Journal of Machine Learning Research 18 (2017) 1-35 Submitted 11/16; Revised 10/17; Published 11/17

A Nonconvex Approach for Phase Retrieval: Reshaped Wirtinger Flow and Incremental Algorithms

Huishuai Zhang HZHAN23@SYR.EDU
Department of EECS
Syracuse University
Syracuse, NY 13244 USA

Yi Zhou ZHOU.1172@OSU.EDU
Yingbin Liang LIANG.889@OSU.EDU
Yuejie Chi CHI.97@OSU.EDU
Department of ECE
The Ohio State University
Columbus, OH 43210 USA

Editor: Benjamin Recht

Abstract

We study the problem of solving a quadratic system of equations, i.e., recovering a vector signal $\boldsymbol{x} \in \mathbb{R}^n$ from its magnitude measurements $y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|, i = 1, ..., m$. We develop a gradient descent algorithm (referred to as RWF for reshaped Wirtinger flow) by minimizing the quadratic loss of the magnitude measurements. Comparing with Wirtinger flow (WF) (Candès et al., 2015), the loss function of RWF is nonconvex and nonsmooth, but better resembles the least-squares loss when the phase information is also available. We show that for random Gaussian measurements, RWF enjoys linear convergence to the true signal as long as the number of measurements is $\mathcal{O}(n)$. This improves the sample complexity of WF ($\mathcal{O}(n \log n)$), and achieves the same sample complexity as truncated Wirtinger flow (TWF) (Chen and Candès, 2015), but without any sophisticated truncation in the gradient loop. Furthermore, RWF costs less computationally than WF, and runs faster numerically than both WF and TWF. We further develop an incremental (stochastic) version of RWF (IRWF) and connect it with the randomized Kaczmarz method for phase retrieval. We demonstrate that IRWF outperforms existing incremental as well as batch algorithms with experiments.

Keywords: gradient descent, phase retrieval, nonconvex optimization, regularity condition, stochastic algorithms

1. Introduction

Many problems in machine learning and signal processing can be reduced to solving a quadratic system of equations. For instance, in phase retrieval applications, i.e., X-ray crystallography and coherent diffraction imaging (Drenth, 2007; Miao et al., 1999, 2008), the structure of an object is to be recovered from the measured far field diffracted intensity when the object is illuminated by a source light. Mathematically, such a problem amounts to recovering the signal from only the magnitudes of its linear measurements. Specifically, the problem is formulated as follows.

Problem 1 *Recover $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ from the measurements given as*

$$y_i = |\langle \boldsymbol{a}_i, \mathbf{x} \rangle|, \quad \text{for } i = 1, \cdots, m, \tag{1}$$

where $\boldsymbol{a}_i \in \mathbb{R}^n/\mathbb{C}^n$ are known design vectors.

Various algorithms have been proposed to solve this problem since the 1970s. The error-reduction methods (Gerchberg, 1972; Fienup, 1982) work well empirically but lack theoretical guarantees. More recently, convex relaxation of the problem has been formulated, for example, via PhaseLift (Candès et al., 2013; Candès and Li, 2014; Chen et al., 2015) and PhaseCut (Waldspurger et al., 2015), and the correspondingly developed algorithms typically come with performance guarantees. The readers can refer to the review paper (Shechtman et al., 2015) to learn more about applications and algorithms of the phase retrieval problem.

Nonetheless, these convex approaches often suffer from high computational complexity particularly when the signal dimension is large. It is therefore desirable to develop more efficient nonconvex approaches that can provably recover the true signal. Netrapalli et al. (2013) proposed the *AltMinPhase* algorithm, which alternates between updates of the phase and the signal, and showed that it converges linearly and recovers the true signal with $\mathcal{O}(n \log^3 n)$ Gaussian measurements, when $\boldsymbol{a}_i$'s are composed of *independent and identically distributed* (i.i.d.) standard Gaussian entries. More recently, Candès et al. (2015) introduced the *Wirtinger flow* (WF) algorithm, which guarantees signal recovery via a simple gradient descent algorithm with only $\mathcal{O}(n \log n)$ Gaussian measurements and attains ϵ-accuracy within $\mathcal{O}(mn^2 \log 1/\epsilon)$ flops[1]. More specifically, WF obtains a good initialization by the spectral method, and then minimizes the following nonconvex loss function based on the quadratic loss of the squared magnitude measurements

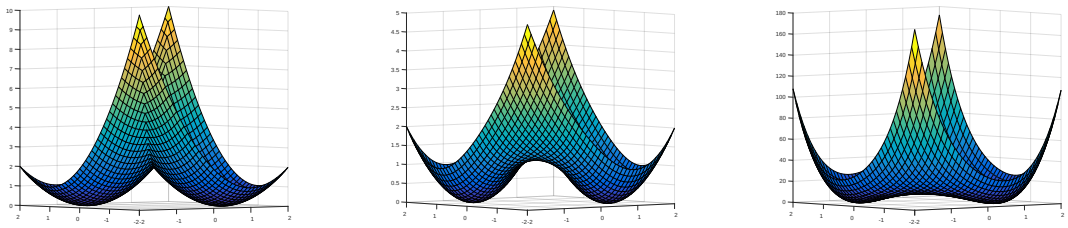
$$\ell_{WF}(\boldsymbol{z}) := \frac{1}{4m} \sum_{i=1}^{m} (|\boldsymbol{a}_i^T \boldsymbol{z}|^2 - |y_i|^2)^2, \tag{2}$$

via the gradient descent scheme.

WF is further improved by *truncated Wirtinger flow* (TWF) (Chen and Candès, 2015), which adopts a Poisson loss function of $|\boldsymbol{a}_i^T \boldsymbol{z}|^2$, and keeps only well-behaved measurements based on carefully designed truncation thresholds for calculating the initialization and the gradient updates. Such truncation assists to achieve a linear convergence rate using a fixed step size; as a consequence, TWF reduces both the sample complexity to $\mathcal{O}(n)$ and the computational complexity to $\mathcal{O}(mn \log 1/\epsilon)$.

Furthermore, incremental/stochastic methods have been proposed to solve Problem 1. Specifically, the randomized Kaczmarz method has been adopted to solve the phase retrieval problem (Wei, 2015; Li et al., 2015), which was shown to have excellent empirical performance, but no nonasymptotic global convergence guarantee was established. Incremental truncated Wirtinger flow (ITWF) (Kolte and Özgür, 2016) is another stochastic algorithm developed based on TWF.

1. WF has computational cost $\mathcal{O}(mn^2 \log(1/\epsilon)$ (Candès et al., 2015, Theorem 3.1) because it needs $\log(1/\epsilon) \cdot \frac{1}{-\log(1-c/n)} \approx \frac{n}{c} \log(1/\epsilon)$ iterations to achieve ϵ-accuracy due to the convergence rate $\left(1 - \frac{c}{n}\right)^k$ and each iteration needs $\mathcal{O}(mn)$ flops. Contrastingly, RWF has computational cost $\mathcal{O}(mn \log 1/\epsilon)$ because it requires only $c \log(1/\epsilon)$ iterations to achieve ϵ-accuracy due to the convergence rate $(1-c)^k$.



(a) Expected loss of LS (b) Expected loss of RWF (c) Expected loss of WF

Figure 1: Surface of the expected loss function of (a) least-squares (mirrored symmetrically), (b) RWF, and (c) WF when $\boldsymbol{x} = [1, -1]^T$.

1.1 Our Contribution

This paper adopts the following loss function

$$\ell(\boldsymbol{z}) := \frac{1}{2m} \sum_{i=1}^{m} \left(|\boldsymbol{a}_i^T \boldsymbol{z}| - y_i \right)^2, \tag{3}$$

which is the quadratic loss of the magnitude measurements. Compared to the loss function (2) of WF that adopts the quadratic loss of $|\boldsymbol{a}_i^T \boldsymbol{z}|^2$, the above loss function adopts the quadratic loss of $|\boldsymbol{a}_i^T \boldsymbol{z}|$ and hence has lower-order variables. While both loss functions are nonconvex in $\boldsymbol{z}$, the WF loss function (2) is a fourth-order smooth function of $\boldsymbol{a}_i^T \boldsymbol{z}$ and our loss function (3) in contrast is nonsmooth.

To minimize such a nonconvex and nonsmooth loss function (3), we develop a gradient descent algorithm, which sets the "gradient" to zero corresponding to nonsmooth samples. We refer to such an algorithm together with an initialization using a new spectral method (different from that employed in TWF or WF) as *reshaped Wirtinger flow* (RWF). We show that the new loss function has great advantage in both statistical and computational efficiency, despite nonsmoothness. In fact, the curvature of such a loss function behaves similarly to that of a least-squares problem with phase information in the neighborhood of global optimizers, and hence yields faster convergence. To provide further insights, consider the standard problem of solving $\boldsymbol{x}$ from *linear measurements* $\langle \boldsymbol{a}_i, \mathbf{x} \rangle$, $i = 1, \cdots, m$, where $\boldsymbol{a}_i$'s are composed of i.i.d. standard Gaussian entries. In this case, it is natural to use the least-squares loss function

$$\ell_{LS}(\boldsymbol{z}) := \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{z} - \boldsymbol{a}_i^T \boldsymbol{x} \right)^2. \tag{4}$$

Examining the expected (with respect to $\boldsymbol{a}_i$'s) loss surface of $\min\{\ell_{LS}(\boldsymbol{z}), \ell_{LS}(-\boldsymbol{z})\}$ (to mimic sign ambiguity), $\ell(\boldsymbol{z})$, and $\ell_{WF}(\boldsymbol{z})$ in Figure 1, whose expressions can be found in Appendix A, it can be seen that the loss of RWF, rather than the loss of WF, has a similar curvature to the quadratic least-squares loss around the global optimizers, which justifies its better performance than WF.

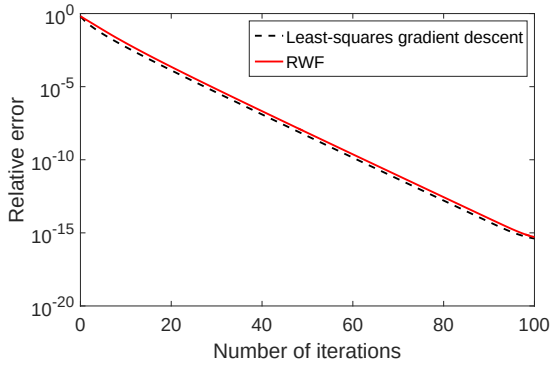


Figure 2: Comparison of convergence behavior between RWF and the least-squares gradient descent with the same initialization, the same parameters $n = 1000$, $m = 6n$, and the same step size $\mu = 0.8$.

The nonsmoothness of the loss function (3) does not negatively impact the performance of RWF because only with negligible probability the algorithm encounters nonsmooth points for some samples, which furthermore are set not to contribute to the gradient direction by RWF. The gradient of the RWF loss (3) is given as

$$\nabla \ell(\boldsymbol{z}) := \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{z} - y_i \cdot \mathrm{sgn}(\boldsymbol{a}_i^T \boldsymbol{z}) \right) \boldsymbol{a}_i, \tag{5}$$

where $\mathrm{sgn}(0) = 0$ by convention. Comparing this with the gradient of the least-squares loss

$$\nabla \ell_{LS}(\boldsymbol{z}) = \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{z} - \boldsymbol{a}_i^T \boldsymbol{x} \right) \boldsymbol{a}_i, \tag{6}$$

one can see that RWF uses estimated phase information $\mathrm{sgn}(\boldsymbol{a}_i^T \boldsymbol{z})$ to generate the gradient updates, and the convergence behavior of RWF is much similar to that of least-squares with phase information if initialized properly. Indeed, Figure 2 illustrates that RWF takes almost the *same* number of iterations for recovering a signal (with only the magnitude information) as the least-squares gradient descent method for recovering a signal (with both the magnitude and the sign information).

We further develop incremental/stochastic versions of RWF using mini-batches of measurements, called incremental RWF (IRWF), and show that IRWF also enjoys the advantageous local curvature of RWF, and achieves excellent statistical and computational performance. Along the way, we establish the performance guarantee of the Kaczmarz-PR update rule by interpreting it as a variant of IRWF. We conduct extensive numerical experiments to demonstrate that IRWF performs better than other competitive incremental algorithms (ITWF and Kaczmarz-PR) as well as batch algorithms (RWF, TWF, WF and AltMinPhase).

We summarize our main results as follows.

- Statistically, we show that RWF recovers the true signal with $\mathcal{O}(n)$ Gaussian measurements, which is order-wise optimal. Thus, RWF improves the sample complexity

$\mathcal{O}(n \log n)$ of WF, and achieves the same sample complexity as TWF but without truncation in the gradient descent step.

- Computationally, RWF converges linearly to the true signal with a constant step size, requiring $\mathcal{O}(mn \log(1/\epsilon))$ flops to reach ϵ-accuracy. Again, without truncation in the gradient descent step, RWF improves the computational cost $\mathcal{O}(mn^2 \log(1/\epsilon))$ of WF and achieves the same computational cost as TWF.

- We also show that RWF is robust to bounded additive noise. The estimation error is shown to diminish at a linear rate with respect to the noise level up to a scaling difference. Experiments on Poisson noise further corroborate the stability guarantee.

- The incremental versions of RWF (IRWF) is shown to shrink the estimation error in expectation for one update, which implies the linear convergence of the algorithm if the iteration path lies in the neighborhood of the true signal. Furthermore, we show that Kaczmarz-PR can be viewed as IRWF under a specific choice of step size, and establish a similar convergence result for Kaczmarz-PR under the same sample complexity.

- Numerically, RWF and its incremental versions require fewer parameters, e.g., truncation thresholds, than TWF in practice. RWF is generally two times faster than TWF and four to six times faster than WF in terms of both the number of iterations and time cost. IRWF also outperforms existing incremental as well as batch algorithms.

Compared to WF and TWF, the nonsmoothness in the loss function requires new bounding techniques in order to establish the theoretical guarantees. On the other hand, our technical proof is much simpler, because the lower-order loss function allows to bypass higher-order moments of variables as well as truncation of samples in the gradient descent steps.

1.2 Related Work

The quadratic loss function of magnitudes (3) was also used in the early literature of phase retrieval (Fienup, 1982) with Fourier magnitude measurements. However, no global convergence guarantee was available in (Fienup, 1982). Along the line of developing nonconvex algorithms with global performance guarantees for the phase retrieval problem, Netrapalli et al. (2013) and Waldspurger (2016) developed alternating minimization algorithms, Candès et al. (2015); Chen and Candès (2015); Zhang et al. (2016); Cai et al. (2016) developed first-order algorithms, and a recent study Sun et al. (2016) characterized the geometric structure of the nonconvex objective and designed a second-order trust-region algorithm. This paper is closely related to (Candès et al., 2015; Chen and Candès, 2015; Zhang et al., 2016), but develops a new gradient descent algorithm based on a lower-order nonsmooth (as well as nonconvex) loss function that yields advantageous statistical and computational efficiency.

Stochastic algorithms were also developed for the phase retrieval problem. Kolte and Özgür (2016) studied the incremental truncated Wirtinger flow (ITWF) and showed that ITWF needs much fewer passes of data than TWF to reach the same accuracy. Wei (2015) adapted the Kaczmarz method to solve the phase retrieval problem and demonstrated its

fast empirical convergence. We show that IRWF is closely related to Kaczmarz-PR, and empirically runs faster than ITWF thanks to the advantageous curvature of the loss function.

After our work was posted on arXiv, an independent work (Wang et al., 2016) was subsequently posted, which also adopts the same loss function but develops a slightly different algorithm called TAF (i.e., truncated amplitude flow). One major difference is that RWF does not require truncation in the gradient loops while TAF still employs truncation. Hence, RWF has fewer parameters to tune, and is easier to implement than TAF in practice. Furthermore, RWF demonstrates the performance advantage of adopting a lower-order loss function even without truncation, which cannot be observed from TAF. The two algorithms also employ different initialization strategies. Moreover, we analyze stochastic algorithms based on the new loss function while Wang et al. (2016) do not.

More generally, various high-dimensional signal estimation problems have been studied by minimizing nonconvex loss functions. For example, a partial list of these studies include matrix completion (Keshavan et al., 2010; Jain et al., 2013; Sun and Luo, 2016; Hardt, 2014; Sa et al., 2015; Zheng and Lafferty, 2016; Jin et al., 2016; Ge et al., 2016), low-rank matrix recovery (Bhojanapalli et al., 2016; Chen and Wainwright, 2015; Tu et al., 2015; Zheng and Lafferty, 2015; Park et al., 2016; Wei et al., 2016; Li et al., 2017), robust PCA (Netrapalli et al., 2014), robust tensor decomposition (Anandkumar et al., 2016), dictionary learning (Arora et al., 2015; Sun et al., 2015), community detection (Bandeira et al., 2016), phase synchronization (Boumal, 2016), blind deconvolution (Lee et al., 2016b; Li et al., 2016), etc.

For minimizing a general nonconvex nonsmooth objective, various algorithms have been proposed, such as gradient sampling (Burke et al., 2005; Kiwiel, 2007) and majorization-minimization (Ochs et al., 2015). These algorithms are often shown to converge to critical points which may be local minimizers or saddle points, without an explicit characterization of convergence rates. In contrast, our algorithm is specifically designed for the phase retrieval problem, and can be shown to converge linearly to the global optimum under an appropriate initialization.

The advantage of using nonsmooth loss functions in our study is analogous in spirit to that of the rectifier activation function (of the form $\max\{0, \cdot\}$) in neural networks. It has been shown that rectified linear unit ($ReLU$) enjoys superb advantage in reducing the training time (Krizhevsky et al., 2012) and promoting sparsity (Glorot et al., 2011) over its counterparts of sigmoid and hyperbolic tangent functions, in spite of non-linearity and non-differentiability at zero. Our results in fact also demonstrate that a nonsmooth but simpler loss function yields improved performance.

1.3 Paper Organization and Notations

The rest of this paper is organized as follows. Section 2 describes the RWF algorithm in detail and establishes its performance guarantee. Section 3 introduces the IRWF algorithm, establishes its performance guarantee and compares it with existing stochastic algorithms. Section 4 compares RWF and IRWF with other competitive algorithms numerically. Finally, Section 5 concludes the paper with comments on future directions.

Throughout the paper, boldface lowercase letters such as $\boldsymbol{a}_i, \boldsymbol{x}, \boldsymbol{z}$ denote vectors, and boldface capital letters such as $\boldsymbol{A}, \boldsymbol{Y}$ denote matrices. For two matrices, $\boldsymbol{A} \prec \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive definite. For a complex matrix and a vector, $\boldsymbol{A}^*$ and $\boldsymbol{z}^*$ denote conjugate

transposes of $\boldsymbol{A}$ and $\boldsymbol{z}$, respectively. For a real matrix and a vector, $\boldsymbol{A}^T$ and $\boldsymbol{z}^T$ denote transposes of $\boldsymbol{A}$ and $\boldsymbol{z}$, respectively. We let $\odot$ denotes element-wise product. The indicator function $\mathbf{1}_A = 1$ if the event A occurs, and $\mathbf{1}_A = 0$ otherwise. We let $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|$ denote the l_1 norm and l_2 norm of a vector $\boldsymbol{x}$, respectively. Moreover, let $\|\boldsymbol{A}\|_F$ and $\|\boldsymbol{A}\|$ denote the Frobenius norm and the spectral norm of a matrix $\boldsymbol{A}$, respectively. We note that the constants c, C, c_0, c_1, c_2 may be different in different equations, for the sake of notational simplicity.

2. Reshaped Wirtinger Flow

Consider the Problem 1 for the complex case. It can be observed that if $\boldsymbol{z}$ is a solution, i.e., satisfying Equation (1), then $\boldsymbol{z}e^{-j\phi}$ is also the solution of the problem where ϕ is an arbitrary phase constant. Therefore, the recovery is up to a phase difference. Thus, we define the Euclidean distance between two complex vectors up to a global phase difference (Candès et al., 2015) as,

$$\operatorname{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min_{\phi \in [0, 2\pi)} \|\boldsymbol{z}e^{-j\phi} - \boldsymbol{x}\|, \tag{7}$$

which is simply $\min \|\boldsymbol{z} \pm \boldsymbol{x}\|$ for the real case.

In this paper, we focus on the real-valued case in our analysis, but the algorithm designed below is applicable to the complex-valued case and the case with other measurement vectors such as *coded diffraction pattern* (CDP) as we demonstrate via numerical experiments in Section 4.

We design RWF (see Algorithm 1) for solving the phase retrieval problem, which contains two stages: spectral initialization and gradient loop. The suggested values for the parameters[2] are given by $\alpha_l = 1, \alpha_u = 5$ and $\mu = 0.8$. The rescaling coefficient in λ_0 and the conjugate transpose $\boldsymbol{a}_i^*$ allow the algorithm readily applicable to the complex and CDP cases. We next describe the two stages of RWF in details in Sections 2.1 and 2.2, respectively, and establish the convergence guarantee in Section 2.3. Finally, we provide the stability guarantee of RWF in Section 2.4.

2.1 Initialization via the Spectral Method

When solving nonconvex problems by iterative algorithms, the starting point is critical. The spectral method is a popular choice in the literature (Keshavan et al., 2010; Netrapalli et al., 2013; Candès et al., 2015; Chen and Candès, 2015), which often provides a good initialization. Different from the spectral initialization used in AltMinPhase (Netrapalli et al., 2013), WF (Candès et al., 2015) and TWF (Chen and Candès, 2015), which are based on the squared magnitudes as the weight of each rank-one matrix $\boldsymbol{a}_i\boldsymbol{a}_i^*$, we propose an alternative initialization in Algorithm 1 that uses the magnitudes instead, and truncates samples that are either too large or too small. We show that such an initialization achieves a smaller sample complexity than WF and the same sample complexity as TWF order-wise, and furthermore, performs better than both WF and TWF numerically.

Our initialization consists of estimation of both the norm and the direction of $\boldsymbol{x}$. The norm estimation of $\boldsymbol{x}$ is given by λ_0 in Algorithm 1. Intuitively, with real-valued Gaussian

2. For the complex Gaussian case, we suggest $\mu = 1.2$.

Algorithm 1 Reshaped Wirtinger Flow

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$;

Parameters: Lower and upper thresholds α_l, α_u for truncation in initialization, step size μ;

Initialization: Let $\boldsymbol{z}^{(0)} = \lambda_0 \tilde{\boldsymbol{z}}$, where $\lambda_0 = \frac{mn}{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1} \cdot \left(\frac{1}{m}\sum_{i=1}^m y_i\right)$ and $\tilde{\boldsymbol{z}}$ is the leading eigenvector of

$$\boldsymbol{Y} := \frac{1}{m}\sum_{i=1}^m y_i \boldsymbol{a}_i \boldsymbol{a}_i^* \mathbf{1}_{\{\alpha_l \lambda_0 < y_i < \alpha_u \lambda_0\}}. \tag{8}$$

Gradient loop: for $t = 0 : T-1$ do

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^* \boldsymbol{z}^{(t)} - y_i \cdot \frac{\boldsymbol{a}_i^* \boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^* \boldsymbol{z}^{(t)}|}\right) \boldsymbol{a}_i. \tag{9}$$

Output $\boldsymbol{z}^{(T)}$.

measurement vectors (i.e., $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I}_{n\times n})$), the scaling coefficient $\frac{mn}{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1} \approx \sqrt{\frac{\pi}{2}}$. Moreover, for $i = 1, \ldots, m$, $y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|$ are independent sub-Gaussian random variables with mean $\sqrt{\frac{2}{\pi}}\|\boldsymbol{x}\|$, and thus $\frac{1}{m}\sum_{i=1}^m y_i \approx \sqrt{\frac{2}{\pi}}\|\boldsymbol{x}\|$. Combining these two facts yields the desired argument.

The direction of $\boldsymbol{x}$ is approximated by the leading eigenvector of $\boldsymbol{Y}$, because $\boldsymbol{Y}$ approaches $\mathrm{E}[\boldsymbol{Y}]$ by concentration of measure arguments, and the leading eigenvector of $\mathrm{E}[\boldsymbol{Y}]$ takes the form $c\boldsymbol{x}$ for some scalar $c \in \mathbb{R}$. We note that (8) involves truncation of samples from both sides, in contrast to truncation only by an upper threshold in TWF (Chen and Candès, 2015). The truncation parameters α_l and α_u are related to the eigenvalue gap between the top two eigenvalues of $\boldsymbol{Y}$, which determines the estimation accuracy of the eigenvector after running k iterations of the power method[3]. The larger the eigenvalue gap, the better the accuracy of the power method given a number of iterations. We note that $\alpha_l = 1, \alpha_u = \infty$ yield the largest eigenvalue gap from the developments in Appendix B. We explicitly set a bounded α_u for the development of the proof and for numerical stability.

We next provide the formal statement that with high probability, the proposed initialization lands in a small neighborhood around the true signal.

Proposition 1 *Fix $\delta > 0, \alpha_l = 1$ and $\alpha_u \gg 1$. The initialization step in Algorithm 1 yields $\boldsymbol{z}^{(0)}$ satisfying dist$(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \le \delta\|\boldsymbol{x}\|$ with the probability at least $1 - \exp(-cm\epsilon^2)$, if $m > C(\delta, \epsilon)n$, where c is some positive constant and C is a positive constant affected by δ and ϵ.*

Proof. See Appendix B. ■

Finally, we numerically compare different initialization methods in Figure 3. It is demonstrated that RWF achieves better initialization accuracy in terms of the relative error $\mathrm{dist}(\boldsymbol{z}^{(0)}, \boldsymbol{x})/\|\boldsymbol{x}\|$ than WF, TWF, as well as the initialization method proposed in TAF (Wang et al., 2016).

3. Numerical linear algebra shows that after k iterations of the power method, the estimation accuracy of the eigenvector is given by $\mathcal{O}((\lambda_2/\lambda_1)^k)$, where λ_1 and λ_2 are the top two eigenvalues.

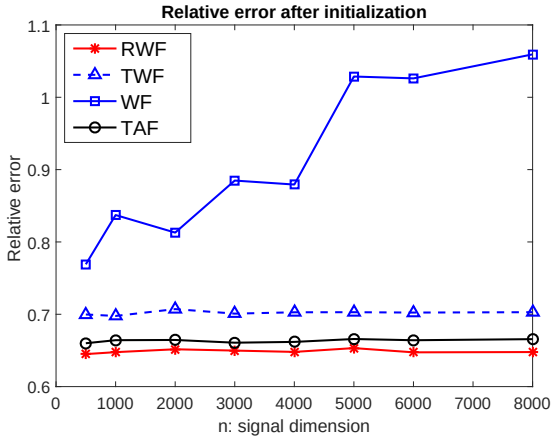


Figure 3: Comparison of different initialization methods with $m = 6n$ and under 50 iterations of power method.

2.2 Gradient Loop

The gradient loop of Algorithm 1 is based on the loss function (3), and the update rule (9) makes itself suitable for the complex case. For the real case the update direction is given as follows:

$$\nabla \ell(\boldsymbol{z}) := \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - y_i \cdot \mathrm{sgn}(\boldsymbol{a}_i^T\boldsymbol{z})\right)\boldsymbol{a}_i = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - y_i \cdot \frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i, \tag{10}$$

where $\mathrm{sgn}(\cdot)$ is the sign function for nonzero arguments. We further set $\mathrm{sgn}(0) = 0$ and $\frac{0}{|0|} = 0$. In fact, $\nabla \ell(\boldsymbol{z})$ equals the gradient of the loss function (3) over the samples that satisfy $\boldsymbol{a}_i^T\boldsymbol{z} \neq 0$. For samples corresponding to the nonsmooth point, i.e., $\boldsymbol{a}_i^T\boldsymbol{z} = 0$, we adopt the Fréchet superdifferential (Kruger, 2003) for nonconvex functions to set the gradient component to be zero (as zero is an element in the Fréchet superdifferential). With abuse of terminology, we still refer to $\nabla \ell(\boldsymbol{z})$ in Equation (10) as the "gradient" for simplicity, which rather represents the update direction in the gradient loop of Algorithm 1.

2.3 Linear Convergence of RWF

We characterize the convergence guarantee of RWF in the following theorem.

Theorem 2 *Consider the problem of solving any given $\boldsymbol{x} \in \mathbb{R}^n$ from a system of equations (1) with Gaussian measurement vectors. There exist some universal constants $\mu_0 > 0$ (μ_0 can be set as 0.8 in practice), $0 < \rho, \nu < 1$ and $c_0, c_1, c_2 > 0$ such that if $m \geq c_0 n$ and $\mu < \mu_0$, then with probability at least $1 - c_1\exp(-c_2 m)$, Algorithm 1 yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \leq \nu(1-\rho)^t\|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N}. \tag{11}$$

Proof We outline the proof here with the details delegated to Appendix C. Compared to WF and TWF, our proof is much simpler due to the lower-order loss function that RWF

relies on. We first introduce a global phase notation for the real case as follows:

$$\Phi_{\boldsymbol{x}}(\boldsymbol{z}) := \begin{cases} 0, & \text{if } \|\boldsymbol{z}-\boldsymbol{x}\| \le \|\boldsymbol{z}+\boldsymbol{x}\|, \\ \pi, & \text{otherwise.} \end{cases} \tag{12}$$

For the sake of simplicity, we let $\boldsymbol{z}$ be $e^{-j\Phi_{\boldsymbol{x}}(\boldsymbol{z})}\boldsymbol{z}$, which indicates that $\boldsymbol{z}$ is always in the neighborhood of $\boldsymbol{x}$.

Here, the central idea is to show that within the neighborhood of the global minimizer, RWF satisfies the *Regularity Condition* $\mathsf{RC}(\mu, \lambda, c)$ (Chen and Candès, 2015), i.e.,

$$\langle \nabla \ell(\boldsymbol{z}), \boldsymbol{z}-\boldsymbol{x} \rangle \ge \frac{\mu}{2} \|\nabla \ell(\boldsymbol{z})\|^2 + \frac{\lambda}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2 \tag{13}$$

for all $\boldsymbol{z}$ obeying $\|\boldsymbol{z}-\boldsymbol{x}\| \le c\|\boldsymbol{x}\|$, where $0 < c < 1$ is some small constant. Then, as shown in (Candès et al., 2015; Chen and Candès, 2015), once the initialization lands into this neighborhood, linear convergence can be guaranteed with a proper choice of the constant step size, i.e.,

$$\mathrm{dist}^2\left(\boldsymbol{z} - \mu\nabla \ell(\boldsymbol{z}), \boldsymbol{x}\right) \le (1-\mu\lambda)\mathrm{dist}^2(\boldsymbol{z}, \boldsymbol{x}), \tag{14}$$

for any $\boldsymbol{z}$ satisfying $\|\boldsymbol{z}-\boldsymbol{x}\| \le c\|\boldsymbol{x}\|$.

Lemmas 6 and 7 in Appendix C yield that for all $\boldsymbol{z}$ satisfying $\|\boldsymbol{z}-\boldsymbol{x}\| \le c\|\boldsymbol{x}\|$,

$$\langle \nabla \ell(\boldsymbol{z}), \boldsymbol{z}-\boldsymbol{x} \rangle \ge (1-0.26-3\epsilon)\|\boldsymbol{z}-\boldsymbol{x}\|^2 = (0.74-3\epsilon)\|\boldsymbol{z}-\boldsymbol{x}\|^2$$

with high probability. Moreover, Lemma 8 in Appendix C further yields that

$$\|\nabla \ell(\boldsymbol{z})\| \le (1+\delta)\cdot 2\|\boldsymbol{z}-\boldsymbol{x}\| \tag{15}$$

with high probability. Therefore, the above two bounds imply that the Regularity Condition (13) holds for μ and λ satisfying

$$0.74 - 3\epsilon \ge \frac{\mu}{2}\cdot 4(1+\delta)^2 + \frac{\lambda}{2} \tag{16}$$

for sufficiently small ϵ and δ. ■

We note that Equation (16) implies an upper bound for the step size $\mu \le \frac{0.74}{2} = 0.37$, by taking ϵ and δ to be sufficiently small. However, in practice, the step size μ can be set much larger than such a bound, say 0.8, while still keeping the algorithm convergent. This is because the coefficients in the proof are set for the convenience of the proof rather than being tightly chosen.

Theorem 2 indicates that RWF recovers the true signal with $\mathcal{O}(n)$ samples, which is order-wise optimal. Such an algorithm improves the sample complexity $\mathcal{O}(n\log n)$ of WF. Furthermore, RWF does not require truncation of samples in the gradient step to achieve the same sample complexity as TWF. This is mainly because RWF benefits from the lower-order loss function given in Equation (3), the curvature of which behaves similarly to the least-squares loss function locally as we explain in the introduction.

Theorem 2 also suggests that RWF converges linearly with a constant step size. To reach ϵ-accuracy, it requires a computational cost of $\mathcal{O}(mn\log(1/\epsilon))$ flops, which is better than WF ($\mathcal{O}(mn^2\log(1/\epsilon))$ and on par with TWF. Numerically, as we demonstrate in Section 4, RWF is two times faster than TWF and four to six times faster than WF in terms of both the iteration counts and the time cost in various examples.

2.4 Stability to Bounded Noise

We have established that RWF guarantees exact recovery at a linear convergence rate for noise-free measurements. We now study RWF in the presence of noise. Suppose the measurements are corrupted by bounded noise, and are given by

$$y_i = |\boldsymbol{a}_i^T\boldsymbol{x}| + w_i, \quad 1 \le i \le m, \tag{17}$$

where $\boldsymbol{w} = \{w_i\}_{i=1}^m$ denote the additive noise. Then the following theorem shows that RWF is robust under bounded noise.

Theorem 3 *Consider the model* (17). *Suppose that the measurement vectors are independently Gaussian, i.e.,* $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I})$ *for* $1 \le i \le m$, *and the noise is bounded, i.e.,* $\|\boldsymbol{w}\|/\sqrt{m} \le c\|\boldsymbol{x}\|$ *where* c *is a positive constant. Then there exist some universal constants* $\mu_0 > 0$ *(*μ_0 *can be set as* 0.8 *in practice),* $0 < \rho < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \ge c_0 n$ *and* $\mu < \mu_0$, *then with probability at least* $1 - c_1\exp(-c_2 m)$, *Algorithm 1 yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|}{\sqrt{m}} + (1-\rho)^t\|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N}. \tag{18}$$

Proof See Appendix D. ■

Theorem 3 shows that under the same sample complexity, RWF converges at a linear rate to a neighborhood around the true signal, whose radius is on the level of the noise. The numerical result under the Poisson noise model in Section 4 further corroborates the stability of RWF.

3. Incremental Reshaped Wirtinger Flow

In large-sample and online scenarios, stochastic algorithms are preferred due to their potential advantage of faster convergence and lower memory requirement. Thus, in this section, we develop stochastic versions of RWF, referred to as incremental reshaped Wirtinger flow (IRWF). We show that IRWF guarantees exact recovery at a linear convergence rate under the same sample complexity. We further draw the connection between IRWF and the randomized Kaczmarz method recently developed for phase retrieval (Wei, 2015; Li et al., 2015; Chi and Lu, 2016), and establish its global convergence as a side product.

3.1 (Mini-batch) IRWF: Algorithm and Convergence

In oder to fully exploit the processing throughput of CPU/GPU, we develop a mini-batch IRWF, described in Algorithm 2. The mini-batch IRWF applies the same initialization step as in RWF, and uses a mini-batch of measurements for each gradient update.

Algorithm 2 Mini-batch Incremental Reshaped Wirtinger Flow (mini-batch IRWF)

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$, mini-batch size k;
Initialization: Same as in RWF (Algorithm 1);
Gradient loop: for $t = 0 : T-1$ do
Choose Γ_t uniformly at random from the subsets of $\{1, 2, \ldots, m\}$ with the cardinality k, and let

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \mu \cdot \boldsymbol{A}_{\Gamma_t}^* \left(\boldsymbol{A}_{\Gamma_t} \boldsymbol{z}^{(t)} - \boldsymbol{y}_{\Gamma_t} \odot \mathrm{Ph}(\boldsymbol{A}_{\Gamma_t} \boldsymbol{z}^{(t)}) \right), \tag{19}$$

where $\boldsymbol{A}_{\Gamma_t}$ is a matrix stacking $\boldsymbol{a}_i^*$ for $i \in \Gamma_t$ as its rows, $\boldsymbol{y}_{\Gamma_t}$ is a vector stacking y_i for $i \in \Gamma_t$ as its elements, and $\mathrm{Ph}(\boldsymbol{z})$ denotes the phase vector of $\boldsymbol{z}$.
Output $\boldsymbol{z}^{(T)}$.

If the gradient update uses only a single sample, i.e., $k = 1$, we refer to Algorithm 2 as IRWF, where the step (19) becomes

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \mu \left(\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)} - y_{i_t} \cdot \frac{\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)}|} \right) \boldsymbol{a}_{i_t}. \tag{20}$$

We characterize the convergence of mini-batch IRWF in the following theorem.

Theorem 4 *Consider the problem of solving any given $\boldsymbol{x} \in \mathbb{R}^n$ from a system of equations (1) with independent Gaussian measurement vectors $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I})$. There exist some universal constants $0 < \rho, \rho_0 < 1$ and $c_0, c_1, c_2 > 0$ such that if $m \geq c_0 n$ and $\mu = \rho_0 / n$ for the update rule (19), then with probability at least $1 - c_1 \exp(-c_2 m)$, we have that*

$$\mathrm{E}_{\Gamma_t} \left[dist^2(\boldsymbol{z}^{(t+1)}, \boldsymbol{x}) \right] \leq \left(1 - \frac{k\rho}{n} \right) \cdot dist^2(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \tag{21}$$

holds for all $\boldsymbol{z}^{(t)}$ satisfying $\frac{dist(\boldsymbol{z}^{(t)}, \boldsymbol{x})}{\|\boldsymbol{z}\|} \leq \frac{1}{10}$.

Proof See Appendix E.1. ■

We suggest that $\rho_0 = 1$ and hence the step size $\mu = \frac{1}{n}$ in practice. Theorem 4 characterizes that the error decays exponentially fast in expectation if the estimate lands into the neighborhood of the global minimizer. For a generic optimization objective, it is not anticipated that incremental/stochastic first-order methods achieve linear convergence due to the variance of stochastic gradients. However, for our specific problem, the variance of stochastic gradients reduces as the estimate approaches the true signal, and hence a fixed step size can be employed and exponential decay can be established. A result similar in spirit was also established for the stochastic algorithm based on TWF (referred to as ITWF) (Kolte and Özgür, 2016). We provide further comparisons between IRWF and ITWF in Section 3.3. On the other hand, it was shown in (Moulines and Bach, 2011; Needell et al., 2016) that stochastic gradient methods yield linear convergence to the minimizer $\boldsymbol{x}_\star$ if the objective $F(\boldsymbol{x}) = \sum_i f_i(\boldsymbol{x})$ is a smooth and strongly convex function and $\boldsymbol{x}_\star$ minimizes

all components $f_i(\boldsymbol{x})$. The summands of our objective (3) also share the same minimizer (although it is neither convex nor smooth), which also helps to explain the convergence property of the IRWF.

3.2 Connection to the Kaczmarz Method for Phase Retrieval

The *Kaczmarz method* was originally developed for solving systems of linear equations (Kaczmarz, 1937). In literature (Wei, 2015; Li et al., 2015), it was adapted to solve the phase retrieval problem, which we refer to as *Kaczmarz-PR*. It has been demonstrated in Wei (2015) that Kaczmarz-PR exhibits better empirical performance than error reduction (ER) (Gerchberg, 1972; Fienup, 1982) and WF (Candès et al., 2015). However, global convergence of Kaczmarz-PR has not been well established yet, although the randomized Kaczmarz method for the least-squares problem is known to converge at a linear rate (Strohmer and Vershynin, 2009; Zouzias and Freris, 2013). For instance, Wei (2015) obtained a bound on the estimation error which can be as large as the signal energy no matter how many iterations are taken. Li et al. (2015) established the asymptotic convergence in the regime when both m and n go to infinity but their ratio is fixed.

In this section, we draw connection between IRWF and Kaczmarz-PR, which enables us to establish the theoretical guarantee of Kaczmarz-PR by adapting that of IRWF. This is analogous to the connection made in Needell et al. (2016) between the Kaczmarz method and the stochastic gradient method for solving the least-squares problem. Here, the connection is made possible due to the lower-order loss function of RWF, which was not evident in previous studies of WF and TWF.

To be more specific, the Kaczmarz-PR (Wei, 2015, Algorithm 3) employs the following update rule

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^*\boldsymbol{z}^{(t)} - y_{i_t}\cdot\frac{\boldsymbol{a}_{i_t}^*\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^*\boldsymbol{z}^{(t)}|}\right)\boldsymbol{a}_{i_t}, \tag{22}$$

where i_t is selected either in a deterministic manner or randomly. We focus on the randomized case where i_t is selected uniformly at random from $\{1,\ldots,m\}$.

Comparing Equation (22) and Equation (20), the update rule of Kaczmarz-PR becomes equivalent to IRWF, if we replace the step size μ by $\frac{1}{\|\boldsymbol{a}_{i_t}\|^2}$. Moreover, these two update rules are close if μ is set as suggested, i.e., $\mu = \frac{1}{n}$, because for Gaussian measurements, $\|\boldsymbol{a}_{i_t}\|^2$ concentrates around n by the law of large numbers. As we demonstrate in the numerical experiments (see Table 1), Kaczmarz-PR and IRWF have similar performance as anticipated. Thus, following the convergence guarantee for IRWF in Theorem 4, we establish the convergence guarantee for the randomized Kaczmarz-PR as follows.

Theorem 5 *Assume the measurement vectors are independent and each* $\boldsymbol{a}_i \sim \mathcal{N}(0,\boldsymbol{I})$. *There exist some universal constants* $0<\rho<1$ *and* $c_0,c_1,c_2>0$ *such that if* $m \geq c_0 n$, *then with probability at least* $1-c_1 m\exp(-c_2 n)$, *the randomized Kaczmarz-PR update rule* (22) *yields*

$$\mathrm{E}_{i_t}\left[dist^2(\boldsymbol{z}^{(t+1)},\boldsymbol{x})\right] \leq \left(1-\frac{\rho}{n}\right)\cdot dist^2(\boldsymbol{z}^{(t)},\boldsymbol{x}) \tag{23}$$

for all $\boldsymbol{z}^{(t)}$ *satisfying* $\frac{dist(\boldsymbol{z}^{(t)},\boldsymbol{x})}{\|\boldsymbol{z}\|} \leq \frac{1}{10}$.

Proof See Appendix E.2. ■

Theorem 5 implies that as long as the iteration sequence $\boldsymbol{z}^{(t)}$ lies in the neighborhood of the true signal, the error decays exponentially fast in expectation. However, Theorem 5 does not guarantee the full trajectory of the estimates to be in the neighborhood of the true signal. After submission of our work, more recent studies (Jeong and Gunturk, 2017; Tan and Vershynin, 2017) established the convergence of full trajectory by employing martingale theory.

Furthermore, Wei (2015) also provided a *block* Kaczmarz-PR (similar to the mini-batch version), whose update rule is given by

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \boldsymbol{A}_{\Gamma_t}^{\dagger}\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)} - \boldsymbol{y}_{\Gamma_t} \odot \mathrm{Ph}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right), \tag{24}$$

where Γ_t is a selected block at iterate t containing row indices, and $\dagger$ represents *Moore-Penrose pseudoinverse*, which is computed as follows:

$$\boldsymbol{A}^{\dagger} = \begin{cases} (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*, & \text{if } \boldsymbol{A} \text{ has linearly independent columns;} \\ \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}, & \text{if } \boldsymbol{A} \text{ has linearly independent rows.} \end{cases} \tag{25}$$

Comparing Equation (24) and the mini-batch IRWF update in Equation (19), these two update rules are similar to each other if $\boldsymbol{A}_{\Gamma_t}\boldsymbol{A}_{\Gamma_t}^*$ approaches $\frac{n}{\rho_0}\boldsymbol{I}_{|\Gamma_t|}$. For the case with Gaussian measurements, $\boldsymbol{A}_{\Gamma_t}$ has linearly independent rows with high probability if $|\Gamma_t| \le n$ and hence $\boldsymbol{A}_{\Gamma_t}\boldsymbol{A}_{\Gamma_t}^*$ is not far from $n\boldsymbol{I}_{|\Gamma_t|}$. Our numerical experiments (see Table 1) further suggest similar convergence rates for these two algorithms with the same block/mini-batch size.

Next, we argue that for the CDP setting, block Kaczmarz-PR is the same as the mini-batch IRWF with $\mu = 1$. The CDP measurements are collected in the following form

$$\boldsymbol{y}^{(l)} = |\boldsymbol{F}\boldsymbol{D}^{(l)}\boldsymbol{x}|, \quad 1 \le l \le L, \tag{26}$$

where $\boldsymbol{F}$ represents the discrete Fourier transform (DFT) matrix, $\boldsymbol{D}^{(l)}$ denotes a diagonal matrix (mask), and L denotes the number of masks. We choose the block size $|\Gamma_t|$ to be the dimension n of the signal for the convenience of Fourier transform. Then $\boldsymbol{A}_{\Gamma_t}$ becomes the Fourier transform composed with $\boldsymbol{D}^{(l)}$ (mask effect) and $\boldsymbol{A}_{\Gamma_t}^*$ becomes $\boldsymbol{D}^{(l)*}$ multiplied by the inverse Fourier transform. Therefore, $(\boldsymbol{A}_{\Gamma_t}\boldsymbol{A}_{\Gamma_t}^*) = \boldsymbol{I}$ if the diagonal elements of $\boldsymbol{D}^{(l)}$ have unit magnitude. Taking the step size $\mu = 1$, the two algorithms are identical.

On the other hand, since the block Kaczmarz-PR needs to calculate the matrix inverse or to solve an inverse problem, the block size cannot be too large. However, mini-batch IRWF works well for a wide range of the mini-batch sizes, which can even grow with the signal dimension n as long as a batch of data is loadable into the memory.

3.3 Comparison with Incremental Truncated Wirtinger Flow (ITWF)

Recently, (Kolte and Özgür, 2016) designed and analyzed an incremental algorithm based on TWF, which is referred to as ITWF. More specifically, ITWF employs the same initialization procedure as TWF and randomly chooses one sample with the index i_t selected uniformly

at random from $\{1, 2, \ldots, m\}$ for the gradient update as follows:

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \mu \cdot \frac{|\boldsymbol{a}_{i_t}^T \boldsymbol{z}|^2 - y_{i_t}^2}{\boldsymbol{a}_{i_t}^T \boldsymbol{z}} \boldsymbol{a}_{i_t} \mathbf{1}_{\mathcal{E}_{1,t}^{i_t} \cap \mathcal{E}_3^{i_t}}, \tag{27}$$

where $\mathbf{1}_{\mathcal{E}_{1,t}^{i_t} \cap \mathcal{E}_3^{i_t}}$ represents the truncation rule determined by the events $\mathcal{E}_{1,t}^{i_t}$ and $\mathcal{E}_3^{i_t}$. As a comparison, the update rule of IRWF is much simpler due to the use of the lower-order loss function and does not require any truncation in the gradient loop. Kolte and Özgür (2016) proved that the update (27) shrinks the estimate error as long as m/n is large enough and the estimate $\boldsymbol{z}$. Compared to ITWF, the IRWF update (19) also shrinks the estimate error, but runs faster than ITWF numerically as demonstrated in Section 4.

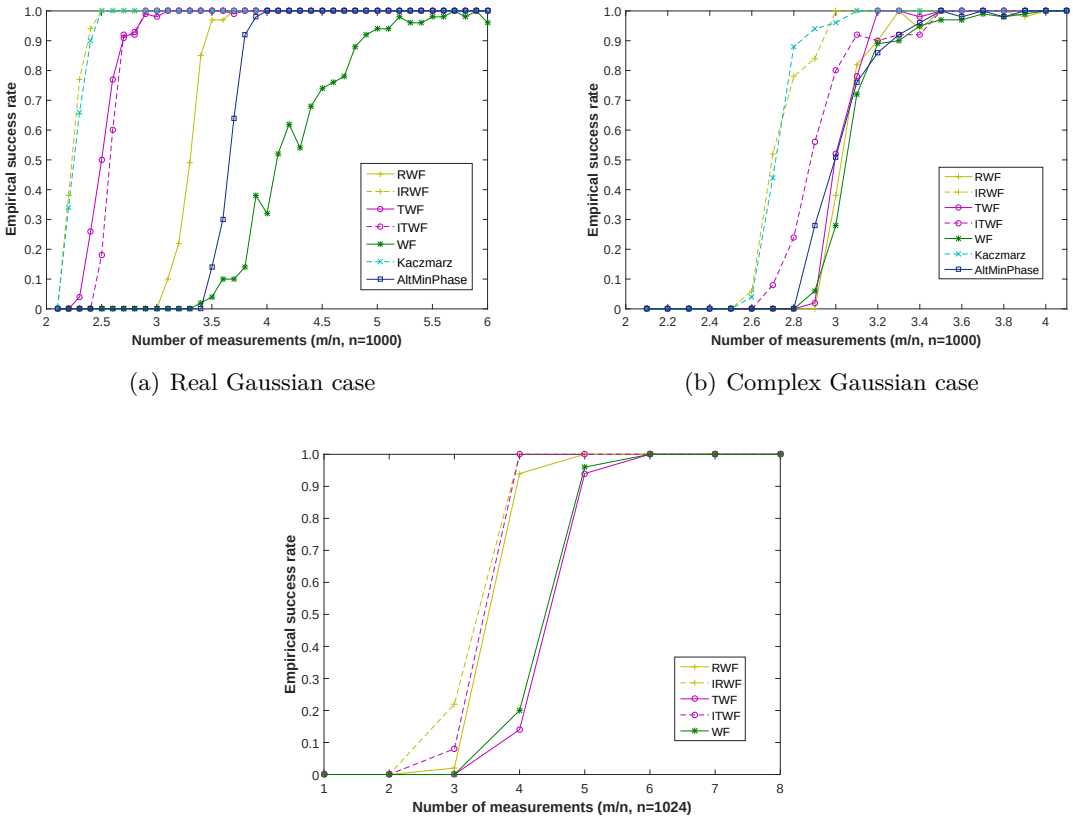
4. Numerical Experiments

In this section, we demonstrate the numerical efficiency of RWF and (mini-batch) IRWF by comparing their performance with other competitive algorithms. Our experiments are conducted not only for the real Gaussian case but also for the complex Gaussian and the CDP cases. All the experiments are implemented in Matlab 2015b and conducted on a computer equipped with Intel Core i7 3.4GHz CPU and 12GB RAM.

We first compare the sample complexity of RWF and IRWF with those of TWF, WF, Kaczmarz-PR and AltMinPhase via the empirical successful recovery rate versus the number of measurements. For RWF, we follow Algorithm 1 with the suggested parameters. For IRWF, we adopt a block size 64 for efficiency and set the step size $\mu = 1/n$. For WF, TWF, we use the code provided in the original papers with the suggested parameters. For ITWF, we also adopt a block size 64 and set the step size $\mu = 0.6/n$ (optimal step size). We conduct the experiments for the real Gaussian, complex Gaussian and CDP cases respectively. For the real and complex cases, we set the signal dimension n to be 1000, and set the ratio m/n to take values from 2 to 6 with a step 0.1. For each m, we run 100 trials and count the number of successful trials. For each trial, we run a maximal number of iterations/passes $T = 10000$ for all algorithms, and a trial is declared to be successful whenever the iterate satisfies $\mathrm{dist}(\boldsymbol{z}^{(T)}, \boldsymbol{x})/\|\boldsymbol{x}\| \leq 10^{-5}$. For the real Gaussian case, we generate the signal $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{n\times n})$, and generate the measurement vectors $\boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{n\times n})$ i.i.d. for $i = 1, \ldots, m$. For the complex Gaussian case, we generate the signal $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_{n\times n}) + j\mathcal{N}(0, \boldsymbol{I}_{n\times n})$ and the measurement vectors $\boldsymbol{a}_i \sim \frac{1}{2}\mathcal{N}(0, \boldsymbol{I}_{n\times n}) + j\frac{1}{2}\mathcal{N}(0, \boldsymbol{I}_{n\times n})$ i.i.d. for $i = 1, \ldots, m$. For the CDP case (26), we set $n = 1024$ for the convenience of FFT and $m/n = L = 1, 2, \ldots, 8$. All other settings are the same as those for the real case.

We note that for the CDP case, the Kaczmarz-PR algorithm is identical to the IRWF with step size $\mu = 1/n$ due to the argument in Section 3.2. Moreover under the CDP case, the AltMinPhase algorithm is identical to the RWF with step size $\mu = 1$ because the inverse of the Fourier measurement matrix is nothing but its conjugate transpose. In the following experiments for the CDP case, we choose the step size $\mu = 1/n$ for the IRWF and $\mu = 1$ for the RWF, under which the Kaczmarz-PR algorithm coincides with the IRWF and the AltMinPhase algorithm coincides with the RWF.

Figure 4 plots the fraction of successful trials out of 100 trials for all algorithms, with respect to m/n. It can be seen that IRWF and Kaczmarz-PR exhibit a similar sample



(a) Real Gaussian case

(b) Complex Gaussian case

(c) CDP case

Figure 4: Comparison of sample complexity among RWF, IRWF, TWF, ITWF, WF, Karzmarz-PR and AltMinPhase.

complexity, which is the best for all three cases, and is close to the theoretical limit (Bandeira et al., 2014). It can also be seen that the two incremental methods (IRWF and ITWF) outperform the batch methods (RWF, TWF, AltMinPhase and WF). This can be due to the inherent noise in incremental methods, which helps to escape bad local minima. This can be extremely helpful in the regime with a small number of samples, where local minima do exist near the global minima. Comparing among the batch methods (RWF, TWF AltMinPhase and WF), it can be seen that although RWF outperforms only WF and AltMinPhase (not TWF) for the real Gaussian case, it has a comparable performance for the complex case and outperforms TWF and WF in the CDP case. An intuitive explanation for the real case is that a substantial number of samples with small $|\boldsymbol{a}_i^T \boldsymbol{z}|$ can deviate the gradient direction so that truncation indeed helps to stabilize the algorithm if the number of measurements is not large.

We next compare the convergence rate of RWF, IRWF with those of TWF, ITWF, WF, Kaczmarcz and AltMinPhase. We run all of the algorithms with the suggested parameters in the original code. We generate the signal and measurements in the same way as those

Table 1: Comparison of iteration count and time cost among algorithms ($n = 5000, m = 8n$).

		Real Gaussian		Complex Gaussian	
		#passes	time(s)	# passes	time(s)
Batch methods	RWF	72	**12.66**	176	**122.4**
	TWF	186	32.36	487	395.1
	WF	319	54.83	932	887.8
	AltMinPhase	**6**	79.58	**159**	9637
Incremental methods	IRWF	9	44.77	21	233.2
	mini-batch IRWF (64)	9	**8.076**	**21**	**48.58**
	mini-batch ITWF (64)	16	37.38	29	149.5
	Kaczmarz-PR	9	50.68	21	248.4
	block Kaczmarz-PR (64)	**8**	28.50	22	89.31

in the first experiment with $n = 5000, m = 8n$. All algorithms are seeded with the RWF initialization. In Table 1, we list the number of passes and the time cost for all the algorithms to achieve a relative error of 10^{-14} averaged over 10 trials. For the incremental methods, one update passes k samples and one pass amounts to m/k updates. Clearly, IRWF with mini-batch size 64 runs the fastest for both the real and complex cases. Moreover, among the batch (deterministic) algorithms, RWF takes much fewer passes as well as runs much faster than TWF and WF. Although RWF takes more iterations than AltMinPhase, it runs much faster than AltMinPhase due to the fact that each iteration of AltMinPhase needs to solve a least-squares problem that takes much longer than a simple gradient update in RWF.

We also compare the performance of the above algorithms on the recovery of a real image from the Fourier intensity measurements (the two dimensional CDP case). The image (see Figure 5) is the Milky Way Galaxy with resolution 1920×1080. Table 2 lists the number of passes and the time cost for the above six algorithms to achieve the relative error of 10^{-15} for one R/G/B channel. All algorithms are seeded with the RWF initialization. To explore the advantage of FFT, we run the incremental/stochastic methods with the mini-batch size equal to the number of pixels for one R/G/B channel. We note that with such a mini-batch size, IRWF is equivalent to block Kaczmarz-PR from the discussion in Section 3.2. It can be seen that in general, the incremental/stochastic methods (IRWF and ITWF) run faster than the batch methods (RWF, TWF, WF). Moreover, among the batch methods, RWF outperforms the other three algorithms in both the number of passes and the computational time. In particular, RWF runs two times faster than TWF and six times faster than WF in terms of both the number of iterations and the computational time.

We next demonstrate the robustness of RWF to noise and compare it with TWF. We consider the phase retrieval problem in imaging applications, where Poisson noise is often used to model the sensor and electronic noise (Fogel et al., 2016). Specifically, the noisy measurements of intensity can be expressed as $y_i = \sqrt{\alpha \cdot \text{Poisson}\left(|\boldsymbol{a}_i^T \boldsymbol{x}|^2/\alpha\right)}$, for

Figure 5: Milky way Galaxy.

Table 2: Comparison of iterations and time cost among algorithms on recovery of Galaxy image (shown in Figure 5), where $L = m/n$ denotes the number of CDP masks.

	Algorithms	RWF	IRWF	TWF	ITWF	WF
$L = 6$	#passes	140	**24**	410	41	fail
	time cost(s)	110	**21.2**	406	43	fail
$L = 12$	#passes	70	**8**	190	12	315
	time cost(s)	107	**13.7**	363.6	25.9	426

$i = 1, 2, ...m$ where α denotes the level of the input noise, and Poisson(λ) denotes a random sample generated by the Poisson distribution with mean λ. It can be observed from Figure 6 that RWF performs better than TWF in terms of the recovery accuracy under two different noise levels.

5. Conclusion

In this paper, we proposed RWF and its incremental version IRWF to recover a signal for the phase retrieval problem, based on a *nonconvex and nonsmooth* quadratic loss function of magnitude measurements. This loss function sacrifices the smoothness but enjoys advantages in statistical and computational efficiency. It has potential to be extended in various scenarios. One interesting direction is to extend such an algorithm to exploit signal structures (e.g., non-negativity, sparsity, etc) to assist the recovery. The lower-order loss function may offer great simplicity to prove the performance guarantee in such cases.

Another interesting direction is to study the convergence of algorithms from random initialization. In the regime with a large sample size ($m \gg n$), the empirical loss surface

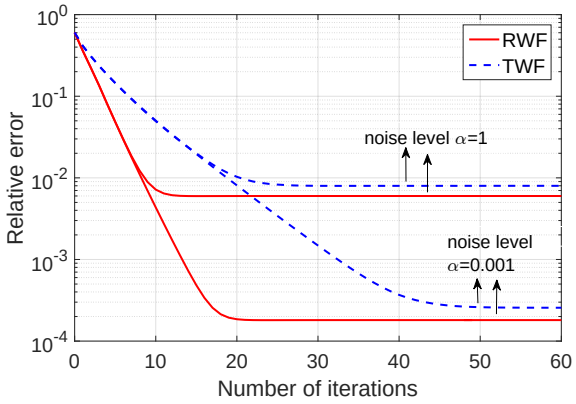


Figure 6: Comparison of relative error under Poisson noise between RWF and TWF.

approaches the asymptotic loss Figure 1(b) and hence has no spurious local minima. Due to the result (Lee et al., 2016a), it is conceivable that the gradient descent algorithm converges from any random starting point. Similar phenomena have been observed in (Sun et al., 2016; Ge et al., 2016). However, under a moderate number of measurements ($m < 10n$), we observe that genuine local minima do exist and often locate not far from the global minima. In such a regime, the batch gradient method often fails with random initialization. As anticipated, stochastic algorithms are efficient in escaping bad local minima or saddle points in nonconvex optimization because of the inherent noise (Ge et al., 2015; Sa et al., 2015). We observe numerically that IRWF and block IRWF from a random starting point still converge to a global minimum even with a very small sample size which is close to the theoretical limit. It is of interest to analyze theoretically why stochastic methods escape these local minima (not just saddle points) efficiently.

Acknowledgments

The work of H. Zhang, Y. Zhou and Y. Liang is supported in part by the grants NSF CCF-1704169, NSF ECCS-1609916 and AFOSR FA9550-16-1-0077. The work of Y. Chi is supported in part by the grants AFOSR under FA9550-15-1-0205, ONR N00014-15-1-2387, NSF CCF-1704245, and ECCS-1650449.

Appendices

Appendix A. Expected Loss Surfaces

It is easy to check that the expectation of the least-squares loss is given by $\mathrm{E}[\ell_{LS}(\boldsymbol{z})] = \|\boldsymbol{z}-\boldsymbol{x}\|^2$. The expectation of the WF loss function (2) is given by (Sun et al., 2016) as

$$\mathrm{E}[\ell_{WF}(\boldsymbol{z})] = \frac{3}{4}\|\boldsymbol{x}\|^4 + \frac{3}{4}\|\boldsymbol{z}\|^4 - \frac{1}{2}\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|^2 - |\boldsymbol{z}^T\boldsymbol{x}|^2. \tag{28}$$

We next show that the expectation of the RWF loss function (3) has the following form:

$$\mathrm{E}[\ell(\boldsymbol{z})] = \frac{1}{2}\|\boldsymbol{x}\|^2 + \frac{1}{2}\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|\|\boldsymbol{z}\| \cdot \mathrm{E}\left[\frac{|\boldsymbol{a}_i^T\boldsymbol{z}|}{\|\boldsymbol{z}\|} \cdot \frac{|\boldsymbol{a}_i^T\boldsymbol{x}|}{\|\boldsymbol{x}\|}\right], \tag{29}$$

where

$$\mathrm{E}\left[\frac{|\boldsymbol{a}_i^T\boldsymbol{z}|}{\|\boldsymbol{z}\|} \cdot \frac{|\boldsymbol{a}_i^T\boldsymbol{x}|}{\|\boldsymbol{x}\|}\right] = \begin{cases} \frac{(1-\rho^2)^{3/2}}{\pi}\int_0^\infty t(e^{\rho t}+e^{-\rho t})K_0(t)dt, & \text{if } |\rho|<1; \\ 1, & \text{if } |\rho|=1; \end{cases} \tag{30}$$

where $\rho = \frac{\boldsymbol{z}^T\boldsymbol{x}}{\|\boldsymbol{x}\|\|\boldsymbol{z}\|}$ and $K_0(\cdot)$ is the modified Bessel function of the second kind.

In order to derive (30), we first define

$$u := \frac{\boldsymbol{a}_i^T\boldsymbol{z}}{\|\boldsymbol{z}\|} \text{ and } v := \frac{\boldsymbol{a}_i^T\boldsymbol{x}}{\|\boldsymbol{x}\|},$$

and it suffices to drive $\mathrm{E}[|uv|]$. Note that $(u,v) \sim \mathcal{N}(0,\Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{and} \quad \rho = \frac{\boldsymbol{z}^T\boldsymbol{x}}{\|\boldsymbol{x}\|\|\boldsymbol{z}\|}.$$

Following (Donahue, 1964), the density function of uv is given by

$$\phi_{uv}(x) = \frac{1}{\pi\sqrt{1-\rho^2}} \exp\left(\frac{\rho x}{1-\rho^2}\right) K_0\left(\frac{|x|}{1-\rho^2}\right), \quad x \neq 0.$$

Thus, the density function of $|uv|$ is given by

$$\psi_{|uv|}(x) = \frac{1}{\pi\sqrt{1-\rho^2}} \left[\exp\left(\frac{\rho x}{1-\rho^2}\right) + \exp\left(-\frac{\rho x}{1-\rho^2}\right)\right] K_0\left(\frac{|x|}{1-\rho^2}\right), \quad x>0, \tag{31}$$

for $|\rho|<1$. Therefore, if $|\rho|<1$, we have

$$\begin{aligned} \mathrm{E}[|uv|] &= \int_0^\infty x \cdot \psi_\rho(x)dx \\ &= \int_0^\infty x \cdot \frac{1}{\pi\sqrt{1-\rho^2}} \left[\exp\left(\frac{\rho x}{1-\rho^2}\right) + \exp\left(-\frac{\rho x}{1-\rho^2}\right)\right] K_0\left(\frac{|x|}{1-\rho^2}\right) dx \\ &= \frac{(1-\rho^2)^{3/2}}{\pi}\int_0^\infty t(e^{\rho t}+e^{-\rho t})K_0(t)dt, \end{aligned}$$

where the last step follows from change of variables.

If $|\rho| = 1$, then $|uv|$ becomes a χ_1^2 random variable, with the density given by

$$\psi_{|uv|}(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} \exp(-x/2), \quad x > 0,$$

and hence $\mathrm{E}[|uv|] = 1$.

Appendix B. Proof of Proposition 1

The idea of using truncation to bound some non-sub-Gaussian sequences has appeared in the literature (Candès et al., 2013, Lemma 2.3) and (Chen and Candès, 2015). Compared to the proof of the initialization for TWF (Chen and Candès, 2015), new technical developments are needed to address the magnitude measurements and truncation from both sides.

We first estimate the norm of $\boldsymbol{x}$ as

$$\lambda_0 = \frac{mn}{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1} \cdot \left(\frac{1}{m} \sum_{i=1}^m y_i \right). \tag{32}$$

Since $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\mathbf{n}\times\mathbf{n}})$, by Hoeffding-type inequality, it can be shown that

$$\left| \frac{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1}{mn} - \sqrt{\frac{2}{\pi}} \right| < \frac{\epsilon}{3} \tag{33}$$

holds with probability at least $1 - 2\exp(-c_1 mn\epsilon^2)$ for some constant $c_1 > 0$.

Moreover, for a fixed $\boldsymbol{x}$, y_i's are independent sub-Gaussian random variables. Thus, by Hoeffding-type inequality, it can be shown that

$$\left| \sqrt{\frac{\pi}{2}} \left(\frac{1}{m} \sum_{i=1}^m y_i \right) - \|\boldsymbol{x}\| \right| < \frac{\epsilon}{3} \|\boldsymbol{x}\| \tag{34}$$

holds with probability at least $1 - 2\exp(-c_1 m\epsilon^2)$ for some constant $c_1 > 0$.

On the event $E_1 = \{\text{both (33) and (34) hold}\}$, it can be argued that

$$|\lambda_0 - \|\boldsymbol{x}\|| < \epsilon \|\boldsymbol{x}\|. \tag{35}$$

Without loss of generality, we let $\|\boldsymbol{x}\| = 1$. Then on the event E_1, the truncation function satisfies the following bounds

$$\mathbf{1}_{\{\alpha_l(1+\epsilon) < |\boldsymbol{a}_i^T \boldsymbol{x}| < \alpha_u(1-\epsilon)\}} \le \mathbf{1}_{\{\alpha_l \lambda_0 < y_i < \alpha_u \lambda_0\}} \le \mathbf{1}_{\{\alpha_l(1-\epsilon) < |\boldsymbol{a}_i^T \boldsymbol{x}| < \alpha_u(1+\epsilon)\}}.$$

Thus, define

$$\begin{aligned} \boldsymbol{Y}_1 :=& \frac{1}{m} \sum_{i=1}^m \boldsymbol{a}_i \boldsymbol{a}_i^T |\boldsymbol{a}_i^T \boldsymbol{x}| \mathbf{1}_{\{\alpha_l(1+\epsilon) < |\boldsymbol{a}_i^T \boldsymbol{x}| < \alpha_u(1-\epsilon)\}} \\ \boldsymbol{Y}_2 :=& \frac{1}{m} \sum_{i=1}^m \boldsymbol{a}_i \boldsymbol{a}_i^T |\boldsymbol{a}_i^T \boldsymbol{x}| \mathbf{1}_{\{\alpha_l(1-\epsilon) < |\boldsymbol{a}_i^T \boldsymbol{x}| < \alpha_u(1+\epsilon)\}}, \end{aligned}$$

and we have $\boldsymbol{Y}_1 \prec \boldsymbol{Y} \prec \boldsymbol{Y}_2$. We further compute the expectations of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ and obtain

$$\mathrm{E}[\boldsymbol{Y}_1] = (\beta_1 \boldsymbol{x}\boldsymbol{x}^T + \beta_2 \boldsymbol{I}), \quad \mathrm{E}[\boldsymbol{Y}_2] = (\beta_3 \boldsymbol{x}\boldsymbol{x}^T + \beta_4 \boldsymbol{I}), \tag{36}$$

where

$$\begin{aligned}
\beta_1 &:= \mathrm{E}[|\xi|^3 \mathbf{1}_{\{\alpha_l(1+\epsilon)<|\xi|<\alpha_u(1-\epsilon)\}}] - \mathrm{E}[|\xi| \mathbf{1}_{\{\alpha_l(1+\epsilon)<|\xi|<\alpha_u(1-\epsilon)\}}], \\
\beta_2 &:= \mathrm{E}[|\xi| \mathbf{1}_{\{\alpha_l(1+\epsilon)<|\xi|<\alpha_u(1-\epsilon)\}}] \\
\beta_3 &:= \mathrm{E}[|\xi|^3 \mathbf{1}_{\{\alpha_l(1-\epsilon)<|\xi|<\alpha_u(1+\epsilon)\}}] - \mathrm{E}[|\xi| \mathbf{1}_{\{\alpha_l(1-\epsilon)<|\xi|<\alpha_u(1+\epsilon)\}}], \\
\beta_4 &:= \mathrm{E}[|\xi| \mathbf{1}_{\{\alpha_l(1-\epsilon)<|\xi|<\alpha_u(1+\epsilon)\}}]
\end{aligned}$$

where $\xi \sim \mathcal{N}(0,1)$. For given α_l and α_u, a small value of ϵ yields arbitrarily close β_1 and β_3, as well as arbitrarily close β_2 and β_4. For example, taking $\alpha_l = 1, \alpha_u = 5$ and $\epsilon = 0.01$, we have $\beta_1 = 0.9678, \beta_2 = 0.4791, \beta_3 = 0.9688, \beta_4 = 0.4888$.

Now applying the standard results on random matrices with non-isotropic sub-Gaussian rows (Vershynin, 2012, equation (5.26)) and noticing that $\boldsymbol{a}_i \boldsymbol{a}_i^T |\boldsymbol{a}_i^T \boldsymbol{x}| \mathbf{1}_{\{\alpha_l(1+\epsilon)<|\boldsymbol{a}_i^T \boldsymbol{x}|<\alpha_u(1-\epsilon)\}}$ can be rewritten as $\boldsymbol{b}_i \boldsymbol{b}_i^T$ for a sub-Gaussian vector $\boldsymbol{b}_i := \boldsymbol{a}_i \sqrt{|\boldsymbol{a}_i^T \boldsymbol{x}|} \mathbf{1}_{\{\alpha_l(1+\epsilon)<|\boldsymbol{a}_i^T \boldsymbol{x}|<\alpha_u(1-\epsilon)\}}$, one can derive

$$\|\boldsymbol{Y}_1 - \mathrm{E}[\boldsymbol{Y}_1]\| \le \delta, \quad \|\boldsymbol{Y}_2 - \mathrm{E}[\boldsymbol{Y}_2]\| \le \delta \tag{37}$$

with probability at least $1 - 4\exp(-c_1(\delta)m)$ for some positive c_1 which is only affected by δ, provided that m/n exceeds a certain constant. Furthermore, when ϵ is sufficiently small, one further has $\|\mathrm{E}[\boldsymbol{Y}_1] - \mathrm{E}[\boldsymbol{Y}_2]\| \le \delta$. Combining the above facts together, one can show that

$$\|\boldsymbol{Y} - (\beta_1 \boldsymbol{x}\boldsymbol{x}^T + \beta_2 \boldsymbol{I})\| \le 3\delta. \tag{38}$$

Let $\tilde{\boldsymbol{z}}^{(0)}$ be the normalized leading eigenvector of $\boldsymbol{Y}$. Following the arguments in (Candès et al., 2015, Section 7.8) and taking δ and ϵ to be sufficiently small, one has

$$\mathrm{dist}(\tilde{\boldsymbol{z}}^{(0)}, \boldsymbol{x}) \le \tilde{\delta}, \tag{39}$$

for a given $\tilde{\delta} > 0$, as long as m/n exceeds a certain constant.

Appendix C. Proof of Theorem 2

The general structure of the proof follows that for WF in (Candès et al., 2015) and TWF in (Chen and Candès, 2015). However, the proof requires the development of new bounds due to the nonsmoothness of the loss function and the magnitude measurements. On the other hand, the proof is much simpler due to the lower-order loss function adopted in RWF.

The idea of the proof is to show that within the neighborhood of the global optimizers, RWF satisfies the *Regularity Condition* $\mathsf{RC}(\mu, \lambda, c)$, i.e.,

$$\langle \nabla \ell(\boldsymbol{z}), \boldsymbol{h} \rangle \ge \frac{\mu}{2} \|\nabla \ell(\boldsymbol{z})\|^2 + \frac{\lambda}{2} \|\boldsymbol{h}\|^2 \tag{40}$$

for all $\boldsymbol{z}$ and $\boldsymbol{h} = \boldsymbol{z} - \boldsymbol{x}$ obeying $\|\boldsymbol{h}\| \le c\|\boldsymbol{x}\|$, where $0 < c < 1$ is some constant. Then, as shown in (Chen and Candès, 2015), once the initialization lands into the neighborhood that satisfies $\|\boldsymbol{z} - \boldsymbol{x}\| \le c\|\boldsymbol{x}\|$, linear convergence can be guaranteed, i.e.,

$$\mathrm{dist}^2\left(\boldsymbol{z} - \mu\nabla\ell(\boldsymbol{z}), \boldsymbol{x}\right) \le (1 - \mu\lambda)\mathrm{dist}^2(\boldsymbol{z}, \boldsymbol{x}), \tag{41}$$

for any $\boldsymbol{z}$ with $\|\boldsymbol{z} - \boldsymbol{x}\| \le c\|\boldsymbol{x}\|$.

To show the regularity condition, we first define a set $\mathcal{S} := \{i : 1 \le i \le m, (\boldsymbol{a}_i^T\boldsymbol{z})(\boldsymbol{a}_i^T\boldsymbol{x}) < 0\}$ (which depends on $\boldsymbol{x}$ and $\boldsymbol{z}$), and then derive the following bound:

$$\begin{aligned}
\langle\nabla\ell(\boldsymbol{z}), \boldsymbol{h}\rangle &= \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - |\boldsymbol{a}_i^T\boldsymbol{x}|\mathrm{sgn}(\boldsymbol{a}_i^T\boldsymbol{z})\right)(\boldsymbol{a}_i^T\boldsymbol{h}) \\
&= \frac{1}{m}\left[\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2 + 2\sum_{i\in\mathcal{S}}(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{h})\right] \\
&\ge \frac{1}{m}\left[\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2 - 2\left|\sum_{i\in\mathcal{S}}(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{h})\right|\right] \\
&\ge \frac{1}{m}\left[\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2 - \sum_{i\in\mathcal{S}}2\left|(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{h})\right|\right].
\end{aligned} \tag{42}$$

The first term in (42) can be bounded using Lemma 3.1 in (Candès et al., 2013), which we state below.

Lemma 6 *For any $0 < \epsilon < 1$, there exist constants $c_0, c_1 > 0$ such that if $m > c_0 n\epsilon^{-2}$, then with probability at least $1 - 2\exp(-c_1\epsilon^2 m)$,*

$$(1 - \epsilon)\|\boldsymbol{h}\|^2 \le \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2 \le (1 + \epsilon)\|\boldsymbol{h}\|^2 \tag{43}$$

holds for all non-zero vectors $\boldsymbol{h} \in \mathbb{R}^n$.

For the second term in (42), we derive

$$\begin{aligned}
\sum_{i\in\mathcal{S}}2\left|\boldsymbol{a}_i^T\boldsymbol{x}\right|\left|\boldsymbol{a}_i^T\boldsymbol{h}\right| &\le \sum_{i\in\mathcal{S}}\left[(\boldsymbol{a}_i^T\boldsymbol{x})^2 + (\boldsymbol{a}_i^T\boldsymbol{h})^2\right] \\
&= \sum_{i=1}^{m}[(\boldsymbol{a}_i^T\boldsymbol{x})^2 + (\boldsymbol{a}_i^T\boldsymbol{h})^2]\cdot\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}} \\
&= \sum_{i=1}^{m}[(\boldsymbol{a}_i^T\boldsymbol{x})^2 + (\boldsymbol{a}_i^T\boldsymbol{h})^2]\cdot\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2+(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{h})<0\}} \\
&\le \sum_{i=1}^{m}[(\boldsymbol{a}_i^T\boldsymbol{x})^2 + (\boldsymbol{a}_i^T\boldsymbol{h})^2]\cdot\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \\
&\le 2\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2\cdot\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}}.
\end{aligned} \tag{44}$$

The right hand side of the above equation can be further upper bounded by the following lemma.

Lemma 7 *For any $\epsilon > 0$, there exist constants $c_0, c_1, C > 0$ such that if $m > c_0 n\epsilon^{-2}\log\epsilon^{-1}$, then with probability at least $1 - C\exp(-c_1\epsilon^2 m)$,*

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2 \cdot \mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le (0.13+\epsilon)\,\|\boldsymbol{h}\|^2 \tag{45}$$

holds for all non-zero vectors $\boldsymbol{h} \in \mathbb{R}^n$ satisfying $\|\boldsymbol{h}\| \le \frac{1}{10}\|\boldsymbol{x}\|$.

Proof See Section C.1. ■

Therefore, combining Lemmas 6 and 7 with (42) yields

$$\langle \nabla\ell(\boldsymbol{z}), \boldsymbol{h}\rangle \ge (1 - 0.26 - 3\epsilon)\|\boldsymbol{h}\|^2 = (0.74 - 3\epsilon)\|\boldsymbol{h}\|^2. \tag{46}$$

We further provide an upper bound on $\|\nabla\ell(\boldsymbol{z})\|$ in the following lemma.

Lemma 8 *Fix $\delta > 0$. There exist constatnts $c_0, c, C > 0$ such that if $m \ge c_0 n$ then with probability at least $1 - C\exp(-cm)$,*

$$\|\nabla\ell(\boldsymbol{z})\| \le (1+\delta)\cdot 2\|\boldsymbol{h}\| \tag{47}$$

holds for all non-zero vectors $\boldsymbol{h}, \boldsymbol{z} \in \mathbb{R}^n$ satisfying $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{h}$ and $\|\boldsymbol{h}\| \le \frac{1}{10}\|\boldsymbol{x}\|$.

Proof See Section C.2. ■

Thus, applying Lemma 8 and (46) to (40), we conclude that *Regularity Condition* holds for μ and λ satisfying

$$0.74 - 3\epsilon \ge \frac{\mu}{2}\cdot 4(1+\delta)^2 + \frac{\lambda}{2}, \tag{48}$$

which concludes the proof.

C.1 Proof of Lemma 7

We first prove bounds for any fixed $\boldsymbol{h} \le \frac{1}{10}\|\boldsymbol{x}\|$, and then develop a uniform bound later on. We introduce a series of auxiliary random Lipschitz functions to approximate the indicator functions. For $i = 1, \ldots, m$, define

$$\chi_i(t) := \begin{cases} t, & \text{if } t > (\boldsymbol{a}_i^T\boldsymbol{x})^2; \\ \frac{1}{\delta}(t - (\boldsymbol{a}_i^T\boldsymbol{x})^2) + (\boldsymbol{a}_i^T\boldsymbol{x})^2, & \text{if } (1-\delta)(\boldsymbol{a}_i^T\boldsymbol{x})^2 \le t \le (\boldsymbol{a}_i^T\boldsymbol{x})^2; \\ 0, & \text{else;} \end{cases} \tag{49}$$

and then $\chi_i(t)$'s are random Lipschitz functions with Lipschitz constant $\frac{1}{\delta}$. We further have

$$|\boldsymbol{a}_i^T\boldsymbol{h}|^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le \chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2) \le |\boldsymbol{a}_i^T\boldsymbol{h}|^2\mathbf{1}_{\{(1-\delta)|\boldsymbol{a}_i^T\boldsymbol{x}|^2<|\boldsymbol{a}_i^T\boldsymbol{h}|^2\}}. \tag{50}$$

For convenience, we denote $\gamma_i := \frac{|\boldsymbol{a}_i^T\boldsymbol{h}|^2}{\|\boldsymbol{h}\|^2}\mathbf{1}_{\{(1-\delta)|\boldsymbol{a}_i^T\boldsymbol{x}|^2<|\boldsymbol{a}_i^T\boldsymbol{h}|^2\}}$ and $\theta := \|\boldsymbol{h}\|/\|\boldsymbol{x}\|$. We next estimate the expectation of γ_i, by the following conditional expectation,

$$\mathrm{E}[\gamma_i] = \int_\Omega \gamma_i d\mathbb{P} = \iint_{-\infty}^{\infty} \mathrm{E}\left[\gamma_i \big| \boldsymbol{a}_i^T\boldsymbol{x} = \tau_1\|\boldsymbol{x}\|, \boldsymbol{a}_i^T\boldsymbol{h} = \tau_2\|\boldsymbol{h}\|\right] \cdot f(\tau_1,\tau_2)d\tau_1 d\tau_2, \tag{51}$$

where $f(\tau_1,\tau_2)$ is the density of two joint Gaussian random variables with the correlation $\rho = \frac{\boldsymbol{h}^T\boldsymbol{x}}{\|\boldsymbol{h}\|\|\boldsymbol{x}\|} \neq \pm 1$. We then continue to derive

$$\begin{aligned}
\mathrm{E}[\gamma_i] &= \iint_{-\infty}^{\infty} \tau_2^2 \cdot \mathbf{1}_{\{\sqrt{1-\delta}|\tau_1|<|\tau_2|\theta\}} \cdot f(\tau_1,\tau_2)d\tau_1 d\tau_2 \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \int_{\frac{-|\tau_2|\theta}{\sqrt{1-\delta}}}^{\frac{|\tau_2|\theta}{\sqrt{1-\delta}}} \exp\left(-\frac{(\tau_1-\rho\tau_2)^2}{2(1-\rho^2)}\right) d\tau_1 d\tau_2 \qquad (52) \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \int_{\frac{-\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}}^{\frac{\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}} \exp\left(-\frac{\tau^2}{2}\right) d\tau d\tau_2 \qquad \text{by change of variables} \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \sqrt{\frac{\pi}{2}} \left(\mathrm{erf}\left(\frac{\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}\right) - \mathrm{erf}\left(\frac{-\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}\right)\right) d\tau_2 \\
&= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \left(\mathrm{erf}\left(\frac{(\frac{\theta}{\sqrt{1-\delta}}-\rho)\tau_2}{\sqrt{1-\rho^2}}\right) + \mathrm{erf}\left(\frac{(\frac{\theta}{\sqrt{1-\delta}}+\rho)\tau_2}{\sqrt{1-\rho^2}}\right)\right) d\tau_2. \qquad (53)
\end{aligned}$$

For $|\rho| < 1$, $\mathrm{E}[\gamma_i]$ is a continuous function of ρ. For $|\rho| = 1$, $\mathrm{E}[\gamma_i] = 0$. The last integral (53) can be calculated numerically. Figure 7 plots $\mathrm{E}[\gamma_i]$ for $\theta = 0.1$ and $\delta = 0.01$ over $\rho \in [-1,1]$. Furthermore, (52) indicates that $\mathrm{E}[\gamma_i]$ is monotonically increasing with both θ and δ. Thus, we obtain a universal bound

$$\mathrm{E}[\gamma_i] \leq 0.13 \quad \text{for } \theta < 0.1 \text{ and } \delta = 0.01, \tag{54}$$

which implies $\mathrm{E}[\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2)] \leq 0.13\|\boldsymbol{h}\|^2$ for $\theta < 0.1$ and $\delta = 0.01$.

Furthermore, $\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2)$'s are sub-exponential with the sub-exponential norm $\mathcal{O}(\|\boldsymbol{h}\|^2)$. By the sub-exponential tail bound (Bernstein type) (Vershynin, 2012), we have

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m \frac{\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2)}{\|\boldsymbol{h}\|^2} > (0.13+\epsilon)\right] < \exp(-cm\epsilon^2), \tag{55}$$

for some universal constant c, as long as $\|\boldsymbol{h}\| \leq \frac{1}{10}\|\boldsymbol{x}\|$.

We have proved so far that the claim holds for a fixed $\boldsymbol{h}$. We next obtain a uniform bound over all $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\| \leq \frac{1}{10}\|\boldsymbol{x}\|$. We first show the claim holds for all $\boldsymbol{h}$ with $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$ and then argue that the claim holds when $\|\boldsymbol{h}\| < \frac{1}{10}\|\boldsymbol{x}\|$ towards the end of the proof. Let $\epsilon' = \epsilon\frac{\|\boldsymbol{x}\|}{10}$ and we construct an ϵ'-net $\mathcal{N}_{\epsilon'}$ covering the sphere with the radius $\frac{1}{10}\|\boldsymbol{x}\|$ in $\mathbb{R}^n$ with the cardinality $|\mathcal{N}_{\epsilon'}| \leq (1+\frac{2}{\epsilon})^n$. Then for any $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$, there exists an $\boldsymbol{h}_0 \in \mathcal{N}_{\epsilon'}$ such that $\|\boldsymbol{h}-\boldsymbol{h}_0\| \leq \epsilon\|\boldsymbol{h}\|$. Taking the union bound for all the points on the net, we claim that

$$\frac{1}{m}\sum_{i=1}^m \chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2\right) \leq (0.13+\epsilon)\|\boldsymbol{h}_0\|^2, \quad \forall \boldsymbol{h}_0 \in \mathcal{N}_{\epsilon'} \tag{56}$$

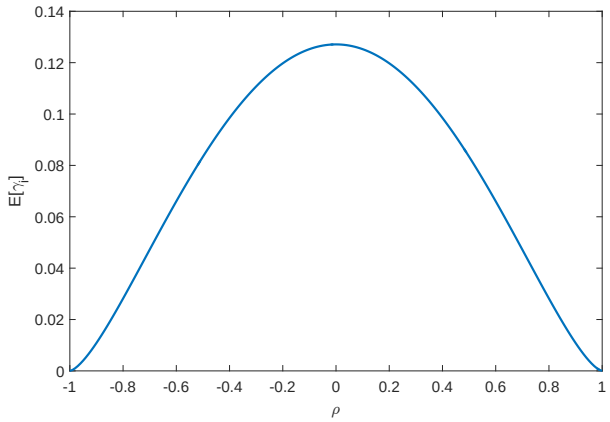


Figure 7: $\mathrm{E}[\gamma_i]$ with respect to ρ.

holds with probability at least $1-(1+2/\epsilon)^n\exp(-cm\epsilon^2)$.

Since $\chi_i(t)$'s are Lipschitz functions with the constant $1/\delta$, we have the following bound

$$\left|\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2)-\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2)\right|\le\frac{1}{\delta}||\boldsymbol{a}_i^T\boldsymbol{h}|^2-|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2|. \tag{57}$$

Moreover, by (Chen and Candès, 2015, Lemma 1), we have

$$\frac{1}{m}\sum_{i=1}^m|\boldsymbol{a}_i^T\boldsymbol{M}\boldsymbol{a}_i|\le c_2\|\boldsymbol{M}\|_F,\qquad\text{for all symmetric rank-2 matrices }\boldsymbol{M}\in\mathbb{R}^{n\times n}, \tag{58}$$

holds with probability at least $1-C\exp(-c_1m)$ as long as $m>c_0n$ for some constants $C,c_0,c_1,c_2>0$. Consequently, on the event that (58) holds, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^m\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}|^2\right)-\frac{1}{m}\sum_{i=1}^m\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2\right)\right|\\
\le&\frac{1}{m}\sum_{i=1}^m\left|\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}|^2\right)-\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2\right)\right|\\
\le&\frac{1}{\delta}\cdot\frac{1}{m}\sum_{i=1}^m\left|\boldsymbol{a}_i^T(\boldsymbol{h}\boldsymbol{h}^T-\boldsymbol{h}_0\boldsymbol{h}_0^T)\boldsymbol{a}_i\right|\qquad\text{because of (57)}\\
\le&\frac{1}{\delta}\cdot c_2\|\boldsymbol{h}\boldsymbol{h}^T-\boldsymbol{h}_0\boldsymbol{h}_0^T\|_F\qquad\text{because of (58)}\\
\le&\frac{1}{\delta}\cdot 3c_2\|\boldsymbol{h}-\boldsymbol{h}_0\|\cdot\|\boldsymbol{h}\|\\
\le&\frac{3c_2\epsilon}{\delta}\|\boldsymbol{h}\|^2,
\end{aligned}$$

where the second to last inequality is due to (Chen and Candès, 2015, Lemma 2).

On the intersection of the events over which (56) and (58) hold respectively, we have

$$\frac{1}{m}\sum_{i=1}^m\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}|^2\right)\le(0.13+\epsilon+3c_2\epsilon/\delta)\,\|\boldsymbol{h}\|^2, \tag{59}$$

for all $\boldsymbol{h}$ with $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$.

For the case when $\|\boldsymbol{h}'\| < \frac{1}{10}\|\boldsymbol{x}\|$, $\boldsymbol{h}' = \omega\boldsymbol{h}$ for some $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$ and $0 < \omega < 1$. By the definition of $\chi_i(\cdot)$, it can be verified that

$$\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}'|^2) = \chi_i(|\boldsymbol{a}_i^T(\omega\boldsymbol{h})|^2) \le \omega^2\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2). \tag{60}$$

Applying (59), on the same event over which (56) and (58) hold, we have

$$\frac{1}{m}\sum_{i=1}^{m}\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}'|^2\right) \le (0.13 + \epsilon + 3c_2\epsilon/\delta)\,\|\boldsymbol{h}'\|^2, \tag{61}$$

for all $\|\boldsymbol{h}'\| < \frac{1}{10}\|\boldsymbol{x}\|$. Since ϵ can be arbitrarily small, the proof is completed.

C.2 Proof of Lemma 8

Denote $v_i := \boldsymbol{a}_i^T\boldsymbol{z} - |\boldsymbol{a}_i^T\boldsymbol{x}|\mathrm{sgn}(\boldsymbol{a}_i^T\boldsymbol{z})$ for $i = 1, \cdots, m$. Then

$$\nabla\ell(\boldsymbol{z}) = \frac{1}{m}\boldsymbol{A}^T\boldsymbol{v}, \tag{62}$$

where $\boldsymbol{A}$ is a matrix with each row being $\boldsymbol{a}_i^T$ and $\boldsymbol{v} = [v_1, \cdots, v_m]^T$. Thus,

$$\|\nabla\ell(\boldsymbol{z})\| = \left\|\frac{1}{m}\boldsymbol{A}^T\boldsymbol{v}\right\| \le \frac{1}{m}\|\boldsymbol{A}\| \cdot \|\boldsymbol{v}\| \le (1+\delta)\frac{\|\boldsymbol{v}\|}{\sqrt{m}} \tag{63}$$

as long as $m \ge c_1 n$ for some sufficiently large $c_1 > 0$, where the spectral norm bound $\|\boldsymbol{A}\| \le \sqrt{m}(1+\delta)$ follows from (Vershynin, 2012, Theorem 5.32).

We next bound $\|\boldsymbol{v}\|$. Let $\boldsymbol{v} = \boldsymbol{v}^{(1)} + \boldsymbol{v}^{(2)}$, where $v_i^{(1)} = \boldsymbol{a}_i^T\boldsymbol{h}$ and $v_i^{(2)} = 2\boldsymbol{a}_i^T\boldsymbol{x}\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{z})(\boldsymbol{a}_i^T\boldsymbol{x})<0\}}$. By the triangle inequality, we have $\|\boldsymbol{v}\| \le \|\boldsymbol{v}^{(1)}\| + \|\boldsymbol{v}^{(2)}\|$. Furthermore, given $m > c_0 n$, following from (Candès et al., 2013, Lemma 3.1), with probability at least $1 - \exp(-cm)$, we have

$$\frac{1}{m}\|\boldsymbol{v}^{(1)}\|^2 = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2 \le (1+\delta)\|\boldsymbol{h}\|^2. \tag{64}$$

By Lemma 7, we have with probability at least $1 - C\exp(-c_1 m)$

$$\frac{1}{m}\|\boldsymbol{v}^{(2)}\|^2 = \frac{1}{m}\sum_{i=1}^{m}4(\boldsymbol{a}_i^T\boldsymbol{x})^2 \cdot \mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}} \le 4(0.13+\epsilon)\|\boldsymbol{h}\|^2. \tag{65}$$

Hence,

$$\frac{\|\boldsymbol{v}\|}{\sqrt{m}} \le \left(\sqrt{1+\delta} + 2\sqrt{0.13+\epsilon}\right)\|\boldsymbol{h}\|. \tag{66}$$

This concludes the proof.

Appendix D. Proof of Theorem 3

The initialization analysis is similar to Appendix B and is hence omitted. To analyze the gradient loop, we consider the following two regimes.

• **Regime 1**: $c_4\|\boldsymbol{z}\| \geq \|\boldsymbol{h}\| \geq c_3\frac{\|\boldsymbol{w}\|}{\sqrt{m}}$. In this regime, the error contraction by each gradient descent step is given by

$$\operatorname{dist}\left(\boldsymbol{z}-\mu\nabla\ell(\boldsymbol{z}),\boldsymbol{x}\right)\leq(1-\rho)\operatorname{dist}(\boldsymbol{z},\boldsymbol{x}). \tag{67}$$

It suffices to justify that $\nabla\ell(\boldsymbol{z})$ satisfies the RC. We first have

$$\begin{aligned}\nabla\ell(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z}-y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i \\ &= \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z}-|\boldsymbol{a}_i^T\boldsymbol{x}|\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i}_{\nabla^{clean}\ell(\boldsymbol{z})} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(w_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i}_{\nabla^{noise}\ell(\boldsymbol{z})}.\end{aligned} \tag{68}$$

All the proofs for Lemmas 6, 7 and 8 are still valid for $\nabla^{clean}\ell(\boldsymbol{z})$, and thus we have

$$\langle\nabla^{clean}\ell(\boldsymbol{z}),\boldsymbol{h}\rangle\geq 0.74\|\boldsymbol{h}\|^2, \tag{69}$$

$$\left\|\nabla^{clean}\ell(\boldsymbol{z})\right\|\leq 2(1+\delta)\|\boldsymbol{h}\|. \tag{70}$$

Next, we analyze the contribution of the noise. Let $\tilde{w}_i = w_i\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}$, and then for sufficiently large m/n, we have

$$\|\nabla^{noise}\ell(\boldsymbol{z})\| = \left\|\frac{1}{m}\boldsymbol{A}^T\tilde{\boldsymbol{w}}\right\| \leq \left\|\frac{1}{\sqrt{m}}\boldsymbol{A}^T\right\|\left\|\frac{\tilde{\boldsymbol{w}}}{\sqrt{m}}\right\| \leq (1+\delta)\frac{\|\tilde{\boldsymbol{w}}\|}{\sqrt{m}} \leq (1+\delta)\frac{\|\boldsymbol{w}\|}{\sqrt{m}}, \tag{71}$$

where the second inequality is due to the spectral norm bound $\|\boldsymbol{A}\|\leq\sqrt{m}(1+\delta)$ (Vershynin, 2012, Theorem 5.32). Given the regime condition $\|\boldsymbol{h}\|\geq c_3\frac{\|\boldsymbol{w}\|}{\sqrt{m}}$, we further have

$$\|\nabla^{noise}\ell(\boldsymbol{z})\|\leq\frac{(1+\delta)}{c_3}\|\boldsymbol{h}\|, \tag{72}$$

$$\left|\left\langle\nabla^{noise}\ell(\boldsymbol{z}),\boldsymbol{h}\right\rangle\right|\leq\left\|\nabla^{noise}\ell(\boldsymbol{z})\right\|\cdot\|\boldsymbol{h}\|\leq\frac{(1+\delta)}{c_3}\|\boldsymbol{h}\|^2. \tag{73}$$

Combining these together, one has

$$\langle\nabla\ell(\boldsymbol{z}),\boldsymbol{h}\rangle\geq\left\langle\nabla^{clean}\ell(\boldsymbol{z}),\boldsymbol{h}\right\rangle-\left|\left\langle\nabla^{noise}\ell(\boldsymbol{z}),\boldsymbol{h}\right\rangle\right|\geq\left(0.74-\frac{(1+\delta)}{c_3}\right)\|\boldsymbol{h}\|^2, \tag{74}$$

and

$$\|\nabla\ell(\boldsymbol{z})\|\leq\left\|\nabla^{clean}\ell(\boldsymbol{z})\right\|+\left\|\nabla^{noise}\ell(\boldsymbol{z})\right\|\leq(1+\delta)\left(2+\frac{1}{c_3}\right)\|\boldsymbol{h}\|. \tag{75}$$

The RC is guaranteed if μ,λ are chosen properly, c_3 is sufficiently large, and δ is sufficiently small.

• **Regime 2**: Once the iterate enters the regime with $\|\boldsymbol{h}\| \le \frac{c_3\|\boldsymbol{w}\|}{\sqrt{m}}$, the gradient descent update may not reduce the estimation error. However, in this regime, the size $\mu\nabla\ell(\boldsymbol{z})$ of each move is at most $\mathcal{O}(\|\boldsymbol{w}\|/\sqrt{m})$. Then the estimation error cannot increase by more than $\|\boldsymbol{w}\|/\sqrt{m}$ with a constant factor. Thus, one has

$$\operatorname{dist}\left(\boldsymbol{z}+\mu\nabla\ell(\boldsymbol{z}),\boldsymbol{x}\right) \le c_5\frac{\|\boldsymbol{w}\|}{\sqrt{m}} \tag{76}$$

for some constant c_5. As long as $\|\boldsymbol{w}\|/\sqrt{m}$ is sufficiently small, it is guaranteed that $c_5\frac{\|\boldsymbol{w}\|}{\sqrt{m}} \le c_4\|\boldsymbol{x}\|$. If the iterate jumps out of *Regime 2*, it falls back into *Regime 1*.

Appendix E. Convergence of Mini-batch IRWF and Kaczmarz-PR

E.1 Proof of Theorem 4

Without loss of generality, we assume $\boldsymbol{z}^{(t)}$ is in the neighborhood of $\boldsymbol{x}$ (otherwise it is in the neighborhood of $-\boldsymbol{x}$). Let $\boldsymbol{h}=\boldsymbol{z}^{(t)}-\boldsymbol{x}$. We follow the notations in Appendix C and let $\mathcal{S}=\{i:(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)})<0\}$. Then we have

$$\begin{aligned}
&\mathrm{E}_{\Gamma_t}\left[\operatorname{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right)\right]\\
&=\mathrm{E}_{\Gamma_t}\left[\left\|\boldsymbol{z}^{(t)}-\mu\boldsymbol{A}_{\Gamma_t}^T\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-\boldsymbol{y}_{\Gamma_t}\odot\operatorname{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)-\boldsymbol{x}\right\|^2\right]\\
&=\|\boldsymbol{h}\|^2-2\mu\mathrm{E}_{\Gamma_t}\left[\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-\boldsymbol{y}_{\Gamma_t}\odot\operatorname{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)^T\boldsymbol{A}_{\Gamma_t}\boldsymbol{h}\right]\\
&\quad+\mu^2\mathrm{E}_{\Gamma_t}\left[\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-\boldsymbol{y}_{\Gamma_t}\odot\operatorname{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)^T\boldsymbol{A}_{\Gamma_t}\boldsymbol{A}_{\Gamma_t}^T\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-\boldsymbol{y}_{\Gamma_t}\odot\operatorname{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)\right]\\
&\overset{(a)}{=}\|\boldsymbol{h}\|^2-\frac{2\mu k}{m}\sum_{i=1}^m\boldsymbol{a}_i^T\boldsymbol{h}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}-y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)+\frac{\mu^2k}{m}\sum_{i=1}^m\|\boldsymbol{a}_i\|^2\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}-y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)^2\\
&=\|\boldsymbol{h}\|^2-\frac{2\mu k}{m}\left(\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{h})^2+\sum_{i\in\mathcal{S}}2(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})\right)\\
&\quad+\frac{\mu^2k}{m}\left(\sum_{i=1}^m\|\boldsymbol{a}_i\|^2(\boldsymbol{a}_i^T\boldsymbol{h})^2+4\sum_{i\in\mathcal{S}}\|\boldsymbol{a}_i\|^2(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)})\right)\\
&\le\|\boldsymbol{h}\|^2-\frac{2\mu k}{m}\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{h})^2+\frac{4\mu k}{m}\sum_{i\in\mathcal{S}}\left|(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})\right|+\frac{\mu^2k}{m}\sum_{i=1}^m\|\boldsymbol{a}_i\|^2(\boldsymbol{a}_i^T\boldsymbol{h})^2,
\end{aligned}\tag{77}$$

where (a) is due to the fact that Γ_t is uniformly chosen from all subsets of $\{1,2,\dots,m\}$ with the cardinality k.

By Lemma 6, we have that if $m\ge c_0\epsilon^{-2}n$ then with probability $1-2\exp(-c_1m\epsilon^2)$

$$(1-\epsilon)\|\boldsymbol{h}\|^2\le\frac{1}{m}\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{h})^2\le(1+\epsilon)\|\boldsymbol{h}\|^2.$$

holds for all vectors $\boldsymbol{h}$. By Lemma 7, we have that with probability $1-C\exp(-c_1m\epsilon^2)$

$$\frac{1}{m}\sum_{i\in\mathcal{S}}\left|(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})\right|\le(0.13+\epsilon)\|\boldsymbol{h}\|^2$$

holds for all $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| \le \frac{1}{10}$.

Define an event $E_1 := \{\max_{1\le i\le m} \|\boldsymbol{a}_i\|^2 \le 6n\}$. It can be shown that E_1 holds with probability $1-m\exp(-1.5n)$. Then on the event E_1, Equation (77) is further upper bounded by

$$\begin{aligned}\mathrm{E}_{\Gamma_t}\left[\mathrm{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right)\right] &\le \left(1-2\mu k(1-\epsilon)+4\mu k(0.13+\epsilon)+\mu^2 k\cdot 6n(1+\epsilon)\right)\|\boldsymbol{h}\|^2 \\ &\le \left(1-2\mu k(0.74-3\epsilon-3n\mu(1+\epsilon))\right)\|\boldsymbol{h}\|^2. \end{aligned} \tag{78}$$

By choosing the step size $\mu \le \frac{0.24}{n}$, the proposition is proved.

E.2 Proof of Theorem 5

Without loss of generality, we assume the $\boldsymbol{z}^{(t)}$ is in the neighborhood of $\boldsymbol{x}$ (otherwise it is in the neighborhood of $-\boldsymbol{x}$). Let $\boldsymbol{h} = \boldsymbol{z}^{(t)} - \boldsymbol{x}$. We follow the notations in Appendix C and let $\mathcal{S} = \{i : (\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}) < 0\}$. Then we have

$$\begin{aligned}&\mathrm{E}_{i_t}\mathrm{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right) \\ &= \mathrm{E}_{i_t}\left\|\boldsymbol{z}^{(t)} - \frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)} - y_{i_t}\cdot\frac{\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}|}\right)\boldsymbol{a}_{i_t} - \boldsymbol{x}\right\|^2 \\ &= \|\boldsymbol{h}\|^2 - 2\mathrm{E}_{i_t}\frac{(\boldsymbol{a}_{i_t}^T\boldsymbol{h})}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)} - y_{i_t}\cdot\frac{\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}|}\right) + \mathrm{E}_{i_t}\frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)} - y_{i_t}\cdot\frac{\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}|}\right)^2 \\ &\overset{(a)}{=} \|\boldsymbol{h}\|^2 - \frac{2}{m}\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})}{\|\boldsymbol{a}_i\|^2}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)} - y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right) + \frac{1}{m}\sum_{i=1}^m\frac{1}{\|\boldsymbol{a}_i\|^2}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)} - y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)^2 \\ &= \|\boldsymbol{h}\|^2 - \frac{2}{m}\left(\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2} + \sum_{i\in\mathcal{S}}\frac{2(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})}{\|\boldsymbol{a}_i\|^2}\right) + \frac{1}{m}\left(\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2} + 4\sum_{i\in\mathcal{S}}\frac{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)})}{\|\boldsymbol{a}_i\|^2}\right) \\ &= \|\boldsymbol{h}\|^2 - \frac{1}{m}\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2} + \frac{4}{m}\sum_{i\in\mathcal{S}}\frac{(\boldsymbol{a}_i^T\boldsymbol{x})^2}{\|\boldsymbol{a}_i\|^2} \end{aligned} \tag{79}$$

where (a) is due to the fact that i_t is sampled uniformly at random from $\{1,2,\cdots,m\}$. By the special case of Lemma 5.20 in (Vershynin, 2012), $\{\sqrt{n}\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|}\}_{i=1}^m$ are independent isotropic random vectors in $\mathbb{R}^n$ and hence

$$\mathrm{E}\left[n\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2}\right] = \|\boldsymbol{h}\|^2.$$

Moreover, $\{\sqrt{n}\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|}\}_{i=1}^m$ are sub-Gaussian and the sub-Gaussian norm is bounded by an absolute constant. Thus, we have that if $m \ge c_0\epsilon^{-2}n$, then with probability $1-2\exp(-c_1m\epsilon^2)$,

$$\frac{1}{m}\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2} \ge \frac{(1-\epsilon)}{n}\|\boldsymbol{h}\|^2.$$

holds for all vectors $\boldsymbol{h}$. By Lemma 7, we have that with probability $1-C\exp(-c_1m\epsilon^2)$

$$\frac{1}{m}\sum_{i\in\mathcal{S}}\left|\boldsymbol{a}_i^T\boldsymbol{x}\right|^2 \le \frac{1}{m}\sum_{i=1}^m\left|\boldsymbol{a}_i^T\boldsymbol{h}\right|^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le (0.13+\epsilon)\|\boldsymbol{h}\|^2$$

holds for all $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| \le \frac{1}{10}$.

Define an event $E_2 := \{\min_{1\le i\le m} \|\boldsymbol{a}_i\|^2 \ge \frac{2}{3}n\}$. It can be shown that $\mathbb{P}\{E_2\} \ge 1 - m\exp(-n/12)$. Then on the event E_2, (79) is further upper bounded by

$$\mathrm{E}_{i_t}\left[\mathrm{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right)\right] \le \left(1 - \frac{1-\epsilon}{n} + \frac{6(0.13+\epsilon)}{n}\right)\|\boldsymbol{h}\|^2 \le \left(1 - \frac{0.22-7\epsilon}{n}\right)\|\boldsymbol{h}\|^2, \quad (80)$$

which concludes the proof.

References

Anima Anandkumar, Prateek Jain, Yang Shi, and Uma Naresh Niranjan. Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations. In *Artificial Intelligence and Statistics*, pages 268–276, 2016.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *COLT*, pages 113–149, 2015.

Afonso S Bandeira, Jameson Cahill, Dustin G Mixon, and Aaron A Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 37(1):106–125, 2014.

Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *29th Annual Conference on Learning Theory*, 2016.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4): 2355–2377, 2016.

James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3): 751–779, 2005.

Tony T Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

E. J. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5): 1017–1026, 2014.

Emmanuel J Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Yuxin Chen and Emmanuel Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

Yuejie Chi and Yue M Lu. Kaczmarz method for solving quadratic equations. *IEEE Signal Processing Letters*, 23(9):1183–1187, 2016.

James D Donahue. Products and quotients of random variables and their applications. Technical report, DTIC Document, 1964.

Jan Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

James R Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982.

Fajwel Fogel, Irène Waldspurger, and Alexandre dAspremont. Phase retrieval for imaging problems. *Mathematical programming computation*, 8(3):311–335, 2016.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT)*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Ralph W Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237, 1972.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Moritz Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE, 2014.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 2013.

Halyun Jeong and C. Sinan Gunturk. Convergence of the randomized Kaczmarz method for phase retrieval. *arXiv preprint arXiv:1706.10291*, 2017.

Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4520–4528, 2016.

Stefan Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de lAcademie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980 –2998, June 2010.

Krzysztof C Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

Ritesh Kolte and Ayfer Özgür. Phase retrieval via incremental truncated Wirtinger flow. *arXiv preprint arXiv:1606.03196*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

A Ya Kruger. On Fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3): 3325–3358, 2003.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016a.

Kiryung Lee, Yanjun Li, Marius Junge, and Yoram Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transactions on Information Theory*, 2016b.

Gen Li, Yuantao Gu, and Yue M Lu. Phase retrieval using iterative projections: Dynamics in the large systems limit. In *The 53rd Annual Allerton Conference on Communication, Control, and Computing*, 2015.

Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.

Yuanxin Li, Yuejie Chi, Huishuai Zhang, and Yingbin Liang. Non-convex low-rank matrix recovery from corrupted random linear measurements. In *Sampling Theory and Applications (SampTA), International Conference on*, 2017.

Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344, 1999.

Jianwei Miao, Tetsuya Ishikawa, Qun Shen, and Thomas Earnest. Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annu. Rev. Phys. Chem.*, 59:387–410, 2008.

Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1-2): 549–573, 2016.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM Journal on Imaging Sciences*, 8(1):331–372, 2015.

Dohyung Park, Anastasios Kyrillidis, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable non-convex projected gradient descent for a class of constrained matrix optimization problems. *arXiv preprint arXiv:1606.01316*, 2016.

Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 2332–2341, 2015.

Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery using nonconvex optimization. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized Kaczmarz: Theoretical guarantees. *arXiv preprint arXiv:1706.09993*, 2017.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, pages 210 – 268, 2012.

Irène Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *arXiv preprint arXiv:1609.03088*, 2016.

Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *arXiv preprint arXiv:1605.08285*, 2016.

Ke Wei. Solving systems of phaseless equations via Kaczmarz methods: a proof of concept study. *Inverse Problems*, 31(12):125008, 2015.

Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Median-truncated nonconvex approach for phase retrieval with outliers. In *Proceedings of International Conference on Machine Learning (ICML)*, 2016.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Anastasios Zouzias and Nikolaos M Freris. Randomized extended Kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34(2):773–793, 2013.