Solving Large-scale Systems of Random Quadratic Equations via Stochastic Truncated Amplitude Flow

Gang Wang*,⋆, Georgios B. Giannakis*, and Jie Chen⋆
* Dept. of ECE and Digital Tech. Center, Univ. of Minnesota, Mpls, MN 55455, USA
⋆ State Key Lab of Intelligent Control and Decision of Complex Systems
Beijing Institute of Technology, Beijing 100081, China
E-mails: {gangwang, georgios}@umn.edu, chenjie@bit.edu.cn

***Abstract*—This work develops a new iterative algorithm, which is called *stochastic truncated amplitude flow* (STAF), to recover an unknown signal $\boldsymbol{x} \in \mathbb{R}^n$ from m "phaseless" quadratic equations of the form $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$, $1 \le i \le m$. This problem is also known as phase retrieval, which is *NP-hard* in general. Building on an amplitude-based nonconvex least-squares formulation, STAF proceeds in two stages: s1) Orthogonality-promoting initialization computed using a stochastic variance reduced gradient algorithm; and, s2) Refinements of the initial point through truncated stochastic gradient-type iterations. Both stages handle a single equation per iteration, therefore lending STAF well to Big Data applications. Specifically for independent Gaussian $\{\boldsymbol{a}_i\}_{i=1}^m$ vectors, STAF recovers exactly any $\boldsymbol{x}$ exponentially fast when there are about as many equations as unknowns. Finally, numerical tests demonstrate that STAF improves upon its competing alternatives.**

***Index Terms*—Phase retrieval, stochastic nonconvex optimization, stochastic variance reduced gradient, linear convergence, global optimum.**

I. Introduction

Consider the problem of recovering a high-dimensional signal from the magnitude-only information, e.g., the modulus of the Fourier transform or any linear transform of the signal. This problem, also known as *phase retrieval*, emerges in many areas of science and engineering such as X-ray crystallography, ptychography, and coherent diffraction imaging [1]. In these settings, optical devices in the far field measure only the (squared) modulus of the Fourier transform of the object, yet the phase of the incident light striking the detector is missing. Nonetheless, very much information is contained in the Fourier phase. It is well known that the Fourier phase of an image encodes often more structural information than its Fourier magnitude [1]. Recovering the phase from modulus-only information is of practical relevance.

Mathematically, phase retrieval boils down to tackling a set of quadratic equations taking the following form

$$\psi_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|, \quad 1 \le i \le m \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is the unknown signal, $\boldsymbol{a}_i \in \mathbb{R}^n$ are given sampling vectors, and $\boldsymbol{\psi} := [\psi_1 \ \cdots \ \psi_m]^{\mathcal{T}}$ collects the measured magnitudes. For concreteness, we will hereafter assume random measurements $\{\boldsymbol{\psi_i}\}$ that are collected from the real Gaussian model (1), with independently and identically distributed (i.i.d.) $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. It has been shown that when $m \ge 2n-1$ generic measurements (e.g., from the Gaussian model) are acquired, the system in (1) dictates uniquely a signal $\boldsymbol{x} \in \mathbb{R}^n$ (up to a global sign); and $m = 2n-1$ is also shown necessary [2]. Postulating that there exists a unique solution $\boldsymbol{x}$, the central goal is to put forward a simple yet effective solver amenable to Big Data implementation: i) that recovers $\boldsymbol{x}$ from a nearly minimal number of quadratic equations as in (1); and ii), that enjoys simultaneously nearly optimal per-iteration and computational complexity as well as linear convergence rate.

Past iterative phase retrieval approaches include mainly the alternating projection algorithms [3], [4], alternating minimization with resampling (AltMinPhase) [5], (truncated) Wirtinger flow (WF/TWF) algorithms [6], [7], truncated amplitude flow (TAF) variants [8]–[11], GESPAR [12], and majorization-minimization [13]. For the real Gaussian model, comparisons of different solvers in terms of the sample and computational complexity to acquire an ϵ-accurate solution are listed in Table I.

Building on our precursor TAF [8], this paper advocates a new algorithm termed *stochastic truncated amplitude flow* (STAF), that operates in two stages: Stage one employs a stochastic variance reduced gradient (SVRG) algorithm to compute an orthogonality-promoting initialization, while the second stage applies truncated stochastic gradient-type iterations to refine the initial estimate. STAF is shown capable of recovering any signal from a nearly minimal number of modulus measurements in linear time. In comparison with TAF, the present work's STAF is tailored to Big Data applications. Being order optimal in sample and computational complexity, STAF also features in $\mathcal{O}(n)$ per-iteration complexity, which not only improves on state-of-the-art approaches that can afford $\mathcal{O}(n^2)$, but is also order optimal. This renders STAF well suited for large-scale imaging applications. As will be shown by our simulated tests, STAF improves upon the state-of-the-art TAF and (T)WF in terms of exact recovery performance and convergence speed.

The work of G. Wang and G. B. Giannakis in this paper was supported in part by NSF grants 1500713 and 1514056.

ISBN 978-0-9928626-7-1

TABLE I: Comparisons of Different Algorithms

Algorithm	Sample complexity m	Computational complexity
AltMinPhase [5]	$\mathcal{O}\big(n\log n(\log^2 n+\log(1/\epsilon)\big)$	$\mathcal{O}\big(n^2\log n(\log^2 n+\log^2(1/\epsilon)\big)$
WF [6]	$\mathcal{O}(n\log n)$	$\mathcal{O}\big(n^3\log n\log(1/\epsilon)\big)$
TAF [8], TWF [7]	$\mathcal{O}(n)$	$\mathcal{O}\big(n^2\log(1/\epsilon)\big)$
This paper	$\mathcal{O}(n)$	$\mathcal{O}\big(n^2\log(1/\epsilon)\big)$

II. Stochastic Truncated Amplitude Flow

To begin, relevant concepts are introduced. If $\boldsymbol{x}\in\mathbb{R}^n$ solves (1), so does $-\boldsymbol{x}$. This prompts the following definition of the Euclidean distance of any estimate $\boldsymbol{z}\in\mathbb{R}^n$ to the solution set of (1): $\mathrm{dist}(\boldsymbol{z},\boldsymbol{x}):=\min\|\boldsymbol{z}\pm\boldsymbol{x}\|$ [6]. Define then the indistinguishable global phase constant as

$$\phi(\boldsymbol{z}):=\begin{cases}0, & \|\boldsymbol{z}+\boldsymbol{x}\|\geq\|\boldsymbol{z}-\boldsymbol{x}\|,\\ \pi, & \text{otherwise.}\end{cases}\tag{2}$$

In the following, with $\boldsymbol{x}$ denoting any solution of the system in (1), we assume $\phi(\boldsymbol{z})=0$; otherwise, $\boldsymbol{z}$ is replaced by $\mathrm{e}^{-j\phi(\boldsymbol{z})}\boldsymbol{z}$, but for brevity, the phase adaptation term $\mathrm{e}^{-j\phi(\boldsymbol{z})}$ shall be dropped whenever clear from the context.

A. Truncated amplitude flow

In this section, the two stages of TAF are first reviewed [8]. Stage one of TAF employs power iterations to obtain an orthogonality-promoting initialization, and the second stage refines the initialization with gradient-type iterations. The orthogonality-promoting initialization advocates approximating the unknown $\boldsymbol{x}$ by $\boldsymbol{z}_0\in\mathbb{R}^n$ that is maximally orthogonal to a carefully chosen subset of sampling vectors $\{\boldsymbol{a}_i\}_{i\in\mathcal{I}_0}$, with $\mathcal{I}_0\subseteq[m]:=\{1,2,\ldots,m\}$. Consider first $\|\boldsymbol{x}\|=1$. Upon computing the squared normalized inner-products $\cos^2\theta_i:=|\langle\boldsymbol{a},\boldsymbol{x}\rangle|^2/(\|\boldsymbol{a}_i\|^2\|\boldsymbol{x}\|^2)$ for all pairs $\{(\boldsymbol{a}_i,\boldsymbol{x})\}_{i=1}^m$, the orthogonality-promoting initialization constructs $\mathcal{I}_0$ by including the indices of $\boldsymbol{a}_i$'s that consist of the smallest $|\mathcal{I}_0|$ normalized inner-products. Precisely, $\boldsymbol{z}_0$ can be found by solving [8]

$$\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathcal{T}}\boldsymbol{Y}_0\boldsymbol{z}:=\min_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathcal{T}}\left(\frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}\right)\boldsymbol{z}\tag{3}$$

with $|\mathcal{I}_0|$ being in the order of n. As proved in [8, Thm. 1], it is sufficient for exact recovery to require $m\geq c_1|\mathcal{I}_0|\geq c_2 n$ to hold for some constants $c_1,\,c_2>0$. Solving (3) entails finding the smallest eigenvalue and the corresponding eigenvector of $\boldsymbol{Y}_0\succeq\boldsymbol{0}$. To bypass the $\mathcal{O}(n^3)$ computational complexity of computing the smallest eigenvector, the concentration result $\sum_{i=1}^m\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}\approxeq\frac{m}{n}\boldsymbol{I}_n$ helps simplifying that to computing the largest eigenvector of

$$\bar{\boldsymbol{Y}}_0:=\frac{1}{|\bar{\mathcal{I}}_0|}\sum_{i\in\bar{\mathcal{I}}_0}\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}\tag{4}$$

where $\bar{\mathcal{I}}_0$ is the complement of $\mathcal{I}_0$ in $[m]$. Upon stacking $\{\boldsymbol{a}_i\}_{i\in\bar{\mathcal{I}}_0}$ into an $n\times|\bar{\mathcal{I}}_0|$ matrix $\boldsymbol{D}$, one can rewrite $\bar{\boldsymbol{Y}}_0=\boldsymbol{D}\boldsymbol{D}^{\mathcal{T}}$ to arrive at the following principal component analysis (PCA) problem

$$\tilde{\boldsymbol{z}}_0:=\arg\max_{\|\boldsymbol{z}\|=1}\ \frac{1}{|\bar{\mathcal{I}}_0|}\boldsymbol{z}^{\mathcal{T}}\boldsymbol{D}\boldsymbol{D}^{\mathcal{T}}\boldsymbol{z}.\tag{5}$$

On the other hand, when $\|\boldsymbol{x}\|\neq1$, the estimate $\tilde{\boldsymbol{z}}_0$ will be scaled by $\sqrt{\sum_{i=1}^m y_i/m}$ to yield [8]

$$\boldsymbol{z}_0:=\sqrt{\frac{1}{m}\sum_{i=1}^m y_i}\,\tilde{\boldsymbol{z}}_0.$$

When the signal dimension n is modest, the solution of (5) can be found exactly by a full singular value decomposition (SVD) of $\boldsymbol{D}$ at runtime of $\mathcal{O}(\min\{n^2\bar{\mathcal{I}}_0,n\bar{\mathcal{I}}_0^2\})$ (or simply $\mathcal{O}(n^3)$ because $|\bar{\mathcal{I}}_0|$ is in the order of n). This clearly grows prohibitively with large dimensions. TAF uses instead the power method to solve (5). Power method, on the other hand, computes a matrix-vector multiplication per iteration, thus incurring per-iteration complexity of $\mathcal{O}(n^2)$ by passing through the data $\{\boldsymbol{a}_i\}_{i\in\bar{\mathcal{I}}_0}$. To reach an ϵ-accurate solution, it has a runtime of $\mathcal{O}(n|\bar{\mathcal{I}}_0|\log(1/\epsilon)/\delta)$ relying on the eigengap $\delta>0$, which is defined to be the gap between the first and the second largest eigenvalues of $\bar{\boldsymbol{Y}}_0$ normalized by the largest one. When δ is small, the runtime $\mathcal{O}(n|\bar{\mathcal{I}}_0|\log(1/\epsilon)/\delta)$ of the power method would be equivalent to many passes over the entire data, and this could be prohibitive for large datasets [14]. So the power method may not be well suited for large-scale imaging applications, particularly those having small eigengaps.

Stage two of TAF relies on truncated gradient-type iterations of the ensuing amplitude-based empirical loss function

$$\underset{\boldsymbol{z}\in\mathbb{R}^n}{\text{minimize}}\ \ \frac{1}{2m}\sum_{i=1}^m\left(\psi_i-|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|\right)^2.\tag{6}$$

With t denoting the iteration number, the truncated gradient stage starts with the initial point $\boldsymbol{z}_0$ found in stage one, and operates iteratively for $t\geq0$:

$$\boldsymbol{z}_{t+1}=\boldsymbol{z}_t-\frac{\mu}{m}\sum_{i\in\mathcal{I}_{t+1}}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t-\psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t|}\right)\boldsymbol{a}_i\tag{7}$$

for appropriately chosen step size $\mu>0$, where the index set accountable for the gradient truncation is given as [8]

$$\mathcal{I}_{t+1}:=\left\{1\leq i\leq m\,\middle|\,\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\geq\frac{1}{1+\gamma}\right\}\tag{8}$$

for some preselected truncation threshold $\gamma>0$.

ISBN 978-0-9928626-7-1

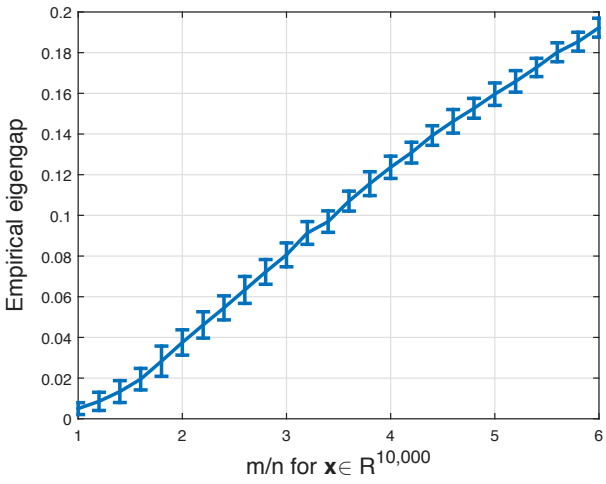

Fig. 1: Eigengaps δ of $\bar{\boldsymbol{Y}}_0$ in (4) averaging over 100 Monte Carlo realizations for $n = 10^4$ fixed and m/n varying by 0.2 from 1 to 6. Real Gaussian model with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.

B. Variance-reducing orthogonality-promoting initialization

This section first presents empirical evidence showing that small eigengaps appear often in our initialization approach. Fig. 1 plots the empirical eigengap of $\bar{\boldsymbol{Y}}_0 \in \mathbb{R}^{n\times n}$ from the real Gaussian model under default parameters of TAF, where $n = 10^4$ is fixed, and m/n the ratio between the number of equations and unknowns increases by 0.2 from 1 to 6. As depicted in Fig. 1, the eigengaps of $\bar{\boldsymbol{Y}}_0$ are rather small particularly for small m/n values approaching the theoretic information limit 2. Effecting power iterations of runtime $\mathcal{O}(n|\bar{\mathcal{I}}_0|\log(1/\epsilon)/\delta)$ thus entails many passes over the selected data due to a large factor of $1/\delta$, which does not scale well to large dimensions in Big Data applications [14]. Effecting recent advances in stochastic optimization [15], a variance-reducing principal component analysis (VR-PCA) algorithm is proposed [14]. It employs cheap stochastic iterations, and has runtime of $\mathcal{O}(n(\bar{\mathcal{I}}_0 + 1/\delta^2)\log(1/\epsilon))$ depending only logarithmically on $\epsilon > 0$. This is in contrast to the standard SGD, whose runtime relies on $1/\epsilon$ due to the large variance of stochastic gradients [14].

To ensure scalability, this paper advocates VR-PCA for computing the initialization from (5). The resulting algorithm is termed the *variance-reducing orthogonality-promoting initialization* (VR-OPI), and summarized in Alg. 1. To be specific, VR-OPI is a double-loop algorithm with a single execution of the inner loop termed an iteration and an execution of the outer loop an epoch. In practice, VR-OPI comprises S epochs, while each epoch runs T (often equal to the data size $|\bar{\mathcal{I}}_0|$) iterations. It is worth mentioning that the full gradient evaluated per execution of the outer loop combined with the stochastic gradients within the inner loop can be shown able to reduce the variance of stochastic gradients [15].

Algorithm 1 Variance-reduced orthogonality-promoting initialization (VR-OPI)

Input: Data matrix $\boldsymbol{D} = \{\boldsymbol{a}_i\}_{i\in\bar{\mathcal{I}}_0}$, step size $\eta = 20/m$, the total number of epochs $S = 100$, and epoch length $T = |\bar{\mathcal{I}}_0|$.
Initialize a unit vector $\tilde{\boldsymbol{u}}_0 \in \mathbb{R}^n$ randomly.
For $s = 0$ **to** $S-1$ **do**
  $\boldsymbol{w} = \frac{1}{|\bar{\mathcal{I}}_0|}\sum_{i\in\bar{\mathcal{I}}_0} \boldsymbol{a}_i(\boldsymbol{a}_i^{\mathcal{T}}\tilde{\boldsymbol{u}}_s)$
  $\boldsymbol{u}_1 = \tilde{\boldsymbol{u}}_s$.
  For $t = 0$ **to** $T-1$ do
    Pick $i_t \in \bar{\mathcal{I}}_0$ uniformly at random
    $\boldsymbol{\nu}_{t+1} = \boldsymbol{u}_t + \eta\left[\boldsymbol{a}_{i_t}\left(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{u}_t - \boldsymbol{a}_{i_t}^{\mathcal{T}}\tilde{\boldsymbol{u}}_s\right) + \boldsymbol{w}\right]$
    $\boldsymbol{u}_{t+1} = \frac{\boldsymbol{\nu}_{t+1}}{\|\boldsymbol{\nu}_{t+1}\|}$.
  End For
  $\tilde{\boldsymbol{u}}_{s+1} = \boldsymbol{u}_T$.
End For
Output: $\tilde{\boldsymbol{z}}_0 = \boldsymbol{u}_S$.

The following results from [14, Thm. 1] establish the linear convergence rate of VR-OPI.

Proposition 1 ([14]): Let $\boldsymbol{v}_1 \in \mathbb{R}^n$ be an eigenvector of $\bar{\boldsymbol{Y}}_0$ corresponding to the largest eigenvalue λ_1. Assume that $\max_{i\in[m]}\|\boldsymbol{a}_i\|^2 \le r := 2.3n$ (which holds with probability at least $1 - me^{-n/2}$), the two largest eigenvalues of $\bar{\boldsymbol{Y}}_0$ are $\lambda_1 > \lambda_2 > 0$ with $\delta = (\lambda_1 - \lambda_2)/\lambda_1$, and that $\langle\tilde{\boldsymbol{u}}_0, \boldsymbol{v}_1\rangle \ge 1/\sqrt{2}$. With any $0 < \epsilon, \xi < 1$, constant step size $\eta > 0$, and epoch length T chosen such that

$$\eta \le \frac{c_0\xi^2}{r^2}\delta \tag{9}$$

$$T \ge \frac{c_1\log(2/\xi)}{\eta\delta} \tag{10}$$

$$T\eta^2 r^2 + r\eta\sqrt{T\log(2/\xi)} \le c_2 \tag{11}$$

for some universal constants $c_0, c_1, c_2 > 0$, successive estimates of VR-OPI (summarized in Alg. 1) after $S = \lceil\log(1/\epsilon)/\log(2/\xi)\rceil$ epochs obey

$$|\langle\tilde{\boldsymbol{u}}_S, \boldsymbol{v}_1\rangle|^2 \ge 1 - \epsilon \tag{12}$$

with probability at least $1 - \lceil\log\epsilon\rceil\xi$.

It is worth stressing that the runtime of VR-OPI is proportional to the time required to scan the entire data once, which improves upon the runtime $\mathcal{O}\big(\frac{1}{\delta}n|\bar{\mathcal{I}}_0|\log(1/\epsilon)\big)$ of the power method. Our simulated tests showcase the effectiveness of VR-OPI over the power method in processing data of large dimensions m and/or n.

C. Stochastic truncated gradient stage

Recall that TAF achieves exact recovery from about $3n$ noiseless equations [8]. It is known that gradient iterations can be trapped in saddle points when dealing with nonconvex optimization. Nevertheless, stochastic iterations are able to escape saddle points, and converge globally to at least a local minimum. Hence, besides the appealing computational advantage, stochastic counterparts of TAF may enjoy further

ISBN 978-0-9928626-7-1 © EURASIP 2017

improved performance. To inherit the merits of TAF, the gradient regularization (8) is also adopted in the new algorithm to lead to our truncated stochastic gradient iterations.

For simplicity, rewrite the cost function as follows

$$\underset{\boldsymbol{z}\in\mathbb{R}^n}{\text{minimize}} \quad \ell(\boldsymbol{z}) = \sum_{i=1}^{m} \ell_i(\boldsymbol{z}) := \frac{1}{2}\sum_{i=1}^{m} \left(\psi_i - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|\right)^2 \quad (13)$$

where the factor $1/2$ is introduced for notational convenience. It is clear that the cost $\ell(\boldsymbol{z})$ or each $\ell_i(\boldsymbol{z})$ is nonconvex and nonsmooth; hence, the optimization in (13) is computationally intractable in general [16]. Along the lines of nonconvex paradigms including WF [6], TWF [7], and TAF [8], our approach to solving the problem at hand amounts to iteratively refining the initial estimate $\boldsymbol{z}_0$ by means of truncated stochastic gradient iterations. This is in contrast to (T)WF and TAF, which rely on (truncated) gradient-type iterations [6]–[8]. STAF processes one datum at a time, and evaluates the generalized gradient of one function $\ell_{i_t}(\boldsymbol{z})$ for some index $i_t \in [m]$ per iteration t. Specifically, STAF successively updates $\boldsymbol{z}_0$ using the following truncated stochastic gradient iterations for all $t \geq 0$

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \mu\partial\ell_{i_t}(\boldsymbol{z}_t)\mathbb{1}_{\left\{\frac{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t|}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}|}\geq\frac{1}{1+\gamma}\right\}} \quad (14)$$

with

$$\partial\ell_{i_t}(\boldsymbol{z}_t) = \left(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t|}\right)\boldsymbol{a}_{i_t} \quad (15)$$

where the constant step size $\mu > 0$ is chosen to be in the order of $1/n$, and the index i_t is sampled uniformly at random from $[m]$, or it simply cycles through $[m]$. Upon fixing $\gamma = 0.7$, the indicator function $\mathbb{1}_{\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t|/|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}|\geq 1/(1+\gamma)\}}$ takes 1, if $|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t|/|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}| \geq 1/(1+\gamma)$ is true; and 0 otherwise. This truncation rule can provably reject "bad" search directions with high probability [8]. Furthermore, this regularization maintains only gradients of sufficiently large $|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|$ values, therefore preventing the objective functional (6) being non-differentiable at point $\boldsymbol{z}^t$ and simplifying the theoretical analysis. The developed STAF scheme is summarized as Alg. 2. Numerical tests showing the performance improvement using STAF will be detailed in Sec. IV.

III. Main Results

In this section, STAF is shown to converge exponentially fast to the globally optimal solution with high probability when the number of equations and unknowns m/n exceeds some constant. Assuming independent data samples $\{(\boldsymbol{a}_i;\psi_i)\}_{i=1}^m$ from the real Gaussian model, the next theorem establishes analytical performance of STAF from noiseless measurements.

*Theorem 1 (***Exact recovery):** Consider the noise-free data $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ with an arbitrary signal $\boldsymbol{x} \in \mathbb{R}^n$, and i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \leq i \leq m$. Provided that

$$m \geq c_0 n \quad \text{and} \quad \mu \leq \frac{\mu_0}{n} \quad (16)$$

Algorithm 2 Stochastic truncated amplitude flow (STAF)

1: **Input:** Data $\{(\boldsymbol{a}_i, \psi_i)\}_{i=1}^m$; maximum number of iterations $T = 500m$; by default, step size $\mu = 0.8/n$, truncation thresholds $|\overline{\mathcal{I}}_0| = \lceil \frac{1}{6}m \rceil$, and $\gamma = 0.7$.

2: **Evaluate** $\overline{\mathcal{I}}_0$ to include the indices of the $|\overline{\mathcal{I}}_0|$ largest quantities among $\{\psi_i/\|\boldsymbol{a}_i\|\}_{i=1}^m$.

3: **Initialize** $\boldsymbol{z}_0$ as $\sqrt{\frac{1}{m}\sum_{i=1}^m \psi_i^2}\tilde{\boldsymbol{z}}_0$, where $\tilde{\boldsymbol{z}}_0$ is computed by Alg. 1 with

$$\bar{\boldsymbol{Y}}_0 := \frac{1}{|\overline{\mathcal{I}}_0|}\sum_{i\in\overline{\mathcal{I}}_0}\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}.$$

4: **Loop:** For $t = 0$ to $T-1$

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \mu\boldsymbol{a}_{i_t}\left(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t|}\right)\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}_t|\geq\psi_{i_t}/(1+\gamma)\right\}}$$

with i_t drawn uniformly at random from $[m]$.

5: **Output:** $\boldsymbol{z}_T$.

then with probability at least $1 - c_1 m\exp(-c_2 n)$, the STAF estimates (tabulated in Alg. 2 with default parameters) obey

$$\mathbb{E}_{\mathcal{P}_t}\left[\text{dist}^2(\boldsymbol{z}_t, \boldsymbol{x})\right] \leq \frac{1}{10}\left(1 - \frac{\nu}{n}\right)^t \|\boldsymbol{x}\|^2, \quad t \geq 0 \quad (17)$$

for some numerical constant $\nu > 0$, where the expectation is taken over the path sequence $\mathcal{P}_t := \{i_0, i_1, \ldots, i_{t-1}\}$, and $c_0, c_1, c_2, \mu_0 > 0$ are universal constants.

The proof of Thm. 1 can be found in our journal version [10]. A few observations regarding Thm. 1 are in order. First, the mean-square distance between the iterate $\boldsymbol{z}_t$ and the solution set $\{\boldsymbol{x}\}$ is reduced by a factor of $(1 - \nu/n)^m$ after one entire pass of the data. Moreover, the expectation $\mathbb{E}_{\mathcal{P}_t}[\cdot]$ is taken over the algorithmic randomness $\mathcal{P}_t$ instead of the data as in general the data may be modeled as deterministic. Interestingly enough, albeit effecting inexpensive stochastic iterations, STAF still enjoys linear convergence rate.

IV. Simulated Tests

This section compares STAF with the state-of-the-art TAF [8] and (T)WF [6], [7]. For fairness, all the parameters pertinent to implementation of each algorithm were set to their default values. The initialization in each scheme was found using a number of iterations equivalent to 100 passes over the entire data, which was refined by a number of iterations corresponding to $1,000$ passes. All estimates were averaged over 100 independent trials. Two performance criteria were used: $\text{Relative error} := \text{dist}(\boldsymbol{z}, \boldsymbol{x})/\|\boldsymbol{x}\|$; and the successful recovery rate among 100 Monte Carlo runs, in which a success is claimed when the returned estimate incurs a relative error less than 10^{-5} [6].

The first experiment compares VR-OPI in Alg. 1 with the power method in computing the orthogonality-promoting initialization from (5). A synthetic data based experiment is carried out based on the noiseless real Gaussian model with $n = 10^4$ under the theoretic information limit number of

ISBN 978-0-9928626-7-1 © EURASIP 2017

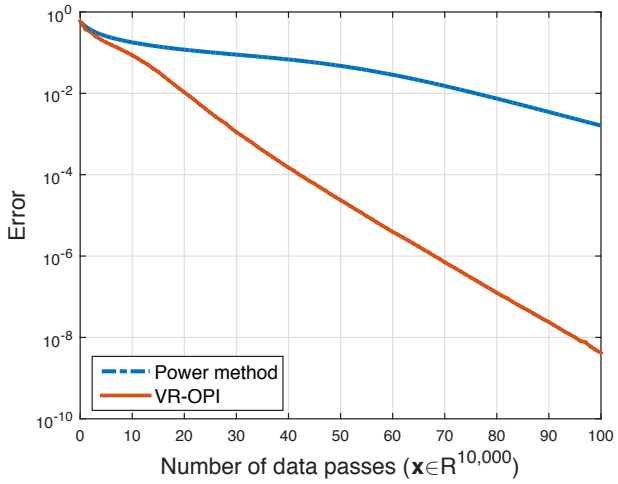

Fig. 2: Error evolution of iterates using: i) power method; and ii) VR-OPI in Alg. 1 for solving (5) with step size $\eta = 1$. Noiseless real Gaussian model with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, where $n = 10^4$, and $m = 2n - 1$.

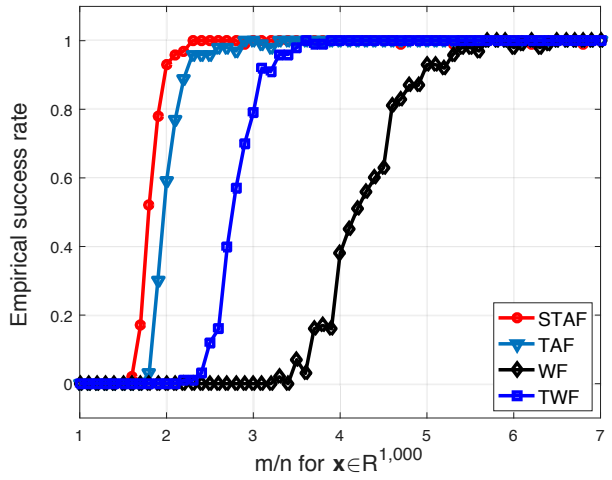

Fig. 3: Empirical success rate for: i) WF [6]; ii) TWF [7]; iii) TAF [8]; and iv) STAF with $n = 1,000$ and m/n varying 0.1 from 1 to 7. Noiseless real Gaussian model with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

measurements $m = 2n - 1$. Fig. 2 plots the error evolution of iterates $\boldsymbol{u}_t$, where the error in logarithmic scale is defined as $\log_{10}(1 - \|\boldsymbol{D}^{\mathcal{T}}\boldsymbol{u}_t\|^2/\|\boldsymbol{D}^{\mathcal{T}}\boldsymbol{v}_0\|^2)$ with the exact principal eigenvector $\boldsymbol{v}_0$ computed from the SVD of $\bar{\boldsymbol{Y}}_0 = \boldsymbol{D}\boldsymbol{D}^{\mathcal{T}}$ in (5). Apparently, the inexpensive stochastic iterations of VR-OPI achieve given solution accuracy with much fewer gradient evaluations or data passes. The second experiment evaluates the exact recovery performance of various schemes with $n = 10^3$ and m/n varying from 1 to 7. Fig. 3 demonstrates improved performance of STAF over its competing alternatives under the noiseless real Gaussian model.

V. Conclusions

This paper developed a new linear-time algorithm abbreviated with STAF to solve systems of quadratic equations. STAF operates in two stages, that first obtains an orthogonality-promoting initialization based on an SVRG algorithm, and subsequently refines the initial estimate by means of truncated stochastic gradient iterations. STAF is shown capable of recovering any signal from about as many equations as unknowns. Relative to past approaches, both stages of our developed STAF achieve order-optimal per-iteration and computational complexity. Numerical tests showcase the merits of STAF over the state-of-the-art approaches.

A few future research directions consist of developing analytical results for STAF in the presence of additive noise, and exploiting the possibility of the orthogonality-promoting initialization in the context of robust phase retrieval and faster semidefinite optimization. Designing inexpensive stochastic solvers for phase retrieval of structured (e.g. sparse) signals, as well as generalization to two-dimensional phase retrieval problems, constitute additional future research directions.

References

[1] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Rev.*, vol. 57, no. 2, pp. 225–251, May 2015.

[2] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Appl. Comput. Harmon. Anal.*, vol. 20, no. 3, pp. 345–356, May 2006.

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

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

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

[6] E. J. Candès, 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.

[7] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Comm. Pure Appl. Math.*, 2016 (to appear).

[8] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *arXiv:1605.08285*, 2016.

[9] G. Wang, L. Zhang, G. B. Giannakis, J. Chen, and M. Akçakaya, "Sparse phase retrieval via truncated amplitude flow," *arXiv:1611.07641*, 2016.

[10] G. Wang, G. B. Giannakis, and J. Chen, "Scalable solvers of random quadratic equations via stochastic truncated amplitude flow," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1961–1974, Apr. 2017.

[11] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen, "Solving almost all systems of random quadratic equations," *aXiv:1705.10407*, 2017.

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

[13] T. Qiu, P. Babu, and D. P. Palomar, "PRIME: Phase retrieval via majorization-minimization," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5174–5186, Oct. 2016.

[14] O. Shamir, "Fast stochastic algorithms for SVD and PCA: Convergence properties and convexity," in *The 33th Proc. of Intl. Conf. on Machine Learning*, New York City, NY, 2016.

[15] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Adv. Neural Inf. Process. Syst.*, 2013, pp. 315–323.

[16] P. M. Pardalos and S. A. Vavasis, "Quadratic programming with one negative eigenvalue is NP-hard," *J. Global Optim.*, vol. 1, no. 1, pp. 15–22, 1991.

ISBN 978-0-9928626-7-1 © EURASIP 2017