Solving Random Systems of Quadratic Equations via Truncated Generalized Gradient Flow

Gang Wang[*,†] **and Georgios B. Giannakis**[†]
[*] ECE Dept. and Digital Tech. Center, Univ. of Minnesota, Mpls, MN 55455, USA
[†] School of Automation, Beijing Institute of Technology, Beijing 100081, China
{gangwang, georgios}@umn.edu

Abstract

This paper puts forth a novel algorithm, termed *truncated generalized gradient flow* (TGGF), to solve for $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ a system of m quadratic equations $y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2$, $i = 1, 2, \ldots, m$, which even for $\{\boldsymbol{a}_i \in \mathbb{R}^n/\mathbb{C}^n\}_{i=1}^m$ random is known to be *NP-hard* in general. We prove that as soon as the number of equations m is on the order of the number of unknowns n, TGGF recovers the solution exactly (up to a global unimodular constant) with high probability and complexity growing linearly with the time required to read the data $\{(\boldsymbol{a}_i;\, y_i)\}_{i=1}^m$. Specifically, TGGF proceeds in two stages: s1) A novel *orthogonality-promoting* initialization that is obtained with simple power iterations; and, s2) a refinement of the initial estimate by successive updates of scalable *truncated generalized gradient iterations*. The former is in sharp contrast to the existing spectral initializations, while the latter handles the rather challenging nonconvex and nonsmooth *amplitude-based* cost function. Empirical results demonstrate that: i) The novel orthogonality-promoting initialization method returns more accurate and robust estimates relative to its spectral counterparts; and, ii) even with the same initialization, our refinement/truncation outperforms Wirtinger-based alternatives, all corroborating the superior performance of TGGF over state-of-the-art algorithms.

1 Introduction

Consider a system of m quadratic equations

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2, \quad i \in [m] := \{1, 2, \ldots, m\} \tag{1}$$

where data vector $\boldsymbol{y} := [y_1 \ \cdots \ y_m]^{\mathcal{T}}$ and feature vectors $\boldsymbol{a}_i \in \mathbb{R}^n/\mathbb{C}^n$, collected in the $m \times n$ matrix $\boldsymbol{A} := [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_m]^{\mathcal{H}}$ are known, whereas vector $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ is the wanted unknown. When $\{\boldsymbol{a}_i\}_{i=1}^m$ and/or $\boldsymbol{x}$ are complex, their amplitudes are given but phase information is lacking; whereas in the real case only the signs of $\{\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\}$ are unknown. Supposing that the system of equations in (1) admits a unique solution $\boldsymbol{x}$ (up to a global unimodular constant), our objective is to reconstruct $\boldsymbol{x}$ from m phaseless quadratic equations, or equivalently, recover the missing signs/phases of $\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle$ in the real-/complex-valued settings. Indeed, it has been established that $m \geq 2n-1$ or $m \geq 4n-4$ generic data $\{(\boldsymbol{a}_i;\, y_i)\}_{i=1}^m$ as in (1) suffice for uniqueness of an n-dimensional real- or complex-valued vector $\boldsymbol{x}$ [1, 2], respectively, and the former with equality has also been shown to be necessary [1].

The problem in (1) constitutes an instance of nonconvex quadratic programming, that is generally known to be *NP-hard* [3]. Specifically for real-valued vectors, this can be understood as a combinatorial optimization since one seeks a series of signs $s_i = \pm 1$, such that the solution to the system of linear equations $\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle = s_i\psi_i$, where $\psi_i := \sqrt{y_i}$, obeys the given quadratic system (1). Concatenating all amplitudes $\{\psi_i\}_{i=1}^m$ to form the vector $\boldsymbol{\psi} := [\psi_1 \ \cdots \ \psi_m]^{\mathcal{T}}$, apparently there are a total of 2^m different combinations of $\{s_i\}_{i=1}^m$, among which only two lead to $\boldsymbol{x}$ up to a global sign.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The complex case becomes even more complicated, where instead of a set of signs $\{s_i\}_{i=1}^m$, one must specify for uniqueness a collection of unimodular complex scalars $\{\sigma_i \in \mathbb{C}\}_{i=1}^m$. In many fields of physical sciences and engineering, the problem of recovering the phase from intensity/magnitude-only measurements is commonly referred to as *phase retrieval* [4, 5]. The plethora of applications include X-ray crystallography, optics, as well as array imaging, where due to physical limitations, optical detectors can record only (squared) modulus of the Fresnel or Fraunhofer diffraction pattern, while losing the phase of the incident light reaching the object [5]. It has been shown that reconstructing a discrete, finite-duration signal from its Fourier transform magnitude is *NP-complete* [6]. Despite its simple form and practical relevance across various fields, tackling the quadratic system (1) under real-/complex-valued settings is challenging and *NP-hard* in general.

1.1 Nonconvex Optimization

Adopting the least-squares criterion, the task of recovering $\boldsymbol{x}$ can be recast as that of minimizing the following *intensity-based* empirical loss

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} \; f(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^m \left(y_i - \left|\boldsymbol{a}_i^{\mathcal{H}}\boldsymbol{z}\right|^2\right)^2 \tag{2}$$

or, the *amplitude-based* one

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} \; \ell(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^m \left(\psi_i - \left|\boldsymbol{a}_i^{\mathcal{H}}\boldsymbol{z}\right|\right)^2. \tag{3}$$

Unfortunately, both cost functions (2) and (3) are nonconvex. Minimizing nonconvex objectives, which may exhibit many stationary points, is in general *NP-hard* [7]. In a nutshell, solving problems of the form (2) or (3) is challenging.

Existing approaches to solving (2) (or related ones using the Poisson likelihood; see, e.g., [8]) or (3) fall under two categories: nonconvex and convex ones. Popular nonconvex solvers include the alternating projection such as Gerchberg-Saxton [9] and Fineup [10], AltMinPhase [11], and (Truncated) Wirtinger flow (WF/TWF) [12, 8], as well as trust-region methods [13]. Convex approaches on the other hand rely on the so-called *matrix-lifting* technique to obtain the solvers abbreviated as PhaseLift [14] and PhaseCut [15].

In terms of sample complexity for Gaussian $\{\boldsymbol{a}_i\}$ designs, convex approaches enable exact recovery from[1] $\mathcal{O}(n)$ noiseless measurements [16], while they require solving a semidefinite program of a matrix variable with size $n \times n$, thus incurring worst-case computational complexity on the order of $\mathcal{O}(n^{4.5})$ [15], that does not scale well with dimensionality n. Upon exploiting the underlying problem structure, $\mathcal{O}(n^{4.5})$ can be reduced to $\mathcal{O}(n^3)$ [15]. Solving for vector variables, nonconvex approaches achieve significantly improved computational performance. Using formulation (3), AltMinPhase adopts a spectral initialization and establishes exact recovery with sample complexity $\mathcal{O}(n\log^3 n)$ under Gaussian $\{\boldsymbol{a}_i\}$ designs with resampling [11]. Concerning formulation (2), WF iteratively refines the spectral initial estimate by means of a gradient-like update [12]. The follow-up TWF improves upon WF through a truncation procedure to separate gradient components of excessively extreme sizes. Likewise, at the initialization stage, since the term $(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x})^2\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{H}}$ responsible for the spectral initialization is heavy-tailed, data $\{y_i\}_{i=1}^m$ are pre-screened in the truncated spectral initialization to yield improved initial estimates [8]. Under Gaussian sampling models, WF allows exact recovery from $\mathcal{O}(n\log n)$ measurements in $\mathcal{O}(mn^2\log(1/\epsilon))$ time/flops to yield an ϵ-accurate solution for any given $\epsilon > 0$ [12], while TWF advances these to $\mathcal{O}(n)$ measurements and $\mathcal{O}(mn\log(1/\epsilon))$ time [8]. Interestingly, the truncation procedure in the gradient stage turns out to be useful in avoiding spurious stationary points in the context of nonconvex optimization. Although for large-scale linear regressions, similar ideas including censoring have been studied [17, 18]. It is worth mentioning that when $m \geq Cn\log^3 n$ for sufficiently large $C > 0$, the objective function in (3) admits benign geometric structure that allows certain iterative algorithms (e.g., trust-region methods) to efficiently find a global minimizer with random initializations [13].

Although achieving a linear (in the number of unknowns n) sample and computational complexity, the state-of-the-art TWF scheme still requires at least $4n \sim 5n$ equations to yield a stable empirical success rate (e.g., $\geq 99\%$) under the real Gaussian model [8, Section 3], which are more than twice the known information-limit of $m = 2n - 1$ [1]. Similar though less obvious results hold also in

[1]The notation $\phi(n) = \mathcal{O}(g(n))$ means that there is a constant $c > 0$ such that $|\phi(n)| \leq c|g(n)|$.

the complex-valued scenario. Even though the truncated spectral initialization improves upon the "plain vallina" spectral initialization, its performance still suffers when the number of measurements is relatively small and its advantage (over the untruncated version) narrows as the number of measurements grows. Further, it is worth stressing that extensive numerical and experimental validation confirms that the *amplitude-based* cost function performs better than the *intensity-based* one; that is, formulation (3) is superior over (2) [19]. Hence, besides enhancing initialization, markedly improved performance in the gradient stage could be expected by re-examining the amplitude-based cost function and incorporating judiciously designed truncation rules.

2 Algorithm: Truncated Generalized Gradient Flow

Along the lines of suitably initialized nonconvex schemes, and building upon the amplitude-based formulation (3), this paper develops a novel linear-time (in both m and n) algorithm, referred to as *truncated generalized gradient flow* (TGGF), that provably recovers $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ exactly from a near-optimal number of noise-free measurements, while also featuring a near-perfect statistical performance in the noisy setup. Our TGGF proceeds in two stages: s1) A novel orthogonality-promoting initialization that relies on simple power iterations to markedly improve upon spectral initialization; and, s2) a refinement of the initial estimate by successive updates of *truncated generalized gradient* iterations. Stages s1) and s2) are delineated next in reverse order. For concreteness, our analysis will focus on the real Gaussian model with $\boldsymbol{x} \in \mathbb{R}^n$ and independently and identically distributed (i.i.d.) design vectors $\boldsymbol{a}_i \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, whereas numerical implementations for the complex Gaussian model having $\boldsymbol{x} \in \mathbb{C}^n$ and i.i.d. $\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n) := \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/2) + j\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/2)$ will be discussed briefly. To start, define the Euclidean distance of any estimate $\boldsymbol{z}$ to the solution set: $\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min \|\boldsymbol{z} \pm \boldsymbol{x}\|$ for real signals, and $\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min_{\phi\in[0,2\pi)} \|\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{i\phi}\|$ for complex ones [12]. Define also the indistinguishable global phase constant in real-valued settings as

$$\phi(\boldsymbol{z}) := \begin{cases} 0, & \|\boldsymbol{z} - \boldsymbol{x}\| \leq \|\boldsymbol{z} + \boldsymbol{x}\|, \\ \pi, & \text{otherwise.} \end{cases} \tag{4}$$

Henceforth, fixing $\boldsymbol{x}$ to be any solution of the given quadratic system (1), we always assume that $\phi(\boldsymbol{z}) = 0$; otherwise, $\boldsymbol{z}$ is replaced by $\mathrm{e}^{-j\phi(\boldsymbol{z})}\boldsymbol{z}$, but for simplicity of presentation, the constant phase adaptation term $\mathrm{e}^{-j\phi(\boldsymbol{z})}$ is dropped whenever it is clear from the context.

Numerical tests comparing TGGF, TWF, and WF will be presented throughout our analysis, so let us first describe our basic test settings. Simulated estimates will be averaged over 100 independent Monte Carlo (MC) realizations without mentioning this explicitly each time. Performance is evaluated in terms of the relative root mean-square error, i.e., $\text{Relative error} := \mathrm{dist}(\boldsymbol{z}, \boldsymbol{x})/\|\boldsymbol{x}\|$, and the success rate among 100 trials, where a success will be claimed for a trial if the resulting estimate incurs relative error less than 10^{-5} [8]. Simulated tests under both noiseless and noisy Gaussian models are performed, corresponding to $\psi_i = \left|\boldsymbol{a}_i^{\mathcal{H}}\boldsymbol{x} + \eta_i\right|$ with $\eta_i = 0$ and $\eta_i \sim \mathcal{N}(0, \sigma^2)$ [11], respectively, with i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ or $\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$.

2.1 Truncated generalized gradient stage

Let us rewrite the amplitude-based cost function in a matrix-vector form as

$$\min_{\boldsymbol{z}\in\mathbb{R}^n} \ \ell(\boldsymbol{z}) = \frac{1}{2m}\big\|\boldsymbol{\psi} - |\boldsymbol{A}\boldsymbol{z}|\,\big\|^2 \tag{5}$$

where $|\boldsymbol{A}\boldsymbol{z}| := \left[|\boldsymbol{a}_1^{\mathcal{T}}\boldsymbol{z}| \ \cdots \ |\boldsymbol{a}_m^{\mathcal{T}}\boldsymbol{z}|\right]^{\mathcal{T}}$. Apart from being nonconvex, $\ell(\boldsymbol{z})$ is nondiffentiable. In the presence of smoothness or convexity, convergence analysis of iterative algorithms relies either on continuity of the gradient (gradient methods) [20], or, on the convexity of the objective functional (subgradient methods) [20]. Although subgradient methods have found widespread applicability in nonsmooth optimization, they are *limited* to the class of convex functions [20, Page 4]. In nonconvex nonsmooth optimization, the so-termed *generalized gradient* broadens the scope of the (sub)gradient to the class of *almost everywhere* differentiable functions [21]. Consider a continuous function $h(\boldsymbol{z}) \in \mathbb{R}$ defined over an open region $\mathcal{S} \subseteq \mathbb{R}^n$.

Definition 1 *[22, Definition 1.1] The generalized gradient of a function h at $\boldsymbol{z}$, denoted by ∂h, is the convex hull of the set of limits of the form $\lim \nabla h(\boldsymbol{z}_k)$, where $\boldsymbol{z}_k \to \boldsymbol{z}$ as $k \to +\infty$, i.e.,*

$$\partial h(\boldsymbol{z}) := \text{conv}\Big\{ \lim_{k\to+\infty} \nabla h(\boldsymbol{z}_k) : \boldsymbol{z}_k \to \boldsymbol{z},\ \boldsymbol{z}_k \notin \mathcal{G}_\ell \Big\}$$

where the symbol 'conv' *signifies the convex hull of a set, and* $\mathcal{G}_\ell$ *denotes the set of points in* $\mathcal{S}$ *at which* h *fails to be differentiable.*

Having introduced the notion of generalized gradient, and with t denoting the iteration number, our approach to solving (5) amounts to iteratively refining the initial guess $\boldsymbol{z}_0$ by means of the ensuing *truncated* generalized gradient iterations

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \mu_t \partial \ell_{\text{tr}}(\boldsymbol{z}_t) \tag{6}$$

where $\mu_t > 0$ is the stepsize, and a piece of the (truncated) generalized gradient $\partial \ell_{\text{tr}}(\boldsymbol{z}_t)$ is given by

$$\partial \ell_{\text{tr}}(\boldsymbol{z}_t) := \sum_{i\in\mathcal{I}_{t+1}} \Big(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t|} \Big) \boldsymbol{a}_i \tag{7}$$

for some index set $\mathcal{I}_{t+1} \subseteq [m]$ to be designed shortly; and the convention $\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t|} := 0$ is adopted, if $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t = 0$. Further, it is easy to verify that the update in (6) monotonically decreases the objective value in (5).

Recall that since they offer descent iterations, the alternating projection variants are guaranteed to converge to a stationary point of $\ell(\boldsymbol{z})$, and any limit point $\boldsymbol{z}^*$ adheres to the following fixed-point equation [23]

$$\boldsymbol{A}^{\mathcal{T}} \Big(\boldsymbol{A}\boldsymbol{z}^* - \boldsymbol{\psi} \odot \frac{\boldsymbol{A}\boldsymbol{z}^*}{|\boldsymbol{A}\boldsymbol{z}^*|} \Big) = \boldsymbol{0} \tag{8}$$

for entry-wise product $\odot$, which may have many solutions. Clearly, if $\boldsymbol{z}^*$ is a solution, so is $-\boldsymbol{z}^*$. Further, both solutions/global minimizers $\boldsymbol{x}$ and $-\boldsymbol{x}$ satisfy (8) due to $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{\psi} \odot \frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} = \boldsymbol{0}$. Considering any stationary point $\boldsymbol{z}^* \neq \pm\boldsymbol{x}$ that has been adapted such that $\phi(\boldsymbol{z}^*) = 0$, one can write $\boldsymbol{z}^* = \boldsymbol{x} + (\boldsymbol{A}^{\mathcal{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathcal{T}} \big[\boldsymbol{\psi} \odot \big(\frac{\boldsymbol{A}\boldsymbol{z}^*}{|\boldsymbol{A}\boldsymbol{z}^*|} - \frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} \big) \big]$. A necessary condition for $\boldsymbol{z}^* \neq \boldsymbol{x}$ is $\frac{\boldsymbol{A}\boldsymbol{z}^*}{|\boldsymbol{A}\boldsymbol{z}^*|} \neq \frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|}$. Expressed differently, there must be sign differences between $\frac{\boldsymbol{A}\boldsymbol{z}^*}{|\boldsymbol{A}\boldsymbol{z}^*|}$ and $\frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|}$ whenever one gets stuck with an undesirable stationary point $\boldsymbol{z}^*$. Building on this observation, it is reasonable to devise algorithms that can detect and separate out the generalized gradient components corresponding to mistakenly estimated signs $\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t|}$ along the iterates $\{\boldsymbol{z}_t\}$. Precisely, if $\boldsymbol{z}_t$ and $\boldsymbol{x}$ lie in different sides of the hyperplane $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z} = 0$, then the sign of $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}_t$ will be different than that of $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}$; that is, $\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \neq \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|}$. Specifically, one can write the i-th generalized gradient component

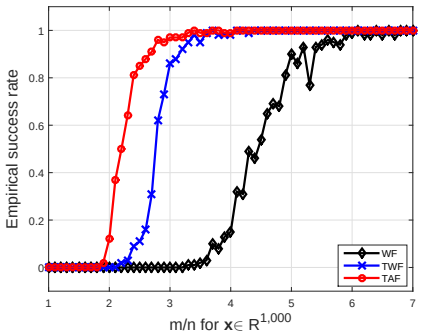

Figure 1: Empirical success rate for WF, TWF, and TGGF with the *same* truncated spectral initialization under the noiseless real Gaussian model.

$$\begin{aligned} \partial \ell_i(\boldsymbol{z}) &= \Big(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z} - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|} \Big) \boldsymbol{a}_i = \Big(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z} - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \Big) \boldsymbol{a}_i + \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} - \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|} \Big) \psi_i \boldsymbol{a}_i \\ &= \boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} + \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} - \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|} \Big) \psi_i \boldsymbol{a}_i \triangleq \boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} + \boldsymbol{r}_i \end{aligned} \tag{9}$$

where $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$. Apparently, the strong law of large numbers (SLLN) asserts that averaging the first term $\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}$ over m instances approaches $\boldsymbol{h}$, which qualifies it as a desirable search direction. However, certain generalized gradient entries involve erroneously estimated signs of $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}$; hence, nonzero $\boldsymbol{r}_i$ terms exert a negative influence on the search direction $\boldsymbol{h}$ by dragging the iterate away from $\boldsymbol{x}$, and they typically have sizable magnitudes. To see why, recall that quantities $\max_{i\in[m]} \psi_i$ and $(1/m)\sum_{i=1}^m \psi_i$ have magnitudes on the order of $\sqrt{m}\|\boldsymbol{x}\|$ and $\sqrt{\pi/2}\|\boldsymbol{x}\|$, respectively, whereas $\|\boldsymbol{h}\| \leq \rho\|\boldsymbol{x}\|$ for some small constant $0 < \rho \leq 1/10$, to be discussed shortly. To maintain a meaningful search direction, those 'bad' generalized gradient entries should be detected and excluded from the search direction.

Nevertheless, it is difficult or even impossible to check whether the sign of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t$ equals that of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}$. Fortunately, when the initialization is accurate enough, most spurious gradient entries (those corrupted by nonzero $\boldsymbol{r}_i$ terms) provably hover around the watershed hyperplane $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t = 0$. For this reason, TGGF includes only those components having $\boldsymbol{z}_t$ sufficiently away from its watershed

$$\mathcal{I}_{t+1} := \Big\{1 \le i \le m \Big| \frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|} \ge \frac{1}{1+\gamma}\Big\}, \quad t \ge 0 \tag{10}$$

for an appropriately selected threshold $\gamma > 0$. It is worth stressing that our novel truncation rule deviates from the intuition behind TWF. Among its complicated truncation procedures, TWF also throws away large-size gradient components corresponding to (10), which is not the case with TGGF. As demonstrated by our analysis, it rarely happens that a generalized gradient component having a large $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t|/\|\boldsymbol{z}_t\|$ yields an incorrect sign of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}$. Further, discarding too many samples (those $i \notin \mathcal{T}_{t+1}$) introduces large bias into $(1/m)\sum_{i\in\mathcal{T}_{t+1}}^{m} \boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_t$, thus rendering TWF less effective when m/n is small. Numerical comparison depicted in Fig. 1 suggests that even starting with the same truncated spectral initialization, TGGF's refinement outperforms those of TWF and WF, corroborating the merits of our novel truncation and update rule over TWF/WF.

2.2 Orthogonality-promoting initialization stage

Leveraging the SLLN, spectral methods estimate $\boldsymbol{x}$ using the (appropriately scaled) leading eigenvector of $\boldsymbol{Y} := \frac{1}{m}\sum_{i\in\mathcal{T}_0} y_i\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}$, where $\mathcal{T}_0$ is an index set accounting for possible truncation. As asserted in [8], each summand $(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x})^2\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}$ follows a heavy-tail probability density function lacking a moment generating function. This causes major performance degradation especially when the number of measurements is limited. Instead of spectral initialization, we shall take another route to bypass this hurdle. To gain intuition for selecting our alternate route, a motivating example is presented first that reveals fundamental characteristics among high-dimensional random vectors.

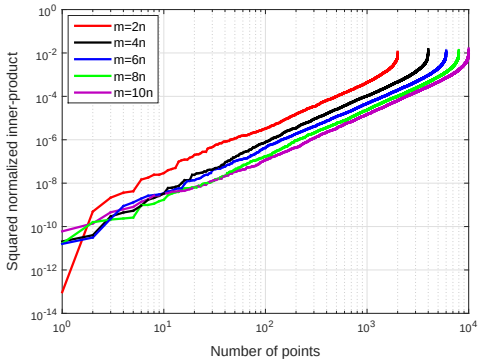

Figure 2: Ordered squared normalized inner-product for pairs $\boldsymbol{x}$ and $\boldsymbol{a}_i$, $\forall i \in [m]$ with m/n varying by 2 from 2 to 10, and $n = 10^3$.

Example: Fixing any nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$, generate data $\psi_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|$ using i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $\forall i \in [m]$, and evaluate the squared normalized inner-product

$$\cos^2\theta_i := \frac{|\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2}{\|\boldsymbol{a}_i\|^2\|\boldsymbol{x}\|^2} = \frac{\psi_i^2}{\|\boldsymbol{a}_i\|^2\|\boldsymbol{x}\|^2}, \ \forall i \in [m] \tag{11}$$

where θ_i is the angle between $\boldsymbol{a}_i$ and $\boldsymbol{x}$. Consider ordering all $\cos^2\theta_i$'s in an ascending fashion, and collectively denote them as $\boldsymbol{\xi} := \big[\cos^2\theta_{[m]} \ \cdots \ \cos^2\theta_{[1]}\big]^{\mathcal{T}}$ with $\cos^2\theta_{[1]} \ge \cdots \ge \cos^2\theta_{[m]}$. Fig. 2 plots the ordered entries in $\boldsymbol{\xi}$ for m/n varying by 2 from 2 to 10 with $n = 10^3$. Observe that almost all $\{\boldsymbol{a}_i\}$ vectors have a squared normalized inner-product smaller than 10^{-2}, while half of the inner-products are less than 10^{-3}, which implies that $\boldsymbol{x}$ is nearly orthogonal to many $\boldsymbol{a}_i$'s.

This example corroborates that random vectors in high-dimensional spaces are almost always nearly orthogonal to each other [24]. This inspired us to pursue an *orthogonality-promoting initialization method*. Our key idea is to approximate $\boldsymbol{x}$ by a vector that is most orthogonal to a subset of vectors $\{\boldsymbol{a}_i\}_{i\in\mathcal{I}_0}$, where $\mathcal{I}_0$ is a set with cardinality $|\mathcal{I}_0| < m$ that includes indices of the smallest squared normalized inner-products $\{\cos^2\theta_i\}$. Since $\|\boldsymbol{x}\|$ appears in all inner-products, its exact value does not influence their ordering. Henceforth, we assume without loss of generality that $\|\boldsymbol{x}\| = 1$.

Using $\{(\boldsymbol{a}_i; \psi_i)\}$, evaluate $\cos^2\theta_i$ according to (11) for each pair $\boldsymbol{x}$ and $\boldsymbol{a}_i$. Instrumental for the ensuing derivations is noticing that the summation of $\cos^2\theta_i$ over indices $i \in \mathcal{I}_0$ is very small, while rigorous justification is deferred to Section 3 and supplementary materials. Thus, a meaningful approximation denoted by $\boldsymbol{z}_0 \in \mathbb{R}^n$ can be obtained by solving

$$\min_{\|\boldsymbol{z}\|=1} \; \boldsymbol{z}^{\mathcal{T}} \left(\frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} \right) \boldsymbol{z} \tag{12}$$

which amounts to finding the smallest eigenvalue and the associated eigenvector of $\frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}$. Yet finding the smallest eigenvalue calls for eigen-decomposition or matrix inversion, each requiring computational complexity $\mathcal{O}(n^3)$. Such a computational burden can be intractable when n grows large. Applying a standard concentration result simplifies greatly those computations next [25].

Since $\boldsymbol{a}_i/\|\boldsymbol{a}_i\|$ has unit norm and is uniformly distributed on the unit sphere, it is uniformly spherically distributed.[2] Spherical symmetry implies that $\boldsymbol{a}_i/\|\boldsymbol{a}_i\|$ has zero mean and covariance matrix $\boldsymbol{I}_n/n$ [25]. Appealing again to the SLLN, the sample covariance matrix $\frac{1}{m} \sum_{i=1}^{m} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}$ approaches $\frac{1}{n} \boldsymbol{I}_n$ as m grows. Simple derivations lead to $\sum_{i \in \mathcal{I}_0} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} = \sum_{i=1}^{m} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} - \sum_{i \in \bar{\mathcal{I}}_0} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} \approxeq \frac{m}{n} \boldsymbol{I}_n - \sum_{i \in \bar{\mathcal{I}}_0} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}$, where $\bar{\mathcal{I}}_0$ is the complement of $\mathcal{I}_0$ in the set $[m]$.

Define $\boldsymbol{S} := [\boldsymbol{a}_1/\|\boldsymbol{a}_1\| \; \cdots \; \boldsymbol{a}_m/\|\boldsymbol{a}_m\|]^{\mathcal{T}} \in \mathbb{R}^{m \times n}$, and form $\bar{\boldsymbol{S}}_0$ by removing the rows of $\boldsymbol{S}$ if their indices do not belong to $\bar{\mathcal{I}}_0$. The task of seeking the smallest eigenvalue of $\boldsymbol{Y}_0 := \frac{1}{|\mathcal{I}_0|} \boldsymbol{S}_0^{\mathcal{T}} \boldsymbol{S}_0$ reduces to computing the largest eigenvalue of $\bar{\boldsymbol{Y}}_0 := \frac{1}{|\bar{\mathcal{I}}_0|} \bar{\boldsymbol{S}}_0^{\mathcal{T}} \bar{\boldsymbol{S}}_0$, namely,

$$\tilde{\boldsymbol{z}}_0 := \arg \max_{\|\boldsymbol{z}\|=1} \; \boldsymbol{z}^{\mathcal{T}} \bar{\boldsymbol{Y}}_0 \boldsymbol{z} \tag{13}$$

which can be efficiently solved using simple power iterations. If, on the other hand, $\|\boldsymbol{x}\| \neq 1$, the estimate $\tilde{\boldsymbol{z}}_0$ from (13) is further scaled so that its norm matches approximately that of $\boldsymbol{x}$ (which is estimated to be $\frac{1}{m} \sum_{i=1}^{m} y_i$), thus yielding $\boldsymbol{z}_0 = \sqrt{\sum_{i=1}^{m} y_i / m} \tilde{\boldsymbol{z}}_0$. It is worth stressing that the constructed matrix $\bar{\boldsymbol{Y}}_0$ does not depend on $\{y_i\}$ explicitly, saving our initialization from suffering heavy-tails of the fourth order of $\{\boldsymbol{a}_i\}$ in spectral initialization schemes.

Fig. 3 compares three initialization schemes showing their relative errors versus the measurement/unknown ratio m/n under the noise-free real Gaussian model, where $\boldsymbol{x} \in \mathbb{R}^{1,000}$ and m/n increases by 2 from 2 to 20. Apparently, all schemes enjoy improved performance as m/n increases. In particular, the proposed initialization method outperforms its spectral alternatives. Interestingly, the spectral and truncated spectral schemes exhibit similar performance when m/n is sufficiently large (e.g., $m/n \geq 14$). This confirms that truncation helps only if m/n is relatively small. Indeed, truncation is effected by discarding measurements of excessively large sizes emerging from the heavy tails of the data distribution. Hence, its advantage over the untruncated one narrows as the number of measurements increases, thus straightening out the heavy tails. On the contrary, the orthogonality-promoting initialization method achieves consistently superior performance over its spectral alternatives.

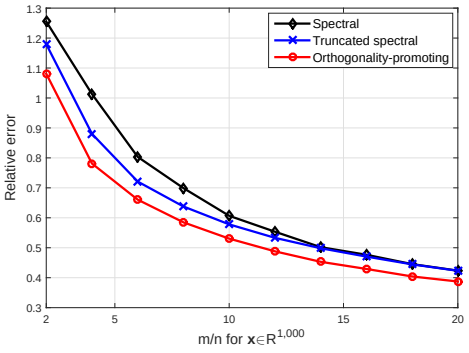

Figure 3: Relative error versus m/n for: i) the spectral method; ii) the truncated spectral method; and iii) our orthogonality-promoting method for noiseless real Gaussian model.

3 Main results

TGGF is summarized in Algorithm 1 with default values set for pertinent algorithmic parameters. Postulating independent samples $\{(\boldsymbol{a}_i; \psi_i)\}$, the following result establishes the performance of our TGGF approach.

[2] A random vector $\boldsymbol{z} \in \mathbb{R}^n$ is said to be spherical (or spherically symmetric) if its distribution does not change under rotations of the coordinate system; that is, the distribution of $\boldsymbol{P}\boldsymbol{z}$ coincides with that of $\boldsymbol{z}$ for any given orthogonal $n \times n$ matrix $\boldsymbol{P}$.

Algorithm 1 Truncated generalized gradient flow (TGGF) solvers

1: **Input:** Data $\{\psi_i\}_{i=1}^m$ and feature vectors $\{\boldsymbol{a}_i\}_{i=1}^m$; the maximum number of iterations $T = 1,000$; by default, take constant step size $\mu = 0.6/1$ for real/complex Gaussian models, truncation thresholds $|\overline{\mathcal{I}}_0| = \lceil \frac{1}{6}m \rceil$ ($\lceil \cdot \rceil$ the ceil operation), and $\gamma = 0.7$.

2: **Evaluate** $\psi_i/\|\boldsymbol{a}_i\|$, $\forall i \in [m]$, and find $\overline{\mathcal{I}}_0$ comprising indices corresponding to the $|\overline{\mathcal{I}}_0|$ largest $(\psi_i/\|\boldsymbol{a}_i\|)$'s.

3: **Initialize** $\boldsymbol{z}_0$ to $\sqrt{\sum_{i=1}^m \psi_i^2/m}\tilde{\boldsymbol{z}}_0$, where $\tilde{\boldsymbol{z}}_0$ is the unit leading eigenvector of $\overline{\boldsymbol{Y}}_0 := \frac{1}{|\overline{\mathcal{I}}_0|}\sum_{i\in\overline{\mathcal{I}}_0} \frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}$.

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

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

where $\mathcal{I}_{t+1} := \big\{1 \le i \le m \big| |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t| \ge \frac{1}{1+\gamma}\psi_i\big\}$.

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

Theorem 1 *Let $\boldsymbol{x} \in \mathbb{R}^n$ be an arbitrary signal vector, and consider (noise-free) measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$, in which $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le i \le m$. Then with probability at least $1 - (m+5)\mathrm{e}^{-n/2} - \mathrm{e}^{-c_0 m} - 1/n^2$ for some universal constant $c_0 > 0$, the initialization $\boldsymbol{z}_0$ returned by the orthogonality-promoting method in Algorithm 1 satisfies*

$$\mathrm{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \le \rho\,\|\boldsymbol{x}\| \tag{14}$$

with $\rho = 1/10$ (or any sufficiently small positive constant), provided that $m \ge c_1|\overline{\mathcal{I}}_0| \ge c_2 n$ for some numerical constants $c_1,\, c_2 > 0$, and sufficiently large n. Further, choosing a constant step size $\mu \le \mu_0$ along with a fixed truncation level $\gamma \ge 1/2$, and starting from any initial guess $\boldsymbol{z}_0$ satisfying (14)*, successive estimates of the TGGF solver (tabulated in Algorithm 1) obey*

$$\mathrm{dist}\left(\boldsymbol{z}_t, \boldsymbol{x}\right) \le \rho\left(1-\nu\right)^t \|\boldsymbol{x}\|, \quad t = 0,\, 1,\, \ldots \tag{15}$$

for some $0 < \nu < 1$, which holds with probability exceeding $1 - (m+5)\mathrm{e}^{-n/2} - 8\mathrm{e}^{-c_0 m} - 1/n^2$. Typical parameters are $\mu = 0.6$, and $\gamma = 0.7$.

Theorem 1 asserts that: i) TGGF recovers the solution $\boldsymbol{x}$ exactly as soon as the number of equations is about the number of unknowns, which is theoretically order optimal. Our numerical tests demonstrate that for the real Gaussian model, TGGF achieves a success rate of 100% when m/n is as small as 3, which is slightly larger than the information limit of $m/n = 2$ (Recall that $m \ge 2n-1$ is necessary for a unique solution); this is a significant reduction in the sample complexity ratio, which is 5 for TWF and 7 for WF. Surprisingly, TGGF enjoys also a success rate of over 50% when m/n is 2, which has not yet been presented for any existing algorithm under Gaussian sampling models and thus, our TGGF bridges the gap; see further discussion in Section 4; and, ii) TGGF converges exponentially fast. Specifically, TGGF requires at most $\mathcal{O}(\log(1/\epsilon))$ iterations to achieve any given solution accuracy $\epsilon > 0$ (a.k.a., $\mathrm{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \le \epsilon\,\|\boldsymbol{x}\|$), with iteration cost $\mathcal{O}(mn)$. Since truncation takes time on the order of $\mathcal{O}(m)$, the computational burden of TGGF per iteration is dominated by evaluating the generalized gradients. The latter involves two matrix-vector multiplications that are computable in $\mathcal{O}(mn)$ flops, namely, $\boldsymbol{A}\boldsymbol{z}_t$ yields $\boldsymbol{u}_t$, and $\boldsymbol{A}^{\mathcal{T}}\boldsymbol{v}_t$ the generalized gradient, where $\boldsymbol{v}_t := \boldsymbol{u}_t - \boldsymbol{\psi} \odot \frac{\boldsymbol{u}_t}{|\boldsymbol{u}_t|}$. Hence, the total running time of TGGF is $\mathcal{O}(mn\log(1/\epsilon))$, which is proportional to the time taken to read the data $\mathcal{O}(mn)$. The proof of Theorem 1 can be found in the supplementary material.

4 Simulated tests and conclusions

Additional numerical tests evaluating performance of the proposed scheme relative to TWF/WF are presented in this section. For fairness, all pertinent algorithmic parameters involved in each scheme are set to their default values. The Matlab implementations of TGGF are available at `http://www.tc.umn.edu/~gangwang/TAF`. The initial estimate was found based on 50 power iterations, and was subsequently refined by $T = 10^3$ gradient-like iterations in each scheme. Left panel in Fig. 4 presents average relative error of three initialization methods on a series of noiseless/noisy *real* Gaussian problems with $m/n = 6$ fixed, and n varying from 500 to 10^4,

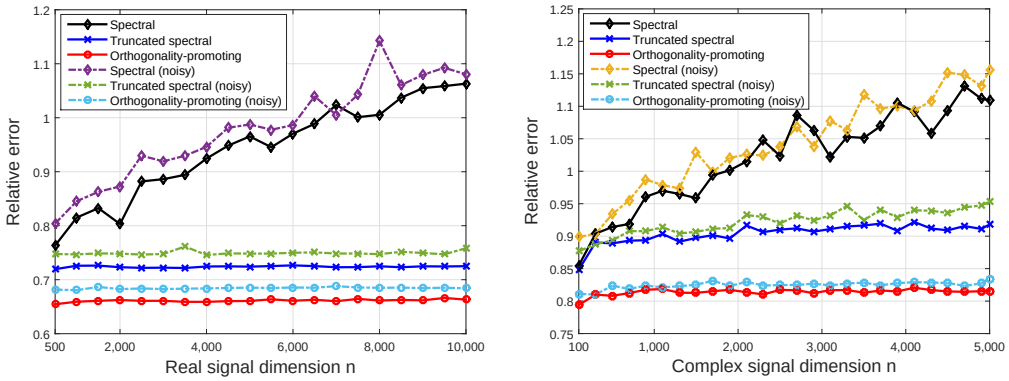

Figure 4: The average relative error using: i) the spectral method [11, 12]; ii) the truncated spectral method [8]; and iii) the proposed orthogonality-promoting method on noise-free (solid) and noisy (dotted) instances with $m/n = 6$, and n varying from $500/100$ to $10,000/5,000$ for real/complex vectors. Left: Real Gaussian model with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and $\sigma^2 = 0.2^2 \|\boldsymbol{x}\|^2$. Right: Complex Gaussian model with $\boldsymbol{x} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$, $\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$, and $\sigma^2 = 0.2^2 \|\boldsymbol{x}\|^2$.

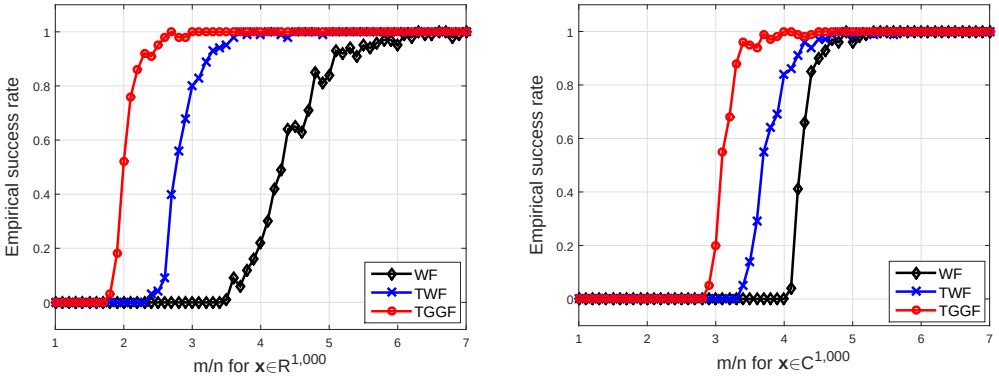

Figure 5: Empirical success rate for WF, TWF, and TGGF with $n = 1,000$ and m/n varying from 1 to 7. Left: Noiseless real Gaussian model with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$; Right: Noiseless complex Gaussian model with $\boldsymbol{x} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$.

while those for the corresponding *complex* Gaussian instances are shown in the right panel. Fig. 5 compares empirical success rate of three schemes under both real and complex Gaussian models with $n = 10^3$ and m/n varying by 1 from 1 to 7. Apparently, the proposed initialization method returns more accurate and robust estimates than the spectral ones. Moreover, for real-valued vectors, TGGF achieves a success rate of over 50% when $m/n = 2$, and guarantees perfect recovery from about $3n$ measurements; while for complex-valued ones, TGGF enjoys a success rate of 95% when $m/n = 3.4$, and ensures perfect recovery from about $4.5n$ measurements. Regarding running times, TGGF converges slightly faster than TWF, while both are markedly faster than WF. Curves in Fig. 5 clearly corroborate the merits of TGGF over Wirtinger alternatives.

This paper developed a linear-time algorithm termed TGGF for solving random systems of quadratic equations. TGGF builds on three key ingredients: a novel orthogonality-promoting initialization, along with a simple yet effective truncation rule, as well as simple scalable gradient-like iterations. Numerical tests corroborate the superior performance of TGGF over state-of-the-art solvers.

Acknowledgements

Work in this paper was supported in part by NSF grants 1500713 and 1514056.

References

[1] 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.

[2] A. Conca, D. Edidin, M. Hering, and C. Vinzant, "An algebraic characterization of injectivity in phase retrieval," *Appl. Comput. Harmon. Anal.*, vol. 38, no. 2, pp. 346–356, Mar. 2015.

[3] 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.

[4] H. A. Hauptman, "The phase problem of X-ray crystallography," *Rep. Prog. Phys.*, vol. 54, no. 11, p. 1427, 1991.

[5] 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.

[6] H. Sahinoglou and S. D. Cabrera, "On phase retrieval of finite-length sequences using the initial time sample," *IEEE Trans. Circuits and Syst.*, vol. 38, no. 8, pp. 954–958, Aug. 1991.

[7] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Math. Prog.*, vol. 39, no. 2, pp. 117–129, 1987.

[8] 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).

[9] 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.

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

[11] 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.

[12] 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.

[13] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv:1602.06664*, 2016.

[14] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Appl. Comput. Harmon. Anal.*, vol. 66, no. 8, pp. 1241–1274, Nov. 2013.

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

[16] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 14, no. 5, pp. 1017–1026, 2014.

[17] G. Wang, D. Berberidis, V. Kekatos, and G. B. Giannakis, "Online reconstruction from big data via compressive censoring," in *IEEE Global Conf. Signal and Inf. Process.*, Atlanta, GA, 2014, pp. 326–330.

[18] D. K. Berberidis, V. Kekatos, G. Wang, and G. B. Giannakis, "Adaptive censoring for large-scale regressions," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process.*, South Brisbane, QLD, Australia, 2015, pp. 5475–5479.

[19] L.-H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller, "Experimental robustness of Fourier ptychography phase retrieval algorithms," *Opt. Express*, vol. 23, no. 26, pp. 33 214–33 240, Dec. 2015.

[20] N. Z. Shor, K. C. Kiwiel, and A. Ruszcayǹski, *Minimization Methods for Non-differentiable Functions*. Springer-Verlag New York, Inc., 1985.

[21] F. H. Clarke, *Optimization and Nonsmooth Analysis*. SIAM, 1990, vol. 5.

[22] ——, "Generalized gradients and applications," *T. Am. Math. Soc.*, vol. 205, pp. 247–262, 1975.

[23] P. Chen, A. Fannjiang, and G.-R. Liu, "Phase retrieval with one or two diffraction patterns by alternating projections of the null vector," *arXiv:1510.07379v2*, 2015.

[24] T. Cai, J. Fan, and T. Jiang, "Distributions of angles in random packing on spheres," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1837–1864, Jan. 2013.

[25] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv:1011.3027*, 2010.