Supplementary information for

Iterative projection meets sparsity regularization: towards practical single-shot quantitative phase imaging with in-line holography

Yunhui Gao[1] and Liangcai Cao[1*]

[1]State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instruments, Tsinghua University, Beijing, 100084, China.

*Corresponding author: clc@tsinghua.edu.cn

1 Proof of duality

The CCTV-regularized denoising problem can be equivalently expressed as a constrained optimization problem as follows:

$$\min_{\boldsymbol{x}} \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{u}\|_1 + I_C(\boldsymbol{x}), \quad \text{subject to} \quad \boldsymbol{u} = \boldsymbol{D}\boldsymbol{x}, \tag{S1}$$

where $\boldsymbol{u} \in \mathbb{C}^{2n}$ is an auxiliary variable. Notice that the problem of Eq. (S1) is a convex optimization problem with an affine equality constraint, for which strong duality holds [1]. The Lagrangian is given by

$$L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w}) = \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{u}\|_1 + I_C(\boldsymbol{x}) + \mathrm{Re}(\langle \boldsymbol{w}, \boldsymbol{D}\boldsymbol{x}-\boldsymbol{u}\rangle), \tag{S2}$$

where $\boldsymbol{w} \in \mathbb{C}^{2n}$ is the dual variable, $\langle\cdot,\cdot\rangle$ denotes the inner product of two vectors, and $\mathrm{Re}(\cdot)$ extracts the real part of a complex number. The Lagrange dual function, by definition, is

$$\begin{aligned}
&\inf_{\boldsymbol{x},\boldsymbol{u}} L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w}) \\
&= \inf_{\boldsymbol{x},\boldsymbol{u}} \left\{ \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{u}\|_1 + I_C(\boldsymbol{x}) + \mathrm{Re}(\langle \boldsymbol{w}, \boldsymbol{D}\boldsymbol{x}-\boldsymbol{u}\rangle) \right\} \\
&= \inf_{\boldsymbol{x}\in C} \left\{ \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \mathrm{Re}(\langle \boldsymbol{w}, \boldsymbol{D}\boldsymbol{x}\rangle) \right\} + \inf_{\boldsymbol{u}} \left\{ \lambda\|\boldsymbol{u}\|_1 - \mathrm{Re}(\langle \boldsymbol{w}, \boldsymbol{u}\rangle) \right\} \\
&\overset{(a)}{=} \inf_{\boldsymbol{x}\in C} \left\{ \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \mathrm{Re}(\langle \boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}, \boldsymbol{x}\rangle) \right\} - I_S(\boldsymbol{w}) \\
&= \inf_{\boldsymbol{x}\in C} \left\{ \frac{1}{2\gamma}\left\|\boldsymbol{x}-(\boldsymbol{v}-\gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w})\right\|_2^2 \right\} + \frac{1}{2\gamma}\|\boldsymbol{v}\|_2^2 - \frac{1}{2\gamma}\left\|\boldsymbol{v}-\gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}\right\|_2^2 - I_S(\boldsymbol{w}) \\
&= \left\|\mathcal{H}_C(\boldsymbol{v}-\gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w})\right\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{v}\|_2^2 - \frac{1}{2\gamma}\left\|\boldsymbol{v}-\gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}\right\|_2^2 - I_S(\boldsymbol{w}),
\end{aligned} \tag{S3}$$

where (a) can be easily derived based on the fact that the convex conjugate of the ℓ_1 norm is the indicator function of $[-1,1]^n$ [1]. The last equality in Eq. (S3), together with strong duality, suggests that the primal optimal solution $\boldsymbol{x}^\star$ is related to the dual optimal solution $\boldsymbol{w}^\star$ via $\boldsymbol{x}^\star = \mathcal{P}_C(\boldsymbol{v}-\gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}^\star)$. The dual problem is to maximize the Lagrange dual function with respect to $\boldsymbol{w}$, which is equivalent to Eq. (7) in the main text.

2 Proof of convergence

2.1 Preliminaries

Since we are primarily dealing with real-valued functions over complex-valued variables, we adopt the CR-calculus as helpful mathematical tool for analysis. The CR-calculus extends the complex derivative to the general non-analytic functions. Readers may refer to Ref. [2] for a detailed introduction. The CR-calculus regards the complex variable $\boldsymbol{x}$ and its conjugate $\bar{\boldsymbol{x}}$ as independent variables. Thus, the fidelity function $F(\boldsymbol{x})$ should be interpreted as a function over the pair of conjugate vectors $\hat{\boldsymbol{x}} = [\boldsymbol{x}^{\mathsf{T}}, \bar{\boldsymbol{x}}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{2n}$. Nevertheless, to keep notations consistent, we still denote the function as $F(\boldsymbol{x})$. The same applies to other functions as well.

The followings are some intermediate results from matrix analysis, which are helpful for proving the convergence theorems below.

Lemma 1 [3] *Given matrices $\boldsymbol{P} \in \mathbb{C}^{n\times n}$, $\boldsymbol{Q} \in \mathbb{C}^{n\times n}$, and $\boldsymbol{R} \in \mathbb{C}^{n\times n}$. The following holds:*

1. $\boldsymbol{P} \succ \boldsymbol{Q} \Rightarrow \boldsymbol{R}^{\mathsf{H}}\boldsymbol{P}\boldsymbol{R} \succ \boldsymbol{R}^{\mathsf{H}}\boldsymbol{Q}\boldsymbol{R}$,
2. $\boldsymbol{P} \succeq \boldsymbol{Q} \succ \boldsymbol{0} \Rightarrow \boldsymbol{Q}^{-1} \succeq \boldsymbol{P}^{-1} \succ \boldsymbol{0}$.

Lemma 2 (Schur Complement [3]) *Given a $2n \times 2n$ Hermitian matrix:*

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} \end{pmatrix}, \tag{S4}$$

where each block is of size $n \times n$, and we have $\boldsymbol{P}_{11}^{\mathsf{H}} = \boldsymbol{P}_{11}$, $\boldsymbol{P}_{22}^{\mathsf{H}} = \boldsymbol{P}_{22}$, and $\boldsymbol{P}_{12}^{\mathsf{H}} = \boldsymbol{P}_{21}$. Then

$$\boldsymbol{P} \succ \boldsymbol{0} \Leftrightarrow \boldsymbol{P}_{11} \succ \boldsymbol{0} \quad \text{and} \quad \boldsymbol{P}_{22} - \boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1}\boldsymbol{P}_{12} \succ \boldsymbol{0}. \tag{S5}$$

Lemma 3 *Given a matrix $\boldsymbol{P} \in \mathbb{C}^{n \times n}$, and a scalar $\varepsilon > 0$,*

$$\boldsymbol{P}\left(\varepsilon\boldsymbol{I} + \boldsymbol{P}^{\mathsf{H}}\boldsymbol{P}\right)^{-1}\boldsymbol{P}^{\mathsf{H}} \prec \boldsymbol{I}. \tag{S6}$$

Proof Suppose the singular value decomposition of $\boldsymbol{P}$ is given by $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{H}}$, where $\boldsymbol{U} \in \mathbb{C}^{n \times n}$ and $\boldsymbol{V} \in \mathbb{C}^{n \times n}$ are unitary matrices, and $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})$ is a real-valued diagonal matrix. Then, we have

$$\varepsilon\boldsymbol{I} + \boldsymbol{P}^{\mathsf{H}}\boldsymbol{P} = \varepsilon\boldsymbol{I} + \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^{\mathsf{H}} = \boldsymbol{V}\operatorname{diag}\left(\varepsilon\boldsymbol{1} + \boldsymbol{\sigma}^2\right)\boldsymbol{V}^{\mathsf{H}}. \tag{S7}$$

That is, $\boldsymbol{P}^{\mathsf{H}}\boldsymbol{P}$ is diagonalizable with real-valued non-negative eigenvalues $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$. $\varepsilon\boldsymbol{I} + \boldsymbol{P}^{\mathsf{H}}\boldsymbol{P}$ is nonsingular and its inverse is given by

$$\left(\varepsilon\boldsymbol{I} + \boldsymbol{P}^{\mathsf{H}}\boldsymbol{P}\right)^{-1} = \boldsymbol{V}\operatorname{diag}\left(\frac{\boldsymbol{1}}{\varepsilon\boldsymbol{1} + \boldsymbol{\sigma}^2}\right)\boldsymbol{V}^{\mathsf{H}}. \tag{S8}$$

Thus, we arrive at the result:

$$\boldsymbol{P}\left(\varepsilon\boldsymbol{I} + \boldsymbol{P}^{\mathsf{H}}\boldsymbol{P}\right)^{-1}\boldsymbol{P}^{\mathsf{H}} = \boldsymbol{U}\operatorname{diag}\left(\frac{\boldsymbol{\sigma}^2}{\varepsilon\boldsymbol{1} + \boldsymbol{\sigma}^2}\right)\boldsymbol{U}^{\mathsf{H}} \prec \boldsymbol{U}\boldsymbol{U}^{\mathsf{H}} = \boldsymbol{I}. \tag{S9}$$

□

2.2 Convergence of the proximal gradient algorithm

The Wirtinger derivatives of $F(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ are given by [4]

$$\frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{1}{2}\left(|\boldsymbol{A}\boldsymbol{x}| - \boldsymbol{y}\right)^{\mathsf{H}}\operatorname{diag}\left(\frac{\overline{\boldsymbol{A}\boldsymbol{x}}}{|\boldsymbol{A}\boldsymbol{x}|}\right)\boldsymbol{A}, \tag{S10}$$

$$\frac{\partial F(\boldsymbol{x})}{\partial \bar{\boldsymbol{x}}} = \frac{1}{2}\left(|\boldsymbol{A}\boldsymbol{x}| - \boldsymbol{y}\right)^{\mathsf{T}}\operatorname{diag}\left(\frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|}\right)\bar{\boldsymbol{A}}. \tag{S11}$$

Let $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m]^{\mathsf{H}}$ where $\boldsymbol{a}_i \in \mathbb{C}^n$ denotes the i-th sampling vector. It should be noted that, the Wirtinger derivatives are not well-defined for $\boldsymbol{x} \in Z$ where Z is defined as

$$Z \stackrel{\text{def}}{=} \left\{\boldsymbol{x} \in \mathbb{C}^n : \exists\, 1 \leq i \leq M, \text{ s.t. } \boldsymbol{a}_i^{\mathsf{H}}\boldsymbol{x} = 0,\ \boldsymbol{a}_i \neq \boldsymbol{0}\right\}. \tag{S12}$$

For any $\boldsymbol{x} \in \mathbb{C}^n \backslash Z$, the complex Hessian is defined as

$$\nabla^2 F(\boldsymbol{x}) = \boldsymbol{H}_{\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}} = \begin{pmatrix} \boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}} & \boldsymbol{H}_{\bar{\boldsymbol{x}}\boldsymbol{x}} \\ \boldsymbol{H}_{\boldsymbol{x}\bar{\boldsymbol{x}}} & \boldsymbol{H}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} \end{pmatrix}, \tag{S13}$$

where the four second-order partial derivatives are calculated as follows:

$$
\begin{aligned}
\boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}} =& \frac{\partial}{\partial \boldsymbol{x}} \left(\frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{x}} \right)^{\mathsf{H}} \\
=& \frac{1}{2} \frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{A}^{\mathsf{H}} \mathrm{diag}\left(\frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} \right) (|\boldsymbol{A}\boldsymbol{x}| - \boldsymbol{y}) \right) \\
=& \frac{1}{2} \frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{A}^{\mathsf{H}} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^{\mathsf{H}} \mathrm{diag}(\boldsymbol{y}) \frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} \right) \\
=& \frac{1}{2} \boldsymbol{A}^{\mathsf{H}} \boldsymbol{A} - \frac{1}{2} \boldsymbol{A}^{\mathsf{H}} \mathrm{diag}(\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{x}} \left(\frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} \right) \\
=& \frac{1}{2} \boldsymbol{A}^{\mathsf{H}} \boldsymbol{A} - \frac{1}{4} \boldsymbol{A}^{\mathsf{H}} \mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|} \right) \boldsymbol{A}, \qquad \text{(S14)}
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{H}_{\bar{\boldsymbol{x}}\boldsymbol{x}} =& \frac{\partial}{\partial \bar{\boldsymbol{x}}} \left(\frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{x}} \right)^{\mathsf{H}} \\
=& \frac{1}{2} \frac{\partial}{\partial \bar{\boldsymbol{x}}} \left(\boldsymbol{A}^{\mathsf{H}} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^{\mathsf{H}} \mathrm{diag}(\boldsymbol{y}) \frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} \right) \\
=& \boldsymbol{0} - \frac{1}{2} \boldsymbol{A}^{\mathsf{H}} \mathrm{diag}(\boldsymbol{y}) \frac{\partial}{\partial \bar{\boldsymbol{x}}} \left(\frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} \right) \\
=& \frac{1}{4} \boldsymbol{A}^{\mathsf{H}} \mathrm{diag}(\boldsymbol{y}) \mathrm{diag}\left(\frac{(\boldsymbol{A}\boldsymbol{x})^2}{|\boldsymbol{A}\boldsymbol{x}|^3} \right) \bar{\boldsymbol{A}}, \qquad \text{(S15)}
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{H}_{\boldsymbol{x}\bar{\boldsymbol{x}}} =& \frac{\partial}{\partial \boldsymbol{x}} \left(\frac{\partial F(\boldsymbol{x})}{\partial \bar{\boldsymbol{x}}} \right)^{\mathsf{H}} = \boldsymbol{H}_{\bar{\boldsymbol{x}}\boldsymbol{x}}^{\mathsf{H}} \\
=& \frac{1}{4} \boldsymbol{A}^{\mathsf{T}} \mathrm{diag}(\boldsymbol{y}) \mathrm{diag}\left(\frac{(\overline{\boldsymbol{A}\boldsymbol{x}})^2}{|\boldsymbol{A}\boldsymbol{x}|^3} \right) \boldsymbol{A}, \qquad \text{(S16)}
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{H}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} =& \frac{\partial}{\partial \bar{\boldsymbol{x}}} \left(\frac{\partial F(\boldsymbol{x})}{\partial \bar{\boldsymbol{x}}} \right)^{\mathsf{H}} = \boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}}^{\mathsf{T}} \\
=& \frac{1}{2} \boldsymbol{A}^{\mathsf{T}} \bar{\boldsymbol{A}} - \frac{1}{4} \boldsymbol{A}^{\mathsf{T}} \mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|} \right) \bar{\boldsymbol{A}}. \qquad \text{(S17)}
\end{aligned}
$$

We now prove that the gradient of $F(\boldsymbol{x})$ is upper Lipschitz bounded by a constant. This is a particularly useful property of the amplitude-based fidelity term, enabling us to use prespecified algorithm parameters while ensuring convergence.

Lemma 4 *For any* $\boldsymbol{x} \in \mathbb{C}^n \backslash Z$*, the Lipschitz constant for the gradient of the data-fidelity function* $\nabla F(\boldsymbol{x})$ *is bounded above by* $(1/2)\rho(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$.

Proof We only need to prove that for any $\tau > (1/2)\rho(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$, we have

$$
\boldsymbol{G} \equiv \tau \boldsymbol{I} - \boldsymbol{H}_{\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}} = \begin{pmatrix} \tau \boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}} & -\boldsymbol{H}_{\bar{\boldsymbol{x}}\boldsymbol{x}} \\ -\boldsymbol{H}_{\boldsymbol{x}\bar{\boldsymbol{x}}} & \tau \boldsymbol{I} - \boldsymbol{H}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} \end{pmatrix} \succ \boldsymbol{0}. \qquad \text{(S18)}
$$

Let $\varepsilon = \tau - (1/2)\rho\left(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\right) > 0$ and denote $\boldsymbol{G}_{11}$, $\boldsymbol{G}_{12}$, $\boldsymbol{G}_{21}$, $\boldsymbol{G}_{22} \in \mathbb{C}^{n\times n}$ as the four block matrices of $\boldsymbol{G}$, we have

$$\begin{aligned}\boldsymbol{G}_{11} =& \left(\tau\boldsymbol{I} - \frac{1}{2}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\right) + \frac{1}{4}\boldsymbol{A}^{\mathsf{H}}\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)\boldsymbol{A}\\ \succ& \varepsilon\boldsymbol{I} + \frac{1}{4}\boldsymbol{A}^{\mathsf{H}}\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)\boldsymbol{A}. \end{aligned} \tag{S19}$$

According to Lemma 1(a) and 1(b), we have

$$\boldsymbol{G}_{21}\boldsymbol{G}_{11}^{-1}\boldsymbol{G}_{12} \prec \boldsymbol{G}_{21}\left(\varepsilon\boldsymbol{I} + \frac{1}{4}\boldsymbol{A}^{\mathsf{H}}\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)\boldsymbol{A}\right)^{-1}\boldsymbol{G}_{12} \tag{S20}$$

Let $\boldsymbol{P} = (1/2)\mathrm{diag}\left((\boldsymbol{y}/|\boldsymbol{A}\boldsymbol{x}|)^{1/2}\right)\boldsymbol{A}$ and use Lemma 1(b), we have

$$\begin{aligned}\boldsymbol{G}_{21}\boldsymbol{G}_{11}^{-1}\boldsymbol{G}_{12} \prec& \frac{1}{4}\boldsymbol{A}^{\mathsf{T}}\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)^{\frac{1}{2}}\mathrm{diag}\left(\frac{(\overline{\boldsymbol{A}\boldsymbol{x}})^2}{|\boldsymbol{A}\boldsymbol{x}|^2}\right)\\ &\times\boldsymbol{P}\left(\varepsilon\boldsymbol{I} + \boldsymbol{P}^{\mathsf{H}}\boldsymbol{P}\right)^{-1}\boldsymbol{P}^{\mathsf{H}}\mathrm{diag}\left(\frac{(\boldsymbol{A}\boldsymbol{x})^2}{|\boldsymbol{A}\boldsymbol{x}|^2}\right)\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)^{\frac{1}{2}}\bar{\boldsymbol{A}}\\ \prec& \frac{1}{4}\boldsymbol{A}^{\mathsf{T}}\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)^{\frac{1}{2}}\mathrm{diag}\left(\frac{(\overline{\boldsymbol{A}\boldsymbol{x}})^2}{|\boldsymbol{A}\boldsymbol{x}|^2}\right)\mathrm{diag}\left(\frac{(\boldsymbol{A}\boldsymbol{x})^2}{|\boldsymbol{A}\boldsymbol{x}|^2}\right)\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)^{\frac{1}{2}}\bar{\boldsymbol{A}}\\ \prec& \frac{1}{4}\boldsymbol{A}^{\mathsf{T}}\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)\bar{\boldsymbol{A}}\\ \prec& \frac{1}{4}\mathrm{diag}\left(\frac{\boldsymbol{y}}{|\boldsymbol{A}\boldsymbol{x}|}\right)\bar{\boldsymbol{A}} + \left(\lambda\boldsymbol{I} - \frac{1}{2}\boldsymbol{A}^{\mathsf{T}}\bar{\boldsymbol{A}}\right)\\ =& \boldsymbol{G}_{22}. \end{aligned} \tag{S21}$$

Therefore, according to Lemma 2, $\boldsymbol{G}$ is positive-definite. This implies that for $\boldsymbol{x} \in \mathbb{C}^n\backslash Z$ the Lipschitz constant of ∇F is upper-bounded by $(1/2)\rho(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$. □

The above Lemma implies that the fidelity function is upper-bounded by a quadratic function for all $\boldsymbol{x} \in \mathbb{C}^n\backslash Z$. The following theorem states that $F(\boldsymbol{x})$ is in fact globally upper-bounded by the same quadratic function for all $\boldsymbol{x} \in \mathbb{C}^n$.

Lemma 5 *Given any $\boldsymbol{z} \in \mathbb{C}^n$, the fidelity function $F(\boldsymbol{x})$ is upper-bounded by a quadratic function $Q(\boldsymbol{x})$:*

$$F(\boldsymbol{x}) \leq Q(\boldsymbol{x}) \overset{\text{def}}{=} F(\boldsymbol{z}) + \langle\nabla F(\boldsymbol{z}), \hat{\boldsymbol{x}} - \hat{\boldsymbol{z}}\rangle + \frac{L}{2}\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{z}}\|_2^2, \tag{S22}$$

where $L = (1/2)\rho(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$.

Proof Let $\Delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{z}$, then either of the two following cases occurs:

1) The line between $\boldsymbol{x}$ and $\boldsymbol{z}$ does not pass through any nonsmooth points, i.e., $\boldsymbol{z} + \alpha\Delta\boldsymbol{x} \in \mathbb{C}\backslash Z, \forall\alpha \in [0,1]$, or $\boldsymbol{x}$ and $\boldsymbol{z}$ lie in the subspace, i.e., $\boldsymbol{z} + \alpha\Delta\boldsymbol{x} \in Z, \forall\alpha \in [0,1]$, the result is obtained directly according to the multivariate Taylor expansion of F:

$$F(\boldsymbol{x}) = F(\boldsymbol{z}) + \langle\nabla F(\boldsymbol{z}), \hat{\boldsymbol{x}} - \hat{\boldsymbol{z}}\rangle + \frac{1}{2}(\hat{\boldsymbol{x}} - \hat{\boldsymbol{z}})^{\mathsf{H}}\nabla^2 F(\boldsymbol{u})(\hat{\boldsymbol{x}} - \hat{\boldsymbol{z}})$$

$$
\begin{aligned}
&\leq F(\boldsymbol{z}) + \langle \nabla F(\boldsymbol{z}), \hat{\boldsymbol{x}} - \hat{\boldsymbol{z}} \rangle + \frac{L}{2}\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{z}}\|_2^2 \\
&= Q(\boldsymbol{x}),
\end{aligned} \tag{S23}
$$

where $\boldsymbol{u}$ is a convex combination of $\boldsymbol{x}$ and $\boldsymbol{z}$.

2) The line between $\boldsymbol{x}$ and $\boldsymbol{z}$ passes through a finite number of nonsmooth points. For simplicity, we consider the case of passing through a single nonsmooth point indexed by j, that is, we have

$$
|\boldsymbol{a}_j^{\mathsf{H}}(\boldsymbol{z} + \alpha^\star \Delta \boldsymbol{x})| = 0, \tag{S24}
$$

for some $0 < \alpha^\star < 1$. According to 1), for any $0 \leq \alpha \leq \alpha^\star$, $F(\boldsymbol{x})$ is upper-bounded by $Q(\boldsymbol{x})$. We now prove that this also holds for any $\alpha^\star < \alpha < 1$. The fidelity function can be written as a function over α for any point that lies on the line between $\boldsymbol{x}$ ad $\boldsymbol{z}$:

$$
\begin{aligned}
g(\alpha) &= F(\boldsymbol{z} + \alpha\Delta\boldsymbol{x}) = \sum_{i=1}^{m} f_i(\boldsymbol{z} + \alpha\Delta\boldsymbol{x}) \\
&= \sum_{i=1, i\neq j}^{m} f_i(\boldsymbol{z} + \alpha\Delta\boldsymbol{x}) + f_j(\boldsymbol{z} + \alpha\Delta\boldsymbol{x}) \\
&= \sum_{i=1, i\neq j}^{m} f_i(\boldsymbol{z} + \alpha\Delta\boldsymbol{x}) + \Big(|\boldsymbol{a}_j^{\mathsf{H}}(\boldsymbol{z} + \alpha\Delta\boldsymbol{x})| - y_j\Big)^2 \\
&= \sum_{i=1, i\neq j}^{m} f_i(\boldsymbol{z} + \alpha\Delta\boldsymbol{x}) + \Big(|\alpha - \alpha^\star||\boldsymbol{a}_j^{\mathsf{H}}\Delta\boldsymbol{x}| - y_j\Big)^2 \\
&\leq \sum_{i=1, i\neq j}^{m} f_i(\boldsymbol{z} + \alpha\Delta\boldsymbol{x}) + \Big((\alpha - \alpha^\star)|\boldsymbol{a}_j^{\mathsf{H}}\Delta\boldsymbol{x}| + y_j\Big)^2 \\
&\leq h(\alpha).
\end{aligned} \tag{S25}
$$

where $f_i(\boldsymbol{x}) \overset{\text{def}}{=} (1/2)(|\boldsymbol{a}_i^{\mathsf{H}}\boldsymbol{x}| - y_i)^2$ and $h(\alpha) \overset{\text{def}}{=} Q(\boldsymbol{z} + \alpha\Delta\boldsymbol{x})$. As a result, we have

$$
F(\boldsymbol{x}) = F(\boldsymbol{z} + \Delta\boldsymbol{x}) = g(1) \leq h(1) = Q(\boldsymbol{x}). \tag{S26}
$$

The above derivation can be easily extended to the case of multiple nonsmooth points.

With this, we conclude that for any $\boldsymbol{x} \in \mathbb{C}^n$, we have

$$
F(\boldsymbol{x}) \leq Q(\boldsymbol{x}), \tag{S27}
$$

which completes the proof. □

We are now ready to prove the main theorem, which establishes the convergence of the basic proximal gradient method.

Theorem 1 *The basic proximal gradient algorithm (with $\beta_t \equiv 0$ in Algorithm 1) for the problem of Eq. (3) converges to a stationary point using a fixed step size γ that satisfies*

$$
\gamma \leq \frac{2}{\rho(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})}. \tag{S28}
$$

Proof The proof is adapted from Ref. [5]. Recall that the proximal update is given by

$$\boldsymbol{x}^{(t+1)} = \operatorname{prox}_{\gamma R}(\boldsymbol{x}^{(t)} - \gamma \nabla_{\boldsymbol{x}} F(\boldsymbol{x}^{(t)})). \tag{S29}$$

According to Lemma 5, we have that

$$\begin{aligned} F(\boldsymbol{x}^{(t+1)}) \leq Q(\boldsymbol{x}^{(t+1)}) = F(\boldsymbol{x}^{(t)}) \\ + \langle \nabla F(\boldsymbol{x}^{(t)}), \hat{\boldsymbol{x}}^{(t+1)} - \hat{\boldsymbol{x}}^{(t)} \rangle + \frac{L}{2} \|\hat{\boldsymbol{x}}^{(t+1)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2. \end{aligned} \tag{S30}$$

By the second prox theorem (Theorem 6.39) in Ref. [5], we have

$$\begin{aligned} \langle \hat{\boldsymbol{x}}^{(t)} - \gamma \nabla F(\boldsymbol{x}^{(t)}) - \hat{\boldsymbol{x}}^{(t+1)}, \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)} \rangle \\ \leq \gamma R(\boldsymbol{x}^{(t)}) - \gamma R(\boldsymbol{x}^{(t+1)}), \end{aligned} \tag{S31}$$

from which it follows that

$$\begin{aligned} \langle \nabla F(\boldsymbol{x}^{(t)}), \hat{\boldsymbol{x}}^{(t+1)} - \hat{\boldsymbol{x}}^{(t)} \rangle \\ \leq R(\boldsymbol{x}^{(t)}) - R(\boldsymbol{x}^{(t+1)}) - \frac{1}{\gamma} \|\hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2. \end{aligned} \tag{S32}$$

Let $J(\boldsymbol{x}) = F(\boldsymbol{x}) + R(\boldsymbol{x})$. Combining Eqs. (S30) and (S32), we arrive at

$$\begin{aligned} J(\boldsymbol{x}^{(t+1)}) \leq & J(\boldsymbol{x}^{(t)}) + \left(\frac{L}{2} - \frac{1}{\gamma} \right) \|\hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2 \\ \leq & J(\boldsymbol{x}^{(t)}) - \frac{L}{2} \|\hat{\boldsymbol{x}}^{(t+1)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2. \end{aligned} \tag{S33}$$

The second inequality holds because $\gamma \leq 1/L$. Thus, the updating step for each iteration is upper-bounded:

$$\|\hat{\boldsymbol{x}}^{(t+1)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 \leq \frac{2}{L} \left(J(\boldsymbol{x}^{(t)}) - J(\boldsymbol{x}^{(t+1)}) \right). \tag{S34}$$

By summing up T iterations, we arrive at

$$\begin{aligned} \sum_{t=0}^{T} \|\hat{\boldsymbol{x}}^{(t+1)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 \leq & \frac{2}{L} \sum_{t=0}^{T} \left(J(\boldsymbol{x}^{(t)}) - J(\boldsymbol{x}^{(t+1)}) \right) \\ \leq & \frac{2}{L} \left(J(\boldsymbol{x}^{(0)}) - J^{\star} \right), \end{aligned} \tag{S35}$$

where $J^{\star} \geq 0$ denotes the global minimum value of the objective function. This implies that

$$\lim_{t \to \infty} \|\hat{\boldsymbol{x}}^{(t+1)} - \hat{\boldsymbol{x}}^{(t)}\|_2 = 0. \tag{S36}$$

That is, the algorithm converges to a stationary point. □

A similar result has been reported in Ref. [6] regarding the Wirtinger gradient descent algorithm for ptychographic phase retrieval. We consider the more general proximal gradient algorithm and present above an alternative proof.

2.3 Convergence of the accelerated gradient projection algorithm

In order to prove the convergence of the denoising algorithm, we first derive an upper Lipschitz bound for the gradient of $G(\boldsymbol{w})$.

Lemma 6 *The Lipschitz constant of the gradient of $G(\boldsymbol{w})$ is upper-bounded by $\gamma^2\rho(\boldsymbol{D}^{\mathsf{H}}\boldsymbol{D})$.*

Proof The proof is adapted from Ref. [7]. Given any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{C}^{2n}$, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{w}} G(\boldsymbol{w}_1) - \nabla_{\boldsymbol{w}} G(\boldsymbol{w}_2)\|_2 \\
&= \gamma\|\boldsymbol{D}\mathcal{P}_C(\boldsymbol{v} - \gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}_1) - \boldsymbol{D}\mathcal{P}_C(\boldsymbol{v} - \gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}_2)\|_2 \\
&\leq \gamma\|\boldsymbol{D}\|_2\|\mathcal{P}_C(\boldsymbol{v} - \gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}_1) - \boldsymbol{D}\mathcal{P}_C(\boldsymbol{v} - \gamma\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}_2)\|_2 \\
&\overset{(b)}{\leq} \gamma^2\|\boldsymbol{D}\|_2\|\boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}_1 - \boldsymbol{D}^{\mathsf{H}}\boldsymbol{w}_2\|_2 \\
&\leq \gamma^2\|\boldsymbol{D}\|_2^2\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 \\
&= \gamma^2\rho(D^{\mathsf{H}}\boldsymbol{D})\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2,
\end{aligned} \tag{S37}$$

where (b) is based on the fact that projection onto convex sets is non-expansive. □

Theorem 2 *Assuming that the constraint set C is closed and convex, the accelerated gradient projection algorithm for the problem of Eq. (7) converges to the global optimum using a fixed step size η that satisfies*

$$\eta \leq \frac{1}{\gamma^2\rho(\boldsymbol{D}^{\mathsf{H}}\boldsymbol{D})}. \tag{S38}$$

Proof The gradient projection algorithm can be viewed as a special case of the proximal gradient algorithm with the nonsmooth term being an indicator function. Because Eq. (6) is a convex optimization problem with a closed and convex C, it is sufficient to prove that the objective function is Lipschitz continuous with a constant no greater than $\gamma^2\rho(\boldsymbol{D}^{\mathsf{H}}\boldsymbol{D})$, which is accomplished by Lemma 6. Based on the convergence results of the accelerated proximal gradient algorithm for convex functions [8], the proof is completed. □

For the particular case of $\boldsymbol{D}$ being the finite difference operator, the Lipschitz bound (and thus the step size η) can be explicitly given according to the following observation.

Lemma 7 *If $\boldsymbol{D}$ represents the finite-difference operator defined by Eq. (4), we have*

$$\rho(\boldsymbol{D}^{\mathsf{H}}\boldsymbol{D}) \leq 8. \tag{S39}$$

Proof Given any $\boldsymbol{x} \in \mathbb{C}^n$, we have

$$\|\boldsymbol{D}\boldsymbol{x}\|_2^2 = \sum_{i=1}^{n_\xi - 1}\sum_{j=1}^{n_v}|X_{i+1,j} - X_{i,j}|^2 + \sum_{i=1}^{n_\xi}\sum_{j=1}^{n_v - 1}|X_{i,j+1} - X_{i,j}|^2$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n_\xi}\sum_{j=1}^{n_\upsilon}|X_{i+1,j}-X_{i,j}|^2+\sum_{i=1}^{n_\xi}\sum_{j=1}^{n_\upsilon}|X_{i,j+1}-X_{i,j}|^2 \\
&\leq 2\sum_{i=1}^{n_\xi}\sum_{j=1}^{n_\upsilon}\left(|X_{i+1,j}|^2+|X_{i,j}|^2\right)+2\sum_{i=1}^{n_\xi}\sum_{j=1}^{n_\upsilon}\left(|X_{i,j+1}|^2+|X_{i,j}|^2\right) \\
&\leq 8\sum_{i=1}^{n_\xi}\sum_{j=1}^{n_\upsilon}|X_{i,j}|^2 \\
&= 8\|\boldsymbol{x}\|_2^2,
\end{aligned}
\tag{S40}
$$

which implies that

$$
\rho(\boldsymbol{D}^{\mathsf{H}}\boldsymbol{D}) = \|\boldsymbol{D}\|_2^2 \leq 8. \tag{S41}
$$

□

References

[1] S. Boyd, S.P. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, 2004)

[2] K. Kreutz-Delgado, The complex gradient operator and the CR-calculus. arXiv preprint arXiv:0906.4835 (2009)

[3] R.A. Horn, C.R. Johnson, *Matrix Analysis* (Cambridge University Press, 2012)

[4] Y. Gao, L. Cao, Generalized optimization framework for pixel super-resolution imaging in digital holography. Optics Express **29**(18), 28,805–28,823 (2021)

[5] A. Beck, *First-Order Methods in Optimization* (SIAM, 2017)

[6] R. Xu, M. Soltanolkotabi, J.P. Haldar, W. Unglaub, J. Zusman, A.F. Levi, R.M. Leahy, Accelerated Wirtinger flow: a fast algorithm for ptychography. arXiv preprint arXiv:1806.05546 (2018)

[7] A. Beck, M. Teboulle, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Transactions on Image Processing **18**(11), 2419–2434 (2009)

[8] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)