Approximate Message Passing with Spectral Initialization for Generalized Linear Models

Marco Mondelli
IST Austria

Ramji Venkataramanan
University of Cambridge

Abstract

We consider the problem of estimating a signal from measurements obtained via a generalized linear model. We focus on estimators based on approximate message passing (AMP), a family of iterative algorithms with many appealing features: the performance of AMP in the high-dimensional limit can be succinctly characterized under suitable model assumptions; AMP can also be tailored to the empirical distribution of the signal entries, and for a wide class of estimation problems, AMP is conjectured to be optimal among all polynomial-time algorithms.

However, a major issue of AMP is that in many models (such as phase retrieval), it requires an initialization correlated with the ground-truth signal and independent from the measurement matrix. Assuming that such an initialization is available is typically not realistic. In this paper, we solve this problem by proposing an AMP algorithm initialized with a spectral estimator. With such an initialization, the standard AMP analysis fails since the spectral estimator depends in a complicated way on the design matrix. Our main contribution is a rigorous characterization of the performance of AMP with spectral initialization in the high-dimensional limit. The key technical idea is to define and analyze a two-phase artificial AMP algorithm that first produces the spectral estimator, and then closely approximates the iterates of the true AMP. We also provide numerical results that demonstrate the validity of the proposed approach.

1 Introduction

We consider the problem of estimating a d-dimensional signal $\boldsymbol{x} \in \mathbb{R}^d$ from n i.i.d. measurements $y_i \sim p(y \mid \langle \boldsymbol{x}, \boldsymbol{a}_i \rangle)$, $i \in \{1, \dots, n\}$, where $\langle \cdot, \cdot \rangle$ is the scalar product, $\{\boldsymbol{a}_i\}_{1 \le i \le n}$ are given sensing vectors, and the (stochastic) output function $p(\cdot \mid \langle \boldsymbol{x}, \boldsymbol{a}_i \rangle)$ is a given probability distribution. This is known as a *generalized linear model* (McCullagh, 2018), and it encompasses many settings of interest in statistical estimation and signal processing (Rangan and Goyal, 2001; Boufounos and Baraniuk, 2008; Yang et al., 2012; Eldar and Kutyniok, 2012). One notable example is the problem of phase retrieval (Fienup, 1982; Shechtman et al., 2015), where $y_i = |\langle \boldsymbol{x}, \boldsymbol{a}_i \rangle|^2 + w_i$, with w_i being noise. Phase retrieval appears in several areas of science and engineering, see e.g. (Fienup and Dainty, 1987; Millane, 1990; Demanet and Jugnon, 2017), and the last few years have witnessed a surge of interest in the design and analysis of efficient algorithms; see the review by Fannjiang and Strohmer (2020) and the discussion at the end of this section.

Here, we consider generalized linear models (GLMs) in the high-dimensional setting where $n, d \to \infty$, with their ratio tending to a fixed constant, i.e., $n/d \to \delta \in \mathbb{R}$. We focus on a family of iterative algorithms known as approximate message passing (AMP). AMP algorithms are derived via approximations of belief propagation on the factor graph representing the estimation problem. AMP algorithms were first proposed for estimation in linear models (Donoho et al., 2009; Bayati and Montanari, 2011), and for estimation in GLMs by Rangan (2011). AMP has since been applied to a wide range of high-dimensional statistical estimation problems including compressed sensing (Krzakala et al., 2012; Bayati and Montanari, 2012; Maleki et al., 2013), low rank matrix estimation (Rangan and Fletcher, 2012; Deshpande and Montanari, 2014; Kabashima et al., 2016), group synchronization (Perry et al., 2018), and specific instances of GLMs such as logistic regression (Sur and Candès, 2019) and phase retrieval (Schniter and Rangan, 2014; Ma et al., 2019; Maillard et al., 2020).

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Starting from an initialization $\boldsymbol{x}^0 \in \mathbb{R}^d$, the AMP algorithm for GLMs produces iteratively refined estimates of the signal, denoted by $\boldsymbol{x}^t$, for $t \geq 1$. An appealing feature of AMP is that, under suitable model assumptions, its performance in the high-dimensional limit can be precisely characterized by a succinct deterministic recursion called *state evolution* (Bayati and Montanari, 2011; Bolthausen, 2014; Javanmard and Montanari, 2013). Specifically, in the high-dimensional limit, the empirical distribution of the estimate $\boldsymbol{x}^t$ converges to the law of the random variable $\mu_t X + \sigma_t W_t$, for $t \geq 1$. Here $X \sim P_X$ (the signal prior), and $W_t \sim \mathsf{N}(0, 1)$ is independent of X. The state evolution recursion specifies how the constants (μ_t, σ_t) can be computed from $(\mu_{t-1}, \sigma_{t-1})$ (see Sec. 3 for details).

Using the state evolution analysis, it has been shown that AMP provably achieves Bayes-optimal performance in some special cases (Donoho et al., 2013; Deshpande and Montanari, 2014; Montanari and Venkataramanan, 2021). Indeed, a conjecture from statistical physics posits that AMP is optimal among all polynomial-time algorithms. The optimality of AMP for generalized linear models is discussed by Barbier et al. (2019).

However, when used for estimation in GLMs, a major issue in current AMP theory is that in many problems (including phase retrieval) we require an initialization $\boldsymbol{x}^0$ that is correlated with the unknown signal $\boldsymbol{x}$ but independent of the sensing vectors $\{\boldsymbol{a}_i\}$. It is often not realistic to assume that such a realization is available. For such GLMs, without a correlated initialization, asymptotic state evolution analysis predicts that the AMP estimates will be uninformative, i.e., their normalized correlation with the signal vanishes in the large system limit. Thus, developing an AMP theory that does not rely on unrealistic assumptions about the initialization is an important open problem.

In this paper, we solve this open problem for a wide class of GLMs by rigorously analyzing the AMP algorithm with a *spectral estimator*. The idea of using a spectral estimator for GLMs was introduced by Li (1992), and its performance in the high-dimensional limit was recently characterized by Lu and Li (2019) and Mondelli and Montanari (2019). It was shown that the normalized correlation of the spectral estimator with the signal undergoes a phase transition, and for the special case of phase retrieval, the threshold for strictly positive correlation with the signal matches the information-theoretic threshold (Mondelli and Montanari, 2019).

Our main technical contribution is a novel analysis of AMP with spectral initialization for GLMs, under the assumption that the sensing vectors $\{\boldsymbol{a}_i\}$ are i.i.d. Gaussian. This yields a rigorous characterization of the performance in the high-dimensional limit (Theorem 1). The analysis of AMP with spectral initialization is far from obvious since the spectral estimator depends in a non-trivial way on the sensing vectors $\{\boldsymbol{a}_i\}$. The existing state evolution analysis for GLMs (Rangan, 2011; Javanmard and Montanari, 2013) crucially depends on the AMP initialization being independent of the sensing vectors, and therefore cannot be directly applied.

At the center of our approach is the design and analysis of an *artificial AMP* algorithm. The artificial AMP operates in two phases: in the first phase, it performs a power method, so that its iterates approach the spectral initialization of the true AMP; in the second phase, its iterates are designed to remain close to the iterates of the true AMP. The initialization of the artificial AMP is correlated with $\boldsymbol{x}$, but independent of the sensing vectors $\{\boldsymbol{a}_i\}$, which allows us to apply the standard state evolution analysis. Note that the initialization of the artificial AMP is impractical (it requires the knowledge of the unknown signal $\boldsymbol{x}$!). However, this is not an issue, since the artificial AMP is employed solely as a proof technique: we prove a state evolution result for the true AMP by showing that its iterates are close to those in the second phase of the artificial AMP.

Initializing AMP with a (different) spectral method has been recently shown to be effective for low-rank matrix estimation (Montanari and Venkataramanan, 2021). However, our proof technique for analyzing spectral initialization for GLMs is different from the approach by Montanari and Venkataramanan (2021). The argument in that paper is specific to the spiked random matrix model and relies on a delicate decoupling argument between the outlier eigenvectors and the bulk. Here, we follow an approach developed by Mondelli et al. (2020), where a specially designed AMP is used to establish the joint empirical distribution of the signal, the spectral estimator, and the linear estimator.

For the case of phase retrieval, Ma et al. (2018) provided a heuristic argument for the validity of spectral initialization, and stated that establishing a rigorous proof is an open problem. Our paper not only solves this open problem, but it also gives a provable initialization method valid for a class of GLMs.

We note that, for some GLMs, AMP does not require a special initialization that is correlated with the signal $\boldsymbol{x}$. In Section 3, we give a condition on the GLM output function that specifies precisely when such a correlated initialization is required (see (3.13)). This condition is satisfied by several popular GLMs, including phase retrieval. It is in these cases that AMP with spectral initialization is most useful.

Other related work. For the problem of phase retrieval, several algorithmic solutions have been proposed and analyzed in recent years. An inevitably non-exhaustive list includes semi-definite programming relaxations (Candès et al., 2013, 2015a,b; Waldspurger et al., 2015), a convex relaxation operating in the natural domain of the signal (Goldstein and Studer, 2018; Bahmani and Romberg, 2017), alternating minimization (Netrapalli et al., 2013), Wirtinger Flow (Candès et al., 2015c; Chen and Candès, 2017; Ma et al., 2020), iterative projections (Li et al., 2015), the Kaczmarz method (Wei, 2015; Tan and Vershynin, 2019), and mirror descent (Wu and Rebeschini, 2020). A generalized AMP (GAMP) algorithm was introduced by Schniter and Rangan (2014), and an AMP to solve the non-convex problem with ℓ_2 regularization was proposed and analyzed by Ma et al. (2019). Most of the algorithms mentioned above require an initialization correlated with the signal $\boldsymbol{x}$ and, to obtain such an initialization, spectral methods are widely employed.

Beyond the Gaussian setting, spectral methods for phase retrieval with random orthogonal matrices are analyzed by Dudeja et al. (2020). Statistical and computational phase transitions in phase retrieval for a large class of correlated real and complex random sensing matrices are investigated by Maillard et al. (2020), and a general AMP algorithm for rotationally invariant matrices is studied by Fan (2020). Emami et al. (2020) characterize the generalization error of GLMs via the multi-layer vector AMP (ML-VAMP) of Fletcher et al. (2018) and Pandit et al. (2020). Thus, the extension of our techniques to more general sensing models represents an interesting avenue for future research.

2 Preliminaries

Notation and definitions. Given $n \in \mathbb{N}$, we use the shorthand $[n] = \{1, \ldots, n\}$. Given a vector $\boldsymbol{x}$, we denote by $\|\boldsymbol{x}\|_2$ its Euclidean norm. The *empirical distribution* of a vector $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathsf{T}}$ is given by $\frac{1}{d}\sum_{i=1}^{d} \delta_{x_i}$, where δ_{x_i} denotes a Dirac delta mass on x_i. Similarly, the empirical joint distribution of vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ is $\frac{1}{d}\sum_{i=1}^{d} \delta_{(x_i, x_i')}$.

Generalized linear models. Let $\boldsymbol{x} \in \mathbb{R}^d$ be the signal of interest, and assume that $\|\boldsymbol{x}\|_2^2 = d$. The signal is observed via inner products with n sensing vectors $(\boldsymbol{a}_i)_{i\in[n]}$, with each $\boldsymbol{a}_i \in \mathbb{R}^d$ having independent Gaussian entries with mean zero and variance $1/d$, i.e., $(\boldsymbol{a}_i) \sim_{\text{i.i.d.}} \mathsf{N}(0, \boldsymbol{I}_d/d)$. Given $g_i = \langle \boldsymbol{x}, \boldsymbol{a}_i \rangle$, the components of the observed vector $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ are independently generated according to a conditional distribution $p_{Y|G}$, i.e., $y_i \sim p_{Y|G}(y_i \mid g_i)$. We stack the sensing vectors as rows to define the $n \times d$ sensing matrix $\boldsymbol{A}$, i.e., $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]^{\mathsf{T}}$. For the special case of phase retrieval, the model is $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2 + \boldsymbol{w}$, where $\boldsymbol{w}$ is a noise vector with independent entries. We consider a sequence of problems of growing dimension d, and assume that, as $d \to \infty$, the sampling ratio $n/d \to \delta$, for some constant $\delta \in (0, \infty)$. We remark that, as $d \to \infty$, the empirical distribution of $\boldsymbol{g} = (g_1, \ldots, g_n)$ converges in Wasserstein distance (W_2) to $G \sim \mathsf{N}(0, 1)$.

Spectral initialization. The spectral estimator $\hat{\boldsymbol{x}}^{\rm s}$ is the principal eigenvector of the $d \times d$ matrix $\boldsymbol{D}_n$, defined as

$$\boldsymbol{D}_n = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{Z}_s \boldsymbol{A}, \text{ with } \boldsymbol{Z}_s = \operatorname{diag}(\mathcal{T}_s(y_1), \ldots, \mathcal{T}_s(y_n)), \tag{2.1}$$

where $\mathcal{T}_s : \mathbb{R} \to \mathbb{R}$ is a preprocessing function. We now review some results from Mondelli and Montanari (2019) and Lu and Li (2019) on the performance of the spectral estimator in the high-dimensional limit.

Let $G \sim \mathsf{N}(0, 1)$, $Y \sim p(\cdot \mid G)$, and $Z_s = \mathcal{T}_s(Y)$. We will make the following assumptions on Z_s.

(A1) $\mathbb{P}(Z_s = 0) < 1$.

(A2) Z_s has bounded support and τ is the supremum of this support, i.e., $\tau = \inf\{z : \mathbb{P}(Z_s \leq z) = 1\}$.

(A3) As λ approaches τ from the right, we have

$$\lim_{\lambda\to\tau^+} \mathbb{E}\left\{\frac{Z_s}{(\lambda - Z_s)^2}\right\} = \lim_{\lambda\to\tau^+} \mathbb{E}\left\{\frac{Z_s \cdot G^2}{\lambda - Z_s}\right\} = \infty. \tag{2.2}$$

For $\lambda \in (\tau, \infty)$ and $\delta \in (0, \infty)$, define

$$\phi(\lambda) = \lambda \mathbb{E}\left\{\frac{Z_s \cdot G^2}{\lambda - Z_s}\right\}, \quad \psi_\delta(\lambda) = \frac{\lambda}{\delta} + \lambda \mathbb{E}\left\{\frac{Z_s}{\lambda - Z_s}\right\}. \tag{2.3}$$

Note that $\phi(\lambda)$ is a monotone non-increasing function and that $\psi_\delta(\lambda)$ is a convex function. Let $\bar{\lambda}_\delta$ be the point at which ψ_δ attains its minimum, i.e., $\bar{\lambda}_\delta = \arg\min_{\lambda\geq\tau} \psi_\delta(\lambda)$. For $\lambda \in (\tau, \infty)$, also define

$$\zeta_\delta(\lambda) = \psi_\delta(\max(\lambda, \bar{\lambda}_\delta)). \tag{2.4}$$

We remark that ζ_δ is an increasing function and, from Lemma 2 by Mondelli and Montanari (2019), we have that the equation $\zeta_\delta(\lambda) = \phi(\lambda)$ admits a unique solution for $\lambda > \tau$.

The following result characterizes the performance of the spectral estimator $\hat{\boldsymbol{x}}^{\rm s}$. Its proof follows directly from Lemma 2 by Mondelli and Montanari (2019).

Lemma 2.1. *Let $\boldsymbol{x}$ be such that $\|\boldsymbol{x}\|_2^2 = d$, $\{\boldsymbol{a}_i\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathsf{N}(\boldsymbol{0}_d, \boldsymbol{I}_d/d)$, and $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $\{y_i\}_{i\in[n]} \sim_{\text{i.i.d.}} p_{Y|G}$. Let $n/d \to \delta$, $G \sim \mathsf{N}(0, 1)$ and* *define $Z_s = \mathcal{T}_s(Y)$ for $Y \sim p_{Y|G}$. Assume that Z_s satisfies the assumptions **(A1)-(A2)-(A3)**. Let $\hat{\boldsymbol{x}}^{\rm s}$ be the principal eigenvector of the matrix $\boldsymbol{D}_n$ defined in* (2.1)*, and let λ_δ^* be the unique solution of $\zeta_\delta(\lambda) = \phi(\lambda)$ for $\lambda > \tau$. Then, as $n \to \infty$,*

$$\frac{|\langle \hat{\boldsymbol{x}}^{\rm s}, \boldsymbol{x}\rangle|^2}{\|\hat{\boldsymbol{x}}^{\rm s}\|_2^2 \, \|\boldsymbol{x}\|_2^2} \xrightarrow{a.s.} a^2 \triangleq \begin{cases} 0, & \text{if } \psi_\delta'(\lambda_\delta^*) \le 0, \\ \dfrac{\psi_\delta'(\lambda_\delta^*)}{\psi_\delta'(\lambda_\delta^*) - \phi'(\lambda_\delta^*)}, & \text{if } \psi_\delta'(\lambda_\delta^*) > 0, \end{cases} \tag{2.5}$$

where ψ_δ' and ϕ' are the derivatives of the respective functions.

Remark 2.1 (Equivalent characterization)**.** Using the definitions (2.3)-(2.4), the conditions $\zeta_\delta(\lambda_\delta^*) = \phi(\lambda_\delta^*)$ and $\psi_\delta'(\lambda_\delta^*) > 0$ are equivalent to

$$\mathbb{E}\left\{\frac{Z_s(G^2-1)}{\lambda_\delta^* - Z_s}\right\} = \frac{1}{\delta}, \text{ and } \mathbb{E}\left\{\frac{Z_s^2}{(\lambda_\delta^* - Z_s)^2}\right\} < \frac{1}{\delta}. \tag{2.6}$$

When these conditions are satisfied, the limit of the normalized correlation in (2.5) can be expressed as

$$a^2 = \frac{\frac{1}{\delta} - \mathbb{E}\left\{\frac{Z_s^2}{(\lambda_\delta^* - Z_s)^2}\right\}}{\frac{1}{\delta} + \mathbb{E}\left\{\frac{Z_s^2(G^2-1)}{(\lambda_\delta^* - Z_s)^2}\right\}}. \tag{2.7}$$

Remark 2.2 (Optimal preprocessing function)**.** Mondelli and Montanari (2019) derived the preprocessing function minimizing the value of δ necessary to achieve weak recovery, i.e., a strictly positive correlation between $\hat{\boldsymbol{x}}^{\rm s}$ and $\boldsymbol{x}$. In particular, let $\delta_{\rm u}$ be defined as

$$\delta_{\rm u} = \left(\int_{\mathbb{R}} \frac{\left(\mathbb{E}_G\left\{p(y \mid G)(G^2-1)\right\}\right)^2}{\mathbb{E}_G\left\{p(y \mid G)\right\}} {\rm d}y\right)^{-1}, \tag{2.8}$$

with $G \sim \mathsf{N}(0,1)$. Furthermore, let us also define

$$\bar{\mathcal{T}}(y) = \frac{\sqrt{\delta_{\rm u}} \cdot \mathcal{T}^*(y)}{\sqrt{\delta} - (\sqrt{\delta} - \sqrt{\delta_{\rm u}})\mathcal{T}^*(y)}, \tag{2.9}$$

where

$$\mathcal{T}^*(y) = 1 - \frac{\mathbb{E}_G\left\{p(y \mid G)\right\}}{\mathbb{E}_G\left\{p(y \mid G) \cdot G^2\right\}}. \tag{2.10}$$

Then, by taking $\mathcal{T}_s = \bar{\mathcal{T}}$, for any $\delta > \delta_{\rm u}$, we almost surely have

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{x}}^{\rm s}, \boldsymbol{x}\rangle|}{\|\hat{\boldsymbol{x}}^{\rm s}\|_2 \, \|\boldsymbol{x}\|_2} > \epsilon, \tag{2.11}$$

for some $\epsilon > 0$. Furthermore, for any $\delta < \delta_{\rm u}$, there is no pre-processing function $\mathcal{T}$ such that, almost surely, (2.11) holds. For a more formal statement of this result, see Theorem 4 of Mondelli and Montanari (2019). The preprocessing function that, at a given $\delta > \delta_{\rm u}$, maximizes the correlation between $\hat{\boldsymbol{x}}^{\rm s}$ and $\boldsymbol{x}$ is also related to $\mathcal{T}^*(y)$ as defined in (2.10), and it is derived in Luo et al. (2019).

3 Generalized Approximate Message Passing with Spectral Initialization

We make the following additional assumptions on the signal $\boldsymbol{x}$, the output distribution $p_{Y|G}$, and the preprocessing function $\mathcal{T}_s$ used for the spectral estimator.

(B1) Let $\hat{P}_{X,d}$ denote the *empirical distribution* of $\boldsymbol{x} \in \mathbb{R}^d$. As $d \to \infty$, $\hat{P}_{X,d}$ converges weakly to a distribution P_X such that $\lim_{d\to\infty} \mathbb{E}_{\hat{P}_{X,d}}\{|X|^2\} = \mathbb{E}_{P_X}\{|X|^2\}$. We note that $\mathbb{E}_{P_X}\{|X|^2\} = 1$, since we assume $\|\boldsymbol{x}\|_2^2 = d$.

(B2) We have $\mathbb{E}\{|Y|^2\} < \infty$, for $Y \sim p_{Y|G}(\,\cdot \mid G)$ and $G \sim \mathsf{N}(0,1)$. Furthermore, there exists a function $q : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and a random variable V independent of G such that $Y = q(G, V)$. More precisely, for any measurable set $A \subseteq \mathcal{Y}$ and almost every g, we have $\mathbb{P}(Y \in A \mid G = g) = \mathbb{P}(q(g, V) \in A)$. We also assume that $\mathbb{E}\{|V|^2\} < \infty$. This is without loss of generality due to the functional representation lemma, see p. 626 of El Gamal and Kim (2011).

(B3) The function $\mathcal{T}_s : \mathbb{R} \to \mathbb{R}$ is bounded and Lipschitz.

Following the terminology of Rangan (2011), we refer to the AMP for generalized linear models as GAMP. In each iteration t, the proposed GAMP algorithm produces an estimate $\boldsymbol{x}^t$ of the signal $\boldsymbol{x}$. The algorithm is defined in terms of a sequence of Lipschitz functions $f_t : \mathbb{R} \to \mathbb{R}$ and $h_t : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, for $t \ge 0$. We initialize using the spectral estimator $\hat{\boldsymbol{x}}^{\rm s}$:

$$\boldsymbol{x}^0 = \sqrt{d}\, \frac{1}{\sqrt{\delta}} \hat{\boldsymbol{x}}^{\rm s}, \tag{3.1}$$

$$\boldsymbol{u}^0 = \frac{1}{\sqrt{\delta}} \boldsymbol{A} f_0(\boldsymbol{x}^0) - \mathsf{b}_0 \frac{\sqrt{\delta}}{\lambda_\delta^*} \boldsymbol{Z}_s \boldsymbol{A} \boldsymbol{x}^0, \tag{3.2}$$

where $\mathsf{b}_0 = \frac{1}{n} \sum_{i=1}^d f_0'(x_i^0)$, the diagonal matrix $\boldsymbol{Z}_s$ is defined in (2.1), and λ_δ^* is given by (2.6). Then, for $t \ge 0$, the algorithm computes:

$$\boldsymbol{x}^{t+1} = \frac{1}{\sqrt{\delta}} \boldsymbol{A}^{\mathsf{T}} h_t(\boldsymbol{u}^t; \boldsymbol{y}) - \mathsf{c}_t f_t(\boldsymbol{x}^t), \tag{3.3}$$

$$\boldsymbol{u}^{t+1} = \frac{1}{\sqrt{\delta}} \boldsymbol{A} f_{t+1}(\boldsymbol{x}^{t+1}) - \mathsf{b}_{t+1} h_t(\boldsymbol{u}^t; \boldsymbol{y}). \tag{3.4}$$

Here the functions f_t and h_t are understood to be applied component-wise, i.e., $f_t(\boldsymbol{x}^t) = (f_t(x_1^t), \ldots, f_t(x_d^t))$ and $h_t(\boldsymbol{u}^t; \boldsymbol{y}) = (h_t(u_1^t; y_1), \ldots, h_t(u_n^t; y_n))$. The scalars $\mathsf{b}_t, \mathsf{c}_t$ are defined as

$$\mathsf{c}_t = \frac{1}{n} \sum_{i=1}^n h_t'(u_i^t; y_i), \qquad \mathsf{b}_{t+1} = \frac{1}{n} \sum_{i=1}^d f_{t+1}'(x_i^{t+1}), \tag{3.5}$$

where $h'_t(\cdot, \cdot)$ denotes the derivative with respect to the first argument.

The asymptotic empirical distribution of the GAMP iterates $\boldsymbol{x}^t, \boldsymbol{u}^t$, for $t \geq 0$, can be succinctly characterized via a deterministic recursion, called *state evolution*. Our main result, Theorem 1, shows that for $t \geq 0$, the empirical distributions of $\boldsymbol{u}^t$ and $\boldsymbol{x}^t$ converge in Wasserstein distance W_2 to the laws of the random variables U_t and X_t, respectively, with

$$X_t \equiv \mu_{X,t} X + \sigma_{X,t} W_{X,t}, \tag{3.6}$$

$$U_t \equiv \mu_{U,t} G + \sigma_{U,t} W_{U,t}, \tag{3.7}$$

where $(G, W_{U,t}) \sim_{\text{i.i.d.}} \mathsf{N}(0,1)$. Similarly, $X \sim P_X$ and $W_{X,t} \sim \mathsf{N}(0,1)$ are independent. The deterministic parameters $(\mu_{U,t}, \sigma_{U,t},\, \mu_{X,t}, \sigma_{X,t})$ are recursively computed as follows, for $t \geq 0$:

$$\begin{aligned}
\mu_{U,t} &= \frac{1}{\sqrt{\delta}} \mathbb{E}\{X f_t(X_t)\}, \\
\sigma_{U,t}^2 &= \frac{1}{\delta} \mathbb{E}\{f_t(X_t)^2\} - \mu_{U,t}^2, \qquad (3.8) \\
\mu_{X,t+1} &= \sqrt{\delta} \mathbb{E}\{G h_t(U_t; Y)\} - \mathbb{E}\{h'_t(U_t; Y)\} \mathbb{E}\{X f_t(X_t)\}, \\
\sigma_{X,t+1}^2 &= \mathbb{E}\{h_t(U_t; Y)^2\}.
\end{aligned}$$

For the spectral initialization in (3.1)-(3.2), with a as defined in (2.5), the recursion is initialized with

$$\mu_{X,0} = a/\sqrt{\delta}, \qquad \sigma_{X,0}^2 = (1 - a^2)/\delta. \tag{3.9}$$

We state the main result in terms of *pseudo-Lipschitz* test functions. A function $\psi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order 2, i.e., $\psi \in \mathrm{PL}(2)$, if there is a constant $C > 0$ such that

$$\|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})\|_2 \leq C(1 + \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2) \, \|\boldsymbol{x} - \boldsymbol{y}\|_2 \,, \tag{3.10}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$. Examples of test functions in PL(2) with $m = 2$ include $\psi(a, b) = (a - b)^2$, $\psi(a, b) = ab$.

Theorem 1. *Let $\boldsymbol{x}$ be such that $\|\boldsymbol{x}\|_2^2 = d$, $\{\boldsymbol{a}_i\}_{i \in [n]} \sim_{\text{i.i.d.}} \mathsf{N}(\boldsymbol{0}_d, \boldsymbol{I}_d/d)$, and $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $\{y_i\}_{i \in [n]} \sim_{\text{i.i.d.}} p_{Y|G}$. Let $n/d \to \delta$, $G \sim \mathsf{N}(0,1)$, and $Z_s = \mathcal{T}_s(Y)$ for $Y \sim p_{Y|G}(\,\cdot \,|\, G)$. Assume that **(A1)-(A2)-(A3)** and **(B1)-(B2)-(B3)** hold. Assume further that $\psi'_\delta(\lambda^*_\delta) > 0$, and let $\hat{\boldsymbol{x}}^{\rm s}$ be the principal eigenvector of $\boldsymbol{D}_n$, defined as in (2.1), with the sign of $\hat{\boldsymbol{x}}^{\rm s}$ chosen so that $\langle \hat{\boldsymbol{x}}^{\rm s}, \boldsymbol{x} \rangle \geq 0$.*

Consider the GAMP iteration in Eqs. (3.4)–(3.3) with initialization in Eqs. (3.1)-(3.2). Assume that for $t \geq 0$, the functions f_t, h_t are Lipschitz with derivatives that are continuous almost everywhere. Then, the following limits hold almost surely for any PL(2) function $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and t such that $\sigma_{X,k}^2$ is strictly positive for $0 \leq k \leq t$:

$$\begin{aligned}
&\lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{d} \psi(x_i, x_i^{t+1}) \qquad (3.11) \\
&\quad = \mathbb{E}\{\psi(X,\, \mu_{X,t+1} X + \sigma_{X,t+1} W_{X,t+1})\}, \\
&\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, u_i^t) = \mathbb{E}\left\{\psi(Y,\, \mu_{U,t} G + \sigma_{U,t} W_{U,t})\right\}. \qquad (3.12)
\end{aligned}$$

The result (3.11) also holds for $(t+1) = 0$. In (3.11) (resp. (3.12)), the expectation is over the independent random variables $X \sim P_X$ and $W_{X,t} \sim \mathsf{N}(0,1)$ (resp. $(G, W_{U,t}) \sim_{\text{i.i.d.}} \mathsf{N}(0,1)$). The scalars $(\mu_{X,t}, \mu_{U,t}, \sigma_{X,t}^2, \sigma_{U,t}^2)_{t \geq 0}$ are given by the recursion (3.8) with the initialization (3.9).

We give a sketch of the proof in Section 5 and defer the technical details to the appendices.

We now comment on some of the assumptions in the theorem. The assumption $\psi'_\delta(\lambda^*_\delta) > 0$ is required to ensure that the spectral initialization $\boldsymbol{x}^0$ has non-zero correlation with the signal $\boldsymbol{x}$ (Lemma 2.1). From Remark 2.2, we also know that for any sampling ratio $\delta > \delta_{\rm u}$ there exists a choice of $\mathcal{T}_s$ such that $\psi'_\delta(\lambda^*_\delta) > 0$. We also note that, for $\delta < \delta_{\rm u}$, GAMP converges to the "un-informative fixed point" (where the estimate has vanishing correlation with signal) even if the initial condition has non-zero correlation with the signal, see Theorem 5 of Mondelli and Montanari (2019).

There is no loss of generality in assuming the sign of $\hat{\boldsymbol{x}}^{\rm s}$ to be such that $\langle \hat{\boldsymbol{x}}^{\rm s}, \boldsymbol{x} \rangle \geq 0$. Indeed, if the sign were chosen otherwise, the theorem would hold with the state evolution initialization in (3.9) being $\mu_{X,0} = -a/\sqrt{\delta}$, $\sigma_{X,0}^2 = (1 - a^2)/\delta$.

The assumption that $\sigma_{X,k}^2$ is positive for $k \leq t$ is natural. Indeed, if $\sigma_{X,k}^2 = 0$, then the state evolution result for iteration k implies that $\|\boldsymbol{x} - \mu_{X,k}^{-1} \boldsymbol{x}^k\|^2/d \to 0$ as $d \to \infty$. That is, we can perfectly estimate $\boldsymbol{x}$ from $\boldsymbol{x}^k$, and thus terminate the algorithm after iteration k.

Let us finally remark that the result in (3.11) is equivalent to the statement that the empirical joint distribution of $(\boldsymbol{x}, \boldsymbol{x}^{t+1})$ converges almost surely in Wasserstein distance (W_2) to the joint law of $(X,\, \mu_{X,t+1} X + \sigma_{X,t+1} W)$. This follows from the fact that a sequence of distributions P_n with finite second moment converges in W_2 to P if and only if P_n converges weakly to P and $\int \|a\|_2^2 \, {\rm d}P_n(a) \to \int \|a\|_2^2 \, {\rm d}P(a)$, see Definition 6.7 and Theorem 6.8 of Villani (2008).

When does GAMP require spectral initialization? For the GAMP to give meaningful estimates, we need either $\boldsymbol{x}^0$ or $\boldsymbol{x}^1$ to have strictly non-zero asymptotic correlation with $\boldsymbol{x}$. To see when this can be arranged without a special initialization, consider the *linear* estimator $\hat{\boldsymbol{x}}^{\mathrm{L}}(\xi) := \boldsymbol{A}^{\mathsf{T}}\xi(\boldsymbol{y})$, for some function $\xi : \mathbb{R} \to \mathbb{R}$ that acts component-wise on $\boldsymbol{y}$. If there exists a function ξ such that the asymptotic normalized correlation between $\hat{\boldsymbol{x}}^{\mathrm{L}}(\xi)$ and $\boldsymbol{x}$ is strictly non-zero, then AMP does not require a special initialization (spectral or otherwise) that is correlated with $\boldsymbol{x}$. Indeed, in this case we can replace the initialization in (3.1)-(3.2) by $\boldsymbol{x}^0 = \boldsymbol{0}$, $\boldsymbol{u}^0 = \boldsymbol{0}$ (by taking $f_0 = 0$), and let $h_0(\boldsymbol{u}^0; \boldsymbol{y}) = \sqrt{\delta}\xi(\boldsymbol{y})$. This gives $\boldsymbol{x}^1 = \boldsymbol{A}^{\mathsf{T}}\xi(\boldsymbol{y}) = \hat{\boldsymbol{x}}^{\mathrm{L}}(\xi)$, which has strictly non-zero asymptotic correlation with $\boldsymbol{x}$. This ensures that $|\mu_{X,1}| > 0$, and the standard AMP analysis (Javanmard and Montanari, 2013) directly yields a state evolution result similar to Theorem 1.

The output function $p_{Y|G}$ determines whether a non-trivial linear estimator exists for the GLM. From Appendix C.1 in Mondelli et al. (2020), we have that, if

$$\int_{\mathbb{R}} \frac{\left(\mathbb{E}_{G\sim\mathsf{N}(0,1)}\left\{G\, p_{Y|G}(y \mid G)\right\}\right)^2}{\mathbb{E}_{G\sim\mathsf{N}(0,1)}\left\{p_{Y|G}(y \mid G)\right\}}\,\mathrm{d}y = 0, \tag{3.13}$$

then the correlation between $\boldsymbol{A}^{\mathsf{T}}\xi(\boldsymbol{y})$ and $\boldsymbol{x}$ will asymptotically vanish for any choice of ξ. The condition (3.13) holds for many output functions of interest, including all distributions $p_{Y|G}$ that are even in G (and, therefore, including phase retrieval). It is for these models that spectral initialization is particularly useful.

We remark that our analysis covers not only the (Wirtinger flow) phase retrieval model $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$, but also the amplitude flow phase retrieval model $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|$. In fact, one can analyze the approximate model $\boldsymbol{y} = \sqrt{|\boldsymbol{A}\boldsymbol{x}|^2 + \boldsymbol{\epsilon}}$ and then let $\|\boldsymbol{\epsilon}\|_2 \to 0$. This is similar to the approach taken e.g. by Ma et al. (2018) and Luo et al. (2020). Since the functions used in each AMP iteration are Lipschitz, state evolution holds as $\|\boldsymbol{\epsilon}\|_2 \to 0$. For other GLMs with non-differentiable output functions, we can use a similar approach to construct a smooth approximation to the output function and obtain the state evolution result.

Bayes-optimal GAMP. Applying Theorem 1 to the PL(2) function $\psi(x, y) = (x - f_t(y))^2$, we obtain the asymptotic mean-squared error (MSE) of the GAMP estimate $f_t(\boldsymbol{x}^t)$. In formulas, for $t \geq 0$, almost surely,

$$\lim_{d\to\infty} \frac{1}{d}\|\boldsymbol{x} - f_t(\boldsymbol{x}^t)\|_2^2 = \mathbb{E}\{(X - f_t(\mu_{X,t}X + \sigma_{X,t}W))^2\}. \tag{3.14}$$

If the limiting empirical distribution P_X of the signal is known, then the choice of f_t that minimizes the MSE in (3.14) is

$$f_t^*(s) = \mathbb{E}\{X \mid \mu_{X,t}X + \sigma_{X,t}W = s\}. \tag{3.15}$$

Similarly, applying the theorem to the PL(2) functions $\psi(x, y) = x f_t(y)$ and $\psi(x, y) = f_t(y)^2$, we obtain the asymptotic normalized correlation with the signal. In formulas, for $t \geq 0$, almost surely

$$\lim_{d\to\infty} \frac{|\langle \boldsymbol{x}, f_t(\boldsymbol{x}^t)\rangle|}{\|\boldsymbol{x}\|_2 \|f_t(\boldsymbol{x}^t)\|_2} = \frac{|\mathbb{E}\{X f_t(\mu_{X,t}X + \sigma_{X,t}W)\}|}{\sqrt{\mathbb{E}\{f_t(\mu_{X,t}X + \sigma_{X,t}W)^2\}}}. \tag{3.16}$$

For fixed $(\mu_{X,t}, \sigma_{X,t}^2)$, the normalized correlation in (3.16) is maximized by taking $f_t = c f_t^*$ for any $c \neq 0$. This choice also maximizes the ratio $\mu_{U,t}^2/\sigma_{U,t}^2$ in (3.8). For $f_t = c f_t^*$, from (3.8) we have

$$\mu_{U,t} = \frac{c}{\sqrt{\delta}}\mathbb{E}\{f_t^*(X_t)^2\}, \quad \sigma_{U,t}^2 = \frac{c}{\sqrt{\delta}}\mu_{U,t} - \mu_{U,t}^2. \tag{3.17}$$

We now specify the choice of $h_t(u; y)$ that maximizes the ratio $\mu_{X,t+1}^2/\sigma_{X,t+1}^2$ for fixed $(\mu_{U,t}, \sigma_{U,t}^2)$.

Proposition 3.1. *Assume the setting of Theorem 1. For a given* $(\mu_{U,t}, \sigma_{U,t}^2)$*, the ratio* $\mu_{X,t+1}^2/\sigma_{X,t+1}^2$ *is maximized when* $h_t(u;\, y) = c\, h_t^*(u;\, y)$ *where* $c \neq 0$ *is any constant, and*

$$h_t^*(u;\, y) \triangleq \frac{\mathbb{E}\{G \mid U_t = u, Y = y\} - \mathbb{E}\{G \mid U_t = u\}}{\mathsf{Var}(G \mid U_t = u)} \tag{3.18}$$

$$= \frac{\mathbb{E}_W\{W p_{Y|G}(y \mid \rho_t u + \sqrt{1-\rho_t\,\mu_{U,t}}\, W)\}}{\sqrt{1-\rho_t\,\mu_{U,t}}\ \mathbb{E}_W\{p_{Y|G}(y \mid \rho_t u + \sqrt{1-\rho_t\,\mu_{U,t}}\, W)\}}, \tag{3.19}$$

where $\rho_t = \mu_{U,t}/(\mu_{U,t}^2 + \sigma_{U,t}^2)$ *and* $W \sim \mathsf{N}(0,1)$*. In* (3.18)*, the random variables* U_t *and* Y *are conditionally independent given* G *with*

$$\begin{aligned} Y &\sim p_{Y|G}(\,\cdot\mid G), \qquad U_t = \mu_{U,t}G + \sigma_{U,t}W_{U,t}, \\ &(G, W_{U,t}) \sim_{\text{i.i.d.}} \mathsf{N}(0,1). \end{aligned} \tag{3.20}$$

The optimal choice for h_t^* in Proposition 3.1 was derived by Rangan (2011) by approximating the belief propagation equations. For completeness, we provide a self-contained proof in Appendix A. The proof also shows that with $h_t = c h_t^*$,

$$\mu_{X,t+1} = c\sqrt{\delta}\,\mathbb{E}\{|h_t^*(U_t; Y)|^2\}, \quad \sigma_{X,t+1}^2 = c\,\frac{\mu_{X,t+1}}{\sqrt{\delta}}.$$

As the choices f_t^*, h_t^* maximize the signal-to-noise ratios $\mu_{U,t}^2/\sigma_{U,t}^2$ and $\mu_{X,t+1}^2/\sigma_{X,t+1}^2$, respectively, we refer to this algorithm as Bayes-optimal GAMP. We note that to apply Theorem 1 to the Bayes-optimal GAMP, we need f_t^*, h_t^* to be Lipschitz. This holds under relatively mild conditions on P_X and $p_{Y|G}$, see Lemma F.1 by Montanari and Venkataramanan (2021).

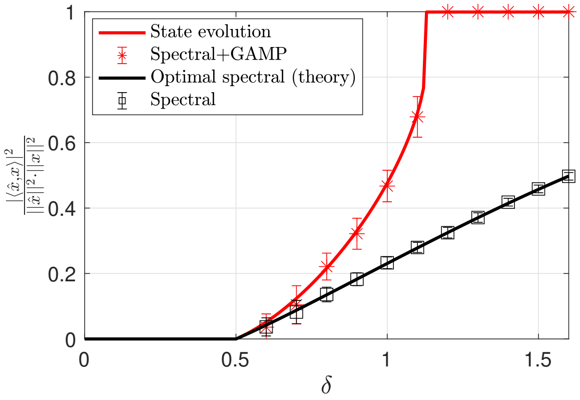
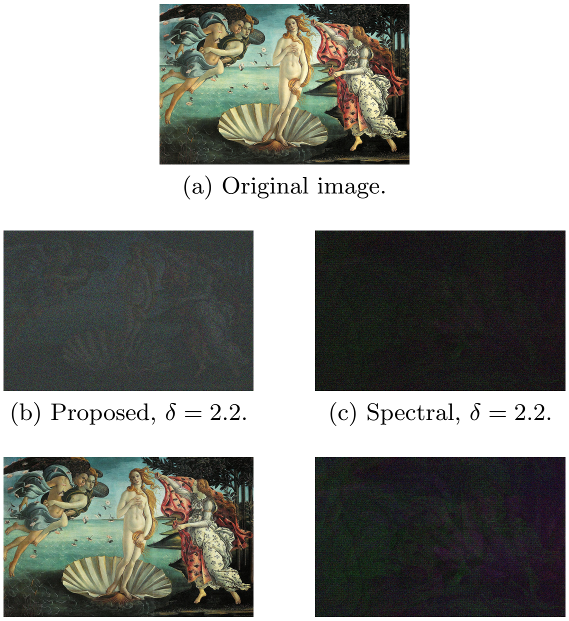


Figure 1: Performance comparison between GAMP with spectral initialization (in red) and the spectral method alone (in black) for a Gaussian prior $P_X \sim \mathsf{N}(0,1)$. The solid lines are the theoretical predictions of Theorem 1 for GAMP with spectral initialization, and of Lemma 2.1 for the spectral method. Error bars indicate one standard deviation around the empirical mean.

4 Numerical Simulations

We now illustrate the performance of the GAMP algorithm with spectral initialization via numerical examples. For concreteness, we focus on noiseless phase retrieval, where $y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2$, $i \in [n]$.

Gaussian prior. In Figure 1, $\boldsymbol{x}$ is chosen uniformly at random on the d-dimensional sphere with radius $\sqrt{d}$ and $\{\boldsymbol{a}_i\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathsf{N}(0, \boldsymbol{I}_d/d)$. Note that, as $d \to \infty$, the limiting empirical distribution P_X of $\boldsymbol{x}$ is a standard Gaussian. We take $d = 8000$, and the numerical simulations are averaged over $n_{\text{sample}} = 50$ independent trials. The performance of an estimate $\hat{\boldsymbol{x}}$ is measured via its normalized squared scalar product with the signal $\boldsymbol{x}$. The black points are obtained by estimating $\boldsymbol{x}$ via the spectral method, using the optimal pre-processing function $\mathcal{T}_s$ reported in Eq. (137) of Mondelli and Montanari (2019). The empirical results match the black curve, which gives the best possible squared correlation in the high-dimensional limit, as given by Theorem 1 of Luo et al. (2019). The red points are obtained by running the GAMP algorithm (3.3)-(3.4) with the spectral initialization (3.1)-(3.2). The function f_t is chosen to be the identity, and $h_t = \sqrt{\delta} h_t^*$, for h_t^* given by Proposition 3.1. The algorithm is run until the normalized squared difference between successive iterates is small. As predicted by Theorem 1, the numerical simulations agree well with the state evolution curve in red, which is obtained by computing the fixed point of the recursion (3.8) initialized with (3.9). We also remark that the threshold for exact recovery can be obtained from the fixed points of state evolution, see e.g. Barbier et al. (2019).



(a) Original image.

(b) Proposed, $\delta = 2.2$.

(c) Spectral, $\delta = 2.2$.

(d) Proposed, $\delta = 2.4$.

(e) Spectral, $\delta = 2.4$.

Figure 2: Visual comparison between the reconstruction of the GAMP algorithm with spectral initialization and that of the spectral method alone for measurements given by coded diffraction patterns.

In Appendix D, we consider a binary-valued prior, and compare the performance of the Bayes-optimal choice f_t^* against f_t equal to the identity.

Coded diffraction patterns. We consider the model of coded diffraction patterns described in Section 7.2 of Mondelli and Montanari (2019). Here the signal $\boldsymbol{x}$ is the image of Figure 2a, and it can be viewed as a $d_1 \times d_2 \times 3$ array with $d_1 = 820$ and $d_2 = 1280$. The sensing vectors are given by

$$a_r(t_1, t_2) = d_\ell(t_1, t_2) \cdot e^{i2\pi k_1 t_1/d_1} \cdot e^{i2\pi k_2 t_2/d_2}, \quad (4.1)$$

where $r \in [n]$, $t_1 \in [d_1]$, $t_2 \in [d_2]$, i denotes the imaginary unit, $a_r(t_1, t_2)$ is the (t_1, t_2)-th component of $\boldsymbol{a}_r \in \mathbb{C}^d$, and the $(d_\ell(t_1, t_2))$'s are i.i.d. and uniform in $\{1, -1, i, -i\}$. The index $r \in [n]$ is associated to a pair (ℓ, k), with $\ell \in [L]$; the index $k \in [d]$ is associated to a pair (k_1, k_2) with $k_1 \in [d_1]$ and $k_2 \in [d_2]$. Thus, $n = L \cdot d$ and, therefore, $\delta = L \in \mathbb{N}$. To obtain non-integer values of δ, we set to 0 a suitable fraction of the vectors $\boldsymbol{a}_r$, chosen uniformly at random.

In this model, the scalar product $\langle \boldsymbol{x}_j, \boldsymbol{a}_r\rangle$ can be computed with an FFT algorithm. Furthermore, in order to evaluate the principal eigenvector for the spectral initialization, we use a power method which stops if either the number of iterations reaches the maximum value of 100000 or the modulus of the scalar product between the estimate at the current iteration T and at the iteration $T - 10$ is larger than $1 - 10^{-7}$.

The GAMP algorithm with spectral initialization for the complex-valued setting is described in Appendix E. Figure 2 shows a visual representation of the results. The improvement achieved by the GAMP algorithm over the spectral estimator is impressive, with GAMP achieving full recovery already at $\delta = 2.4$. A numerical comparison of the performance of the two methods is given in Figure 5 in Appendix E. We emphasize that the state evolution result of Theorem 1 is only valid for Gaussian sensing matrices. Extending it to structured matrices such as coded diffraction patterns is an interesting direction for future work.

5 Sketch of the Proof of Theorem 1

We give an outline of the proof here, and provide the technical details in the appendices.

The artificial GAMP algorithm. We construct an artificial GAMP algorithm, whose iterates are denoted by $\tilde{\boldsymbol{x}}^t, \tilde{\boldsymbol{u}}^t$, for $t \geq 0$. Starting from an initialization $(\tilde{\boldsymbol{x}}^0, \tilde{\boldsymbol{u}}^0)$, for $t \geq 0$ we iteratively compute:

$$\tilde{\boldsymbol{x}}^{t+1} = \frac{1}{\sqrt{\delta}} \boldsymbol{A}^{\mathsf{T}} \tilde{h}_t(\tilde{\boldsymbol{u}}^t; \boldsymbol{y}) - \tilde{\mathsf{c}}_t \tilde{f}_t(\tilde{\boldsymbol{x}}^t), \tag{5.1}$$

$$\tilde{\boldsymbol{u}}^{t+1} = \frac{1}{\sqrt{\delta}} \boldsymbol{A} \tilde{f}_{t+1}(\tilde{\boldsymbol{x}}^{t+1}) - \tilde{\mathsf{b}}_{t+1} \tilde{h}_t(\tilde{\boldsymbol{u}}^t; \boldsymbol{y}). \tag{5.2}$$

For $t \geq 0$, the functions $\tilde{f}_t : \mathbb{R} \to \mathbb{R}$ and $\tilde{h}_t : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are Lipschitz, and will be specified below. The scalars $\tilde{\mathsf{c}}_t$ and $\tilde{\mathsf{b}}_{t+1}$ are defined as

$$\tilde{\mathsf{c}}_t = \frac{1}{n} \sum_{i=1}^{n} \tilde{h}'_t(\tilde{u}_i^t; y_i), \qquad \tilde{\mathsf{b}}_{t+1} = \frac{1}{n} \sum_{i=1}^{d} \tilde{f}'_{t+1}(\tilde{x}_i^{t+1}), \tag{5.3}$$

where $\tilde{h}'_t$ denotes the derivative with respect to the first argument. The iteration is initialized as follows. Choose any $\alpha \in (0, 1)$, and a standard Gaussian vector $\boldsymbol{n} \sim \mathsf{N}(0, \boldsymbol{I}_d)$ that is independent of $\boldsymbol{x}$ and $\boldsymbol{A}$. Then,

$$\tilde{\boldsymbol{x}}^0 = \alpha\, \boldsymbol{x} + \sqrt{1 - \alpha^2}\, \boldsymbol{n}, \qquad \tilde{\boldsymbol{u}}^0 = \frac{1}{\sqrt{\delta}} \boldsymbol{A} \tilde{f}_0(\tilde{\boldsymbol{x}}^0). \tag{5.4}$$

The artificial GAMP is divided into two phases. In the first phase, which lasts up to iteration T, the functions $\tilde{f}_t, \tilde{h}_t$ for $0 \leq t \leq (T-1)$, are chosen such that as $T \to \infty$, the iterate $\tilde{\boldsymbol{x}}^T$ approaches the initialization $\boldsymbol{x}^0$ of the true GAMP algorithm defined in (3.1). In the second phase, the functions $\tilde{f}_t, \tilde{h}_t$ for $t \geq T$, are chosen to match those of the true GAMP. The key observation is that a state evolution result for the artificial GAMP follows directly from the standard analysis of GAMP (Javanmard and Montanari, 2013) since the initialization $\tilde{\boldsymbol{x}}^0$ is independent of $\boldsymbol{A}$. By showing that as $T \to \infty$, the iterates and the state evolution parameters of the artificial GAMP approach the corresponding quantities of the true GAMP, we prove that the state evolution result of Theorem 1 holds.

We now specify the functions used in the artificial GAMP. For $0 \leq t \leq (T-1)$, we set

$$\tilde{f}_t(x) = \frac{x}{\beta_t}, \qquad \tilde{h}_t(x; y) = \sqrt{\delta}\, x \, \frac{\mathcal{T}_s(y)}{\lambda^*_\delta - \mathcal{T}_s(y)}, \tag{5.5}$$

where $\mathcal{T}_s$ is the pre-processing function used for the spectral estimator, λ^*_δ is the unique solution of $\zeta_\delta(\lambda) = \phi(\lambda)$ for $\lambda > \tau$ (also given by (2.6)), and $(\beta_t)_{t \geq 0}$ are constants coming from the state evolution recursion defined below. Furthermore, for $t \geq T$, we set

$$\tilde{f}_t(x) = f_{t-T}(x), \qquad \tilde{h}_t(x; y) = h_{t-T}(x; y). \tag{5.6}$$

With these choices of $\tilde{f}_t, \tilde{h}_t$, the coefficients $\tilde{\mathsf{c}}_t$ and $\tilde{\mathsf{b}}_t$ in (5.3) become:

$$\tilde{\mathsf{c}}_t = \frac{\sqrt{\delta}}{n} \sum_{i=1}^{n} \frac{\mathcal{T}_s(y_i)}{\lambda^*_\delta - \mathcal{T}_s(y_i)}, \; \tilde{\mathsf{b}}_t = \frac{1}{\delta \beta_t}, \quad 0 \leq t \leq (T-1),$$

$$\tilde{\mathsf{c}}_t = \frac{1}{n} \sum_{i=1}^{n} h'_{t-T}(\tilde{u}_i^t; y_i), \; \tilde{\mathsf{b}}_t = \frac{1}{n} \sum_{i=1}^{d} f'_{t-T}(\tilde{x}_i^t), \quad t \geq T. \tag{5.7}$$

Since the initialization $\tilde{\boldsymbol{x}}^0$ in (5.4) is independent of $\boldsymbol{A}$, the state evolution result of Javanmard and Montanari (2013) can be applied to the artificial GAMP. This result, formally stated in Proposition B.1 in Appendix B.1, implies that for $t \geq 0$, the empirical distributions of $\tilde{\boldsymbol{x}}^t$ and $\tilde{\boldsymbol{u}}^t$ converge in W_2 distance to the laws of the random variables $\tilde{X}_t$ and $\tilde{U}_t$, respectively, with

$$\tilde{X}_t \equiv \mu_{\tilde{X},t} X + \sigma_{\tilde{X},t} W_{\tilde{X},t}, \quad \tilde{U}_t \equiv \mu_{\tilde{U},t} G + \sigma_{\tilde{U},t} W_{\tilde{U},t}. \tag{5.8}$$

Here $W_{\tilde{X},t}, W_{\tilde{U},t}$ are standard normal and independent of X and G, respectively. The state evolution recursion defining the parameters $(\mu_{\tilde{X},t}, \sigma_{\tilde{X},t}, \mu_{\tilde{U},t}, \sigma_{\tilde{U},t}, \beta_t)$ has the same form as (3.8), except that we use the functions defined in (5.5) for $0 \leq t \leq (T-1)$, and the functions in (5.6) for $t \geq T$. The detailed expressions are given in Appendix B.1.

Analysis of the first phase. The first phase of the artificial GAMP is designed so that its output vectors after T iterations $(\tilde{\boldsymbol{x}}^T, \tilde{\boldsymbol{u}}^T)$ are close to the initialization $(\boldsymbol{x}^0, \boldsymbol{u}^0)$ of the true GAMP algorithm given by (3.1)-(3.2). This part of the algorithm is similar to the GAMP used in Mondelli et al. (2020) to approximate the spectral estimator $\hat{\boldsymbol{x}}^{\rm s}$. In particular, the state evolution recursion of the first phase (given in (B.2)) converges as $T \to \infty$ to the following fixed point:

$$\lim_{T \to \infty} \mu_{\tilde{X},T} = \frac{a}{\sqrt{\delta}}, \qquad \lim_{T \to \infty} \sigma^2_{\tilde{X},T} = \frac{1 - a^2}{\delta}, \tag{5.9}$$

where a is the limit (normalized) correlation between the spectral estimator $\hat{\boldsymbol{x}}^{\rm s}$ and the signal, see (2.7). Furthermore, the GAMP iterate $\tilde{\boldsymbol{x}}^T$ approaches $\hat{\boldsymbol{x}}^{\rm s}$, i.e.,

$$\lim_{T\to\infty}\lim_{d\to\infty}\frac{\|\sqrt{d}\,\hat{\boldsymbol{x}}^{\rm s}-\sqrt{\delta}\,\tilde{\boldsymbol{x}}^T\|^2}{d}=0 \quad \text{a.s.} \tag{5.10}$$

These results are formally stated in Lemma B.2 and B.3, respectively, contained in Appendix B.2.

Analysis of the second phase. The second phase of the artificial GAMP is designed so that its iterates $\tilde{\boldsymbol{x}}^{T+t}, \tilde{\boldsymbol{u}}^{T+t}$ are close to $\boldsymbol{x}^t, \boldsymbol{u}^t$, respectively for $t \geq 0$, and the corresponding state evolution parameters are also close. In particular, in order to prove Theorem 1, we first analyze a slightly modified version of the true GAMP algorithm in (3.3)-(3.4) where the 'memory' coefficients b_t and c_t in (3.5) are replaced by deterministic values obtained from state evolution. The iterates of this modified GAMP, denoted by $\hat{\boldsymbol{x}}^t, \hat{\boldsymbol{u}}^t$, are as follows. Start with the initialization

$$\hat{\boldsymbol{x}}^0=\boldsymbol{x}^0=\sqrt{d}\,\frac{1}{\sqrt{\delta}}\hat{\boldsymbol{x}}^{\rm s}, \tag{5.11}$$

$$\hat{\boldsymbol{u}}^0=\frac{1}{\sqrt{\delta}}\boldsymbol{A}f_0(\hat{\boldsymbol{x}}^0)-\bar{\mathsf{b}}_0\frac{\sqrt{\delta}}{\lambda_\delta^*}\boldsymbol{Z}_s\boldsymbol{A}\hat{\boldsymbol{x}}^0, \tag{5.12}$$

where $\bar{\mathsf{b}}_0=\frac{1}{\delta}\mathbb{E}\{f_0'(X_0)\}$. Then, for $t \geq 0$:

$$\hat{\boldsymbol{x}}^{t+1}=\frac{1}{\sqrt{\delta}}\boldsymbol{A}^{\mathsf{T}}h_t(\hat{\boldsymbol{u}}^t;\boldsymbol{y})-\bar{\mathsf{c}}_tf_t(\hat{\boldsymbol{x}}^t), \tag{5.13}$$

$$\hat{\boldsymbol{u}}^{t+1}=\frac{1}{\sqrt{\delta}}\boldsymbol{A}f_{t+1}(\hat{\boldsymbol{x}}^{t+1})-\bar{\mathsf{b}}_{t+1}h_t(\hat{\boldsymbol{u}}^t;\boldsymbol{y}). \tag{5.14}$$

Here, for $t \geq 0$, the deterministic memory coefficients $\bar{\mathsf{b}}_t$ and $\bar{\mathsf{c}}_t$ are

$$\bar{\mathsf{c}}_t=\mathbb{E}\{h_t'(U_t;Y)\}, \qquad \bar{\mathsf{b}}_t=\mathbb{E}\{f_t'(X_t)\}/\delta, \tag{5.15}$$

where X_t, U_t are defined in (3.6)-(3.7).

Let us now summarize our approach. We have defined three different GAMP iterations: *(i)* the *true GAMP* with iterates $(\boldsymbol{x}^t, \boldsymbol{u}^t)$ given by (3.3)-(3.4) and initialization $(\boldsymbol{x}^0, \boldsymbol{u}^0)$ given by (3.1)-(3.2), *(ii)* the *modified GAMP* with iterates $(\hat{\boldsymbol{x}}^t, \hat{\boldsymbol{u}}^t)$ given by (5.13)-(5.14) and initialization $(\hat{\boldsymbol{x}}^0, \hat{\boldsymbol{u}}^0)$ given by (5.11)-(5.12), and *(iii)* the *artificial GAMP* with iterates $(\tilde{\boldsymbol{x}}^t, \tilde{\boldsymbol{u}}^t)$ given by (5.1)-(5.2) and initialization $(\tilde{\boldsymbol{x}}^0, \tilde{\boldsymbol{u}}^0)$ given by (5.4). We recall that the *true GAMP* is the algorithm with spectral initialization that is actually implemented and whose performance we want to study. As the true GAMP is initialized with the spectral estimator $\hat{\boldsymbol{x}}^{\rm s}$ which depends on $\boldsymbol{A}$, its performance cannot be characterized using the existing theory. To solve this problem, we introduce the *artificial GAMP* purely as a proof technique. In fact, the initialization of the artificial GAMP assumes knowledge of the signal, which makes it impractical. Finally, the *modified GAMP* is a slight modification of the true GAMP to simplify the proof.

Lemma B.5 in Appendix B.3 proves that, for each $t \geq 0$, *(i)* the iterates $(\tilde{\boldsymbol{x}}^{t+T}, \tilde{\boldsymbol{u}}^{t+T})$ are close to $(\hat{\boldsymbol{x}}^t, \hat{\boldsymbol{u}}^t)$ for sufficiently large T, and *(ii)* the corresponding state evolution parameters are also close. We then use this lemma to prove Theorem 1 by showing that the iterates of the *true GAMP* have the same asymptotic empirical distribution as those of the *modified GAMP*. In particular, we show in in Appendix B.4 that, almost surely,

$$\begin{aligned}\lim_{d\to\infty}\frac{1}{d}\sum_{i=1}^d\psi(x_i,x_i^t)&=\lim_{d\to\infty}\frac{1}{d}\sum_{i=1}^d\psi(x_i,\hat{x}_i^t)\\&=\mathbb{E}\left\{\psi(X,\,\mu_{X,t}X+\sigma_{X,t}\,W)\right\}.\end{aligned} \tag{5.16}$$

6 Discussion

A major shortcoming in existing AMP theory for GLMs, like phase retrieval, is the unrealistic assumption that the initialization of the algorithm is correlated with the ground-truth signal and, at the same time, independent of the measurement matrix. This paper solves this problem by providing a rigorous analysis of AMP with a spectral initialization. Spectral initializations have been widely studied in recent years, and have two attractive features. First, for phase retrieval, they meet the information theoretic threshold for weak recovery (Mondelli and Montanari, 2019). This means that, when the spectral initialization fails, no other method can work. Second, for a large class of GLMs, if the spectral method is unsuccessful, then AMP has an attractive fixed point at 0, see Theorem 5 in Mondelli and Montanari (2019). This is a strong indication that, when the spectral initialization fails, the problem is computationally hard. An interesting future direction is to analyze the fixed points of AMP with spectral initialization, and compare with those of other algorithms that can be initialized with a spectral estimator, e.g., gradient descent.

Our analysis is based on an artificial AMP that first closely approximates the spectral estimator and then the true AMP algorithm. This technical tool is versatile and could be used beyond GLMs with Gaussian sensing matrices. Examples include more general measurement models (Fan, 2020; Emami et al., 2020), other message passing algorithms, e.g., Vector AMP (Schniter et al., 2016), or the design of an artificial AMP that leads to a different estimator. We also highlight that the AMP analyzed here is rather general, and it includes as special cases both the Bayes-optimal AMP for GLMs and AMPs designed to optimize objective functions tailored to the signal prior.

Acknowledgements

The authors would like to thank Andrea Montanari for helpful discussions. M. Mondelli was partially supported by the 2019 Lopez-Loreta Prize. R. Venkataramanan was partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Anderson, G. W., Guionnet, A., and Zeitouni, O. (2009). *An introduction to random matrices*. Cambridge University Press.

Bahmani, S. and Romberg, J. (2017). Phase retrieval meets statistical learning theory: A flexible convex relaxation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 252–260.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

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

Bayati, M. and Montanari, A. (2012). The LASSO risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58:1997–2017.

Bolthausen, E. (2014). An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366.

Boufounos, P. T. and Baraniuk, R. G. (2008). 1-bit compressive sensing. In *Conference on Information Sciences and Systems (CISS)*, pages 16–21.

Candès, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. (2015a). Phase retrieval via matrix completion. *SIAM Review*, 57(2):225–251.

Candès, E. J., Li, X., and Soltanolkotabi, M. (2015b). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299.

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

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

Chen, Y. and Candès, E. J. (2017). Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883.

Demanet, L. and Jugnon, V. (2017). Convex recovery from interferometric measurements. *IEEE Transactions on Computational Imaging*, 3(2):282–295.

Deshpande, Y. and Montanari, A. (2014). Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2197–2201.

Donoho, D. L., Javanmard, A., and Montanari, A. (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464.

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message Passing Algorithms for Compressed Sensing. *Proceedings of the National Academy of Sciences*, 106:18914–18919.

Dudeja, R., Bakhshizadeh, M., Ma, J., and Maleki, A. (2020). Analysis of spectral methods for phase retrieval with random orthogonal matrices. *IEEE Transactions on Information Theory*, 66(8):5182–5203.

El Gamal, A. and Kim, Y.-H. (2011). *Network information theory*. Cambridge university press.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: Theory and applications*. Cambridge University Press.

Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S., and Fletcher, A. (2020). Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, pages 2892–2901. PMLR.

Fan, Z. (2020). Approximate message passing algorithms for rotationally invariant matrices. arXiv:2008.11892.

Fannjiang, A. and Strohmer, T. (2020). The numerics of phase retrieval. arXiv:2004.05788.

Fienup, C. and Dainty, J. (1987). Phase retrieval and image reconstruction for astronomy. *Image recovery: theory and application*, 231:275.

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

Fletcher, A. K., Rangan, S., and Schniter, P. (2018). Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1884–1888. IEEE.

Goldstein, T. and Studer, C. (2018). Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689.

Javanmard, A. and Montanari, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference*, pages 115–144.

Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A., and Zdeborová, L. (2016). Phase transitions and sample complexity in Bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. (2012). Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009.

Li, G., Gu, Y., and Lu, Y. M. (2015). Phase retrieval using iterative projections: Dynamics in the large systems limit. In *Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1114–1118.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039.

Lu, Y. M. and Li, G. (2019). Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference*.

Luo, Q., Wang, H., and Lin, S. (2020). Phase retrieval via smoothed amplitude flow. *Signal Processing*, 177:107719.

Luo, W., Alghamdi, W., and Lu, Y. M. (2019). Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9):2347–2356.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632.

Ma, J., Xu, J., and Maleki, A. (2018). Approximate message passing for amplitude based optimization. In *International Conference on Machine Learning (ICML)*, pages 3371–3380.

Ma, J., Xu, J., and Maleki, A. (2019). Optimization-based amp for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Transactions on Information Theory*, 65(6):3600–3629.

Maillard, A., Loureiro, B., Krzakala, F., and Zdeborová, L. (2020). Phase retrieval in high dimensions: Statistical and computational phase transitions. arXiv:2006.05228.

Maleki, A., Anitori, L., Yang, Z., and Baraniuk, R. G. (2013). Asymptotic analysis of complex lasso via complex approximate message passing (CAMP). *IEEE Transactions on Information Theory*, 59(7):4290–4308.

McCullagh, P. (2018). *Generalized linear models*. Routledge.

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

Mondelli, M. and Montanari, A. (2019). Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19:703–773.

Mondelli, M., Thrampoulidis, C., and Venkataramanan, R. (2020). Optimal combination of linear and spectral estimators for generalized linear models. arXiv:2008.03326.

Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *Annals of Statistics*, 45(1):321–345.

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

Pandit, P., Sahraee-Ardakan, M., Rangan, S., Schniter, P., and Fletcher, A. K. (2020). Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*, 1(1):336–347.

Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. (2018). Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322.

Rangan, S. (2011). Generalized Approximate Message Passing for Estimation with Random Linear Mixing. In *IEEE International Symposium on Information Theory (ISIT)*.

Rangan, S. and Fletcher, A. K. (2012). Iterative estimation of constrained rank-one matrices in noise. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1246–1250.

Rangan, S. and Goyal, V. K. (2001). Recursive consistent estimation with bounded noise. *IEEE Transactions on Information Theory*, 47(1):457–464.

Schniter, P. and Rangan, S. (2014). Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055.

Schniter, P., Rangan, S., and Fletcher, A. K. (2016). Vector approximate message passing for the generalized linear model. In *50th Asilomar Conference on*

Signals, Systems and Computers, pages 1525–1529. IEEE.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015). Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Tan, Y. S. and Vershynin, R. (2019). Phase retrieval via randomized kaczmarz: Theoretical guarantees. *Information and Inference: A Journal of the IMA*, 8(1):97–123.

Villani, C. (2008). *Optimal transport: Old and new*, volume 338. Springer Science & Business Media.

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

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

Wu, F. and Rebeschini, P. (2020). A continuous-time mirror descent approach to sparse phase retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 20192–20203.

Yang, F., Lu, Y. M., Sbaiz, L., and Vetterli, M. (2012). Bits from photons: Oversampled image acquisition using binary Poisson statistics. *IEEE Transactions on Image Processing*, 21(4):1421–1436.