerchberg and W. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972).
4. J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. **3**, 27–29 (1978).
5. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).
6. H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Hybrid projection–reflection method for phase retrieval," JOSA A **20**, 1025–1034 (2003).
7. V. Elser, "Solution of the crystallographic phase problem by iterated projections," Acta Crystallogr. Sect. A: Foundations Crystallogr. **59**, 201–209 (2003).
8. D. R. Luke, "Relaxed averaged alternating reflections for diffraction imaging," Inverse Probl. **21**, 37–50 (2005).
9. S. Marchesini, "Invited Article: A unified evaluation of iterative projection algorithms for phase retrieval," Rev. Sci. Instruments **78**, 011301 (2007).
10. R. A. Dilanian, G. J. Williams, L. W. Whitehead, D. J. Vine, A. G. Peele, E. Balaur, I. McNulty, H. M. Quiney, and K. A. Nugent, "Coherent diffractive imaging: a new statistically regularized amplitude constraint," New J. Phys. **12**, 093042 (2010).
11. J. A. Rodriguez, R. Xu, C.-C. Chen, Y. Zou, and J. Miao, "Oversampling smoothness: an effective algorithm for phase retrieval of noisy diffraction intensities," J. Appl. Crystallogr. **46**, 312–318 (2013).
12. F. Soulez, e. Thiébaut, A. Schutz, A. Ferrari, F. Courbin, and M. Unser, "Proximity operators for phase retrieval," Appl. Opt. **55**, 7412–7421 (2016).
13. T. Latychevskaia and H.-W. Fink, "Solution to the Twin Image Problem in Holography," Phys. Rev. Lett. **98**, 233901 (2007).
14. M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, "Compressive phase retrieval," in *Wavelets XII,* vol. 6701 (International Society for Optics and Photonics, 2007), p. 670120.
15. S. Mukherjee and C. S. Seelamantula, "An iterative algorithm for phase retrieval with sparsity constraints: application to frequency domain optical coherence tomography," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (2012), pp. 553–556.
16. Y. Rivenson, Y. Wu, H. Wang, Y. Zhang, A. Feizi, and A. Ozcan, "Sparsity-based multi-height phase recovery in holographic microscopy," Sci. Reports **6**, 37862 (2016).
17. F. Jolivet, F. Momey, L. Denis, L. Méès, N. Faure, N. Grosjean, F. Pinston, J.-L. Marié, and C. Fournier, "Regularized reconstruction of absorbing and phase objects from a single in-line hologram, application to fluid mechanics and micro-biology," Opt. Express **26**, 8923–8940 (2018).
18. A. Berdeu, O. Flasseur, L. Méès, L. Denis, F. Momey, T. Olivier, N. Grosjean, and C. Fournier, "Reconstruction of in-line holograms: combining model-based and regularized inversion," Opt. Express **27**, 14951–14968 (2019).
19. F. Soulez, L. Denis, E. Thiébaut, C. Fournier, and C. Goepfert, "Inverse problem approach in particle digital holography: out-of-field particle detection made possible," JOSA A **24**, 3708–3716 (2007).

20. J. W. Goodman, *Introduction to Fourier Optics* (Roberts and Company Publishers, 2005). Google-Books-ID: ow5xs_Rtt9AC.
21. E. Wolf, "Three-dimensional structure determination of semi-transparent objects from holographic data," Opt. Commun. **1**, 153–156 (1969).
22. G. Mie, "Beiträge zur optik trüber medien, speziell kolloidaler metallösungen," Annalen der physik **330**, 377–445 (1908).
23. C. Fournier, F. Jolivet, L. Denis, N. Verrier, E. Thiebaut, C. Allier, and T. Fournel, "Pixel super-resolution in digital holography by regularized reconstruction," Appl. Opt. **56**, 69–77 (2017).
24. O. Flasseur, F. Jolivet, F. Momey, L. Denis, and C. Fournier, "Improving color lensless microscopy reconstructions by self-calibration," in *Unconventional Optical Imaging,* vol. 10677 (International Society for Optics and Photonics, 2018), p. 106771A.
25. Y. Cotte, F. Toy, P. Jourdain, N. Pavillon, D. Boss, P. Magistretti, P. Marquet, and C. Depeursinge, "Marker-free phase nanoscopy," Nat. Photonics **7**, 113–117 (2013).
26. J. Bailleul, B. Simon, M. Debailleul, L. Foucault, N. Verrier, and O. Haeberlé, "Tomographic diffractive microscopy: Towards high-resolution 3-D real-time data acquisition, image reconstruction and display of unlabeled samples," Opt. Commun. **422**, 28–37 (2018).
27. S. P. Hau-Riege, H. Szoke, H. N. Chapman, A. Szoke, S. Marchesini, A. Noy, H. He, M. Howells, U. Weierstall, and J. C. H. Spence, "SPEDEN: reconstructing single particles from their diffraction patterns," Acta Crystallogr. Sect. A: Foundations Crystallogr. **60**, 294–305 (2004).
28. S. Sotthivirat and J. A. Fessler, "Penalized-likelihood image reconstruction for digital holography," JOSA A **21**, 737–750 (2004).
29. L. Denis, D. Lorenz, E. Thiébaut, C. Fournier, and D. Trede, "Inline hologram reconstruction with sparsity constraints," Opt. Lett. **34**, 3475–3477 (2009).
30. D. J. Brady, K. Choi, D. L. Marks, R. Horisaki, and S. Lim, "Compressive Holography," Opt. Express **17**, 13040–13049 (2009).
31. Y. Rivenson, A. Stern, and B. Javidi, "Compressive Fresnel Holography," J. Disp. Technol. **6**, 506–509 (2010).
32. Y. Shechtman, A. Beck, and Y. C. Eldar, "GESPAR: Efficient Phase Retrieval of Sparse Signals," IEEE Transactions on Signal Process. **62**, 928–938 (2014).
33. A. Repetti, E. Chouzenoux, and J. Pesquet, "A nonconvex regularized approach for phase retrieval," in *2014 IEEE International Conference on Image Processing (ICIP),* (2014), pp. 1753–1757.
34. A. Drémeau and F. Krzakala, "Phase recovery from a Bayesian point of view: The variational approach," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (2015), pp. 3661–3665.
35. A. M. Tillmann, Y. C. Eldar, and J. Mairal, "Dictionary learning from phaseless measurements," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (2016), pp. 4702–4706.
36. J. Song, C. L. Swisher, H. Im, S. Jeong, D. Pathania, Y. Iwamoto, M. Pivovarov, R. Weissleder, and H. Lee, "Sparsity-Based Pixel Super Resolution for Lens-Free Digital In-line Holography," Sci. Reports **6**, 24681 (2016).
37. A. Berdeu, F. Momey, B. Laperrousaz, T. Bordy, X. Gidrol, J.-M. Dinten, N. Picollet-D'hahan, and C. Allier, "Comparative study of fully three-dimensional reconstruction algorithms for lens-free microscopy," Appl. Opt. **56**, 3939–3951 (2017).
38. F. Soulez, L. Denis, C. Fournier, E. Thiébaut, and C. Goepfert, "Inverse-problem approach for particle digital holography: accurate location based on local optimization," JOSA A **24**, 1164–1171 (2007).
39. O. Flasseur, L. Denis, C. Fournier, and E. Thiébaut, "Robust object characterization from lensless microscopy videos," in *2017 25th European Signal Processing Conference (EUSIPCO),* (IEEE, 2017), pp. 1445–1449.
40. B. Liu and N. C. Gallagher, "Convergence of a Spectrum Shaping Algorithm," Appl. Opt. **13**, 2470–2471 (1974).
41. H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization," JOSA A **19**, 1334–1345 (2002).
42. D. Noll and A. Rondepierre, "On Local Convergence of the Method of Alternating Projections," Foundations Comput. Math. **16**, 425–455 (2016).
43. J. Miao, D. Sayre, and H. N. Chapman, "Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects," JOSA A **15**, 1662–1669 (1998).
44. J. Miao, P. Charalambous, J. Kirz, and D. Sayre, "Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens," Nature **400**, 342–344 (1999).
45. Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase Retrieval with Application to Optical Imaging: A contemporary overview," IEEE Signal Process. Mag. **32**, 87–109 (2015).
46. H. N. Chapman and K. A. Nugent, "Coherent lensless X-ray imaging," Nat. Photonics **4**, 833–839 (2010).
47. W. Bishara, T.-W. Su, A. F. Coskun, and A. Ozcan, "Lensfree on-chip microscopy over a wide field-of-view using pixel super-resolution," Opt. Express **18**, 11181–11191 (2010).
48. Y. Wu and A. Ozcan, "Lensless digital holographic microscopy and its applications in biomedicine and environmental monitoring," Methods **136**, 4–16 (2018).
49. T. Latychevskaia, "Iterative phase retrieval in coherent diffractive imaging: practical issues," Appl. Opt. **57**, 7187–7197 (2018).
50. A. Levi and H. Stark, "Image restoration by the method of generalized projections with application to restoration from magnitude," JOSA A **1**, 932–943 (1984).

51. R. Horisaki, Y. Ogura, M. Aino, and J. Tanida, "Single-shot phase imaging with a coded aperture," Opt. Lett. **39**, 6466–6469 (2014).
52. Z. Wang, Q. Dai, D. Ryu, K. He, R. Horstmeyer, and A. Katsaggelos, "Dictionary-based phase retrieval for space-time super resolution using lens-free on-chip holographic video," in *Imaging and Applied Optics 2017 (3D, AIO, COSI, IS, MATH, pcAOP) (2017), paper CTu2B.3,* (Optical Society of America, 2017), p. CTu2B.3.
53. F. Eilenberger, S. Minardi, D. Pliakis, and T. Pertsch, "Digital holography from shadowgraphic phase estimates," Opt. Lett. **37**, 509–511 (2012).
54. O. Flasseur, L. Denis, E. Thiébaut, T. Olivier, and C. Fournier, "ExPACO: detection of an extended pattern under nonstationary correlated noise by patch covariance modeling," in *EUSIPCO 2019,* (Coruna, Spain, 2019).
55. A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation* (SIAM, 2005).
56. A. Ribes and F. Schmitt, "Linear inverse problems in imaging," IEEE Signal Process. Mag. **25**, 84–99 (2008).
57. J. A. Fessler, "Penalized weighted least-squares image reconstruction for positron emission tomography," IEEE Transactions on Med. Imaging **13**, 290–300 (1994).
58. L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Phys. D: Nonlinear Phenom. **60**, 259–268 (1992).
59. P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," IEEE Transactions on Image Process. **6**, 298–311 (1997).
60. N. Parikh, S. Boyd *et al.*, "Proximal algorithms," Foundations Trends Optim. **1**, 127–239 (2014).
61. A. Beck and M. Teboulle, "Fast Gradient-Based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems," IEEE Transactions on Image Process. **18**, 2419–2434 (2009).
62. I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," Commun. on Pure Appl. Math. A J. Issued by Courant Inst. Math. Sci. **57**, 1413–1457 (2004).
63. G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," IEEE Transactions on Inf. Theory **64**, 773–794 (2017).
64. J. Nocedal, "Updating quasi-Newton matrices with limited storage," Math. Comput. **35**, 773–782 (1980).
65. E. Thiebaut, "Optimization issues in blind deconvolution algorithms," in *Astronomical Data Analysis II,* vol. 4847 (International Society for Optics and Photonics, 2002), pp. 174–183.
66. M. Defrise, F. Noo, R. Clackdoyle, and H. Kudo, "Truncated Hilbert transform and image reconstruction from limited tomographic data," Inverse Probl. **22**, 1037 (2006).
67. M. Unser, E. Soubies, F. Soulez, M. McCann, and L. Donati, "GlobalBioIm: A Unifying Computational Framework for Solving Inverse Problems," in *Imaging and Applied Optics 2017 (3D, AIO, COSI, IS, MATH, pcAOP) (2017), paper CTu1B.1,* (Optical Society of America, 2017), p. CTu1B.1.
68. E. Soubies, F. Soulez, M. T. McCann, T.-a. Pham, L. Donati, T. Debarre, D. Sage, and M. Unser, "Pocket guide to solve inverse problems with GlobalBioIm," Inverse Probl. **35**, 104006 (2019).