Robust Phase Retrieval via Reverse Kullback-Leibler Divergence and Wirtinger Flow

Nazia Afroz Choudhury∗†, Bariscan Yonel∗†, Member, IEEE and Birsen Yazici∗†, Fellow, IEEE

Abstract—Robustness to noise and outliers is a desirable trait in phase retrieval algorithms for many applications in imaging and signal processing. In this paper, we develop a novel robust phase retrieval algorithm based on the minimization of reverse Kullback-Leibler divergence (RKLD) within the Wirtinger Flow (WF) framework. We use RKLD over intensity-only measurements in two distinct ways: i) to design a novel initial estimate based on minimum distortion design of spectral estimates, and ii) as a loss function for iterative refinement based on WF. The RKLD-based loss function offers implicit regularization by processing data at the logarithmic scale and provides the following benefits: suppressing the influence of large magnitude errors and promoting projections orthogonal to noise subspace. We present three algorithms based on RKLD minimization, including two with truncation schemes to enhance the robustness to significant contamination. Our numerical study demonstrates the advantages of our algorithms in terms of sample efficiency, convergence speed, and robustness to outliers over the state-of-the-art techniques using both synthetic and real optical imaging data.

Index Terms—Phase retrieval, reverse Kullback-Leibler divergence, Wirtinger flow, phaseless imaging, robustness

I. INTRODUCTION

A. Motivation and Overview of Our Approach

Phase retrieval [1], [2] is a problem of great interest in signal processing [3], [4] and imaging [5], [6]. The problem refers to the recovery of a real or complex unknown signal from its intensity only measurements. In many applications, the measurements are modeled as

\[ y_m = |\langle a_m, x \rangle|^2, \quad m = 1, 2, \cdots, M \] (1)

where \( x \in \mathbb{C}^N \) is the unknown signal and \( a_m \in \mathbb{C}^N \) is the \( m \)-th sampling or measurement vector. Thus, the phase retrieval problem requires solving a system of \( M \) quadratic equations, which, in general, is known to be NP-hard [7].

In many practical applications, the measurements are contaminated with statistical and additive outliers and noise. Hence, robustness in phase retrieval algorithms is a desirable trait. In this paper, we introduce novel, robust algorithms based on the minimization of reverse Kullback-Leibler Divergence (RKLD) to address phase retrieval in the presence of noise and outliers. Conventionally, the choice of a particular loss function for an inference problem is motivated through the underlying noise distribution in the measurement process. In the context of phase retrieval, our rationale to study RKLD, an information-theoretic measure of dissimilarity, stems from having strictly non-negative measurements. Due to the strict non-negativity of the search in the range of the model in (1), we approach the problem from the perspective of probability density estimation, where RKLD minimization becomes similar to the process described in [6]. Most importantly, RKLD minimization generates robust estimates in the presence of additive outliers due to its logarithmic processing of the measurements, because logarithmic minimization i) suppresses the influence of large magnitudes in residuals and ii) promotes projections orthogonal to noise subspace.

In our approach, we use the Wirtinger Flow (WF) framework [9] to minimize the RKLD-loss and develop three different algorithms for phase retrieval, namely the RKLD-based Wirtinger Flow (RKLD-WF), the median truncated Wirtinger Flow (RKLD-MTWF), and the gradient truncated Wirtinger Flow (RKLD-GTWF). All these algorithms have the following key features: i) the RKLD-based minimization instead of the classical \( \ell_2 \) minimization, and ii) the RKLD-based spectral initialization based on the method we developed in [10]. We further investigate the enhancements to our method by adapting the sample truncation schemes used in the phase retrieval literature for the Gaussian sampling model. We present numerical simulations with synthetic and real data [11] to compare the performance of the RKLD-loss with those of \( \ell_2 \) and Poisson-loss as well as other state-of-the-art algorithms. Our experimental results demonstrate improvements over the state-of-the-art in terms of robustness to additive noise and outliers, sample efficiency, and convergence speed.

B. Related Work and Advantages of Our Approach

Early studies of phase retrieval developed heuristic error reduction algorithms based on alternating projection methods [12][14]. In the last decade, several convex and non-convex algorithms, with exact recovery guarantees, have been proposed [9], [15][21]. Most commonly, convex approaches rely on semidefinite relaxation to recover the unknown vector in the form of a rank-one positive definite matrix in the lifted domain. In [18] and [22], linear programming-based approaches that solve phase retrieval in the original signal space, are presented. Moreover, in [23] and [24], convex methods that are robust to additive noise and outliers, are
developed. The theoretical analysis in \[23\] and \[24\] shows that these algorithms can recover the unknown exactly from an optimal number of Gaussian measurements, when a constant fraction of the measurements is corrupted by arbitrary outliers. Recently, computationally efficient non-convex approaches, specifically WF \[9\] and its variants \[7, 25–33\], have gained popularity. Despite the non-convexity of the problem, the WF-based algorithms successfully deploy a gradient descent based procedure by initially fielding an accurate estimate sufficiently close to the true solution based on the classical spectral methods \[34\]. However, the corresponding recovery guarantees in \[9, 18, \] and \[34\] rely on certain statistical measurement models, namely, the Gaussian sampling, or coded-diffraction patterns. This limitation is mitigated in the analysis of the Wirtinger Flow (WF) framework in \[35\], which identifies a sufficient condition on the lifted forward map to guarantee exact recovery.

The WF-inspired variants control the impact of the additive outliers as well as the statistical ones in the measurements, arising from the sampling vectors \[7, 25–29\]. To this end, the existing algorithms can be grouped under two major categories: \(i\) sample truncation methods to eliminate the contribution of statistical outliers in model mismatch \[7, 25, 27, 30\], \(ii\) “reshaping” of the least-squares formulation by using the amplitude measurements, i.e., \(|\langle a_m, x \rangle|\), as the argument of the \(\ell_2\)-loss function to suppress the outliers in the data arising from the sampling vectors \[7, 26–29\]. However, the amplitude-based least-squares formulation is not smooth and susceptible to converging to local minima compared to the intensity-based formulation as observed in \[36\] and \[37\].

The standard WF algorithm \[9\] is based on the least-squares formulation of the problem, i.e., minimization of the \(\ell_2\)-loss over the intensity-only measurements defined as follows:

\[
\min_z f(z) := \frac{1}{2M} \sum_{m=1}^{M} (y_m - |\langle a_m, z \rangle|^2)^2. \tag{2}
\]

It is well-known that a fundamental limitation of least-squares regression is its susceptibility to biasing estimates towards outliers \[38\]. This susceptibility stems from the squared penalization of the estimation error, which effectively increases the weight of contaminated samples, thereby degrading the quality of the estimate. Even with ideal data satisfying \(1\), this susceptibility has been observed as a source of degradation in computational and sample efficiency of the WF method under the Gaussian sampling model due to the presence of statistical outliers in the measurement vectors \[9\]. In \[25\], the Poisson-loss function, equivalent to forward KL-divergence (FKLD) over the synthesized and the observed measurements, is studied within the WF framework. For \(x \in \mathbb{R}^N\) and real Gaussian measurement vectors, the gradient of this loss function may become too large due to small values of \(|a_m^* x|\) in the denominator and, as a result, the iterates may leave the basin of attraction around the true solution. Therefore, a sample truncation method is employed to eliminate the unfavorable measurement vectors. However, truncating small \(|a_m^* z|\)’s prevents this method to promote projections orthogonal to noise subspace and, hence, the statistical bias of the gradient increases slightly. Unlike FKLD-based approach \[25\] and the standard WF algorithm \[9\], RKLD-based minimization derives stable gradients without the requirement of sample truncation for real-valued data. The gradient of RKLD-based loss function processes the observed and synthesized intensities on a logarithmic scale, thereby suppresses the effect of large magnitudes inherently. Moreover, as observed empirically, unlike the \(\ell_2\) and Poisson-loss schemes, RKLD-based algorithms do not require an adaptive step-size to converge in the gradient descent iterations while requiring fewer iterations than the ones in \[9, 26, \] and \[27\].

Minimizing the \(\ell_2\)-loss function implicitly assumes that the measurements follow a Gaussian noise model. Similarly, minimizing the FKLD implicitly assumes Poisson noise model \[39\]. Unlike the \(\ell_2\)-loss and FKLD-loss, RKLD-loss is agnostic to the statistical distribution of measurements. This provides a desirable flexibility for inference using data with a variety of imperfections, such as those due to dynamic imaging environments, multiple scattering, or interference. Furthermore, the RKLD minimization-based approach can pursue data fit around additive outliers and potential model mismatches due to its distribution-agnostic formulation. In \[36\] and \[40\], FKLD is considered as a loss function within the classical Gerchberg-Saxton algorithm \[12, 41\]. This approach reportedly provides improved robustness to additive mixed Gaussian and Poisson noise, especially at low signal-to-noise-ratios (SNRs). Recently, \[42\] studied Bregman divergences for audio signal recovery and demonstrated improved performance of both FKLD and RKLD-based loss functions over the \(\ell_2\)-loss in low SNR conditions.

Our truncation techniques are inspired from the Truncated Wirtinger Flow (TWF) \[25\] and the Median Truncated Wirtinger Flow (median-TWF) algorithms \[27\] which utilize the Poisson-loss function. The TWF algorithm \[25\] uses a sample mean-based truncation mechanism. However, in the presence of strong additive outliers, it fails to recover the phase information due to high magnitude samples. The reshaped \(\ell_2\)-loss based algorithms such as Reshaped Wirtinger Flow (RWF) \[26\], and Truncated Amplitude Flow (TAF) \[7\] yield stable recovery in additive noise-settings but fail to perform successfully in the presence of large magnitude additive outliers. In \[26\], the properties of median estimator are exploited to develop the median-TWF and median-RWF algorithms using the Poisson and the reshaped \(\ell_2\)-loss functions, respectively. These algorithms demonstrate improved robustness against additive noise and outliers than those of other truncation-based approaches \[7, 25\]. However, these algorithms require suboptimal sample complexity for exact recovery in the presence of sparse outliers. Our numerical study using both synthetic and real data demonstrates that the RKLD-WF and RKLD-MTWF algorithms have superior sample efficiency than those in \[9\] and \[27\]. In \[28\], the Robust Wirtinger Flow (Robust-WF) algorithm is developed. This algorithm avoids truncation by applying a hard-thresholding technique; however, it requires a priori knowledge of the fraction of corrupted measurements. Unlike Robust-WF, our approach does not require such a priori information.
II. PROBLEM FORMULATION

A. Measurement Model

In practice, the measurements typically contain additive noise and arbitrary outliers [27], [43]. Therefore, we modify (1) and express the measurement model in the following form:

\[
y = |Ax|^2 + \eta + w,
\]

where \( y = [y_1, \ldots, y_M]^T \in \mathbb{R}^M \) is the vector containing the measured data, \( A = [a_m^1, \ldots, a_m^M] \in \mathbb{C}^{M \times N} \) is the sampling matrix where \((\cdot)^*\) denotes the conjugate transpose of \( a_m \) described in (1). Here, \( \eta \in \mathbb{R}^M \) denotes the outlier vector which is assumed to be sparse, i.e., its \( \ell_0 \)-norm \( ||\eta||_0 \) satisfies \( ||\eta||_0 \leq pM \), where \( 0 < p < 1 \) is an arbitrary constant denoting the fraction of measurements corrupted by outliers. In addition, we assume that the elements of \( \eta \) take arbitrary values and satisfy \( ||\eta||_\infty \leq \theta ||x||_2^2 \) where \( ||\cdot||_\infty \) denotes the \( \ell_\infty \)-norm of \( \eta \), \( ||\cdot||_2 \) denotes the \( \ell_2 \)-norm of \( x \) and \( \theta > 0 \) is some arbitrary constant. Here, \( w \in \mathbb{R}^M \) denotes the noise vector that, we assume, is bounded in magnitude and satisfies \( ||w||_\infty \leq \sigma ||x||_2^2 \) for some constant \( \sigma > 0 \).

Phase retrieval is an ill-posed problem since both \( x \) and \( xe^{i\phi} \), \( \phi \in [0, 2\pi) \), result in the same magnitude measurements. Hence, we can recover the unknown only up to a global phase factor.

B. The RKLD-Based Loss Function

KLD, on the \( M \)-standard simplex, is a measure of dissimilarity between two probability distributions. More generally, beyond \( M \)-simplex, KLD measures the dissimilarity between two vectors, \( p, q \in \mathbb{R}_+^M \) as follows [10]:

\[
D_{KL}(q, p) = \sum_{m=1}^M q_m \log \frac{q_m}{p_m} - \sum_{m=1}^M (q_m - p_m),
\]

where \( p_m \) and \( q_m \) are the \( m \)-th elements of the vectors \( p \) and \( q \), respectively. KLD is noncommutative in its input arguments, which implies that FKLD, denoted by \( D_{KL}(p, q) \), and RKLD, denoted by \( D_{KL}(q, p) \), are not equal.

In this paper, we derive three algorithms and a spectral initialization scheme based on RKLD. Therefore, substituting \( q \) and \( p \) with \( |Az|^2 \) and \( y \), respectively, in (3), we define the mismatch between the estimates and the measurements as follows:

\[
D_{KL}(|Az|^2, y) = \sum_{m=1}^M |a_m^*z|^2 \log \frac{|a_m^*z|^2}{y_m} - \sum_{m=1}^M (|a_m^*z|^2 - y_m).
\]

The RKLD defined in (3), is a non-convex function which is nonnegative everywhere and has global minima when the estimate is equal to the true solution.

When the measurements are Poisson distributed, the corresponding negative log-likelihood function is equivalent to the FKLD, which has been studied in truncation based phase retrieval algorithms [25], [27]. In FKLD, the log-mismatch term, \( \log \frac{|a_m^*z|^2}{y_m} \), is weighted with the corresponding actual measurement. However, in RKLD, we use the estimated measurements as the weighting factors, which are being refined during each iteration. Thus, the optimization landscape provided by RKLD is significantly different from that of FKLD. Fig. 1 illustrates the dissimilarity between the loss functions derived from these two divergences for \( z \in \mathbb{R}^2 \) and \( A \in \mathbb{C}^{1000 \times 2} \).

The key features of our interest in RKLD stem from the logarithmic processing of the underlying measurements \( \{y_m\}_{m=1}^M \). Firstly, the logarithmic processing naturally suppresses outliers in the data \( \{y_m\}_{m=1}^M \) by shrinking the dynamic range of measurements. This prevents strong outliers to heavily influence the gradient computation, where the multiplicative factors reduce to additive residuals over the logarithm. Another interesting property is the fact that logarithmic processing promotes orthogonality. For \( y_m \to 0 \), at any estimate that does not satisfy \( |a_m^*z|^2 \to 0 \), the loss \( f \) diverges. Essentially, \( y_m = 0 \) inserts a hard-constraint for the estimates, as the set of sampling vectors \( \{a_m\}_{I(y=0)} \) provide an orthogonal subspace.
to the solution, where \( I(y = 0) \) denotes the index-set of the measurements that are equal to 0.

Note that this orthogonal subspace projection is pursued inherently, and it is in approximate effect for \( y_m \approx 0 \). We can express the loss function in an alternative form to bring out its orthogonality promoting nature as follows:

\[
\tilde{f}(z) = \tilde{f}(z) + i (z \perp \{a_m\}_{m \in I(y = 0)}),
\]

where \( i(\cdot) \) is the indicator function that assigns \( \infty \) to estimates that do not reside in the constraint set (the loss is identical if the index set consists of \( y_m = 0 \) precisely) and \( \tilde{f}(z) \) takes the following form:

\[
\tilde{f}(z) = \sum_{m \notin I(y = 0)} \left( |a_m^* z|^2 \log \frac{|a_m^* z|^2}{y_m} - |a_m^* z|^2 \right).
\]

This constraint on the search space provides further robustness against outliers present in the data, as their impact within the update step is cascaded by the projection operation which can mitigate their influence on the iterative estimates.

In the next two sections, we introduce a novel initialization and three algorithms based on RKLD.

III. INITIALIZATION BASED ON RKLD

Accuracy of the initial estimates provided by spectral methods is known to facilitate the exact recovery guarantees for non-convex iterative approaches [9], [25], [34], [35], under the Gaussian model and coded diffraction patterns. The classical spectral method involves setting the leading eigenvector of the following empirical scatter matrix:

\[
Y = \frac{1}{M} \sum_{m=1}^{M} y_m a_m a_m^*
\]

which is coupled with a norm estimate of the signal to initialize the algorithm.

In our work, we obtain an initial estimate of the unknown signal using the spectral estimation framework that we have recently introduced in [10]. In [10], we consider the classical formulation of the spectral method as an approximate minimization of the \( \ell_2 \)-loss function, which is pursued with correlation in the range of the lifted forward model. The accuracy of the spectral methods are interpreted under a certain restricted isometry-type property related to the sufficient conditions of the WF-based exact phase retrieval algorithm [35]. This observation is then generalized to Bregman divergences, yielding the sample processing function \( h \) that can be obtained by solving the following optimization problems:

\[
(P1) : \min_{q \in X \subset S} d(q, p) = -\langle q, h(p, q) \rangle
\]

\[
(P2) : \max_{\|v\| = 1} \langle h(p, q), a_m a_m^* \rangle v,
\]

Here, \( P1 \) represents the approximate minimization of the Bregman divergence over \( q = \|a_m, v\|^2 \), where \( v \) is the spectral search variable. This approximation is facilitated by a sample processing function, \( h \), which is designed to introduce minimal distortion to the original objective using the Bregman representation property [44], denoted by \( \hat{q} \). Hence, \( P2 \) represents an alternative to (6), such that the Bregman divergence minimization is approximated by a tractable spectral method. The two problems are complementary in the sense that the maximization in \( P2 \) sufficiently approximates the minimization of the loss in \( P1 \) through the derivation of \( h \).

Constructing the spectral estimate for the KL-divergence loss we consider:

\[
D_{KL}(q, p) = \sum_{m=1}^{M} q_m \log \frac{q_m}{p_m} - q_m + p_m
\]

constrained on the M-simplex. Then, setting \( p = y/\|y\|_1 \), the RKLD minimizing spectral method is formulated under minimum distortion as follows:

\[
D_{KL}(q, p) \approx -\sum_{m=1}^{M} q_m \log \left( \frac{y_m}{\|y\|_1} \sum_{i=1}^{M} \|a_i\|^2 \right) - \sum_{m=1}^{M} q_m |a_m|^2
\]

\[
z_0 := \arg\max_{\|v\| = 1} \langle \sum_{m=1}^{M} h_m a_m a_m^* \rangle v. \]

Algorithm 1 summarizes the pseudo-code to obtain the RKLD based spectral initialization.

Algorithm 1 RKLD Initialization

**Input:** The measurements, \( \{y_m\}_{m=1}^{M} \), the sampling vectors, \( \{a_m\}_{m=1}^{M} \)

**Step 1.** Compute the sample processing function

\[
h_m = \log \left( \frac{y_m/\|y\|_1}{\|a_m\|^2/\sum_{i=1}^{M} \|a_i\|^2} \right)
\]

**Step 2.** \( z_0 := \arg\max_{\|v\| = 1} \langle \sum_{m=1}^{M} h_m a_m a_m^* \rangle v. \)

**Output:** The leading eigenvector, \( z_0 \)

IV. PHASE RETRIEVAL BASED ON RKLD & THE WF FRAMEWORK

In this section, we introduce and describe the corresponding WF algorithms.

A. RKLD-WF Algorithm

We formulate the following loss function, \( f(z) \), for the RKLD-WF algorithm:

\[
f(z) := \sum_{m=1}^{M} |a_m^* z|^2 \log \frac{|a_m^* z|^2}{y_m} - \sum_{m=1}^{M} |a_m^* z|^2.
\]

Since \( f(z) \) is a real-valued function of a complex variable, it is non-holomorphic and not complex differentiable [45]. Hence, we compute the Wirtinger derivative of \( f(z) \) [9],

\[
\nabla f(z) = \left( \frac{\partial f}{\partial \bar{z}} \right)^*,
\]

which results in,

\[
\nabla f(z) = \sum_{m=1}^{M} \left[ \log \left( |a_m^* z|^2 \right) - \log (y_m) \right] a_m^* z a_m.
\]
can be compactly expressed in terms of matrix-vector products as follows:
\[
\nabla f(z) = A^*[Az \odot (\log(|Az|^2) - \log(y))],
\]
where, \(A^*\) denotes the conjugate transpose of \(A\), and \(\odot\) denotes the element-wise multiplication of two vectors.

Due to the logarithmic processing of the quantities in (17), the Wirtinger gradient is not defined when one or more components of \(Az\) are zero (or close to zero). This motivates the use of a regularized loss function. Therefore, instead of (14), we introduce the following alternative loss function:
\[
f_\lambda(z) := \sum_{m=1}^{M} |a^*_m z|^2 + \lambda \log \frac{|a^*_m z|^2}{y_m + \lambda} - \sum_{m=1}^{M} |a^*_m z|^2 + \lambda, \quad \text{(18)}
\]
where \(0 < \lambda \ll 1\). The corresponding Wirtinger gradient expression is given as follows:
\[
\nabla f_\lambda(z) = A^*[Az \odot (\log(|Az|^2 + \lambda 1) - \log(y + \lambda 1))], \quad \text{(19)}
\]
where \(1\) is an all-ones vector of length \(M\).

For a fixed step-size, \(\mu\), we write the update equation for the steepest descent iterations as follows:
\[
z_{k+1} = z_k - \mu \nabla f_\lambda(z_k). \quad \text{(20)}
\]
Algorithm 2 summarizes the pseudo-code for the regularized RKLD-WF algorithm.

Algorithm 2 Regularized RKLD-WF
\[
\text{Input: } \{y\}_{m=1}^{M}, \{a_m\}_{m=1}^{M}, \text{ Step size, } \mu, \text{ regularization parameter, } 0 < \lambda \ll 1.
\]
\[\text{Step 0. Initialize } z_0 \text{ following Algorithm 1}.
\]
\[\text{for } k = 0 \text{ to } (K - 1) \text{ do}
\]
\[\text{Step 1. Compute } \nabla f_\lambda(z^{(t)}) \text{ as in (19)}.
\]
\[\text{Step 2. } z_{k+1} = z_k - \mu \nabla f_\lambda(z_k)
\]
\[\text{end for}
\]
\[\text{Output: } z_K
\]

B. Residual Truncation with RKLD
To further improve the performance of the RKLD-WF algorithm, we propose to control the aggregate gradient to enable the iterative procedure to operate within the local convex region around the true solution of the objective function. For this purpose, we regularize the weights of the gradient components leading to the truncation of the sampling vectors which do not satisfy a predetermined statistical criteria. Thus, instead of using all \(M\) samples, we use an adaptive collection of suitable \(M'_k \leq M\) sampling vectors to compute the total loss at each iteration. This motivates the development of the RKLD-based algorithms, which utilizes the following truncated loss function:
\[
f^t(z) := \sum_{m=1}^{M} \{ |a^*_m z|^2 \log \frac{|a^*_m z|^2}{y_m} - |a^*_m z|^2 \} I_{T^m}, \quad \text{(21)}
\]
where \(I\) denotes the \(1/0\) indicator function for the index set \(T^m \subset \{1, 2, \cdots, M\}\) determined by the truncation scheme.

For the choice of the index set, \(T^m\), several alternatives that induce robustness exist [7], [25], [27]. The most straight forward approaches are based on the absolute residuals, i.e., the model mismatch at a given estimate \(z\), \(|y_m - |a^*_m z|^2|\) [25], [27]. The truncation set \(T^m\) can be constructed with respect to the mean or the median of the absolute residuals. This data trimming mechanism is employed in [25], and [27], respectively, for the Poisson-loss where the ensemble loss is modified by carefully choosing from the sampling vectors so that the selected sampling vectors satisfy certain statistical criteria conditioned upon the residuals. In the TWF algorithm [25], the truncation scheme is based on the following index set, \(T^m\), that is adaptively modified at each iteration:
\[
T^m := \{ m \mid |y_m - |a^*_m z|^2| \leq \gamma^c K_t \}, \quad \text{(22)}
\]
\[
K_t := \text{median} \{ |y_m - |a^*_m z|^2| \}_{m=1}^{M}, \quad \text{(23)}
\]
where \(\gamma^c\) is a predetermined threshold. In essence, the mean serves as a statistic for the residual level of the estimate \(z\), obtained from the \(\ell_2\) minimizer on the residuals.

Using the sample mean of absolute residuals as a truncation criteria can be considered analogous to conducting an \(\ell_2\)-norm estimate for the residual level. This is equivalent to fitting a Gaussian distribution to \(\{|a^*_m z|^2 - y_m\}_{m=1}^{M}\), such that \(K_t\) becomes the mean absolute deviation. However, this is not an accurate characterization, even if the two components were to be approximated as independent. Indeed, assuming Gaussianity on the sampling vectors, if the two components were independent, the residuals follow a Laplace distribution. This promotes the median of the residuals as a more appropriate statistic for choosing the residual level,
\[
K_t := \text{median} \{ |y_m - |a^*_m z|^2| \}_{m=1}^{M}, \quad \text{(24)}
\]
which results in the scheme used in [27] to determine the index set \(T^m\) at each iteration. We summarize the pseudo-code for the RKLD-based median truncated WF (RKLD-MTWF) algorithm in Algorithm 3.

The median is notably the estimator using the \(\ell_1\) fit on the residuals. This naturally provides further robustness against outliers beyond that of the sample mean. In the presence of statistical outliers, the advantages of the median truncation can be equivalently understood from its derivation from the \(\ell_1\) norm of the residuals. The \(\ell_1\)-norm penalizes large residuals less aggressively than the quadratic loss under which the mean statistic is derived.

C. Gradient Truncation with RKLD
Residual-based truncation schemes are natural for the general class of bowl-shaped functions, which are conventional in inference and defined directly as norms of residuals. The residual term then immediately relates to the influence of a sampling vectors in updating the estimates, whereas the truncation criteria on residuals is directly related to applying pruning on the weights that appear in the synthesis of the gradient at each iteration.
Algorithm 3 RKLD-MTWF

**Input:** \( \{y_m\}_{m=1}^M, \{a_m\}_{m=1}^M \), step size, \( \mu_{mtr} \), truncation parameters, \( \gamma_{ub} \), and \( \gamma_{tr} \).

**Step 0.** Initialize \( z_0 \) following Algorithm \( \text{[1]} \)

for \( k = 0 \) to \( (K - 1) \) do

**Step 1.** Compute
\[
K_i := \text{median}(\{||y_m - |a_m^*z_k|^2||\}_{m=1}^M)\]

**Step 2.** Compute \( I_k = [I_{T_1}, I_{T_2}, \ldots, I_{T_M}]^T \) where,
\[
T_1^m := \left\{ m : |a_m^*z_k| \leq \gamma_{ub} \right\}.
\]
\[
T_2^m := \left\{ m : |y_m - |a_m^*z_k|^2| \leq \gamma_{tr}K_i |a_m^*z_k| / ||z_k||_2 \right\}.
\]

**Step 3.** Compute, \( A_k^i = A \odot I_k \) and \( y_k^i = y \odot I_k \)

**Step 4.** Compute,
\[
\nabla f^i(z_k) = A_k^i \odot [\log(|A_k^i z_k|^2) - \log(y_k^i)]
\]

**Step 5.** \( z_{k+1} = z_k - \mu_{mtr} \nabla f^i(z_k) \)
end for

**Output:** \( z_K \)

Equivalently, the objective of truncation is to control the impact of certain directions in updating the estimates, where certain directions are potentially biased via the outliers that contaminate the measurements. With RKLD, the residual is not the argument of the loss function, hence the residual-based truncation does not equivalently translate to controlling weights of sampling vectors within the gradient update. To address this fundamental mismatch to the residual-based truncation schemes, we take inspiration from the Poisson-loss implementation of TWF \( \text{[25]} \), which directly controls the impact of sampling vectors within generated estimates. Unlike the gradient of Poisson-loss, the measurement mismatch in RKLD is not over the residual term, instead it appears under a non-linear logarithmic processing in the gradient expression. This corresponds to truncation schemes using a *residual of logarithms*, i.e., \( \log(|y_m|) - \log(|a_m^*z_k|^2) \), to control the weights that are featured within the gradient computation. Thus, for the gradient truncation, we develop the following one-sided truncation scheme:

\[
T_3^m := \left\{ m : \log(|y_m|) - \log(|a_m^*z_k|^2) \leq \gamma_{tr}D \right\}, \tag{25}
\]

where,
\[
D := \text{median}(\{\log(y_m) - \log(|a_m^*z_k|^2)\}_{m=1}^M), \tag{26}
\]

with \( \gamma_{tr} \) denoting a predetermined threshold parameter and \( D \) denoting the median of the *residual of logarithms*.

The truncation scheme in \( \text{(25)} \) is inspired by its ability to conserve the sampling vectors demonstrating the orthogonality promoting properties of RKLD. Assume \( y_m = 0 \), if the model prediction \( |a_m^*z|^2 \neq 0 \), the ideal action is to enforce such orthogonality by the update using the gradient, which would promote an orthogonal projection as discussed in Section 2. However, by using a two-sided truncation scheme on the residual of logarithm term, one effectively removes orthogonality information from the synthesis, as \( |\log(|a_m^*z|^2) - \log(y_m)| \to \infty \) when \( y_m \to 0 \), and \( |a_m^*z| \neq 0 \). As a result, a large residual of logarithms does not necessarily indicate an outlier in the data; on the contrary, it potentially indicates key information on the unknown of interest which we wish to leverage by the use of RKLD.

Notably, the absolute residual corresponds to:
1) \( \log(q_m - \log y_m) \text{ if } q_m > y_m \), or
2) \( \log y_m - \log q_m \text{ if } y_m > q_m \).

Clearly, the natural outliers that result from \( y_m = 0 \) are included in the first case \( (q_m \geq 0 \text{ by definition}) \). This is a case that we aim to preserve despite implementing a truncation scheme. On the other hand, when data is contaminated with outliers, the second case is the instance we expect to observe detrimental influence in generating estimates. We can therefore apply a *single-sided* scheme, where only the residual of logarithms in the form of case (2) are truncated. This truncation mechanism leads to the following gradient expression:

\[
\nabla f^i(z) = (A^i)^* [A^i z \odot (\log(|A^i z|^2) - \log(y^i))], \tag{27}
\]

where \( A^i \) and \( y^i \) are the selected measurement matrix and the measurements, respectively, containing components corresponding to the indices in \( T_3^m \) which gets updated at each iteration.

The pseudo-code for the RKLD-based gradient truncated WF (RKLD-GTWF) algorithm is summarized in Algorithm \( \text{[4]} \).

Algorithm 4 RKLD-GTWF with one-sided truncation

**Input:** \( \{y_m\}_{m=1}^M, \{a_m\}_{m=1}^M \), step size, \( \mu_{gtr} \), truncation parameter, \( \gamma_{h} \).

**Step 0.** Initialize, \( z_0 \) following Algorithm \( \text{[1]} \)

for \( k = 0 \) to \( (K - 1) \) do

**Step 1.** Compute
\[
D := \text{median}(\{\log(y_m) - \log(|a_m^*z_k|^2)\}_{m=1}^M),
\]

**Step 2.** Compute \( I_k = [I_{T_1}, \ldots, I_{T_M}]^T \) where,
\[
T_3^m := \left\{ m : \log(|y_m|) - \log(|a_m^*z_k|^2) \leq \gamma_{h}D \right\},
\]

**Step 3.** Compute \( A_k^i = A \odot I_k \) and \( y_k^i = y \odot I_k \)

**Step 4.** Compute,
\[
\nabla f^i(z_k) = A_k^i \odot [\log(|A_k^i z_k|^2) - \log(y_k^i)]
\]

**Step 5.** \( z_{k+1} = z_k - \mu_{gtr} \nabla f^i(z_k) \)
end for

**Output:** \( z_K \)

D. Computational Complexity

We now discuss the computational complexity of the RKLD-WF, RKLD-MTWF and RKLD-GTWF algorithms.
Gradient computation is the most expensive step in each algorithm, which performs two matrix multiplications, i.e., \( A^*v \), and \( v = Az \). Hence, the per-iteration computational complexity of each algorithm is \( O(MN) \), which is equivalent to the per-iteration computational complexity of the state-of-the-art methods \([7, 25, 27]\).

The initialization procedure entails the computation of the leading eigenvector, which has \( O(N^3) \) complexity, identical to the computational complexity of the classical spectral method.

V. NUMERICAL EXPERIMENTS

We conducted numerical simulations to evaluate the performance of our RKLD-WF, RKLD-MTWF, and RKLD-GTWF algorithms for synthetic data in the statistical setting, as well as for real optical imaging data. We compare the efficiency of these algorithms with their \( \ell_2 \) and Poisson-loss based counterparts.

A. Performance Evaluation

Our experiments utilize Monte-Carlo (MC) simulations for performance evaluation. We generate \( S \) many independent and uncorrelated instances of the measurement matrix, \( A \), for the MC simulations. Thus, all algorithms run \( S \) times independently for a fixed unknown \( x \). We use the average relative error (ARE) as a figure of merit to evaluate the quality of reconstructed signal using synthetic data. We compute an empirical average of the relative reconstruction error over the MC simulations which is defined as follows:

\[
\text{ARE} := \frac{1}{S} \sum_{s=1}^{S} \frac{\|\hat{x}_s - x\|_2}{\|x\|_2},
\]

where \( \hat{x}_s \) denotes the recovered signal at the \( s \)-th MC simulation after \( K \) iterations. Here, \( \|\hat{x}_s - x\|_2 \) is the Euclidean distance between two vectors \( x \) and \( \hat{x}_s \), which is defined as follows:

\[
\|\hat{x}_s - x\|_2 := \min_{\phi \in [0, 2\pi]} \|xe^{i\phi} - \hat{x}_s\|_2.
\]

To demonstrate the sample efficiency of our algorithms, we use the empirical probability of successful recovery as the evaluation metric. We vary the oversampling factor, \( \alpha \), and run MC simulations for each instance of \( \alpha \). We consider the outcome of the \( s \)-th MC simulation to be successful if it satisfies \( \|\hat{x}_s - x\|_2 < 10^{-5} \). The empirical probability of success is defined as follows:

\[
\text{Empirical probability} := \frac{\text{No. of successful simulations}}{\text{Total no. of simulations}}.
\]

We investigate the performance of algorithms at different levels of signal-to-noise ratio (SNR) of the measured data. We define the SNR as follows:

\[
\text{SNR} := 20 \log_{10} \left( \frac{\text{var}_y}{\text{var}_w} \right),
\]

where \( \text{var}_y \) and \( \text{var}_w \) are the variances in the data and the additive noise, respectively.

For optical imaging data, we evaluate the quality of the reconstructed images via the average correlation coefficient (ACC) over MC simulations, i.e.,

\[
\text{ACC}_i(x, \hat{x}) := \frac{1}{S} \sum_{s=1}^{S} \frac{\|x_i - \hat{x}_i\|_2}{\|x_i\|_2} \quad (32)
\]

where \( \hat{x}_i \) is the unit norm reconstruction of \( x_i \) at the \( s \)-th MC simulation and for the \( i \)-th image, where \( i \in \{1, 2, 3, 4, 5\} \) denotes the index of the 5 images in the optical imaging dataset \([11]\). We also evaluate the empirical sampling efficiency for the real data. As a figure of merit, we use ACC\( \alpha \), which is the average of ACC\( i \)'s over the 5 images. We define the average correlation coefficient, with respect to \( \alpha \), as following:

\[
\text{ACC}_\alpha(x, \hat{x}) := \frac{1}{5} \sum_{i=1}^{5} \text{ACC}_i \quad (33)
\]

B. Simulation Set-up and Corresponding Parameters

1) Signal and Measurement Model: Our simulation set-up generates a random ground truth signal, \( x \in \mathbb{C}^N \) which has elements, \( x = r + jxi \), where \( r, x_i \sim \mathcal{N}(0,1) \). We also use a real signal, \( x \in \mathbb{R}^N \) due to its physical significance and prevalence in various imaging modalities. For the real case, we set \( x \sim \mathcal{N}(0,1) \). For the measurement model, we generate the complex sampling vectors \( a_m \in \mathbb{C}^N \) under the standard Gaussian distribution, i.e., \( a_m \sim \mathcal{C}\mathcal{N}(0, I) \) for \( m = 1, \cdots , M \). The elements of \( a_m \)'s are sampled from the Gaussian distribution independently and identically.

Besides the Gaussian sampling model, our simulation setup also includes the coded diffraction model \([16]\), which is inspired from the physically realistic applications that emerge in X-ray imaging and other related disciplines. The coded diffraction patterns (CDP) are a collection of intensity data, \( y \), which are modeled as the squared magnitude of the Fourier transform of the modulated unknown at the detectors. Mathematically,

\[
y_m = \left| \sum_{n=0}^{N-1} x[n]d_l[n]e^{-j2\pi kn/N} \right|^2, \quad (34)
\]

where \( d_l[n] \in \mathbb{C}_{n=1,\cdots ,N}; m = (l, k) \) for \( 0 \leq k \leq N - 1 \), and \( 1 \leq l \leq L \). Here, \( L \) denotes the number of modulating patterns. Thus, for \( L \) diffraction patterns, we get the measurements, \( M = LN \). Here, \( d_l \)'s are independently and identically drawn from a distribution, known as the octanary pattern \([16]\).

2) Noise and Outlier Model: We assume the noise is uniformly distributed, i.e., each entry of \( w \) is independently and identically drawn from the uniform distribution \( \mathcal{U}(0, \omega_{\text{max}}) \), where \( \omega_{\text{max}} \) is the largest element of \( w \) that satisfies \( \omega_{\text{max}} \leq \sigma \|x\|_2 \). Here, \( \sigma > 0 \) is an arbitrary constant. In addition, we generate a sparse outlier vector which is assumed to be uniformly distributed, i.e., \( \mathcal{U}(0, \eta_{\text{max}}) \). Here, \( \eta_{\text{max}} \) denotes the largest magnitude of the outlier vector so that it satisfies \( \eta_{\text{max}} \leq \theta \|x\|_2 \) where \( \theta > 0 \) is an arbitrary constant. We vary the values of \( \sigma \) and \( \theta \) to generate different levels of additive noise and outliers, respectively.
C. Performance of the RKLD-WF algorithm

We compare the performance of the RKLD-WF algorithm with that of the standard WF algorithm \cite{9} using the $\ell_2$-loss and Poisson-loss functions, and the RWF algorithm \cite{26} using the reshaped $\ell_2$-loss function. For this purpose, we employed the codes provided with the original papers, keeping their simulation parameters unchanged.

First, we assess the quality of reconstruction by the RKLD-WF algorithm. We fixed the signal-length $N = 500$ and $\alpha = 6$ for the Gaussian sampling. For the CDP case, we set $N = 512$ and the number of modulating patterns $L = 8$. The step size, $\mu$, was chosen to be 0.6 for the Gaussian case and 0.4 for the CDP case. The regularization parameter was set to be $\lambda = 10^{-8}$. We fixed the number of iterations $K = 500$ and run $S = 100$ MC simulations. We then computed the average reconstruction error using (28).

Figures 2(a) and (b) present the results of the RKLD-WF, WF-$\ell_2$ \cite{9}, WF-Poisson and RWF \cite{26} algorithms for both Gaussian and CDP models. Figures 2(a) and 2(b) show ARE vs. iteration count using noiseless data. (c)-(d) depict the ARE vs. iteration count using noisy data. (c) shows the results for the Gaussian model for $N = 500$. (d) shows the results for the CDP model for $N = 512$. (e) shows the empirical probability of success w.r.t. oversampling factors ($\alpha$) with $N = 100$ for the Gaussian model. (f) shows the empirical probability of success w.r.t. the number of modulating pattern ($L$) with $N = 256$ for the CDP model.

1https://viterbi-web.usc.edu/soltanol/WFcode.html, https://github.com/hubevan/reshaped-Wirtinger-flow
Fig. 3. Comparison among the RKLD-MTWF (Algorithm 3), RKLD-GTWF (Algorithm 4), median-TWF and median-RWF [26] algorithms in terms of the empirical probability of success vs. the oversampling factor, in the presence of only additive outliers. (a) We used a complex signal and complex Gaussian model with signal length $N = 100$. (b) We used a complex signal and the CDP model with signal length $N = 128$. For both plots, we use $\theta = 5$ which implies $\eta_{\max} \leq 5 \parallel x \parallel^2$.

Fig. 4. Comparison of performance among the RKLD-MTWF (Algorithm 3), RKLD-GTWF (Algorithm 4), median-TWF and median-RWF [26] algorithms in the presence of both additive outliers and noise. (a), (b), and (c) demonstrate the performance using Gaussian measurements with $N = 100$ and the oversampling factor, $\alpha = 8$. (a) shows the ARE vs. the iteration index. (b) shows the ARE vs. the outlier magnitude parameter, $\theta$ with a fixed fraction of outliers in data $\rho = 0.1$. (c) shows the empirical probability of success only vs. the fraction of outliers in the data, $s$ and no noise and a fixed $\theta = 5$. Similarly, (d), (e), and (f) demonstrate the performance using Gaussian measurements with $N = 128$ and the number of modulating patterns, $L = 8$.

of 1. We run 100 MC simulations with $K = 1000$ for each choice of $\alpha$ and $L$. Figures 2(e) and (f) shows the empirical probability of success vs. the oversampling factors, for the Gaussian and CDP models, respectively. For the Gaussian case, it can be clearly seen that the required number of samples for our RKLD-WF method is significantly lower than those of the other algorithms. This implies a higher empirical sample efficiency of the RKLD-WF algorithm. For the CDP case also, RKLD-WF has a sample efficiency notably higher than those of WF-$\ell_2$ and WF-Poisson.

D. Performance of the RKLD-MTWF and RKLD-GTWF Algorithms

We perform numerical experiments to evaluate the performance of the RKLD-MTWF and the RKLD-GTWF algorithms, in the presence of additive noise and outliers. We compare the outcomes of these algorithms with those of the median-TWF\footnote{https://github.com/hubevan/Median-TWF} and median-RWF algorithms in [26].

In Figure 3, we present the empirical probability of success vs. the oversampling factor $\alpha$ to evaluate the sample efficiency.
Here, we set $N = 100$ for the Gaussian model and $N = 128$ for the CDP case. We fixed the outlier magnitude parameter, $\theta = 5$ so that $\eta_{\text{max}} \leq \theta \|x\|_2^2$ and the fraction of outliers in data $\rho = 0.1$. Figure 3(a) depicts the performance of the Gaussian model. As we can see, our proposed RKLD-MTWF algorithm requires lower number of samples for 100% recovery than the other algorithms and thus, demonstrates superior sample efficiency over the other algorithms. The sample efficiency of RKLD-GTWF is close to those of median-TWF and median-RWF. Figure 3(b) shows the empirical probability of success vs. the number of modulating patterns $L$ for the CDP model. We see that RKLD-MTWF and RKLD-GTWF both show superior sample efficiency than median-TWF and median-RWF, in the presence of outliers.

Next we assess the performance of our algorithms in the presence of both additive noise and outliers. Figure 4 demonstrates the quality of reconstruction as we vary different parameters of the noise and outliers. For this experiment, we fix the signal length $N = 100$ and $\alpha = 8$ for the complex Gaussian model and $N = 128$ and $L = 8$ for the CDP model. Figures 4(a), 4(b), and 4(c) present the results for the Gaussian model and Figures 4(d), 4(e), and 4(f) shows the results for the CDP model. In Figures 4(a) and 4(d), we plot the ARE vs. the iteration index for $\theta = 10$, $\rho = 0.1$ and the noise magnitude parameter, $\sigma = 0.01$. We clearly see that for both Gaussian and CDP models, our proposed algorithms, RKLD-MTWF and RKLD-GTWF achieve a significantly lower ARE, on the order of $10^{-3}$, than those of the median-TWF and median-RWF algorithms [26]. We also note that RKLD-GTWF obtains slightly lower error level than RKLD-MTWF for the Gaussian model but significantly lower error level in case of the CDP model.

In Figures 4(b) and 4(e), we plot the ARE vs. the outlier magnitude parameter, $\theta$ for a fixed number of outliers in data indicated by $\rho = 0.1$. Here, the noise magnitude parameter is $\sigma = 0.01$. We see that our proposed algorithms can deal with high magnitude outliers more robustly than other median-truncated algorithms. Figures 4(c) and 4(f) present the empirical probability of success vs. the fraction of outliers $\rho$ to show the reconstruction quality of our algorithms against varying number of outliers. Here, we fix the outlier magnitude parameter $\theta = 5$ so that $\eta_{\text{max}} \leq 5\|x\|_2^2$. Figure 4(c) presents the result for the Gaussian model. We clearly see that RKLD-GTWF and RKLD-MTWF can tolerate a considerably higher number of outliers in the data than the median-TWF and median-RWF algorithms. Similarly, in Figure 4(f), for CDP case, our algorithms perform better in the presence of an increased number of outliers.

Figure 5 shows ARE vs. SNR (dB) for the Gaussian model. For this experiment, we chose $x \in \mathbb{R}^N$ and $A \in \mathbb{C}^{M \times N}$, where $N = 100$ and $M = 8N$. The elements of the noise vector were sampled from a uniform distribution, $U(0, \sigma_{\text{max}})$ where $w_{\text{max}} \leq \sigma \|x\|_2^2$. We generated different levels of SNR by varying the noise magnitude parameter $\sigma$. We ran 50 MC simulations to generate the ARE. When the SNR level increases, the ARE decreases for all the algorithms. The RKLD-MTWF, RKLD-GTWF and median-TWF algorithms have the same ARE for all SNR levels. However, the ARE for the median-RWF algorithm is considerably higher than our algorithms. Next, besides the noise, we added a sparse outlier vector to the data and run the experiment for the same SNR levels. The elements of the outliers were generated from $U(0, \eta_{\text{max}})$ where $\eta_{\text{max}} \leq 10\|x\|_2^2$ and the fraction of outliers in data $\rho = 0.1$. We present the results in Figure 6. We observe that the ARE of the median-TWF algorithm diverges very quickly whereas the AREs for the other algorithms decrease consistently with the increasing SNR. We note that, similar to the previous case, the RKLD-GTWF and RKLD-MTWF algorithms achieve a consistent lower error than the median-RWF algorithm. It can be clearly seen that the RKLD-based algorithms demonstrates superior performance than the other algorithms.

### E. Real Optical Imaging Dataset

We evaluate the performance of our RKLD-based algorithms in the optical imaging setup, using the real measurements provided in [11]. The setup employs a spatial light
modulator (SLM) that modulates a coherent light using some imaging pattern, \(x \in \mathbb{R}^N\), i.e., the ground truth image. The modulated light then passes through a diffruser and lands on a sensor that measures only the intensity of the form, \(y = |Ax + \epsilon|^2\), where \(\epsilon\) denotes noise. In this setup, the transmission matrix, \(A\), characterizes the linear mapping from the image pattern to the measurements as light passes through the scattering medium. First, a system of equations is formed to estimate \(A\) by sending a sufficient number of calibration patterns through the medium. Then, a phase retrieval technique is employed to learn \(A\) from the equations.

\[
\begin{align*}
\|Ax\|^2 &= \|\epsilon\|^2 + 2Re(A^*\epsilon) \\
&= 2Re(A^*\epsilon) \\
&= 2Re(A^T\epsilon) \\
&= 2Re(\epsilon^T A) \\
&= 2Re(y) \\
&= 2Re(|Ax|^2) \\
&= 2|Ax|^2 \\
&= 2|Ax + \epsilon|^2
\end{align*}
\]

We consider the \(16 \times 16\) reconstruction task for our simulation, where measurements have an estimated SNR of 22dB. The full dataset contains \(M = 256^2\) measurements for each image, where the signal dimension is \(N = 256\) and the oversampling factor is \(\alpha = 256\). For our experiment, we used randomly sampled subsets of measurements from the full dataset at varying oversampling factors.

First, we performed reconstruction using RKLD-WF (Alg. [2], WF-\(\ell_2\) [9] using \(\ell_2\)-loss function, TWF [25] using the Poisson-loss function and RWF [20] using reshaped \(\ell_2\)-loss function, for all 5 test images. We used the PhasePack Library [46] to implement the WF-\(\ell_2\), TWF and RWF algorithms. We ran 50 MC simulations for each image. The average absolute value images obtained from the MC experiments at \(\alpha = 5\) are shown in Figure 7 without outliers, and in Figure 8 with additive sparse outliers with \(\theta = 1\). For all methods iterations are stopped at the equal number of iterations required by the RKLD-WF method. The corresponding ACC vs. image index are provided in Figures 9(a) and 9(b), respectively. Most strikingly, in the presence of outliers, our RKLD-WF method, in this experiment, does not employ a truncation scheme whereas the Poisson-loss incorporates a truncation mechanism (the TWF algorithm [25]) to provide robustness, with standard WF serving as a control without truncation. As can be observed, although truncation and reshaping provide improvements over the WF-\(\ell_2\) in the presence of outliers, RKLD without any truncation is capable of superior reconstruction performance at identical total computational complexity. This highlights the natural robustness of the RKLD minimization approach, which supports our motivations and main arguments.

Figure 10 shows the ACC vs. the oversampling factor \(\alpha\), using (33), for all four algorithms. We perform reconstructions by randomly sampling measurements for different choices of \(\alpha \in \{5, 10, 20, 25, 40\}\). It can be seen that the RKLD-WF algorithm has a higher correlation coefficient than those of the other three algorithms for low oversampling factors. This highlights the fact that RKLD-WF provides recovery that is more effective in case of data scarcity. Furthermore, we provide Table 1 to demonstrate the iteration count for these algorithms. The number of iterations was averaged over the images and recorded for each value of \(\alpha\). All four algorithms ran for a maximum of 2000 iterations. A tolerance value of \(10^{-10}\) is set as the early stopping criteria. When the norm of the difference between two consecutive updates falls below the tolerance value, the iterative process was terminated. We see that the RKLD-WF converges much faster than the other three algorithms. This result also agrees with our arguments and illustrates the advantage of the RKLD minimization approach.

**VI. Conclusion**

In this paper, we approach phase retrieval from a novel perspective and present an RKLD minimization method to attain robustness to outliers beyond a level offered by state-of-the-art. The robust inference is crucial for the practical applicability
of the phase retrieval algorithms to the physical sensing and imaging problems, as measurements are commonly collected under model imperfections, dynamic propagation media, and challenging acquisition conditions which can considerably degrade the quality of estimates. The RKLD provides a robust measure of dissimilarity for processing intensity-only data and promotes a model agnostic approach in formulating the optimization problem. We assessed the feasibility of the RKLD-based methods empirically using synthetic statistical sampling data and real optical imaging data, both of which demonstrate improvements over the state-of-the-art methods in phase retrieval literature.

Our phase retrieval algorithm exploits strict non-negativity of the measurements and performs well in sparse contamination. Therefore, we expect that the RKLD minimization approach will be useful in developing robust algorithms for specific wave-based imaging problems. Our future work will also focus on the theoretical analysis of the RKLD minimization approach to establish recovery guarantees. Another potential direction will be the design and study of the stochastic version of RKLD minimization since stochastic methods are known to achieve faster convergence than the standard approaches [7], [47].

Fig. 9. (a) Average correlation coefficient over the 50 Monte-Carlo instances with no added outlier, at \( \alpha = 5 \), for each of the 5 test images. (b) Average correlation coefficient over the 50 Monte-Carlo instances with outlier with \( \theta = 1 \), at \( \alpha = 5 \), for each of the 5-test images.

Fig. 10. Average correlation coefficient (\( ACC_\alpha \)) vs. the oversampling factor over the 5 images in the optical imaging dataset [11]. We show the results for the RKLD-WF, WF-\( \ell_2 \) TWF, and RWF algorithms using 50 MC simulations for each choice \( \alpha \).

**TABLE I**

The average iteration count over 50 MC simulations required by the RKLD-WF, WF-\( \ell_2 \), TWF, and RWF algorithms for oversampling factors \( \alpha \in \{2, 5, 10, 20, 40\} \).

| oversampling Factor, \( \alpha \) | 2   | 5   | 10  | 20  | 40  |
|---------------------------------|-----|-----|-----|-----|-----|
| RKLD-WF                         | 222 | 23.2| 15.6| 13  | 7.6 |
| WF-\( \ell_2 \)                 | 2000| 616.6| 211.6| 120.8| 103.2|
| TWF                            | 1317.2| 1395.4| 1651| 1706.2| 702.4|
| RWF                            | 2000| 2000| 1855.4| 1456.6| 1083.2|

References

[1] J. R. Fienup, “Phase retrieval algorithms: a personal tour,” *Applied optics*, vol. 52, no. 1, pp. 45–56, 2013.
[2] M. V. Klibanov, P. E. Sacks, and A. V. Tikhonravov, “The phase retrieval problem,” *Inverse problems*, vol. 11, no. 1, p. 1, 1995.
[3] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345–356, 2006.
[4] Y. Shechtman, A. Beck, and Y. C. Eldar, “Gespar: Efficient phase retrieval of sparse signals,” *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 926–938, Feb 2014.
[5] F. Fogel, I. Waldspurger, and A. d’Aspremont, “Phase retrieval for imaging problems,” *Mathematical programming computation*, vol. 8, no. 3, pp. 311–335, 2016.
[6] 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 processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.
[7] G. Wang, G. B. Giannakis, and Y. C. Eldar, “Solving systems of random quadratic equations via truncated amplitude flow,” *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 773–794, 2017.
[8] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.
[9] E. J. Candès, X. Li, and M. Soltanolkotabi, “Phase retrieval via wirtinger flow: Theory and algorithms,” *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.
[10] B. Yonel and B. Yazıcı, “A spectral estimation framework for phase retrieval via bregman divergence minimization,” *arXiv preprint arXiv:2012.01632*, 2020.
[11] C. A. Metzler, M. K. Sharma, S. Nagesh, R. G. Baraniuk, O. Cossairt, and A. Veeraraghavan, “Coherent inverse scattering via transmission matrices: Efficient phase retrieval algorithms and a public dataset,” in *2017 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2017, pp. 1–16.
[12] R. W. Gerchberg, “A practical algorithm for the determination of phase from image and diffraction plane pictures,” Optik, vol. 35, pp. 237–246, 1972.

[13] J. R. Fienup, “Reconstruction of an object from the modulus of its fourier transform,” Optics letters, vol. 3, no. 1, pp. 27–29, 1978.

[14] ——, “Phase retrieval algorithms: a comparison,” Applied optics, vol. 21, no. 15, pp. 2758–2769, 1982.

[15] E. Candes and T. Strohmer, “Phaselift: Exact and stable recovery from magnitude measurements via pure programming,” Comm Pure Appl Math, vol. 66, pp. 1241–1274, 2013.

[16] E. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval from coded diffraction patterns,” Applied and Computational Harmonic Analysis, vol. 39, no. 2, pp. 277–299, 2015.

[17] I. Waldspurger, A. d’Aspremont, and S. Mallat, “Phase recovery, maxcut and complex semidefinite programming,” Mathematical Programming, vol. 149, no. 1, pp. 47–81, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10107-013-0738-9

[18] E. Candes, Y. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” SIAM Journal on Imaging Sciences, vol. 6, no. 1, pp. 199–225, 2013.

[19] E. J. Candès and X. Li, “Solving quadratic equations via phaselift when there are about as many equations as unknowns,” Foundations of Computational Mathematics, vol. 14, no. 5, pp. 1017–1026, 2014.

[20] L. Demanet and P. Hand, “Stable optimizationless recovery from phaseless linear measurements,” Journal of Fourier Analysis and Applications, vol. 20, no. 1, pp. 199–221, 2014. [Online]. Available: http://dx.doi.org/10.1007/s00041-013-9305-2

[21] P. Hand and V. Voroninski, “An elementary proof of convex phase retrieval in the natural parameter space via the linear program phasemax,” arXiv preprint arXiv:1611.09395, 2016.

[22] ——, “Corruption robust phase retrieval via linear programming,” arXiv preprint arXiv:1612.03547, 2016.

[23] P. Hand, “Phaselift is robust to a constant fraction of arbitrary errors,” Applied and Computational Harmonic Analysis, vol. 42, no. 3, pp. 550–562, 2017.

[24] Y. Chen and E. J. Candes, “Solving random quadratic systems of equations is nearly as easy as solving linear systems,” Communications on Pure and Applied Mathematics, vol. 70, no. 5, pp. 822–883, 2017.

[25] H. Zhang and Y. Liang, “Reshaped wirtinger flow for solving quadratic systems of equations,” stat, vol. 1050, p. 25, 2016.

[26] H. Zhang, Y. Chi, and Y. Liang, “Median-truncated nonconvex approach for phase retrieval with outliers,” IEEE Transactions on Information Theory, vol. 64, no. 11, pp. 7287–7310, 2018.

[27] J. Chen, L. Wang, X. Zhang, and Q. Gu, “Robust wirtinger flow for phase retrieval with arbitrary corruption,” arXiv preprint arXiv:1704.06256, 2017.

[28] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen, “Phase retrieval via reweighted amplitude flow,” IEEE Transactions on Signal Processing, vol. 66, no. 11, pp. 2818–2833, 2018.

[29] R. Kolte and A. Özgür, “Phase retrieval via incremental truncated wirtinger flow,” arXiv preprint arXiv:1606.03196, 2016.

[30] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, “Reshaped wirtinger flow and incremental algorithms for solving quadratic systems of equations,” 2017, preprint.

[31] G. Wang, G. B. Giannakis, and J. Chen, “Solving large-scale systems of random quadratic equations via stochastic truncated amplitude flow,” in 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017, pp. 1420–1424.

[32] Z. Li, K. Lange, and J. A. Fessler, “Poisson phase retrieval with wirtinger flow,” in 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 2828–2832.

[33] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in Advances in Neural Information Processing Systems, 2013, pp. 2796–2804.

[34] B. Yonel and B. Yazici, “A deterministic theory for exact non-convex phase retrieval,” IEEE Transactions on Signal Processing, vol. 68, pp. 4612–4626, 2020.

[35] R. W. Deming, “Phase retrieval from intensity-only data by relative entropy minimization,” JOSA A, vol. 24, no. 11, pp. 3666–3679, 2007.

[36] J. Fienup and C. Wackerman, “Phase-retrieval stagnation problems and solutions,” JOSA A, vol. 3, no. 11, pp. 1897–1907, 1986.

[37] P. J. Huber, Robust statistics. John Wiley & Sons, 2004, vol. 523.

[38] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with bregman divergences,” Journal of machine learning research, vol. 6, no. Oct, pp. 1705–1749, 2005.