On Fienup Methods for Sparse Phase Retrieval

Edouard Jean Robert Pauwels 🆔 , Amir Beck, Yonina C. Eldar, *Fellow, IEEE*, and Shoham Sabach

Abstract—Alternating minimization, or Fienup methods, have a long history in phase retrieval. We provide new insights related to the empirical and theoretical analysis of these algorithms when used with Fourier measurements and combined with convex priors. In particular, we show that Fienup methods can be viewed as performing alternating minimization on a regularized nonconvex least-squares problem with respect to amplitude measurements. Furthermore, we prove that under mild additional structural assumptions on the prior (semialgebraicity), the sequence of signal estimates has a smooth convergent behavior toward a critical point of the nonconvex regularized least-squares objective. Finally, we propose an extension to Fienup techniques, based on a projected gradient descent interpretation and acceleration using inertial terms. We demonstrate experimentally that this modification combined with an ℓ_1 prior constitutes a competitive approach for sparse phase retrieval.

Index Terms—Non-convex optimization, iterative algorithms, phase retrieval, sparse signal processing, Fourier measurements.

I. INTRODUCTION

PHASE retrieval is an old and fundamental problem in a variety of areas within engineering and physics [1], [2]. Many applications of the phase retrieval problem involve estimation of a signal from the modulus of its Fourier measurements. This problem is ill posed in general, so that uniqueness and recovery typically require prior knowledge on the input, particularly in one-dimensional problems. Here we focus on the estimation of real sparse signals from their Fourier magnitude, a problem which has been treated in several recent works [3]–[6].

A longstanding line of algorithms to tackle the phase retrieval problem involve application of the alternating minimization method which alternate between the constraints in time and the Fourier magnitude constraints [7]–[9]. These methods were pioneered by the work of Gerchberg and Saxton and later extended by Fienup; see [10] for an optimization point of view

E. J. R. Pauwels is with the Informatics Department, IRIT, Université Toulouse III Paul Sabatier, Toulouse 31062, France (e-mail: epauwels@irit.fr).

A. Beck is with the School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997801, Israel (e-mail: becka@tauex.tau.ac.il).

Y. C. Eldar is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel (e-mail: yonina@ee.technion.ac.il).

S. Sabach is with the Department of Industrial Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel (e-mail: ssabach@ie.technion.ac.il).

on these techniques and a rich historical perspective. Relations with saddle point optimization and alternating direction methods of multipliers were also considered in [11], [12]. Alternating minimization approaches have also been recently applied to phase retrieval from random measurements [13], [14]. The main advantage of this class of algorithms is their simplicity and scalability.

A more recent approach to phase retrieval is to formulate the recovery as a smooth nonconvex least-squares estimation problem and use dedicated techniques to estimate the signal using continuous optimization algorithms that guarantee convergence to critical points. The GESPAR algorithm [4] is an example of this approach which is based on the Gauss-Newton method coupled with sparsity priors. For phase retrieval with random measurements, gradient descent methods have been proposed and analyzed such as Wirtinger flow [15] and truncated amplitude flow [16]. Both treat least-squares objectives where Wirtinger flow measures the loss with respect to the squared-magnitude of the measurements while the amplitude flow approach performs a truncated gradient descent on an amplitude objective. Another line of work suggests the use of matrix lifting and semidefinite programming based relaxations [6], [17]–[20]. These techniques are limited by the size of problems that can be tackled using available numerical solvers.

Our main contribution is to propose a new look at alternating minimization algorithms for phase retrieval in the context of Fourier measurements and convex priors. We refer collectively to these techniques as Fienup methods. The use of Fourier measurements is less flexible than general measurements and is less suited for statistical analysis due to its strong structure. This renders the phase retrieval problem from Fourier measurements harder to solve than from random unstructured measurements. On the other hand, the strong structure of the Fourier transform can be exploited for richer algorithmic construction and analysis.

As a first step we provide two new interpretations of Fienup algorithms. First we show that these techniques are naturally linked to a nonsmooth nonconvex least-squares problem with respect to an amplitude objective. Fienup approaches can then be understood as majorization-minimization methods for solving this problem. Second, we demonstrate that Fienup algorithms can be viewed as a projected gradient descent scheme to minimize a smooth convex objective function over a nonconvex constraint set. This observation allows to characterize the behavior of the algorithm and develop extensions based on known ideas for accelerating gradient methods using inertial terms [21]. We then specialize these results to the case of ℓ_0 and ℓ_1 priors, leading to a new inertial gradient scheme, which we refer to as FISTAPH: FISTA for PHase retrieval.

On the theoretical side, we show that if the convex prior is well structured (semi-algebraic or more generally representable), then the sequence of signal estimates produced by Fienup has

a smooth convergence behavior, independently of the initialization. Recall that, broadly speaking, an object is said to be semi-algebraic if it can be represented by systems of polynomial inequalities. The notion of smooth convergence is a very desirable property, even more in nonconvex settings where it is usually not possible to obtain global convergence estimates. We emphasize that this result goes beyond more usual properties of the set of accumulation points as it states that the set of accumulation points is actually a singleton. Our convergence analysis follows well established techniques from tame optimization [22], [23]. These techniques build upon the Kurdyka-Łojasiewicz (KL) property which holds for many classes of functions [24]–[27]. We then provide numerical experiments based on synthetic problems to compare Fienup with ℓ_0 and ℓ_1 priors, GESPAR [4], Wirtinger flow (or gradient) methods [15] with ℓ_0 and ℓ_1 priors and FISTAPH. Numerical results suggest that the latter combined with an ℓ_1 prior constitutes a very competitive alternative for sparse phase retrieval.

The rest of the paper is organized as follows. Section II introduces our notation and states the problem of interest more formally. We also introduce several mathematical definitions that are required for the rest of the paper and review the numerical algorithms that are used in subsequent sections. Section III describes our characterization of Fienup methods in the context of phase retrieval from Fourier measurements with convex priors. We detail the relation of Fienup with a nonsmooth nonconvex least-squares problem as well as its interpretation as projected gradient descent. Our main convergence result and FISTAPH algorithm are presented in Section IV. Simulation results are provided in Section V.

II. PROBLEM FORMULATION AND MATHEMATICAL BACKGROUND

A. Notation

Throughout the paper vectors are denoted by boldface letters. For a vector $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x}[i]$ is the i-th entry of $\mathbf{x}$, $i = 1, 2, \ldots, n$ and $\mathrm{supp}(\mathbf{x})$ is the support of $\mathbf{x}$, namely, the set $\{i = 1, 2, \ldots n; \mathbf{x}[i] \neq 0\}$. Furthermore, $\|\mathbf{x}\|_0$ counts the number of nonzero entries of the vector $\mathbf{x}$: $\|\mathbf{x}\|_0 = |\mathrm{supp}(\mathbf{x})|$ and $\|\mathbf{x}\|_p$ denotes the ℓ_p norm of $\mathbf{x}$ for $p \in \mathbb{R}_+$. The notations $|\cdot|$, $\mathrm{Re}(\cdot), \mathrm{Im}(\cdot)$ and $\bar{\cdot}$ describe the modulus, real part, imaginary part and complex conjugate, respectively, defined over the field of complex numbers. If their argument is a vector, then they should be understood component-wise. Similarly, basic operations, e.g. powers, are taken component-wise when applied to vectors. For $\mathbf{x} \in \mathbb{C}^n$ and $N \in \mathbb{N}$, $\mathcal{F}(\mathbf{x}, N) \in \mathbb{C}^N$ is the vector composed of the N first coefficients of the discrete Fourier transform of $\mathbf{x}$ (obtained by zero padding if $n < N$). For simplicity, we use the shorthand notation $\mathcal{F}(\mathbf{x}) = \mathcal{F}(\mathbf{x}, n)$ to denote the standard discrete Fourier transform of $\mathbf{x} \in \mathbb{C}^n$ and $\mathcal{F}^{-1}$ to denote its inverse. For a set S, $\delta_S : S \to \mathbb{R} \cup \{+\infty\}$ is the indicator function of S (0 if its argument is in S, $+\infty$ otherwise) and P_S denotes the orthogonal projection onto the set S.

B. Phase Retrieval

Given $\mathbf{x}_0 \in \mathbb{R}^n$, we consider measurements

$$\mathbf{c} = |\mathcal{F}(\mathbf{x}_0)| + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is an unknown vector of errors. In the rest of the paper, we assume that $\mathbf{c}$ has positive entries (it is always possible to set the potential negative entries of $\mathbf{c}$ to zero). The phase retrieval problem consists of producing an estimate $\hat{\mathbf{x}} \in \mathbb{R}^n$ of $\mathbf{x}_0$ based solely on the knowledge of $\mathbf{c}$ given by (1).

As mentioned in the introduction, phase retrieval of one-dimensional vectors from Fourier measurements requires the use of prior knowledge. We focus on support and sparsity inducing priors. For $J \subseteq \{1, 2, \ldots, n\}$, we define the set $X_J = \{\mathbf{x} \in \mathbb{R}^n; \mathrm{supp}(\mathbf{x}) \subseteq J\}$. The prior function that we use will be denoted by $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. We focus on the following two priors (for a given support J):
- $g : \mathbf{x} \mapsto \|\mathbf{x}\|_0 + \delta_{X_J}(\mathbf{x})$, or ℓ_0-based nonconvex prior.
- $g : \mathbf{x} \mapsto \|\mathbf{x}\|_1 + \delta_{X_J}(\mathbf{x})$, or ℓ_1-based convex prior.

In the experimental section, we compare between these two classes of priors. The algorithmic derivations in this paper will be made under the assumption that g is proper and lower semi-continuous, and the main convergence result (c.f. Theorem 4.1) will require in addition the convexity of g. In order to efficiently implement the proposed algorithm, we need to focus on priors for which the proximity operator [28] is easy to compute. We provide several examples of such priors in Section II-D.

In the rest of the paper, $\mathbf{c} \in \mathbb{R}_+^n$ denotes modulus measurements which are assumed to be given, fixed and obtained through (1). Given $\mathbf{c} \in \mathbb{R}_+^n$, we define $Z_{\mathbf{c}} = \{\mathbf{z} \in \mathbb{C}^n; |\mathcal{F}(\mathbf{z})| = \mathbf{c}\}$ as the set of vectors $\mathbf{z}$ that could have produced $\mathbf{c}$ (ignoring the noise). To estimate $\mathbf{x}_0$, we consider the regularized least-squares problem

$$\min_{\mathbf{x} \in \mathbb{R}^n, \mathbf{z} \in Z_{\mathbf{c}}} \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + g(\mathbf{x}), \tag{2}$$

where g encodes the prior knowledge. Our algorithmic approach consists of employing an alternating minimization method, or one of its variants, to solve the above formulation. Note that if $g = 0$, then problem (2) has multiple solutions.

C. Prior Algorithms for Phase Retrieval

We briefly review several existing algorithms for phase retrieval that will be used in our experiments in Section V.

One approach to sparse phase retrieval is the GESPAR algorithm which is based on the damped Gauss-Newton method in conjunction with an ℓ_0 prior [4]. Damped Gauss-Newton allows to solve smooth, nonlinear least-squares problems. The work of [15] is based on the notion of Wirtinger derivatives to treat the same smooth least-squares problem as GESPAR. The notion of Wirtinger derivative is needed since the objective is not differentiable (holomorphic) as a function of complex variables (see [15] for details). In the case of real valued functions of real variables, the Wirtinger derivative reduces to a standard gradient (up to a constant multiplicative factor). An obvious extension of these types of methods is the use of proximal decomposition, or forward-backward methods which consist of alternating a gradient step on the smooth part of the objective with a proximal step on the nonsmooth part [21], [29]. This is the approach that we use in the numerical experiments to treat phase retrieval with priors.

Finally, we consider alternating minimization methods that are the main focus of this work. This approach consists of solving (2) by applying the alternating minimization algorithm. The special structure of the problem allows to perform each partial minimization efficiently. In particular, the projection onto $Z_{\mathbf{c}}$ is

easy, as described below in (5). These types of methods are also referred to as Fienup algorithms. A deeper interpretation of this approach is given in Section III.

D. Tools from Convex and Nonsmooth Analysis

Throughout the paper, our results will be based on tools from convex and nonsmooth analysis which we review here.

The gradient of a differentiable function f is denoted by ∇f. This concept admits extensions to nonsmooth analysis; the subgradient of a nonsmooth function g is denoted by ∂g. For convex functions, subgradients correspond to tangent affine lower bounds. This definition no longer holds for nonconvex functions. In this case, the proper understanding of subgradients involves much more machinery which will not be discussed here. We only consider the notion of a Fréchet critical point which generalizes classical first order criticality for differentiable functions (see [30]).

Definition 1 (Fréchet critical point): Let $S \subseteq \mathbb{R}^n$ be a closed set and $f : \mathbb{R}^n \to \mathbb{R}$ be a lower semicontinuous function. We say that $\bar{\mathbf{x}} \in S$ is a Fréchet critical point of the problem

$$\min_{\mathbf{x} \in S} f(\mathbf{x})$$

if

$$\liminf_{\substack{\mathbf{x} \to \bar{\mathbf{x}} \\ \mathbf{x} \neq \bar{\mathbf{x}} \\ \mathbf{x} \in S}} \frac{f(\mathbf{x}) - f(\bar{\mathbf{x}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \geq 0.$$

In other words, the negative variations of f in S around $\bar{x}$ are negligible at the first order.

We will also heavily use the notion of the proximity operator of a function.

Definition 2 (Proximity operator): For a nonsmooth function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the (potentially multivalued) proximity operator is denoted by prox_g and defined by

$$\mathrm{prox}_g(\mathbf{x}) \equiv \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + g(\mathbf{y}) \right\}. \quad (3)$$

Note that when g is proper lower semicontinuous and convex, this operator is single valued.

We next provide a few examples of such functions with their proximity operators; many more can be found, for example, in [31].

Example 1 (Proximity operators):
- **Support prior:** If $C \subseteq \mathbb{R}^n$ is a closed convex set, then prox_{δ_C} is the Euclidean projection onto C. This can be used for example to encode knowledge about the support of the signal $\mathbf{x}_0$ by choosing $C = X_J$ for some $J \subseteq \{1, 2, \ldots, n\}$. In this case, the projection simply consists in setting the coefficients $\mathbf{x}[i]$ to 0 for $i \notin J$.
- **Sparsity prior:** If g is the ℓ_1 norm, then the proximal operator is the soft thresholding operator.[1] This can be combined with support information prior by first setting the coefficients outside of the support to 0 and then applying the soft thresholding operator.

- **Change of basis:** Suppose that D is an $n \times n'$ real matrix such that its columns form an orthonormal family, that is $D^T D$ is the identity in $\mathbb{R}^{n'}$. Suppose that $\tilde{g} : \mathbb{R}^{n'} \to \mathbb{R}$ is a lower semicontinuous convex function and let $g(\mathbf{x}) = \tilde{g}(D^T \mathbf{x})$. In this case, we have $\mathrm{prox}_g(\mathbf{x}) = \mathbf{x} + D \left(\mathrm{prox}_{\tilde{g}}(D^T \mathbf{x}) - D^T \mathbf{x} \right)$ (see [31, Table 1]). This allows to express priors in different orthonormal bases, such as wavelets for example.
- **Trivial prior:** When $g = 0$ the proximity operator is simply the identity map.

It is also worth mentioning that the proximity operator is efficiently computable for some nonconvex priors. For example, if $g = \delta_C$ where $C = \{\mathbf{x} \in \mathbb{R}^n; \ \|\mathbf{x}\|_0 \leq k\}$, then the proximity operator is obtained by setting the $n - k$ lowest coefficients (in absolute value) to 0. This can also be combined with support information.

III. Fienup, Majorization-Minimization and Projected Gradient

In this section we expand on the alternating minimization approach to (2) leading to the Fienup family of algorithms. For this section, the prior term g in (2) is taken to be a general proper lower semicontinuous function. We begin by describing the algorithm and then provide two interpretations of it.

A. Alternating Minimization Algorithm

The alternating minimization algorithm applied to problem (2) is explicitly written below.

Alternating Minimization (Fienup)
Initialization. $\mathbf{x}^0 \in \mathbb{R}^n$
General Step. For $k \in \mathbb{N}$,

$$\mathbf{z}^{k+1} \in \operatorname*{argmin}_{\mathbf{z} \in Z_\mathbf{c}} \frac{1}{2} \|\mathbf{x}^k - \mathbf{z}\|_2^2,$$

$$\mathbf{x}^{k+1} \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{x} - \mathbf{z}^{k+1}\|_2^2 + g(\mathbf{x}). \quad (4)$$

The main interest in this scheme is that both partial minimization steps in (4) can be carried out efficiently whenever g is "proximable", meaning that its prox (or a member in its prox) is easily computed. First consider, in (4), the partial minimization in $\mathbf{z}$ with $\mathbf{x} \in \mathbb{R}^n$ being arbitrary but fixed. This minimization amounts to computing $P_{Z_\mathbf{c}}(\mathbf{x}^k)$, the orthogonal projection of $\mathbf{x}^k$ onto $Z_\mathbf{c}$. For a given $\mathbf{x} \in \mathbb{R}^n$, all the members in $P_{Z_\mathbf{c}}(\mathbf{x})$ are of the form $\mathbf{z} = \mathcal{F}^{-1}(\hat{\mathbf{z}})$, where for $j = 1, 2, \ldots, n$, we have ($i = \sqrt{-1}$ in the equation below)

$$\hat{\mathbf{z}}[j] = \begin{cases} \mathbf{c}[j] \frac{\mathcal{F}(\mathbf{x})[j]}{|\mathcal{F}(\mathbf{x})[j]|}, & \text{if } |\mathcal{F}(\mathbf{x})[j]| \neq 0, \\ \mathbf{c}[j] e^{i\theta_j}, & \text{for an arbitrary } \theta_j \text{ otherwise.} \end{cases} \quad (5)$$

Next, we treat the subproblem in (4) of minimizing with respect to $\mathbf{x}$ where $\mathbf{z} \in \mathbb{C}^n$ is arbitrary but fixed. The partial minimization in $\mathbf{x}$ is given by the expression

$$\operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + g(\mathbf{x}) \right\} = \mathrm{prox}_g(\mathrm{Re}(\mathbf{z})), \quad (6)$$

---
[1]The soft thresholding operator is given by $\mathcal{T}_\alpha(\mathbf{x})_i = \mathrm{sgn}(x_i) \max\{|x_i| - \alpha, 0\}$. If $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ for some $\lambda > 0$, then $\mathrm{prox}_g(\mathbf{x}) = \mathcal{T}_\lambda(\mathbf{x})$.

where Re is the real part taken component-wise. We have used the definition of the proximity operator of g given in (3). When this operator is easy to compute, each step of the algorithm can be carried out efficiently.

The iterations of the alternating minimization method are summarized as follows:

$$\mathbf{z}^{k+1} \in P_{Z_\mathbf{c}}(\mathrm{prox}_g(\mathrm{Re}(\mathbf{z}^k))),$$
$$\mathbf{x}^{k+1} \in \mathrm{prox}_g(\mathrm{Re}(P_{Z_\mathbf{c}}(\mathbf{x}^k))). \qquad (7)$$

We now consider several special cases of (7):

- If $g = 0$, then prox_g is the identity and we recover the original algorithm from Fienup [9], or alternating projection [10].

Fienup
Initialization. $\mathbf{x}^0 \in \mathbb{R}^n$.
General Step. For $k \in \mathbb{N}$,

$$\mathbf{x}^{k+1} = \mathrm{Re}(P_{Z_\mathbf{c}}(\mathbf{x}^k)).$$

The convergence result given in Theorem 4.1 also holds in this case since constant functions are convex and continuous.

- If $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ for some $\lambda > 0$, then $\mathrm{prox}_g = \mathcal{T}_\lambda$, where $\mathcal{T}_\lambda$ is the soft thresholding operator (see footnote on page 8). We refer to the resulting algorithm as "AM L1".

AM L1
Initialization. $\mathbf{x}^0 \in \mathbb{R}^n$, $\lambda > 0$
General Step. For $k \in \mathbb{N}$,

$$\mathbf{x}^{k+1} = \mathcal{T}_\lambda(\mathrm{Re}(P_{Z_\mathbf{c}}(\mathbf{x}^k))).$$

- If $g = \delta_{C_K}$, where C_K is the set of all K-sparse vectors, $C_K = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \le K\}$, then $\mathrm{prox}_g = P_{C_K}$ is the so-called hard thresholding operator. This operator outputs a vector which is all zeros except for the largest K components (in absolute values) of its input vector which are kept the same. The hard thresholding operator is multivalued and the resulting algorithm, which we term "AM L0" picks an arbitrary point in its range.

AM L0
Initialization. $\mathbf{x}^0 \in \mathbb{R}^n$, $K \in \mathbb{N}$
General Step. For $k \in \mathbb{N}$,

$$\mathbf{x}^{k+1} \in P_{C_K}(\mathrm{Re}(P_{Z_\mathbf{c}}(\mathbf{x}^k))).$$

B. Majorization-Minimization Interpretation

In this section, we focus on partial minimization in $\mathbf{z}$. We show that the value of this partial minimization leads to a least-squares objective. This allows to interpret the Fienup algorithm as a majorization-minimization process on this least-squares function. For the rest of this section, for any $\mathbf{x} \in \mathbb{R}^n$, we denote by $\mathbf{z}(\mathbf{x})$ an arbitrary but fixed member of $P_{Z_\mathbf{c}}(\mathbf{x})$.

1) Partial Minimization in $\mathbf{z}$: The following lemma provides a connection between partial minimization in $\mathbf{z}$ and the evaluation of a nonsmooth least-squares objective.

Lemma 3.1: For any $\mathbf{x} \in \mathbb{R}^n$, we have

$$\min_{\mathbf{z} \in Z_\mathbf{c}} \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 = \frac{1}{2n}\||\mathcal{F}(\mathbf{x})| - \mathbf{c}\|_2^2. \qquad (8)$$

Proof: An optimal solution of the minimization problem is given by $\mathbf{z} = \mathcal{F}^{-1}(\hat{\mathbf{z}})$ where $\hat{\mathbf{z}}$ has the form (5). Now,

$$\min_{\mathbf{z} \in Z_\mathbf{c}} \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 = \frac{1}{2}\|\mathbf{x} - \mathcal{F}^{-1}(\hat{\mathbf{z}})\|_2^2$$
$$= \frac{1}{2}\|\mathcal{F}^{-1}(\mathcal{F}(\mathbf{x}) - \hat{\mathbf{z}})\|_2^2$$
$$= \frac{1}{2n}\|\mathcal{F}(\mathbf{x}) - \hat{\mathbf{z}}\|_2^2.$$

Using the expression of $\hat{\mathbf{z}}$ in (5), we have for all $j = 1, 2, \ldots, n$,

$$|\mathcal{F}(\mathbf{x})[j] - \hat{\mathbf{z}}[j]| = \begin{cases} ||\mathcal{F}(\mathbf{x})[j]| - \mathbf{c}[j]|, & \text{if } |\mathcal{F}(\mathbf{x})[j]| \neq 0, \\ \mathbf{c}[j], & \text{otherwise.} \end{cases}$$

Putting everything together,

$$\min_{\mathbf{z} \in Z_\mathbf{c}} \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 = \frac{1}{2n}\|\mathcal{F}(\mathbf{x}) - \hat{\mathbf{z}}\|_2^2 = \frac{1}{2n}\||\mathcal{F}(\mathbf{x})| - \mathbf{c}\|_2^2, \qquad (9)$$

which completes the proof. ∎

A direct consequence of Lemma 3.1 is the following corollary that connects between problem (2) and a regularized nonlinear least-squares problem.

Corollary 3.2: The pair $(\mathbf{x}, \mathbf{z})$ is an optimal solution of problem (2) if and only if $\mathbf{x}$ is an optimal solution of

$$\min\left\{ F(\mathbf{x}) \equiv \frac{1}{2n}\||\mathcal{F}(\mathbf{x})| - \mathbf{c}\|_2^2 + g(\mathbf{x}) \right\}, \qquad (10)$$

and $\mathbf{z} = \mathcal{F}^{-1}(\hat{\mathbf{z}})$, where $\hat{\mathbf{z}}$ is of the form given in (5).

Note that in (10), the least-squares objective is defined with respect to the amplitude $|\mathcal{F}(\mathbf{x})|$ and not the magnitude-squared $|\mathcal{F}(\mathbf{x})|^2$. For random measurements, it has been shown in [16] that the amplitude objective leads to superior performance over the standard magnitude-squared approach.

2) Fienup as Majorization-Minimization: In order to understand further the connection with the Fienup algorithm, we define the following function:

$$h(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{2}\|\mathbf{y} - \mathbf{z}(\mathbf{x})\|_2^2 + g(\mathbf{y}).$$

Now, for any $\mathbf{x} \in \mathbb{R}^n$, using Lemma 3.1, we have the following two properties (recalling the definition of F in (10)):

$$h(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{y} - \mathbf{z}(\mathbf{x})\|_2^2 + g(\mathbf{y}) \qquad (11)$$
$$\ge \frac{1}{2}\|\mathbf{y} - \mathbf{z}(\mathbf{y})\|_2^2 + g(\mathbf{y}) = F(\mathbf{y}), \quad \forall \mathbf{y} \in \mathbb{R}^n,$$

and

$$h(\mathbf{x}, \mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{z}(\mathbf{x})\|_2^2 + g(\mathbf{x}) = F(\mathbf{x}).$$

In other words, using the convexity of g, we have that $h(\mathbf{x}, \cdot)$ is a 1-strongly convex global upper bound on the objective F.

Computing this upper bound amounts to performing partial minimization over $\mathbf{z}$ in (2). Minimizing the upper bound $h(\mathbf{x}, \mathbf{y})$ in $\mathbf{y}$ corresponds to partial minimization over $\mathbf{x}$ in (2). The upper bound is tight in the sense that we recover the value of the objective at the current point, $h(\mathbf{x}, \mathbf{x}) = F(\mathbf{x})$. Therefore the alternating minimization algorithm is actually a majorization-minimization method for the nonsmooth least-squares problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2n} \||\mathcal{F}(\mathbf{x})| - \mathbf{c}\|_2^2 + g(\mathbf{x}). \tag{12}$$

The steps presented in (4) can then be summarized as follows:

$$\begin{aligned}
\mathbf{x}^{k+1} &= \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^n} h(\mathbf{x}^k, \mathbf{y}) \\
&= \operatorname{prox}_g(\operatorname{Re}(\mathbf{z}(\mathbf{x}^k))) \\
&= \operatorname{prox}_g(\operatorname{Re}(P_{Z_{\mathbf{c}}}(\mathbf{x}^k))),
\end{aligned}$$

which is exactly the mapping given in (7). Note that there is a strong connection between proximal methods and majorization minimization methods [32]. Finally, $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ is the sequence which corresponds to signal estimates and the sequence $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$ is auxiliary in the majorization minimization interpretations.

C. Projected Gradient Descent Interpretation

We now provide an additional interpretation of the alternating minimization algorithm as a projected gradient method for an optimization problem related to (2) which consists of a smooth convex objective and a nonconvex constraint set. This interpretation is valid whenever g is assumed to be proper, lower semicontinuous and convex.

For any $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{C}^n$, we can write (2) as $\|\mathbf{x} - \mathbf{z}\|_2^2 = \|\mathbf{x} - \operatorname{Re}(\mathbf{z})\|_2^2 + \|\operatorname{Im}(\mathbf{z})\|_2^2$. To move from complex numbers to real numbers, we set $\mathbf{w}_1 = \operatorname{Re}(\mathbf{z})$ and $\mathbf{w}_2 = \operatorname{Im}(\mathbf{z})$. Defining a new constraint set $\tilde{Z}_{\mathbf{c}} = \{(\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^n \times \mathbb{R}^n : \mathbf{w}_1 + i\mathbf{w}_2 \in Z_{\mathbf{c}}\}$, problem (2) is equivalent to

$$\min_{\mathbf{x} \in \mathbb{R}^n, (\mathbf{w}_1, \mathbf{w}_2) \in \tilde{Z}_{\mathbf{c}}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{w}_1\|_2^2 + \frac{1}{2} \|\mathbf{w}_2\|_2^2 + g(\mathbf{x}) \right\}. \tag{13}$$

Minimizing first w.r.t. $\mathbf{x}$, (13) reduces to the following minimization problem in $\mathbf{w}_1, \mathbf{w}_2$:

$$\min_{(\mathbf{w}_1, \mathbf{w}_2) \in \tilde{Z}_{\mathbf{c}}} \left\{ H(\mathbf{w}_1, \mathbf{w}_2) \equiv G(\mathbf{w}_1) + \frac{1}{2} \|\mathbf{w}_2\|^2 \right\}, \tag{14}$$

where

$$G(\mathbf{w}_1) \equiv \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{w}_1 - \mathbf{x}\|_2^2 + g(\mathbf{x}) \right\}.$$

The next result allows us to relate the gradient of H to the optimization primitives used in the alternating minimization method.

Lemma 3.3: Assume that g is proper, lower semicontinuous and convex. Then the function H is continuously differentiable, its gradient is 1-Lipschitz and can be expressed as

$$\nabla H(\mathbf{w}_1, \mathbf{w}_2) = \left(\mathbf{w}_1 - \operatorname{prox}_g(\mathbf{w}_1), \mathbf{w}_2\right). \tag{15}$$

Proof: From Moreau [28, Proposition 7.d], we know that G is differentiable and $\nabla G(\mathbf{x}) = \mathbf{x} - \operatorname{prox}_g(\mathbf{x}) = \operatorname{prox}_{g^*}(\mathbf{x})$, where g^* is the conjugate function of g, which is convex. The

computation of the gradient of H is then immediate. We can use the fact that proximal mappings of convex functions are nonexpansive [28, Proposition 5.b] to verify that ∇H is 1-Lipschitz. Indeed, for any $(\mathbf{w}_1, \mathbf{w}_2)$ and $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2)$, we have

$$\begin{aligned}
&\|\nabla H(\mathbf{w}_1, \mathbf{w}_2) - \nabla H(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2)\|_2^2 \\
&= \|\operatorname{prox}_{g^*}(\mathbf{w}_1) - \operatorname{prox}_{g^*}(\tilde{\mathbf{w}}_1)\|_2^2 + \|\mathbf{w}_2 - \tilde{\mathbf{w}}_2\|_2^2 \\
&\le \|\mathbf{w}_1 - \tilde{\mathbf{w}}_1\|_2^2 + \|\mathbf{w}_2 - \tilde{\mathbf{w}}_2\|_2^2 \\
&= \|(\mathbf{w}_1, \mathbf{w}_2) - (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2)\|_2^2,
\end{aligned}$$

completing the proof. ∎

Consider applying projected gradient descent to solve (14). From Lemma 3.3, we can use a step size of magnitude 1. In this case, taking into account the form of the gradient given in (15), we obtain that the general update step takes the form

$$\begin{aligned}
&(\mathbf{w}_1^{k+1}, \mathbf{w}_2^{k+1}) \\
&= P_{\tilde{Z}_{\mathbf{c}}}((\mathbf{w}_1^k, \mathbf{w}_2^k) - \nabla H(\mathbf{w}_1^k, \mathbf{w}_2^k)) \\
&= P_{\tilde{Z}_{\mathbf{c}}}((\mathbf{w}_1^k, \mathbf{w}_2^k) - (\mathbf{w}_1^k - \operatorname{prox}_g(\mathbf{w}_1^k), \mathbf{w}_2^k)) \\
&= P_{\tilde{Z}_{\mathbf{c}}}(\operatorname{prox}_g(\mathbf{w}_1^k), 0).
\end{aligned}$$

We now go back to the complex domain by setting $\mathbf{z} = \mathbf{w}_1 + i\mathbf{w}_2$. Note that projecting $(\mathbf{w}_1, \mathbf{w}_2)$ onto $\tilde{Z}_{\mathbf{c}}$ is equivalent to projecting $\mathbf{z}$ onto $Z_{\mathbf{c}}$. With this notation, the iterations of projected gradient descent can be summarized by the following iteration mapping (on complex numbers):

$$\mathbf{z}^{k+1} = P_{Z_{\mathbf{c}}}(\operatorname{prox}_g(\operatorname{Re}(\mathbf{z}^k))),$$

which is exactly the same as (7). Therefore, the Fienup algorithm is equivalent to projected gradient descent with unit stepsize applied to the formulation (14). Note that from the point of view of nonsmooth analysis, problem (14) is much better behaved than (12).

IV. CONSEQUENCES AND EXTENSIONS

We now use the interpretations of Section III to analyze the convergence of alternating minimization applied to problem (2), and to offer several extensions.

A. Convergence Analysis

Our main convergence result is given in the following theorem. Recall that a function is semi-algebraic if its graph can be defined by combining systems of polynomial equalities and inequalities (for example, the ℓ_1 norm is semi-algebraic).

Theorem 4.1: Assume that g is proper, lower semicontinuous, convex and semi-algebraic. Then the sequence $\{\mathbf{x}^k, \mathbf{z}^k\}_{k \in \mathbb{N}}$ generated by the alternating minimization algorithm satisfies the following:

i) It holds that $\sum_{k \ge 0} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2 < +\infty$ and the sequence $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ converges to a point $\mathbf{x}^* \in \mathbb{R}^n$.

ii) For any accumulation point $\mathbf{z}^*$ of $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$, $(\mathbf{x}^*, \mathbf{z}^*)$ is a Fréchet critical point of problem (2) and $(\mathbf{w}_1^*, \mathbf{w}_2^*) = (\operatorname{Re}(\mathbf{z}^*), \operatorname{Im}(\mathbf{z}^*))$ is a Fréchet critical point of problem (14).

The proof is quite technical and is given in Appendix A. The semi-algebraic assumption on g can be relaxed to

representability in o-minimal structures over the real field, see [33]. Therefore, the proposed result actually applies to much more general regularizers. For example, using boundedness of the feasible set in (14), the same result holds if g is analytic (see the dicussion in [34, Section 5]). The arguments build upon a nonsmooth variant of the celebrated Kurdyka-Łojasiewicz (KL) property [24]–[27]. Note that direct application of the results of [22], [23] to projected gradient descent or those of [34], [35] to majorization minimization is not possible here. Indeed, [22], [23] require step sizes strictly smaller than $1/L$ where L is the Lipschitz constant of the gradient and convergence does not hold in this case. Furthermore, the approach of [34], [35] requires the local model $h(\mathbf{x}, \mathbf{y})$ to be continuous with respect to both its arguments which is not the case here due to the arbitrary choice of $\mathbf{z}(\mathbf{x})$. These added difficulties require separating the analysis with respect to $\mathbf{x}$ and $\mathbf{z}$ as we propose.

The most important implication of Theorem 4.1 is that the sequence of estimated signals converges smoothly to a point which satisfies certain optimality conditions related to problems (2) and (14). This limit is a candidate solution to the phase retrieval problem and the sequence $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$ allows to describe stationarity properties of this possible solution. On the other hand, our result does not characterize the limit uniquely so that different initializations could still produce different limits, and hence possible solutions.

This approach is a departure from standard convergence results that are only able to guarantee that accumulation points of the generated sequence of iterates satisfy certain optimality conditions. It is important to underline that the result is global: it holds for any initialization of the algorithm and does not require any regularity assumption beyond semi-algebraicity and convexity of g. This is in contrast with local convergence results which are typical for alternating projection methods [36], [37] that are applicable when the prior term g is an indicator function.

B. Acceleration and Momentum Term

A benefit of the interpretation of alternating minimization as a projected gradient method is that it allows proposing new variants inspired by known extensions of projected gradient algorithms. In this section we focus on the incorporation of an inertial term that results in an alternating minimization scheme that includes a momentum term. This line of research has a long history in optimization, starting with the development of the heavy-ball method [38] which inspired an optimal first order scheme for convex optimization developed by Nesterov [39], and its extension to convex composite problems with the FISTA method [21]. Although this last technique was proposed and analyzed only in the context of convex optimization, we consider its application in our nonconvex constrained problem since it empirically exhibits good performances. The resulting algorithm is referred to as FISTAPH, and is described as follows.

FISTAPH: FISTA for Phase retrieval
Initialization. $\mathbf{z}^0 \in Z_\mathbf{c}$ and $\alpha^k \in [0, 1)$ for all $k \in \mathbb{N}$. Set $\mathbf{y}^0 = \mathbf{z}^0$ and $\mathbf{z}^{-1} = \mathbf{z}^0$.
General Step. For $k \in \mathbb{N}$,
- $\mathbf{z}^{k+1} \in P_{Z_\mathbf{c}}(\mathrm{prox}_g(\mathrm{Re}(\mathbf{y}^k)))$.
- $\mathbf{y}^{k+1} = \mathbf{z}^k + \alpha^k (\mathbf{z}^k - \mathbf{z}^{k-1})$.

If $\mathbf{z}^m$ is the last produced iteration, then the output of the algorithm is $\hat{\mathbf{x}} = \mathrm{prox}_g(\mathrm{Re}(\mathbf{z}^m))$. A typical choice for the weight sequence is $\alpha^k = \frac{k-1}{k+2}$. The question of convergence of the iterates produced by this method in nonconvex settings is an interesting topic to explore in future research. We may also further consider monotone variants of similar types of methods, see e.g. [29].

In the numerical experiments we employ FISTAPH in the setting where $g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$ for some $\lambda > 0$. In this case, $\mathrm{prox}_g = \mathcal{T}_\lambda$ with $\mathcal{T}_\lambda$ being the soft thresholding operator with parameter λ (see footnote on page 3).

V. EXPERIMENTS AND NUMERICAL RESULTS

In this section, we describe experiments and numerical results comparing the different algorithms introduced in Section II-C on the task of phase retrieval.

A. Experimental Setup

Given measurements $\mathbf{c}$ as in (1), our problem consists of finding the corresponding $\mathbf{x}_0$. We focus on the setting in which $\mathbf{x}_0$ is known to be sparse. We vary the signal size n (with $J = \{1, 2, \dots, n/2\}$), the sparsity level K and the signal to noise ratio (SNR). In the following discussion, we will refer to a recovery method $\mathcal{M}$ which can be seen as a black box which takes as input a vector of measurements $\mathbf{c} \in \mathbb{R}^n_+$, support information J, sparsity level K, an initial estimate $\mathbf{x}$ and outputs an estimate $\hat{\mathbf{x}} \in \mathbb{R}^n$ with $\mathrm{supp}(\hat{\mathbf{x}}) \subseteq J$ and $\|\mathbf{x}\|_0 \leq K$. One recovery experiment consists of the following:

- Fix a recovery method $\mathcal{M}$, a signal length n, a support information set $J = \{1, 2, \dots, n/2\}$, a sparsity level K and an SNR.
- Generate $\mathbf{x}_0 \in \mathbb{R}^n$ by the following procedure:
  – Choose K coordinates among J uniformly at random.
  – Set these coordinate values at random in $[-4, -3] \cup [3, 4]$.
  – Set all other coordinates to be 0.
- Generate the measurements $\mathbf{c}^2 = |\mathcal{F}(\mathbf{x}_0)|^2 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is white Gaussian noise according to the chosen SNR. Set negative entries of $\mathbf{c}^2$ to be 0 in order to take square root.
- Call method $\mathcal{M}$ 100 times with data $(\mathbf{c}, J, K)$ and randomly generate initial estimates to get 100 candidate solutions $\{\hat{\mathbf{x}}_{it}\}_{it=1,2,\dots,100}$.
- Compute the best estimate $\hat{\mathbf{x}}_{best}$ with $best = \mathrm{argmin}_{it=1,2\dots,100}\{\|\mathcal{F}(\hat{\mathbf{x}}_{it})| - \mathbf{c}\|_2^2\}$.
- Compare $\hat{\mathbf{x}}_{best}$ and $\mathbf{x}_0$ (modulo Fourier invariances) with the following metric (sign is understood coordinatewise with $\mathrm{sign}(0) = 0$):

$$recovery(\hat{\mathbf{x}}_{best}, \mathbf{x}_0) = \begin{cases} 1, & \mathrm{sign}(\hat{\mathbf{x}}_{best}) = \mathrm{sign}(\mathbf{x}_0) \\ 0, & \text{otherwise.} \end{cases}$$

This procedure was repeated 100 times. That is, for each method, signal length, sparsity level and SNR, we have 100 signal recovery experiments, each one associated with a support recovery status. We aggregate these results by considering the recovery probability (average of $recovery(\hat{\mathbf{x}}_{best}, \mathbf{x}_0)$) and the median CPU usage for a single simulation (100 calls to the method with different initialization estimates). We use the same initialization for all methods by careful initialization of random seeds. All the

experiments were performed on a desk station with two 3.2 GHz Quad Core Intel Xeon processors and 8 GB of RAM.

B. Implementation Details

In our numerical implementation, we used the following stopping criterion.

- For alternating minimization and Wirtinger methods: the difference in successive objective value less than 10^{-8}.
- For GESPAR: no swap improvement.
- For FISTAPH: the norm of the gradient mapping less than 10^{-8}.

The tuning of these criteria allows to balance accuracy and computational time to some extent.

The ℓ_1 penalized problem includes a prior sparsity inducing term of the form $g(\cdot) = \lambda \| \cdot \|_1$. It is necessary to tune the λ parameter in order to obtain meaningful results. We considered the following strategies for different methods.

- For alternating minimization, we set $\lambda = 0.2$ in all experiments.
- For Wirtinger based method, λ is tuned a posteriori as a function of n and k. The experiment was conducted for $\lambda = 1, 2.15, 4.64, 10, 21.5, 46.4, 100, 215, 464$, and we report only the best experiment for each setting.

An interesting feature of alternating minimization based methods is that in our experiments, recovery performance was very consistent for different values of λ in different settings. As a result, we chose a single value of λ for all experiments. The tuning of λ for Wirtinger based algorithms is practically much more difficult. In particular, we found that the best λ was a highly dependant function of the sparsity level K.

Finally, we note that ℓ_0 based priors have the sparsity level of the estimate, K, as a parameter. On the other hand, ℓ_1 based priors will not necessarily produce K-sparse estimates. We therefore use truncation and keep the K largest entries in absolute value of the last iteration.

C. Numerical Results

The performance in terms of support recovery are presented in Fig. 1 with the corresponding algorithm run time in Fig. 2. Each point in these plots is an average over 100 simulations of the recovery process, each simulation consisting of 100 random initializations of the method considered. AM corresponds to Fienup methods with different priors, FISTA is the accelerated variant, and WIRT stands for Wirtinger.

We make the following observations from the numerical results:

- For alternating minimization, there is a consistent increase in recovery performance by switching from ℓ_0 to ℓ_1 based regularization priors.
- The ℓ_1 prior degrades the performance of Wirtinger based methods compared to the ℓ_0 prior.
- FISTAPH consistently provides the best performance and is significantly faster than its competitors.
- Fienup with ℓ_0 prior leads to lower performance compared to GESPAR, which was already reported in [4].

In order to evaluate the benefits of sparsity in reducing the number of measurements, we compared Wirtinger flow with ℓ_0 prior, GESPAR and FISTAPH on the same recovery problems ($N = 256$ and SNR $= 20$) with the difference that we only keep
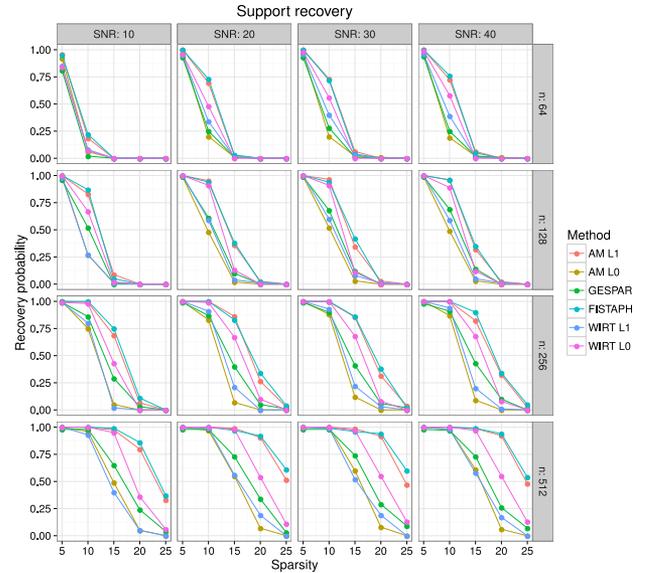

Fig. 1. Support recovery comparison. For each point, the probability is estimated based on 100 simulations. AM stands for alternating minimization and WIRT for WIRTINGER. FISTAPH is described in Section IV, and GESPAR is the method presented in [4].
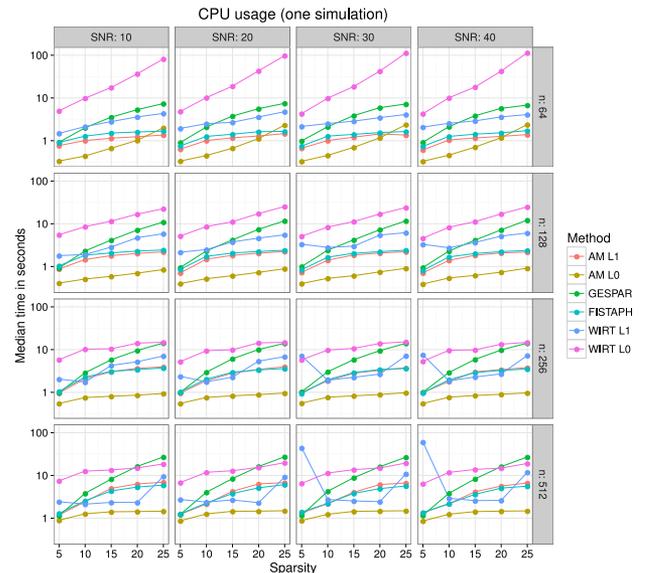

Fig. 2. Timing comparison. The ordinate axis is displayed in logarithmic scale. Each point is the median over 100 simulations, each simulation consisting of 100 random initializations for every method. AM stands for alternating minimization and WIRT for WIRTINGER. FISTAPH is described in Section IV, and GESPAR is the method presented in [4].

the first measurements (for other experiments, the number of measurements is equal to the size of the signal). The results are presented in Fig. 3. We observe that for all three methods, using half the number of measurements allows to maintain a level of recovery rate close to the setting where all measurements are used. FISTAPH benefits the most from this effect. In the simulations, the choice of the regularization parameter was scaled proportionally to the number of measurements. The phase transition curve is roughly linear until a certain sparsity
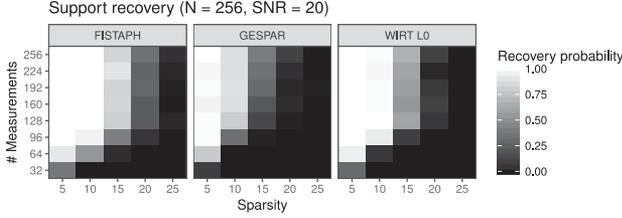
Fig. 3. Recovery as a function of the number of measurements and the sparsity level. For each point, the probability is estimated based on 100 simulations. WIRT stands for WIRTINGER. FISTAPH is described in Section IV, and GESPAR is the method presented in [4].
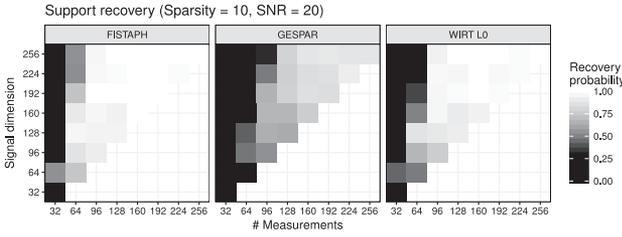

Fig. 4. Recovery as a function of the number of measurements and the dimension. The sparsity level is fixed to 10. For each point, the probability is estimated based on 100 simulations. WIRT stands for WIRTINGER. FISTAPH is described in Section IV, and GESPAR is the method presented in [4]. We only consider the upper diagonal part for each window.

threshold is reached, after which, recovery does not occur even if the number of measurements increases. For the same three algorithms, we considered a fixed sparsity level and SNR ratio (10 and 20 respectively) and varied the dimension and number of measurements. The results are presented in Fig. 4. This experiment suggests that the main factor explaining if recovery occurs or not is the number of measurements rather than signal dimension. Proposing theoretical ground for these empirical observations will be the subject of future work.

As described in the experimental section, we added noise on the squared measurements rather than on the measurements themselves. This noise model is closer to the optimization formulation considered for GESPAR and Wirtinger flow than model (2) which is related to problem (12). We tried changing the noise model on a subset of experiments (additive noise on the measurements rather than squared measurements), however, the performance of the different methods was very similar. Therefore, we only report results related to the squared-measurement noise model.

VI. CONCLUSION

The main theoretical contribution of this work is to provide a strong theoretical basis to the fact that Fienup-type methods, when used with Fourier transforms and convex priors, lead to smoothly converging sequences of estimates. This result holds under minimal assumptions and in particular, it holds globally, independently of the initialization point. Furthermore, we characterize the properties of the limiting point as Fréchet critical points of different optimization problems. These results shed light on properties of one of the most well known algorithms used in the context of phase retrieval. Furthermore, based on an interpretation as a projected gradient method, we proposed

a new variant of Fienup with the incorporation of a momentum term which we call FISTAPH.

On the practical side, we demonstrated based on numerical simulations that FISTAPH with ℓ_1 regularization constitutes a very competitive alternative to other methods in the context of sparse phase retrieval. Although this observation is only based on empirical evidence, it potentially illustrates the fact that momentum based methods are able to avoid bad local minima and saddle points in complicated nonconvex landscapes.

APPENDIX
PROOF OF THEOREM 4.1

The proof of Theorem 4.1 involves many notions of nonsmooth analysis which can be found in [30]. Throughout the proof, we only consider subgradients of subdifferentially regular functions. Each subgradient can be interpreted as a Fréchet subgradient and the subgradient set valued mapping is closed. We adopt the notation of Section III-C, letting $\mathbf{z} = \mathbf{w}_1 + i\mathbf{w}_2$ for two real vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ and consider the constraint set $\tilde{Z}_{\mathbf{c}} = \{(\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^n \times \mathbb{R}^n; \; \mathbf{w}_1 + i\mathbf{w}_2 \in Z_{\mathbf{c}}\}$. We let $K(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2) = \frac{1}{2}\|\mathbf{x} - \mathbf{w}_1\|_2^2 + \frac{1}{2}\|\mathbf{w}_2\|_2^2 + g(\mathbf{x})$ be the objective function of problem (2) which with this notation becomes

$$\min_{\mathbf{x} \in \mathbb{R}^n, (\mathbf{w}_1, \mathbf{w}_2) \in \tilde{Z}_{\mathbf{c}}} K(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2). \tag{16}$$

We will denote by $\delta_{\tilde{Z}_{\mathbf{c}}}$, the indicator function of the set $\tilde{Z}_{\mathbf{c}}$ (0 on the set and $+\infty$ outside). We set $\tilde{K}(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2) = K(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2) + \delta_{\tilde{Z}_{\mathbf{c}}}(\mathbf{w}_1, \mathbf{w}_2)$ so that problem (16) is equivalent to the (unconstrained) minimization of $\tilde{K}$.

Proof of (i)**:** Using [30, Proposition 10.5 and Exercise 10.10], the subgradient of this nonsmooth function is of the form

$$\partial \tilde{K}(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2) = \begin{pmatrix} \partial_{\mathbf{x}} \tilde{K}(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2) \\ \partial_{(\mathbf{w}_1, \mathbf{w}_2)} \tilde{K}(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2) \end{pmatrix} \tag{17}$$

$$= \begin{pmatrix} \mathbf{x} - \mathbf{w}_1 + \partial g(\mathbf{x}) \\ \begin{pmatrix} \mathbf{w}_1 - \mathbf{x} \\ \mathbf{w}_2 \end{pmatrix} + \partial \delta_{\tilde{Z}_{\mathbf{c}}}(\mathbf{w}_1, \mathbf{w}_2) \end{pmatrix}.$$

Partial minimization over iterations the following

$$0 \in \mathbf{x}^{k+1} - \mathbf{w}_1^k + \partial g(\mathbf{x}^{k+1}) \tag{18}$$

$$0 \in \begin{pmatrix} \mathbf{w}_1^k - \mathbf{x}^k \\ \mathbf{w}_2^k \end{pmatrix} + \partial \delta_{\tilde{Z}_{\mathbf{c}}}(\mathbf{w}_1^k, \mathbf{w}_2^k). \tag{19}$$

Combining (17) (18) and (19), we have

$$\begin{pmatrix} 0 \\ \begin{pmatrix} \mathbf{w}_1^k - \mathbf{x}^{k+1} \\ \mathbf{w}_2^k \end{pmatrix} + \partial \delta_{\tilde{Z}_{\mathbf{c}}}(\mathbf{w}_1^k, \mathbf{w}_2^k) \end{pmatrix} \tag{20}$$

$$\subset \partial \tilde{K}(\mathbf{x}^{k+1}, \mathbf{w}_1^k, \mathbf{w}_2^k).$$

Using (19),

$$\begin{pmatrix} 0 \\ \mathbf{x}^k - \mathbf{x}^{k+1} \\ 0 \end{pmatrix} \in \partial \tilde{K}(\mathbf{x}^{k+1}, \mathbf{w}_1^k, \mathbf{w}_2^k). \tag{21}$$

Finally, from strong convexity of $\tilde{K}$ with respect to its first argument, we have

$$\tilde{K}\big(\mathbf{x}^{k+1}, \mathbf{w}_1^k, \mathbf{w}_2^k\big) + \frac{1}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2$$
$$\leq \tilde{K}\big(\mathbf{x}^k, \mathbf{w}_1^k, \mathbf{w}_2^k\big)$$
$$\leq \tilde{K}\big(\mathbf{x}^k, \mathbf{w}_1^{k-1}, \mathbf{w}_2^{k-1}\big). \qquad (22)$$

Since g is semi-algebraic, $\tilde{K}$ is also semi-algebraic. Any semi-algebraic function satisfies the nonsmooth Kurdyka-Łojasievicz property [27]. We can now use the well established recipe [22, Section 2.3] [23, Section 3.2] with the two conditions (21) and (22) to obtain that the sequence $\big\{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2\big\}_{k\in\mathbb{N}}$ is summable. This proves statement (i) (convergence holds by Cauchy criterion).

Proof of (ii)**:** Using the fact that $\tilde{K}$ has compact sublevel sets, the sequence $\big\{(\mathbf{x}^{k+1}, \mathbf{w}_1^k, \mathbf{w}_2^k)\big\}_{k\in\mathbb{N}}$ is bounded and hence has a converging subsequence. We fix an accumulation point $(\mathbf{x}^*, \mathbf{w}_1^*, \mathbf{w}_2^*)$ of the sequence (note that $\mathbf{x}^*$ is given by (i)). We remark that, thanks to (20) and the fact that $\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \to 0$, any accumulation point of the sequence is a critical point of $\tilde{K}$. Furthermore, since $\mathbf{x}^k \to \mathbf{x}^*$, we have using (20) that

$$-\begin{pmatrix} \mathbf{w}_1^* - \mathrm{prox}_g\big(\mathbf{w}_1^*\big) \\ \mathbf{w}_2^* \end{pmatrix} \in \partial \delta_{\tilde{Z}_{\mathrm{c}}}\big(\mathbf{w}_1^*, \mathbf{w}_2^*\big).$$

This is actually the criticality condition for problem (14) which proves statement (ii).

REFERENCES

[1] N. E. Hurt, *Phase Retrieval and Zero Crossings: Mathematical Methods in Image Reconstruction*, vol. 52. New York, NY, USA: Springer, 2001.

[2] 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.*, vol. 32, no. 3, pp. 87–109, May 2015.

[3] K. Jaganathan, S. Oymak, and B. Hassibi, "Recovery of sparse 1-D signals from the magnitudes of their Fourier transform," in *Proc. IEEE Int. Sympos. Inf. Theory*, Cambridge, MA, 2012, pp. 1473–1477.

[4] Y. Shechtman, A. Beck, and Y. Eldar, "GESPAR: Efficient phase retrieval of sparse signals," *IEEE Trans. Signal Process.*, vol. 62, no. 4, pp. 928–938, Feb. 2014.

[5] H. Ohlsson and Y. C. Eldar, "On conditions for uniqueness in sparse phase retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1841–1845.

[6] A. S. Y. Shechtman, Y. C. Eldar, and M. Segev, "Sparsity based subwavelength imaging with partially incoherent light via quadratic compressed sensing," *Opt. Express*, vol. 19, no. 16, pp. 14 807–14 822, 2011.

[7] R. Gerchberg and W. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[8] J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," *Opt. Lett.*, vol. 3, no. 1, pp. 27–29, 1978.

[9] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Appl. Opt.*, vol. 21, no. 15, pp. 2758–2769, 1982.

[10] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization," *J. Opt. Soc. Amer. A*, vol. 19, no. 7, pp. 1334–1345, 2002.

[11] S. Marchesini, "Phase retrieval and saddle-point optimization," *J. Opt. Soc. Amer. A*, vol. 24, no. 10, pp. 3289–3296, 2007.

[12] Z. Wen, C. Yang, X. Liu, and S. Marchesini, "Alternating direction methods for classical and ptychographic phase retrieval," *Inverse Probl.*, vol. 28, no. 11, 2012, Art. no. 115010.

[13] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4814–4826, Sep. 2015.

[14] I. Waldspurger, "Phase retrieval with random gaussian sensing vectors by alternating projections," arXiv:1609.03088, 2016.

[15] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[16] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Trans. Inf. Theory*, preprint, doi: 10.1109/TIT.2017.2756858.

[17] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, MaxCut and complex semidefinite programming," *Math. Program.*, vol. 149, no. 1–2, pp. 47–81, 2015.

[18] H. Ohlsson, A. Yang, R. Dong, and S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," in *Proc. 16th IFAC Symp. Syst. Identif.*, 2012, pp. 89–94.

[19] K. Jaganathan, S. Oymak, and B. Hassibi, "Recovery of sparse 1-d signals from the magnitudes of their Fourier transform," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 1473–1477.

[20] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imag. Sci.*, vol. 6, no. 1, pp. 199–225, 2013.

[21] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[22] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods," *Math. Program.*, vol. 137, no. 1–2, pp. 91–129, 2013.

[23] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, no. 1–2, pp. 459–494, 2014.

[24] S. Łojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," in *Les Équations aux Dérivées Partielles*, Éditions du Centre National de la Recherche Scientifique, 1963, vol. 117, pp. 87–89.

[25] K. Kurdyka, "On gradients of functions definable in o-minimal structures," *Annales de l'Institut Fourier*, vol. 48, no. 3, pp. 769–783, 1998.

[26] J. Bolte, A. Daniilidis, and A. S. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM J. Optim.*, vol. 17, no. 4, pp. 1205–1223, 2007.

[27] J. Bolte, A. Daniilidis, A. S. Lewis, and M. Shiota, "Clarke subgradients of stratifiable functions," *SIAM J. Optim.*, vol. 18, no. 2, pp. 556–572, 2007.

[28] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299, 1965.

[29] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery," *Convex Optimization in Signal Processing and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[30] R. T. Rockafellar and R. Wets, *Variational Analysis*, vol. 317. New York, NY, USA: Springer, 1998.

[31] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. New York, NY, USA: Springer, 2011, pp. 185–212.

[32] C. Byrne and J. S. Lee, "Alternating minimization, proximal minimization and optimization transfer are equivalent," *J. Nonlinear and Convex Anal.*, to be published.

[33] L. V. D. Dries and C. Miller, "Geometric categories and o-minimal structures," *Duke Math. J.*, vol. 84, no. 2, pp. 497–540, Aug. 1996.

[34] J. Bolte and E. Pauwels, "Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs," *Math. Oper. Res.*, vol. 41, no. 2, pp. 442–465, 2016.

[35] E. Pauwels, "The value function approach to convergence analysis in composite optimization," *Oper. Res. Lett.*, vol. 44, no. 6, pp. 790–795, 2016.

[36] A. S. Lewis, D. R. Luke, and J. Malick, "Local linear convergence for alternating and averaged nonconvex projections," *Found. Comput. Math.*, vol. 9, no. 4, pp. 485–513, 2009.

[37] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang, "Restricted normal cones and the method of alternating projections: Applications," *Set-Valued Variational Anal.*, vol. 21, no. 3, pp. 475–501, 2013.

[38] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[39] Y. Nesterov, "A method of solving a convex programming problem with convergence rate O(1/k2)," *Sov. Math. Doklady*, vol. 27, no. 2, 1983, pp. 372–376.

Edouard Jean Robert Pauwels received the Ph.D. degree in November 2013 at Center for Computational Biology, Mines ParisTech, Paris, France, working on the application of machine learning techniques in biological contexts. During a Postdoc in MAC team with LAAS-CNRS, Toulouse, France, he studied the application of polynomial optimization to direct and inverse optimal control. From 2014 to 2015, he was a Postdoctoral Researcher in the Operation Research Group, Faculty of Industrial Engineering and Management in Technion, Haifa, Israel, working on large scale convex optimization. He is currently an Assistant Professor (Maître de conférence) with the University of Toulouse III Paul Sabatier, Toulouse, France. He is affiliated to the Informatics Department (IRIT, Adria team). His research interest focuses on numerical optimization with applications to machine learning, robotics, and computational biology.

Amir Beck is currently a Professor with the School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel. He has authored or co-authored numerous papers and given invited lectures at international conferences. His research interest focuses on continuous optimization, including theory, algorithmic analysis, and applications. He is currently an Associate Editor for *Mathematics of Operations Research*, *Mathematical Programming Series A*, and the *Journal of Optimization Theory and Applications*, and an Area Editor for optimization in *Operations Research*. His research has been supported by various funding agencies, including the Israel Science Foundation, the German–Israeli Foundation, the Binational US–Israel Foundation, the Israeli Science and Energy Ministries, and the European Community. He was the recipient of the Salomon Simon Mani Award for Excellence in Teaching and the Henry Taub Research Prize.

Yonina C. Eldar (S'98–M'02–SM'07–F'12) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering in 1995 and 1996, respectively, both from Tel Aviv University, Tel Aviv, Israel, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2002.

She is currently a Professor with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel, where she holds the Edwards Chair in Engineering. She is also a Research Affiliate with the Research Laboratory of Electronics, MIT, an Adjunct Professor with Duke University, Durham, NC, USA, and was a Visiting Professor with Stanford University, Stanford, CA, USA. She is the author of the book *Sampling Theory: Beyond Bandlimited Systems* and the coauthor of the books *Compressed Sensing* and *Convex Optimization Methods in Signal Processing and Communications*, all published by Cambridge University Press. Her research interests include the broad areas of statistical signal processing, sampling theory and compressed sensing, optimization methods, and their applications to biology and optics.

Dr. Eldar is a member of the Israel Academy of Sciences and Humanities (elected 2017) and a EURASIP Fellow. She was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. She was a member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education. She is the Editor-in-Chief of *Foundations and Trends in Signal Processing*, a member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. In the past, she was a Signal Processing Society Distinguished Lecturer, member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the EURASIP JOURNAL OF SIGNAL PROCESSING, the SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS, and the SIAM JOURNAL ON IMAGING SCIENCES. She was the Cochair and Technical Cochair of several international conferences and workshops. He was the recipient of many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), the IEEE Kiyo Tomiyasu Award (2016), the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair at the Technion, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion's Award for Excellence in Teaching (two times). She was also the recipient of several best paper awards and best demo awards together with her research students and colleagues including the SIAM outstanding Paper Prize, the UFFC Outstanding Paper Award, the Signal Processing Society Best Paper Award, and the IET Circuits, Devices and Systems Premium Award. She was selected as one of the 50 most influential women in Israel.

Shoham Sabach received the B.Sc. degree in mathematics (*summa cum laude*) and the M.Sc. degree in mathematics (*magna cum laude*) from the University of Haifa, Haifa, Israel, in 2004 and 2008, respectively, and the Ph.D. degree in mathematics from the Technion—Israel Institute of Technology, Haifa, Israel, in 2012. From October 2012 to September 2013, he was a Postdoctoral Fellow with the School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel. From October 2013 to September 2014, he was a Humboldt Postdoctoral Research Fellow with the Institute for Numerical and Applied Mathematics, University Gottingen, Gottingen, Germany. He is currently an Assistant Professor with the Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology. His research interest focuses on continuous optimization, including theory, algorithmic analysis, and its applications.