# Approximate message passing for amplitude based optimization

**Junjie Ma** [* 1]   **Ji Xu** [* 2]   **Arian Maleki** [3]

## Abstract

We consider an $\ell_2$-regularized non-convex optimization problem for recovering signals from their noisy phaseless observations. We design and study the performance of a message passing algorithm that aims to solve this optimization problem. We consider the asymptotic setting $m, n \to \infty$, $m/n \to \delta$ and obtain sharp performance bounds, where $m$ is the number of measurements and $n$ is the signal dimension. We show that for complex signals the algorithm can perform accurate recovery with only $m = \left( \frac{64}{\pi^2} - 4 \right) n \approx 2.5n$ measurements. Also, we provide sharp analysis on the sensitivity of the algorithm to noise. We highlight the following facts about our message passing algorithm: (i) Adding $\ell_2$ regularization to the non-convex loss function can be beneficial even in the noiseless setting; (ii) spectral initialization has marginal impact on the performance of the algorithm.

## 1. Motivation

Phase retrieval refers to the task of recovering a signal $\boldsymbol{x}_* \in \mathbb{C}^{n \times 1}$ from its $m$ phaseless linear measurements:

$$y_a = \left| \sum_{i=1}^n A_{ai} x_{*,i} \right| + w_a, \quad a = 1, 2, \ldots, m, \quad (1)$$

where $x_{*,i}$ is the $i$th component of $\boldsymbol{x}_*$, $A_{ai} \overset{\text{i.i.d}}{\sim} \mathcal{CN}(0, \frac{1}{m})$ and $w_a \sim \mathcal{N}(0, \sigma_w^2)$ a Gaussian noise. The recent surge of interest has led to a better understanding of the theoretical aspects of this problem. Early theoretical results on phase retrieval, such as PhaseLift (Candès et al., 2013) and

PhaseCut (Waldspurger et al., 2015), are based on semidefinite relaxations. For random Gaussian measurements, a variant of PhaseLift can recover the signal exactly (up to global phase) in the noiseless setting using $O(n)$ measurements (Candès & Li, 2014). A different convex optimization approach for phase retrieval was proposed in Goldstein & Studer (2016) and Bahmani & Romberg (2016). This method does not involve lifting and is computationally more attractive than its SDP-based counterparts. Apart from these convex relaxation approaches, non-convex optimization approaches have recently raised intensive research interests. These algorithms typically consist of a carefully designed initialization step (usually accomplished via a spectral method (Netrapalli et al., 2013)) followed by low-cost iterations such as alternating minimization algorithm (Netrapalli et al., 2013) or gradient descent variants like Wirtinger flow (Candès et al., 2015; Ma et al., 2017), truncated Wirtinger flow (Chen & Candès, 2017), amplitude flow (Wang et al., 2016; Zhang & Liang, 2016), incremental reshaped Wirtinger flow (Zhang et al., 2017) and reweighted amplitude flow (Wang et al., 2017a). Other approaches include Kaczmarz method (Wei, 2015; Chi & Lu, 2016; Tan & Vershynin, 2017; Jeong & Güntürk, 2017), trust region method (Sun et al., 2016), coordinate decent (Zeng & So, 2017), prox-linear (Duchi & Ruan, 2017), Polyak subgradient (Davis et al., 2017), block coordinate decent (Barmherzig & Sun, 2017).

Thanks to such research we now have access to several algorithms, inspired by different ideas, that are theoretically guaranteed to recover $\boldsymbol{x}_*$ exactly in the noiseless setting. Despite all these progresses, there is still a gap between the theoretical understanding of the recovery algorithms and what practitioners would like to know. For instance, for many algorithms, including Wirtinger flow and amplitude flow, the exact recovery is guaranteed with either $cn \log n$ or $cn$ measurements, where $c$ is often a fixed but large constant that does not depend on $n$. In both cases, it is often claimed that the large value of $c$ or the existence of $\log n$ is an artifact of the proving technique and the algorithm is expected to work with $cn$ for a reasonably small value of $c$. Such claims have left many users wondering

Q.1 Which algorithm should we use? The theoretical analyses may not be sharp and many factors may have im-

*Equal contribution [1]Department of Statistics, Columbia University, New York, USA [2]Department of Computer Science, Columbia University, New York, USA [3]Department of Statistics, Columbia University, New York, USA. Correspondence to: Junjie Ma <jm4520@columbia.edu>, Ji Xu <jixu@cs.columbia.edu>, Arian Maleki <arian@stat.columbia.edu>.

pact on the simulations including the distribution of the noise, the true signal $\boldsymbol{x}_*$, and the number of measurements.

Q.2 When can we trust the performance of these algorithms in the presence of noise?

Q.3 What is the impact of initialization schemes, such as spectral initialization?

Researchers have developed certain intuition based on a combination of theoretical and empirical results, to give heuristic answers to these questions. However, as demonstrated in a series of papers in the context of compressed sensing, such folklores are sometimes inaccurate (Zheng et al., 2017). To address Question Q.1, several researchers have adopted the asymptotic framework $m, n \rightarrow \infty$, $m/n \rightarrow \delta$, and provided sharp analyses for the performance of several algorithms (Dhifallah & Lu, 2017; Dhifallah et al., 2017; Abbasi et al., 2017). This line of work studies recovery algorithms that are based on convex optimization. In this paper, we adopt the same asymptotic framework and study the following popular non-convex problem, known as amplitude-based optimization (Zhang & Liang, 2016; Wang et al., 2016):

$$\min_{\boldsymbol{x}} \quad \sum_{a=1}^{m} \left( y_a - |(\boldsymbol{Ax})_a| \right)^2 + \frac{\mu}{2} \|\boldsymbol{x}\|_2^2. \qquad (2)$$

where $(\boldsymbol{Ax})_a$ denotes the $a$-th entry of $\boldsymbol{Ax}$. Note that compared to them, (2) has an extra $\ell_2$-regularizer. Regularization is known to reduce the variance of an estimator and hence is expected to be useful when $\boldsymbol{w} \neq \boldsymbol{0}$. However, as we will clarify later in Section 2, since the loss function $\sum_{a=1}^{m} \left( y_a - |(\boldsymbol{Ax})_a| \right)^2$ is non-convex, regularization can help the iterative algorithm that aims to solve (2) even in the noiseless settings. To answer Q.1 to Q.3, we study a message passing algorithm that aims to solve (2). As a result of our studies, we present sharp characterization of the mean square error (even the constants are sharp) in both noiseless and noisy settings. Furthermore, in simulation section (Section 4.3), we compare our algorithm with other existing methods and present a quantitative characterization of the gain that spectral initialization can offer to our algorithms.

For phase retrieval, a Bayesian GAMP algorithm has been discussed in Schniter & Rangan (2015); Barbier et al. (2017). However, they did not provide rigorous performance analysis, particularly, how they handle the difficulty related to initialization, for which we will provide a solution in this paper. Further, the algorithm in Barbier et al. (2017) is based on the Bayesian framework, and performance analyses of Bayesian algorithms are often very challenging under "non-ideal" situations which the algorithms

are not designed for. This paper considers an AMP algorithm referred as AMP.A for solving the popular optimization problem (2). Contrary to the Bayesian GAMP, the asymptotic performance of AMP.A does not depend on the signal and noise distributions except for their second moments. Further, given the fact that the most popular schemes in practice are iterative algorithms derived for solving non-convex optimization problems, the detailed analyses of AMP.A presented in our paper may also shed light on the performance of these algorithms and suggest new ideas to improve their performances.

## 2. AMP.A Algorithm

Our algorithm is based on the approximate message passing (AMP) framework (Donoho et al., 2009; Bayati & Montanari, 2011), in particular the generalized approximate message passing (GAMP) algorithm developed and analyzed in Rangan (2011) and Javanmard & Montanari (2013). Following the steps proposed in Rangan (2011), we obtain the following algorithm called, *Approximate Message Passing for Amplitude-based optimization* (AMP.A) (the derivation is shown in Appendix B in supplementary). Starting from an initial estimate $\boldsymbol{x}^0 \in \mathbb{C}^{n \times 1}$, AMP.A proceeds as follows for $t \geq 0$:

$$\boldsymbol{p}^t = \boldsymbol{Ax}^t - \frac{\lambda_{t-1}}{\delta} \cdot \frac{g(\boldsymbol{p}^{t-1}, \boldsymbol{y})}{-\text{div}_p(g_{t-1})}, \qquad (3a)$$

$$\boldsymbol{x}^{t+1} = \lambda_t \cdot \left( \boldsymbol{x}^t + \boldsymbol{A}^{\text{H}} \frac{g(\boldsymbol{p}^t, \boldsymbol{y})}{-\text{div}_p(g_t)} \right). \qquad (3b)$$

In these iterations

$$g(p, y) = y \cdot \frac{p}{|p|} - p,$$

and

$$\lambda_t = \frac{-\text{div}_p(g_t)}{-\text{div}_p(g_t) + \mu \left( \tau_t + \frac{1}{2} \right)}, \qquad (3c)$$

$$\tau^t = \frac{1}{\delta} \frac{\tau^{t-1} + \frac{1}{2}}{-\text{div}_p(g_{t-1})} \cdot \lambda_{t-1}. \qquad (3d)$$

In the above, $p/|p|$ at $p = 0$ can be any fixed number and does not affect the performance of AMP.A. Further, the "divergence" term $\text{div}_p(g_t)$ is defined as

$$
\begin{aligned}
\text{div}_p(g_t) &\triangleq \frac{1}{m} \sum_{a=1}^{m} \frac{1}{2} \left( \frac{\partial g(p_a^t, y_a)}{\partial p_a^R} - \mathrm{i} \frac{\partial g(p_a^t, y_a)}{\partial p_a^I} \right) \\
&= \frac{1}{m} \sum_{a=1}^{m} \frac{y_a}{2|p_a^t|} - 1,
\end{aligned}
\qquad (4)
$$

where $p_a^R$ and $p_a^I$ denote the real and imaginary parts of $p_a^t$ respectively (i.e., $p_a^t = p_a^R + \mathrm{i} p_a^I$).

The first point that we would like to discuss here is the benefits of the regularization on AMP.A. Since the optimization problem in (2) is non-convex, iterative algorithms intended to solve it can get stuck at bad local minima. In this regard, regularization can still help AMP.A to escape bad local minima through continuation concept even in the noiseless setting. Continuation is popular in convex optimization for improving the convergence rate of iterative algorithms (Hale et al., 2008), and has been applied to the phase retrieval problem in (Balan, 2016). In continuation we start with a value of $\mu$ for which AMP.A is capable of finding the global minimizer of (2). Then, once AMP.A converges we gradually change $\mu$ towards the target value of $\mu$ for which we want to solve the problem and use the previous fixed point of AMP.A as the initialization for the new AMP.A. We continue this process until we reach the value of $\mu$ we are interested in. For instance, if we would like to solve the noiseless phase retrieval problem then $\mu$ should eventually go to zero so that we do not introduce unnecessary bias.

A more general version of the continuation idea we discussed above is to let $\mu$ change at every iteration (denoted as $\mu^t$), and set $\lambda_t$ according to $\mu^t$:

$$\lambda_t = \frac{-\mathrm{div}_p(g_t)}{-\mathrm{div}_p(g_t) + \mu^t\left(\tau_t + \frac{1}{2}\right)}, \tag{5}$$

This way not only we can automate the continuation process, but also let AMP.A decide which choice of $\mu$ is appropriate at a given stage of the algorithm. Our discussion so far has been heuristic. It is not clear whether and how much the generalized continuation can benefit the algorithm. To give a partial answer to this question, we focus on the following particular continuation strategy: $\mu^t = \frac{1+2\mathrm{div}_p(g_t)}{1+2\tau_t}$ and obtain the following version of AMP.A:

$$\boldsymbol{p}^t = \boldsymbol{A}\boldsymbol{x}^t - \frac{2}{\delta}g(\boldsymbol{p}^{t-1}, \boldsymbol{y}), \tag{6a}$$

$$\boldsymbol{x}^{t+1} = 2\left[-\mathrm{div}_p(g_t) \cdot \boldsymbol{x}^t + \boldsymbol{A}^{\mathrm{H}}g(\boldsymbol{p}^t, \boldsymbol{y})\right]. \tag{6b}$$

Note that this choice of $\mu_t$ removes $\mathrm{div}_p(g_t)$ from the denominator of (3), stabilizes the algorithm, and significantly improves the convergence behavior of AMP.A. A key property of AMP (including GAMP) is that its asymptotic behavior can be characterized exactly via the state evolution platform (Donoho et al., 2009; Bayati & Montanari, 2011; Rangan, 2011). Based on a standard asymptotic framework developed in Bayati & Montanari (2011) we can analyze the state evolution (SE), that captures the performance of AMP.A under the asymptotic framework. We assume that the sequence of instances $\{\boldsymbol{x}_*(n), \boldsymbol{A}(n), \boldsymbol{w}(n)\}$ is a converging sequence defined in Bayati & Montanari (2011). Further, without loss of generality, we assume $\frac{1}{n}\|\boldsymbol{x}_*(n)\|^2 \to \kappa = 1$. Then, roughly

speaking, the estimate $\boldsymbol{x}^t$ can be modeled as $\alpha_t\boldsymbol{x}_* + \sigma_t\boldsymbol{h}$, where $\boldsymbol{h}$ behaves like an iid standard complex normal noise. Further, the scaling constant $\alpha_t$ and the noise standard deviation $\sigma_t$ evolve according to a known deterministic rule, called the state evolution (SE), defined below.

**Definition 1.** *Starting from fixed* $(\alpha_0, \sigma_0^2) \in \mathbb{C} \times \mathbb{R}_+\backslash(0,0)$, *the sequences* $\{\alpha_t\}_{t\geq 1}$ *and* $\{\sigma_t^2\}_{t\geq 1}$ *are generated via the following recursion:*

$$\begin{aligned}\alpha_{t+1} &= \psi_1(\alpha_t, \sigma_t^2), \\ \sigma_{t+1}^2 &= \psi_2(\alpha_t, \sigma_t^2; \delta, \sigma_w^2),\end{aligned} \tag{7}$$

*where* $\psi_1 : \mathbb{C} \times \mathbb{R}_+ \mapsto \mathbb{C}$ *and* $\psi_2 : \mathbb{C} \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ *are respectively given by (with* $\theta_\alpha$ *being the phase of* $\alpha$*):*

$$\psi_1(\alpha, \sigma^2) = e^{i\theta_\alpha} \cdot \int_0^{\frac{\pi}{2}} \frac{|\alpha|\sin^2\theta}{\left(|\alpha|^2\sin^2\theta + \sigma^2\right)^{\frac{1}{2}}}\mathrm{d}\theta, \tag{8a}$$

$$\begin{aligned}\psi_2(\alpha, \sigma^2; \delta, \sigma_w^2) = &\frac{4}{\delta}\left(|\alpha|^2 + \sigma^2 + 1\right) \\ &- \frac{4}{\delta}\int_0^{\frac{\pi}{2}} \frac{2|\alpha|^2\sin^2\theta + \sigma^2}{\left(|\alpha|^2\sin^2\theta + \sigma^2\right)^{\frac{1}{2}}}\mathrm{d}\theta + 4\sigma_w^2.\end{aligned} \tag{8b}$$

The state evolution framework for generalized AMP (GAMP) algorithms (Rangan, 2011) was formally proved in Javanmard & Montanari (2013). To apply the results in (Rangan, 2011; Javanmard & Montanari, 2013) to AMP.A, however, we need two generalizations. First, we need to extend the results to complex-valued models. This is straightforward by applying a complex-valued version of the conditioning lemma introduced in Rangan (2011); Javanmard & Montanari (2013). Second, existing results in Rangan (2011) and Javanmard & Montanari (2013) require the function $g$ to be smooth. Our simulation results in Section 4 show that SE predicts the performance of AMP.A despite the fact that $g$ is not smooth. For theoretical purpose, we use the smoothing idea discussed in Zheng et al. (2017) to prove the connection between the SE equations presented in (7) and the iterations of AMP.A in (6) rigorously. Let $\epsilon > 0$ be a small fixed number,

$$\boldsymbol{p}^t = \boldsymbol{A}\boldsymbol{x}_\epsilon^t - \frac{2}{\delta}g_\epsilon(\boldsymbol{p}^{t-1}, \boldsymbol{y}), \tag{9a}$$

$$\boldsymbol{x}_\epsilon^{t+1} = 2\left[-\mathrm{div}_p(g_{t,\epsilon}) \cdot \boldsymbol{x}_\epsilon^t + \boldsymbol{A}^{\mathrm{H}}g_\epsilon(\boldsymbol{p}^t, \boldsymbol{y})\right], \tag{9b}$$

where $g_\epsilon(\boldsymbol{p}^{t-1}, \boldsymbol{y})$ refers to a vector produced by applying $g_\epsilon : \mathbb{C} \times \mathbb{R}_+ \mapsto \mathbb{C}$ below component-wise:

$$g_\epsilon(p, y) \triangleq y \cdot h_\epsilon(p) - p,$$

where for $p = p_1 + ip_2$, $h_\epsilon(p)$ is defined as $h_\epsilon(p) \triangleq \frac{p_1 + ip_2}{\sqrt{p_1^2 + p_2^2 + \epsilon}}$. Note that as $\epsilon \to 0$, $g_{t,\epsilon} \to g_t$ and hence we expect the iterations of smoothed-AMP.A converge to the iterations of AMP.A.

**Theorem 1** (asymptotic characterization). *Let* $\{\boldsymbol{x}_*(n), \boldsymbol{A}(n), \boldsymbol{w}(n)\}$ *be a converging sequence of instances. For each instance, let* $\boldsymbol{x}^0(n)$ *be an initial estimate independent of* $\boldsymbol{A}(n)$. *Assume that the following hold almost surely*

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{x}_*^{\mathrm{H}}\boldsymbol{x}^0 = \alpha_0 \quad and \quad \lim_{n\to\infty} \frac{1}{n}\|\boldsymbol{x}^0\|^2 = \sigma_0^2 + |\alpha_0|^2.$$

*Let* $\boldsymbol{x}_\epsilon^t(n)$ *be the estimate produced by the smoothed AMP.A initialized by* $\boldsymbol{x}^0(n)$ *(which is independent of* $\boldsymbol{A}(n)$*) and* $\boldsymbol{p}^{-1}(n) = \boldsymbol{0}$. *Let* $\epsilon_1, \epsilon_2, \ldots$ *denote a sequence of smoothing parameters for which* $\epsilon_i \to 0$ *as* $i \to \infty$ *Then, for any iteration* $t \geq 1$, *the following holds almost surely*

$$\lim_{j\to\infty}\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n |x_{\epsilon_j,i}^t(n) - e^{\mathrm{i}\theta_t}x_{*,i}|^2 \tag{10}$$
$$= \mathbb{E}\left[|X^t - e^{\mathrm{i}\theta_t}X_*|^2\right] = \left|1 - |\alpha_t|\right|^2 + \sigma_t^2,$$

*where* $\theta_t = \angle\alpha_t$, $X^t = \alpha_t X_* + \sigma_t H$ *and* $X_* \sim p_X$ *is independent of* $H \sim \mathcal{CN}(0,1)$. *Further,* $\{\alpha\}_{t\geq 1}$ *and* $\{\sigma_t^2\}_{t\geq 1}$ *are determined by* (7) *with initialization* $\alpha_0$ *and* $\sigma_0^2$.

The proof of theorem can be found in Appendix A.2 in supplementary.

## 3. Main results for SE mapping

### 3.1. Convergence of the SE for noiseless model

We now analyze the dynamical behavior of the SE. Before we proceed, we point out that in phase retrieval, one can only hope to recover the signal up to global phase ambiguity (Netrapalli et al., 2013; Candès et al., 2013; 2015), for generic signals without any structure. In light of (10), AMP.A is successful if $|\alpha_t| \to 1$ and $\sigma_t^2 \to 0$ as $t \to \infty$. By analyzing the SE, i.e, the update rule for $(\alpha_t, \sigma_t^2)$ in (8), the following two values of $\delta$ will play critical roles in the analysis:

$$\delta_{\mathrm{AMP}} \triangleq \frac{64}{\pi^2} - 4 \approx 2.48 \quad and \quad \delta_{\mathrm{global}} \triangleq 2.$$

The importance of $\delta_{\mathrm{AMP}}$ and $\delta_{\mathrm{global}}$ is revealed by the following two theorems (proofs can be found in Appendix A.3 and A.4 in supplementary file respectively):

**Theorem 2** (convergence of SE). *Consider the noiseless model where* $\sigma_w^2 = 0$. *If* $\delta > \delta_{\mathrm{AMP}}$, *then for any* $0 < |\alpha_0| \leq 1$ *and* $\sigma_0^2 \leq 1$, *the sequences* $\{\alpha_t\}_{t\geq 1}$ *and* $\{\sigma_t^2\}_{t\geq 1}$ *defined in* (7) *converge to*

$$\lim_{t\to\infty}|\alpha_t| = 1 \quad and \quad \lim_{t\to\infty}\sigma_t^2 = 0.$$

**Theorem 3** (local convergence of SE). *When* $\sigma_w^2 = 0$, *then* $(\alpha, \sigma^2) = (1,0)$ *is a fixed point of the SE in* (8). *Furthermore, if* $\delta > \delta_{\mathrm{global}}$, *then there exist two constants* $\epsilon_1 > 0$
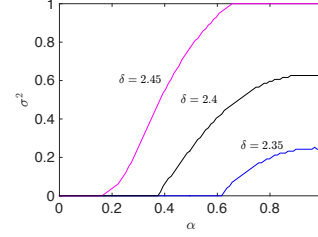


*Figure 1.* The regions below the curves exhibit the basin of attraction of $(\alpha, \sigma^2) = (1,0)$ for different values of $\delta$ respectively (left to right: $\delta = 2.45, 2.4, 2.35$). The results are obtained by running the state evolution (SE) of AMP.A (complex-valued version) with $\alpha_0$ and $\sigma_0^2$ chosen from $100 \times 100$ values equispaced in $[0,1] \times [0,1]$.

*and* $\epsilon_2 > 0$ *such that the SE converges to this fixed point for any* $\alpha_0 \in (1 - \epsilon_1, 1)$ *and* $\sigma_0^2 \in (0, \epsilon_2)$. *On the other hand if* $\delta < \delta_{\mathrm{global}}$, *then the SE cannot converge to* $(1,0)$ *except when initialized there.*

There are a couple of points that we would like to emphasize here:

1. $\alpha_0 \neq 0$ is essential for the success of AMP.A. This can be seen from the fact that $\alpha = 0$ is always a fixed point of $\psi_1(\alpha, \sigma^2)$ for any $\sigma^2 > 0$. From our definition of $\alpha_0$ in Theorem 1, $\alpha_0 = 0$ is equivalent to $\lim_{n\to\infty} \frac{1}{n}\langle\boldsymbol{x}_*, \boldsymbol{x}^0\rangle = 0$. This means that the initial estimate $\boldsymbol{x}^0$ cannot be orthogonal to the true signal vector $\boldsymbol{x}_*$, otherwise there is no hope to recover the signal no matter how large $\delta$ is. This will be discussed in more details in Section 4.1.

2. Fig. 1 exhibits the basin of attraction of $(\alpha, \sigma^2) = (1,0)$ as a function of $\delta$. As expected, the basin of attraction shrinks as $\delta$ decreases. According to Theorem 3, if SE is initialized in the basin of attraction of $(\alpha, \sigma) = (1,0)$, then it still converges to $(\alpha, \sigma^2)$ even if $\delta_{\mathrm{global}} < \delta < \delta_{\mathrm{AMP}}$. However, there are two points we should emphasize here: (i) we find that when $\delta < \delta_{\mathrm{AMP}}$, standard initialization techniques, such as the spectral method, do not help AMP.A much. Again details are discussed in Section 4 . Hence, the question of finding initialization in the basin of attraction of $(\alpha, \sigma^2) = (1,0)$ (when $\delta < \delta_{\mathrm{AMP}}$) remains open for future research. (ii) As $\delta$ decreases from $\delta_{\mathrm{AMP}}$ to $\delta_{\mathrm{global}}$ the basin of attraction of $(\alpha, \sigma^2) = (1,0)$ shrinks.

### 3.2. Noise sensitivity

So far we have only discussed the performance of AMP.A in the ideal setting where the noise is not present in the

measurements. In general, one can use (7) to calculate the asymptotic MSE (AMSE) of AMP.A as a function of the variance of the noise and $\delta$. However, as our next theorem demonstrates it is possible to obtain an explicit and informative expression for AMSE of AMP.A in the high signal-to-noise ratio (SNR) regime.

**Theorem 4** (noise sensitivity). *Suppose that $\delta > \delta_{\mathrm{AMP}} = \frac{64}{\pi^2} - 4$ and $0 < |\alpha_0| \le 1$ and $\sigma_0^2 < 1$. Then, in the high SNR regime, the asymptotic MSE defined by ($\theta_t \triangleq \angle \frac{\boldsymbol{x}_*^{\mathrm{H}} \boldsymbol{x}^t}{n}$)*

$$\mathrm{AMSE}(\delta, \sigma_w^2) \triangleq \lim_{t \to \infty} \frac{\|\boldsymbol{x}^t - e^{\mathrm{i}\theta_t} \boldsymbol{x}_*\|_2^2}{n},$$

*behaves as*

$$\lim_{\sigma_w^2 \to 0} \frac{\mathrm{AMSE}(\sigma_w^2, \delta)}{\sigma_w^2} = \frac{4}{1 - \frac{2}{\delta}}.$$

The proof of this theorem can be found in Appendix E.4 in supplementary. Note that as intuitively expected, as $\delta$ decreases the sensitivity of the algorithm to noise increases. Hence, one should set the number of measurements according to the accepted noise level in the recovered signal.

## 4. Initialization and Simulations

### 4.1. Initialization

As shown in Section 3.1, to achieve successful reconstruction, the initial estimate $\boldsymbol{x}^0$ cannot be orthogonal to the true signal $\boldsymbol{x}_*$, namely,

$$\alpha_0 = \lim_{n \to \infty} \frac{1}{n} \boldsymbol{x}_*^{\mathrm{H}} \boldsymbol{x}^0 \ne 0. \tag{11}$$

In many important applications (e.g., astronomic imaging and crystallography (Millane, 1990)), the signal is known to be real and nonnegative. In such cases, the following initialization of AMP.A meets the non-orthogonality requirement:

$$\boldsymbol{x}^0 = \rho \boldsymbol{1}, \quad \rho \ne 0.$$

(At the same time, we set $g(\boldsymbol{p}^{-1}, \boldsymbol{y}) = \boldsymbol{0}$.) However, finding initializations that satisfy (11) is not straightforward for generic complex-valued signals. Also, random initialization does not necessarily work either, since asymptotically speaking a random vector will be orthogonal to $\boldsymbol{x}_*$. One promising direction to alleviate this issue is the spectral initialization method that was introduced in (Netrapalli et al., 2013) for phase retrieval and subsequently studied in Candès et al. (2015); Chen & Candès (2017); Wang et al. (2016); Lu & Li (2017); Mondelli & Montanari (2017). Specifically, the "direction" of the signal is estimated by the principal eigenvector $\boldsymbol{v}$ ($\|\boldsymbol{v}\|^2 = n$) of the following matrix:

$$\boldsymbol{D} \triangleq \boldsymbol{A}^{\mathrm{H}} \mathrm{diag}\{\mathcal{T}(y_1), \ldots, \mathcal{T}(y_m)\} \boldsymbol{A}, \tag{12}$$

where $\mathcal{T} : \mathbb{R}_+ \to \mathbb{R}$ is a nonlinear processing function, and $\mathrm{diag}\{a_1, \ldots, a_m\}$ is a diagonal matrix with diagonal entries given by $\{a_1, \ldots, a_m\}$. The exact asymptotic performance of the spectral method was characterized in Lu & Li (2017) under some regularity assumptions on $\mathcal{T}$. The analysis in Lu & Li (2017) reveals a phase transition phenomenon: the spectral estimate is not orthogonal to the signal vector $\boldsymbol{x}_*$ (i.e., (11) holds) if and only if $\delta$ is larger than a threshold $\delta_{\mathrm{weak}}$. Later, Mondelli & Montanari (2017) derived the optimal nonlinear processing function $\mathcal{T}$ (in the sense of minimizing $\delta_{\mathrm{weak}}$) and showed that the minimum weak threshold is $\delta_{\mathrm{weak}} = 1$ for the complex-valued model.

The above discussions suggest that the spectral method can provide the required non-orthogonal initialization for AMP.A. However, the naive combination of the spectral estimate with AMP.A will not work. As shown in Figure 2, the performance of AMP.A that is initialized with the spectral method does not follow the state evolution. This is due to the fact that $\boldsymbol{x}^0$ is heavily correlated with the matrix $\boldsymbol{A}$ and violates the assumptions of SE. A trivial remedy is data splitting, i.e, we generate initialization and apply AMP.A on two separate sets of measurements (Netrapalli et al., 2013). However, this simple solution is sub-optimal in terms of sample complexity. To avoid such loss, we propose the following modification to the spectral initialization method, that we call decoupled spectral initialization:

**Decoupled spectral initialization:** Let $\delta > 2$. Set $\boldsymbol{v}$ to be the eigenvector of $\boldsymbol{D}$ corresponding to the largest eigenvalue defined in (12). Let $\boldsymbol{x}^0 = \rho \cdot \boldsymbol{v}$, where $\rho$ is a fixed number which will be discussed later. Define

$$\boldsymbol{p}^0 = (1 - 2\tau \mathcal{T}(\boldsymbol{y})) \circ \boldsymbol{A}\boldsymbol{x}^0, \tag{13}$$

where $\circ$ denotes entry-wise product and $\tau$ is the unique solution of[1]

$$\varphi_1(\delta, \tau) = \frac{1}{\delta}, \quad \tau \in (0, \tau^\star), \tag{14}$$

and $\tau^\star$ is the unique solution of

$$\varphi_2(\delta, \tau^\star) = \frac{1}{\delta}, \quad \tau^\star \in (0, \tau_{\max}), \tag{15}$$

where

$$\varphi_1(\delta, \tau) \triangleq \mathbb{E}\left[(\delta |Z|^2 - 1) \frac{2\tau \mathcal{T}(Y)}{1 - 2\tau \mathcal{T}(Y)}\right], \tag{16a}$$

$$\varphi_2(\delta, \tau) \triangleq \mathbb{E}\left[\left(\frac{2\tau \mathcal{T}(Y)}{1 - 2\tau \mathcal{T}(Y)}\right)^2\right]. \tag{16b}$$

---

[1]The uniqueness of solution in (14) and (15) is guaranteed by our choice of $\mathcal{T}(y)$ in (17)(Lu & Li, 2017; Mondelli & Montanari, 2017). Yet, in noisy case, (14) and (15) can only be calculated precisely if we know the variance of the noise.

The expectations above are over $Z \sim \mathcal{CN}(0, 1/\delta)$ and $Y = |Z| + W$, where $W \sim \mathcal{CN}(0, \sigma_w^2)$ is independent of $Z$.

Now we use $\boldsymbol{x}^0$ and $\boldsymbol{p}^0$ as the initialization for AMP.A. So far, we have not discussed how we can set $\rho$ and $\mathcal{T}$. In this paper, we use the following $\mathcal{T}(y)$ derived by Mondelli & Montanari (2017):

$$\mathcal{T}(y) \triangleq \frac{\delta y^2 - 1}{\delta y^2 + \sqrt{\delta} - 1}. \qquad (17)$$

Note that our initial estimate is given by $\boldsymbol{x}^0 = \rho \cdot \boldsymbol{v}$ (where $\|\boldsymbol{v}\| = \sqrt{n}$). Recall from Theorem 2 that we require $0 < |\alpha_0| < 1$ and $0 \le \sigma_0^2 < 1$ for $\delta > \delta_{\mathrm{AMP}}$. To satisfy this condition, we can simply set $\rho = \|\boldsymbol{y}\|/\sqrt{n}$, which is an accurate estimate of $\|\boldsymbol{x}_*\|/\sqrt{n}$ in the noiseless setting (Lu & Li, 2017)[2]. Under this choice, we have $|\alpha_0|^2 + \sigma_0^2 = \rho^2 = 1$. Hence, as long as $\alpha_0 \ne 0$, we have $0 < |\alpha_0| < 1$ and $0 \le \sigma_0^2 < 1$. The choice we have picked for $\rho$ is not necessarily optimal. We will discuss the optimal spectral initialization and what it can offer to AMP.A in Section 4.3.

In summary, our initialization in (13) intuitively satisfies "enough independency" requirement such that the SE for AMP.A still holds and this is supported by our numerical results in Section 4.3. We have clarified this intuition in Section 4.2. Our numerical experiments in Section 4.3 show that the estimate $\boldsymbol{x}^0$ behaves as if it is independent of the matrix $\boldsymbol{A}$. Our finding is summarized below.

**Finding 1.** *Let $\boldsymbol{x}^0$ and $\boldsymbol{p}^0$ be generated according to (13), and $\{\boldsymbol{x}^t\}_{t \ge 1}$ and $\{\boldsymbol{p}^t\}_{t \ge 1}$ generated by the AMP.A algorithm as described in (6). The AMSE converges to*

$$\lim_{n \to \infty} \frac{1}{n} \|\boldsymbol{x}^t - e^{\mathrm{i}\theta_t} \boldsymbol{x}_*\|_2^2 = (1 - |\alpha_t|)^2 + \sigma_t^2,$$

*where $\theta_t = \angle(\boldsymbol{x}_*^{\mathrm{H}}, \boldsymbol{x}_t)$, $\{|\alpha_t|\}_{t \ge 1}$ and $\{\sigma_t^2\}_{t \ge 1}$ are generated according to (7) and*

$$|\alpha_0|^2 = \frac{1 - \delta\varphi_2(\delta, \tau)}{1 + \delta\varphi_3(\delta, \tau)} \quad \text{and} \quad \sigma_0^2 = 1 - |\alpha_0|^2, \quad (18)$$

*where $\tau$ is the solution to (13) and $\varphi_3$ are defined as ($\varphi_2$ is defined in (16))*

$$\varphi_3(\delta, \tau) \triangleq \mathbb{E}\left[ (\delta|Z|^2 - 1)\left( \frac{2\tau\mathcal{T}(Y)}{1 - 2\tau\mathcal{T}(Y)} \right)^2 \right], \quad (19)$$

*where $Y = |Z| + W$.*

We expect to provide a rigorous proof of this finding in a forthcoming paper.

## 4.2. Intuition of our initialization

Note that in conventional AMP.A, we set initial $g(\boldsymbol{p}^{-1}, \boldsymbol{y}) = \boldsymbol{0}$ and therefore $\boldsymbol{p}^0 = \boldsymbol{A}\boldsymbol{x}^0$. Hence, our modification in (13) appears to be a rescaling procedure of $\boldsymbol{p}^0$. Note that solving the principal eigenvector of $\boldsymbol{D}$ in (12) is equivalent to the following optimization problem:

$$\boldsymbol{v} = \operatorname*{argmin}_{\|\boldsymbol{x}\| = \sqrt{n}} -\sum_{a=1}^m \mathcal{T}(y_a) \cdot \left|(\boldsymbol{A}\boldsymbol{x})_a\right|^2. \qquad (20)$$

Following the derivations proposed in Rangan (2011), we obtain the following approximate message passing algorithm for spectral method (denoted as AMP.S):

$$\hat{\tau}^t = \frac{1}{\delta} \frac{1}{\mathrm{div}_p(h_{t-1})} \cdot \frac{\sqrt{n}}{\|\hat{\boldsymbol{r}}_{t-1}\|}, \qquad (21\mathrm{a})$$

$$\hat{\boldsymbol{p}}^t = \boldsymbol{A}\hat{\boldsymbol{x}}^t - \frac{1}{\delta}\frac{h(\hat{\boldsymbol{p}}^{t-1}, \boldsymbol{y}, \hat{\tau}^{t-1})}{\mathrm{div}_p(h_{t-1})} \cdot \frac{\sqrt{n}}{\|\hat{\boldsymbol{r}}_{t-1}\|}, \qquad (21\mathrm{b})$$

$$\hat{\boldsymbol{r}}^t = \hat{\boldsymbol{x}}^t - \frac{\boldsymbol{A}^{\mathrm{H}} h(\hat{\boldsymbol{p}}^t, \boldsymbol{y}, \hat{\tau}^t)}{\mathrm{div}_p(h_t)}, \qquad (21\mathrm{c})$$

$$\hat{\boldsymbol{x}}^{t+1} = -\frac{\sqrt{n}}{\|\hat{\boldsymbol{r}}_t\|} \cdot \hat{\boldsymbol{r}}^t, \qquad (21\mathrm{d})$$

where we defined $h(\hat{p}, y, \hat{\tau}) \triangleq \frac{2\mathcal{T}(y)}{1 - 2\hat{\tau}\mathcal{T}(y)} \cdot \hat{p}$. The optimizer $\boldsymbol{v}$ of (20) can be regarded as the limit of the estimate $\hat{\boldsymbol{x}}^t$ under correct initialization of AMP.S. Note that AMP.S acts as a proxy and we do not intend to use it for the eigenvector calculations. (There are standard numerical recipes for that purpose.) But, the correction term used in (13) is suggested by the Onsager correction term in AMP.S. To see that let $\hat{\boldsymbol{p}}^\infty$, $\hat{\boldsymbol{x}}^\infty$, $\hat{\tau}^\infty$ represent the limits of $\hat{\boldsymbol{p}}^t$, $\hat{\boldsymbol{x}}^t$, $\hat{\tau}^t$ respectively. Then, from (21a) and (21b), we obtain the following equation

$$\hat{\boldsymbol{p}}^\infty \overset{(a)}{=} \boldsymbol{A}\hat{\boldsymbol{x}}^\infty - \hat{\tau}^\infty h(\hat{\boldsymbol{p}}^\infty, \boldsymbol{y}, \hat{\tau}^\infty),$$

$$\overset{(b)}{=} \boldsymbol{A}\hat{\boldsymbol{x}}^\infty - \hat{\tau}^\infty \underbrace{\frac{2\mathcal{T}(\boldsymbol{y})}{1 - 2\hat{\tau}^\infty \mathcal{T}(\boldsymbol{y})} \circ \hat{\boldsymbol{p}}^\infty}_{\text{Onsager term}} \qquad (22)$$

By solving (22), we obtain (13) with rescaling of $\frac{\|\boldsymbol{y}\|}{\sqrt{n}}$ (since $\hat{\boldsymbol{x}}^\infty = \sqrt{n}\boldsymbol{v}$ and $\boldsymbol{x}^0 = \|\boldsymbol{y}\|\boldsymbol{v}$). Further, (14) and (15) that determine the value of $\hat{\tau}^\infty$ can be simplified through solving the fixed point of the following state evolution of AMP.S:

$$\hat{\alpha} = \frac{\hat{\alpha}\,\varphi_1(\delta, \hat{\tau})}{\sqrt{\hat{\alpha}^2\,\varphi_1^2(\delta, \hat{\tau}) + \frac{1}{\delta}\varphi_2(\delta, \hat{\tau}) + \frac{\hat{\alpha}^2}{\delta}\varphi_3(\delta, \hat{\tau})}}, \qquad (23\mathrm{a})$$

$$1 = \frac{1}{\delta}\frac{1}{\sqrt{\hat{\alpha}^2\,\varphi_1^2(\delta, \hat{\tau}) + \frac{1}{\delta}\varphi_2(\delta, \hat{\tau}) + \frac{\hat{\alpha}^2}{\delta}\varphi_3(\delta, \hat{\tau})}}, \qquad (23\mathrm{b})$$

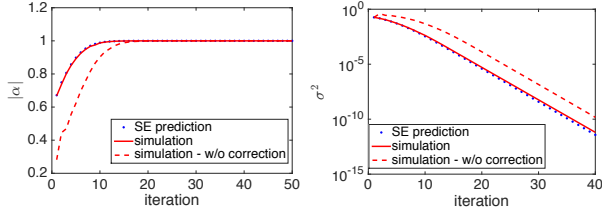where $\varphi_1, \varphi_2$ are defined in (16) and $\varphi_3$ is defined in (19).

*Figure 2.* State evolution prediction for AMP.A with spectral initialization in the noiseless setting. **Left:** predicted and simulated results of $|\alpha|$. **Right:** predicted and simulated results of $\sigma^2$. The solid curves show the simulation results for the proposed initialization, and the dashed curves show the results for a naive approach without the proposed correction (namely, we set $\boldsymbol{p}^0 = \boldsymbol{A}\boldsymbol{x}^0$). In these experiments, $n = 5000$ and $m = 20000$. The optimal $\mathcal{T}$ in (17) is employed.

## 4.3. Simulation results

We now provide simulation results to verify our analysis and compare AMP.A in (6) with existing algorithms. Notice that our analysis of the SE is based on a smoothing idea. Our simulation results in this section show that, for the complex-valued settting, the SE predicts the performance of AMP.A even without smoothing $g$.

### 1) Accuracy of state evolution

We first consider the noiseless setting. Fig. 2 verifies the accuracy of SE predictions of AMP.A together with the proposed initialization (i.e., (13)). The true signal is generated as $\boldsymbol{x}_* \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$. We measure the following two quantities (averaged over 10 runs):

$$\hat{\alpha}_t = \frac{\boldsymbol{x}_*^{\mathrm{H}} \boldsymbol{x}^t}{\|\boldsymbol{x}_*\|^2} \quad \text{and} \quad \hat{\sigma}_t^2 = \frac{\|\boldsymbol{x}^t - \hat{\alpha}_t \boldsymbol{x}_*\|^2}{\|\boldsymbol{x}_*\|^2}.$$

We expect $\hat{\alpha}_t$ and $\hat{\sigma}_t^2$ to converge to their deterministic counterparts $\alpha_t$ and $\sigma_t^2$ (as described in Finding 1). Indeed, Fig. 2 shows that the match between the simulated $\hat{\alpha}_t$ and $\hat{\sigma}_t^2$ (solid curves) and the SE predictions (dotted curves) is precise. For reference, we also include the simulation results for the "blind approach" where the spectral initialization is incorporated into AMP.A without applying the proposed correction (i.e., we use $\boldsymbol{p}^0 = \boldsymbol{A}\boldsymbol{x}^0$ instead of (13)). From Fig. 2, we see that this blind approach deviates significantly from the SE predictions. Note that the blind approach still recovers the signal correctly for the current experiment, albeit $\hat{\sigma}_t^2$ deviates from theoretical predictions. However, we found that (results are not shown here) the blind approach is unstable, and can perform rather poorly for other popular choices of $\mathcal{T}$ (such as the orthogonality-promoting method proposed in (Wang et al., 2016)).

We next consider a noisy setting. In Fig. 3, we plot the simulated MSE and the corresponding SE predictions for two different cases. For the figure on the left panel, the true signal is generated as $\boldsymbol{x}_* \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, and the decoupled spectral initialization discussed in Section 4.1 is used. For the figure on the right panel, the signal is nonnegative and we use the initialization $\boldsymbol{x}^0 = \boldsymbol{1}$ and $\boldsymbol{g}(\boldsymbol{p}^{-1}, \boldsymbol{y}) = \boldsymbol{0}$. The nonnegative signal is generated in the following way: we set 90% of the entries to be zero and remaining 10% to be constants. (Note that the signal is sparse, but the sparsity information is not exploited in the AMP.A algorithm.) The signal-to-noise ratio (SNR) is defined to be $\mathbb{E}[\|\boldsymbol{A}\boldsymbol{x}\|^2]/\mathbb{E}[\|\boldsymbol{w}\|^2]$. The figure displays the following MSE performance:

$$\mathrm{MSE} = \inf_{\theta \in [0, 2\pi)} \frac{\|\boldsymbol{x}^t - e^{\mathrm{i}\theta} \boldsymbol{x}_*\|^2}{\|\boldsymbol{x}_*\|^2}.$$

The SE prediction of the above MSE is given by $(1 - |\alpha_t|)^2 + \sigma_t^2$. Again, we see from Fig. 3 that simulated MSE matches the SE predictions reasonably well. Further, the right figure exhibits larger fluctuations. This is mainly due to the fact that in our experiment the initialization for the right figure is less accurate than that adopted for the left figure.

### 2) Basin of attraction of AMP.A and spectral initialization

In this Section, we aim to address Q.3 we raised in the introduction. As discussed in Section 4.1, the spectral method can provide the required non-orthogonal estimate for AMP.A. Besides that, as discussed in Q.3 in Section 1, it is interesting to see if the spectral method can help AMP.A for $\delta < \delta_{\mathrm{AMP}}$. To answer this, we need to examine whether $(\alpha_0, \sigma_0^2)$ produced by the spectral estimate can fall into the attraction basin of the good fixed point $(\alpha, \sigma^2) = (1, 0)$. Currently, the basin of attraction cannot be analytically characterized, but it can be conveniently computed via SE. Specifically, for a given $(\alpha_0, \sigma_0^2)$, we run the SE for a sufficiently large number of iterations and see if it converges to $(1, 0)$ (up to a pre-defined tolerance).

Fig. 4 plots the basin of attraction of the fixed point $(\alpha, \sigma) = (1, 0)$ for $\delta = 2.4$ or 2.41 (indicated by the blue curve). The straight line is obtained in the following way: From (Lu & Li, 2017), for a given $\delta$ and $\mathcal{T}$, the ratio $\sigma_0/\alpha_0$ can be computed by solving a set of fixed point equations, and this ratio determines a straight line $\sigma/\alpha = \sigma_0/\alpha_0$ in the $\alpha - \sigma$ plane. The red line in Fig. 4 is obtained using $\mathcal{T}$ in (17). The region above the red line can be potentially achieved by certain choices of $\mathcal{T}$ together with linear scaling. On the other hand, no known $\mathcal{T}$ can achieve the region below the red line. As we see in this figure, the spectral estimate cannot fall into the basin of attraction in the current example for $\delta = 2.4$ (left subfigure). The smallest $\delta$ such that two curves intersect is numerically found to be around $\delta = 2.41$ (right subfigure) which is quite close to $\delta_{\mathrm{AMP}} \approx 2.48$. Notice that for $\delta > \delta_{\mathrm{AMP}}$, AMP.A works for any $\alpha_0 \neq 0$. This means that the spectral method can-
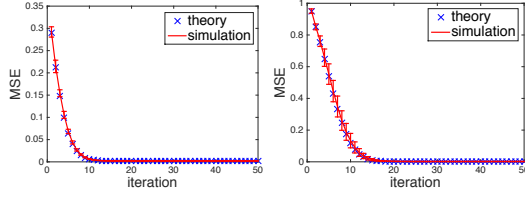
Figure 3. Simulated MSE and SE predictions in noisy settings. The solid curves show the average MSE over 10 runs. The error bars show one standard deviation.
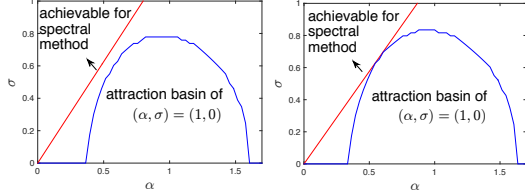


Figure 4. Plot of the attraction basin of AMP.A and the achievable region of the spectral method. **Left:** $\delta = 2.40$. **Right:** $\delta = 2.41$. In this figure, the vertical axis is $\sigma$ instead of $\sigma^2$.

not help AMP.A much besides providing an estimate not orthogonal to the true signal.

### 3) Comparison with existing methods

Fig. 5 displays the success recovery rate of AMP.A and the Gerchberg-Saxton algorithm (GS) (Gerchberg, 1972), truncated Wirtinger flow (TWF) (Chen & Candès, 2017), truncated amplitude flow (TAF) (Wang et al., 2017b), incremental reshaped Wirtinger flow (IRWF) (Zhang et al., 2017) and reweighted amplitude flow (RAF) (Wang et al., 2017a). Notice that the GS algorithm involves solving a least squares problem in each iteration and is thus computationally more expensive than other algorithms. For the figure in the left panel, the signal is $\boldsymbol{x}_* \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ and the initialization is generated via the spectral method with $\mathcal{T}$ defined in (17). For the right panel, the signal is nonnegative (generated in the same way as that in Fig. 3) and the initial estimate is $\boldsymbol{x}^0 = \boldsymbol{1}$ for all algorithms.

We see that AMP.A outperforms all other algorithms except at $\delta = 2.7$ for the figure in the left panel. Based on simulation results not shown in this paper, we find that AMP.A outperforms IRWF consistently for a larger problem size (say $n = 2000$). However, we adopt the current setting where $n = 1000$ for ease of comparison (Chen & Candès, 2017; Wang et al., 2017b; Zhang et al., 2017; Wang et al., 2017a). Comparing the two figures in Fig. 5, we see that all algorithms are quite sensitive to the quality of initialization except for AMP.A. Notice that in the asymptotic setting where $n \to \infty$, AMP.A is able to recover the signal for all $\delta > \delta_{\text{AMP}} \approx 2.48$ based on our SE analysis.
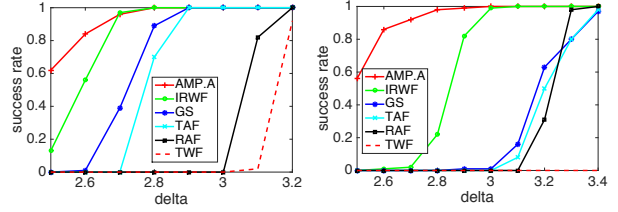


Figure 5. Recovery performance of various algorithms. We fix $n = 1000$ and vary $\delta$. All algorithms have run 1000 iterations. Reconstruction is considered successful if the final AMSE is smaller than $10^{-10}$. The success rates are measured in 100 independent realizations of $\boldsymbol{A}$ and $\boldsymbol{x}_*$. **Left:** spectral initialization with random Gaussian signal. **Right:** $\boldsymbol{x}^0 = \boldsymbol{1}$ and $\boldsymbol{p}^0 = \boldsymbol{A}\boldsymbol{x}^0$. The signal is nonnegative.
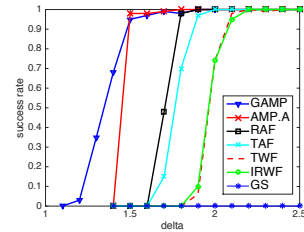


Figure 6. Recovery performance of various algorithms in real-valued case.

Finally, we present simulation results for the real-valued case in Fig. 6. Due to the lack of space, a thorough discussion of the real-valued AMP.A and its state evolution will be reported in a later paper. Yet, in this paper, we want to emphasize two points through Fig. 6. First, we see that AMP.A outperforms the competing algorithms with a clear phase transition between $\delta = 1.4$ and $\delta = 1.5$. This is consistent with our analysis ($\delta_{\text{AMP}} = \frac{\pi^2}{4} - 1$ in real value case). Second, we notice that the IRWF algorithm (which performs best next to AMP.A in Fig. 5) is outperformed by RAF in this case. For reference, we also included the performance of the Bayesian GAMP algorithm Schniter & Rangan (2015); Barbier et al. (2017) (in conjunction with our own proposed decouple initialization to get the best performance of the Bayesian GAMP), under the assumption that the signal distribution (in this case, Gaussian) is perfectly known. As discussed in Section 1, this assumption can be unrealistic in practice. Nevertheless, the performance of Bayesian GAMP algorithm is a meaningful benchmark and hence included in Fig. 5.[3]

---

[3] We also carried out simulations of Bayesian GAMP for the complex-valued case. However, we found that its performance is not competitive under the setting of Fig. 5: its recovery rate is less than 95% at $\delta = 3.5$, even when the MSE threshold is set to $10^{-6}$. (Note that the MSE threshold is $10^{-10}$ for the curves in Fig. 5).

## Acknowledgements

## References

Abbasi, E., Salehi, F., and Hassibi, B. Performance of real phase retrieval. In *International Conference on Sampling Theory and Applications (SampTA)*, July 2017.

Bahmani, S. and Romberg, J. Phase retrieval meets statistical learning theory: A flexible convex relaxation. *arXiv preprint arXiv:1610.04210*, 2016.

Balan, R. Reconstruction of signals from magnitudes of redundant representations: The complex case. *Foundations of Computational Mathematics*, 16(3):677–721, 2016.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. Phase transitions, optimal errors and optimality of message-passing in generalized linear models. *arXiv preprint arXiv:1708.03395*, 2017.

Barmherzig, D. and Sun, J. A local analysis of block coordinate descent for Gaussian phase retrieval. *arXiv preprint arXiv:1712.02083*, 2017.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, Feb 2011.

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

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

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

Chen, Y. and Candès, E. J. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70:822–883, May 2017.

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

Davis, D., Drusvyatskiy, D., and Paquette, C. The non-smooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

Dhifallah, O. and Lu, Y. M. Fundamental limits of PhaseMax for phase retrieval: A replica analysis. *arXiv preprint arXiv:1708.03355*, 2017.

Dhifallah, O., Thrampoulidis, C., and Lu, Y. M. Phase retrieval via linear programming: Fundamental limits and algorithmic improvements. *arXiv preprint arXiv:1710.05234*, 2017.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Duchi, J. C. and Ruan, F. Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. *arXiv preprint arXiv:1705.02356*, 2017.

Gerchberg, R. W. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

Goldstein, T. and Studer, C. PhaseMax: Convex phase retrieval via basis pursuit. *arXiv preprint arXiv:1610.07531*, 2016.

Hale, E. T., Yin, W., and Zhang, Y. Fixed-point continuation for $\ell_1$-minimization: methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

Javanmard, A. and Montanari, A. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115, 2013.

Jeong, H. and Güntürk, C. S. Convergence of the randomized Kaczmarz method for phase retrieval. *arXiv preprint arXiv:1706.10291*, 2017.

Lu, Y. M. and Li, G. Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *arXiv preprint arXiv:1702.06435*, 2017.

Ma, C., Wang, K., Chi, Y., and Chen, Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

Millane, R. P. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

Mondelli, M. and Montanari, A. Fundamental limits of weak recovery with applications to phase retrieval. *arXiv preprint arXiv:1708.05932*, 2017.

Netrapalli, P., Jain, P., and Sanghavi, S. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pp. 2796–2804, 2013.

Rangan, S. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Information Theory Proceedings*, pp. 2168–2172, July 2011.

Schniter, P. and Rangan, S. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2015.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2379–2383, July 2016.

Tan, Y. S. and Vershynin, R. Phase retrieval via randomized Kaczmarz: Theoretical guarantees. *arXiv preprint arXiv:1706.09993*, 2017.

Waldspurger, I., d'spremont, A., and Mallat, S. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

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

Wang, G., Giannakis, G., Saad, Y., and Chen, J. Solving most systems of random quadratic equations. In *Advances in Neural Information Processing Systems*, pp. 1865–1875, 2017a.

Wang, S., Weng, H., and Maleki, A. Which bridge estimator is optimal for variable selection? *arXiv preprint arXiv:1705.08617*, 2017b.

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

Zeng, W.-J. and So, H. Coordinate descent algorithms for phase retrieval. *arXiv preprint arXiv:1706.03474*, 2017.

Zhang, H. and Liang, Y. Reshaped Wirtinger flow for solving quadratic system of equations. In *Advances in Neural Information Processing Systems*, pp. 2622–2630, 2016.

Zhang, H., Zhou, Y., Liang, Y., and Chi, Y. A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18(141):1–35, 2017.

Zheng, L., Maleki, A., Weng, H., Wang, X., and Long, T. Does $\ell_p$-minimization outperform $\ell_1$-minimization? *IEEE Transactions on Information Theory*, PP(99):1–1, 2017.