PNAS RESEARCH ARTICLE | STATISTICS

Approximate message passing from random initialization with applications to Z_2 synchronization

Gen Li[a], Wei Fan[a], and Yuting Wei[a,1]

Edited by Marco Mondelli, Institute of Science and Technology Austria, Klosterneuburg, Austria; received February 20, 2023; accepted June 24, 2023 by Editorial Board Member David A. Weitz

This paper is concerned with the problem of reconstructing an unknown rank-one matrix with prior structural information from noisy observations. While computing the Bayes optimal estimator is intractable in general due to the requirement of computing high-dimensional integrations/summations, Approximate Message Passing (AMP) emerges as an efficient first-order method to approximate the Bayes optimal estimator. However, the theoretical underpinnings of AMP remain largely unavailable when it starts from random initialization, a scheme of critical practical utility. Focusing on a prototypical model called Z_2 synchronization, we characterize the finite-sample dynamics of AMP from random initialization, uncovering its rapid global convergence. Our theory—which is nonasymptotic in nature—in this model unveils the non-necessity of a careful initialization for the success of AMP.

approximate message passing j random initialization j nonasymptotic analysis j spiked Wigner model j global convergence

The problem of estimating an unknown low-rank matrix, when given access to highly noisy observations, has been the subject of considerable studies, shedding light on a diverse array of contexts including collaborative filtering, synchronization and alignment, localization, and causal panel data, to name just a few (1–8). While low-rank estimators are not in short supply, the quest for algorithms that can work all the way to the information-theoretic limits continues to inspire theoretical and algorithmic development.

1. Motivation and An Informal Overview

In this paper, we focus on how to reconstruct a structured signal $v^\star \in \mathbb{R}^n$ (or equivalently, $v^\star v^{\star\top}$) from noisy data:

$$M = v^\star v^{\star\top} + W \in \mathbb{R}^{n\times n} \qquad \text{with } \lambda > 0: \tag{1}$$

This classical model is commonly referred to as a deformed Gaussian Wigner model or spiked Gaussian Wigner model when the entries of the noise matrix $W = [W_{ij}]_{1\le i,j\le n}$ are independently drawn from Gaussian distributions—more precisely, $W_{ii} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and $W_{ji} = W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n})$ for $i < j$—which serves as a prototypical model toward understanding the feasibility and fundamental limits of low-rank matrix estimation.

The spectral properties of the observed matrix M have been extensively studied (see, e.g. refs. 9–13), motivating the design of spectral methods when there is no structural information associated with (1, 3, 14–16). In practice, there is no shortage of applications where additional structural information about $v^\star$ is available a priori, examples including finite-group structure (17), cone constraints (18, 19), and sparsity (20, 21), among others. The presence of prior structure further exacerbates the nonconvexity issue when computing the maximum likelihood estimate or Bayes optimal estimate, thereby presenting a pressing need for the search of algorithms that can be executed efficiently.

Remarkably, the approximate message passing (AMP) algorithm emerges as an efficient nonconvex paradigm that rises to the aforementioned challenge (22, 23). Originally proposed in the context of compressed sensing, AMP has served as not only a family of first-order iterative algorithms that enjoy rapid convergence (24–28) but also a powerful statistical machinery that assists in determining the performance limits of other statistical procedures in high-dimensional asymptotics (29–38). Over the past two decades, AMP has also received widespread adoption in a variety of engineering and science applications, including but not limited to imaging, wireless communications, signal processing, and deep learning (see, e.g., refs. 39–43 and references therein).

Significance

Approximate Message Passing (AMP) serves as both a family of efficient first-order algorithms and a powerful theoretical machinery for high-dimensional data analysis, which has found applications in a diverse array of problems such as sparse regression, generalized linear models, and low-rank matrix and tensor estimation. While the existing suite of AMP theory covers a wealth of applications, the theoretical guarantees for AMP remain mostly unavailable when it starts from random initialization, limiting its applicability. To address this issue, our paper delivers a nonasymptotic characterization of AMP when initialized randomly, justifying and advocating the use of random initialization in practice. In other words, a carefully designed initialization is completely unnecessary for the success of AMP.

Author affiliations: [a]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

Author contributions: G.L. and Y.W. designed research; G.L., W.F., and Y.W. performed research; G.L. and W.F. analyzed data; and G.L., W.F., and Y.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. M.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: ytwei@wharton.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2302930120/-/DCSupplemental.

Published July 25, 2023.

Inadequacy of Prior AMP Theory. Nevertheless, while the existing suite of AMP theory covers a wealth of applications, it remains inadequate in at least two aspects. To begin with, a dominant fraction of existing AMP theory is asymptotic in nature, in the sense that it predicts the AMP dynamics in the large-n limit for any fixed iteration t. For this reason, prior AMP theory falls short of describing how AMP evolves after a growing number of iterations (even when it is run for only $\log n$ iterations), which stands in contrast to other optimization-based procedures that often come with nonasymptotic analysis accommodating a large number of iterations (3, 6, 44). Another issue that further complicates matters stems from the requirement of an informative initialization, that is, existing AMP theory for low-rank estimation often requires starting from a point that already enjoys nonvanishing correlation with the true signal (45–47). While an informative initial estimate like spectral initialization is sometimes plausible and analyzable, this requirement presents a hurdle to understanding the effectiveness of other widely adopted alternatives like random initialization. This motivates the following natural questions that remain by and large open:

> *Is a warm start like spectral initialization necessary for the success of AMP? Can we start with a simpler initialization scheme but still work equally well as spectral initialization?*

Thus far, there has been no rigorous evidence precluding one from starting randomly and uninformatively. As shall be made clear shortly, tackling this issue necessitates a different and powerful nonasymptotic framework for AMP, due to the difficulty of tracking the AMP dynamics when the iterates exhibit only extremely weak correlation with the truth.

Inspired by the aforementioned issues, there has been growing interest in understanding the finite-sample performance of AMP. A seminal work by Rush and Venkataramanan (48) [see also its follow-up work (49)], studied AMP for sparse regression and permitted the total number of iterations to be as large as $o\left(\frac{\log n}{\log\log n}\right)$. This order of iteration number, however, is still highly insufficient in understanding randomly initialized AMP, as at least an order of $\log n$ iterations might be required for AMP to achieve nontrivial correlation with the truth. A recent work by Li and Wei (50) developed a nonasymptotic framework for the spiked Gaussian Wigner model, which characterized the AMP behavior for up to $O\left(\frac{n}{\mathrm{poly}(\log n)}\right)$ iterations. Although the theory therein is well suited to the studies of spectrally initialized AMP, it remains largely elusive whether it is capable of accommodating random initialization, a circumstance whose resultant initial stage is far more challenging and subtle to track.

This Paper: Randomly Initialized AMP for $\mathbb{Z}_2$ Synchronization. In this work, we take a step toward addressing the above challenges by studying a concrete model called $\mathbb{Z}_2$ synchronization. To be precise, $\mathbb{Z}_2$ synchronization is a special case of the spiked Gaussian Wigner model when the ground truth is known to have a discrete structure obeying $v^\natural \in \{\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\}^n$. Here and throughout, we impose a prior distribution on $v^\natural = [v_i^\natural]_{1\le i\le n}$ such that $v_i^\natural \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\left\{\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\right\}$, $1 \le i \le n$.

The goal is to reconstruct $v^\natural$ on the basis of the measurements M (Eq. **1**). This problem can be viewed as a basic example of a more general problem—synchronization over compact groups (1, 2, 17, 51–53)—and has an intimate connection to stochastic block models (35, 54).

The AMP Algorithm. Note that it is in general intractable to calculate the Bayes optimal solution directly due to computational difficulty in computing high-dimensional integrations/summations. A common alternative is to resort to the variational inference approximation, while the computational challenge still remains due to the nonconvexity nature of the variational inference objective. This motivates the search for computationally feasible alternatives, for which AMP emerges as a natural and successful option (46, 50, 54, 55). More concretely, given the initialization $x_0, x_1 \in \mathbb{R}^n$, AMP tailored to $\mathbb{Z}_2$ synchronization adopts the following update rule:

$$x_{t+1} = M f_t(x_t) - f_t'(x_t) f_{t-1}(x_{t-1}), \qquad t \ge 1, \tag{2}$$

where we denote $\langle x\rangle := \frac{1}{n}\sum_{i=1}^n x_i$ for any vector $x = [x_i]_{1\le i\le n}$, and the denoising function is given by*

$$f_t(x) = \gamma_t \tanh(\beta_t x), \quad \text{for } t \ge 1$$
$$\text{with } \beta_t := \sqrt{n}\,\frac{\max(|\langle x\rangle|, 1) - 1}{1 - \langle x\rangle^2}, \quad \text{and } \gamma_t := \|\tanh(\beta_t x_t)\|_2^{-1}. \tag{3}$$

Here, it is understood that the functions $f_t(\cdot)$, $f_t'(\cdot)$ and $\tanh(\cdot)$ are applied entrywise if the input argument is a vector.

Thus far, there have been two strategies to accommodate a growing number of iterations in the most challenging regime (i.e., when κ is above but very close to the information-theoretic threshold 1). One attempt was made by Celentano et al. (46), which proposed a three-stage hybrid algorithm that runs spectrally initialized AMP followed by natural gradient descent (NGD). It was conjectured therein that the third stage (i.e., NGD) is unnecessary. Recently, Li and Wei (50) put forward another strategy to address this conjecture, showing that a third refinement stage is indeed not needed as long as spectral initialization is adopted. Despite the nonconvex nature of the underlying optimization problem, AMP with spectral initialization is nearly Bayes optimal.

The Effect of Random Initialization. As alluded to previously, all existing AMP theory for this problem (45, 46, 50, 56) requires informative initialization obtained by, for example, spectral methods. By contrast, one initialization strategy that enjoys widespread adoption is to initialize AMP randomly; for instance,

$$x_1 \sim \mathcal{N}\left(0, \frac{1}{n} I_n\right) \quad \text{(independent of } M\text{)} \qquad \text{and} \quad f_0(x_0) = 0. \tag{4}$$

In order to investigate whether a warm start is required for AMP to be effective, let us first conduct a series of numerical experiments using Eq. **4**, as reported in Fig. 1. Encouragingly, AMP with random initialization seems to work surprisingly well: it only takes several tens of iterations to achieve nearly the same performance as spectrally initialized AMP (note that spectral initialization also consists of several tens of power iterations). Such encouraging numerical results motivate us to pursue in-depth theoretical understanding about the effect of random initialization upon AMP convergence, which was previously unavailable in the literature.

*Note that for ease of analysis, we adopt a slightly different scaling from that of ref. 54, but they are equivalent up to global scaling.

Fig. 1. The correlation of $_t(x_t)$ and $v^?$ (i.e., $\frac{\mathrm{jh}_t(x_t),v^? \mathrm{ij}}{\mathrm{k}\,\mathrm{k}}$) vs. iteration count t for AMP with both random and spectral initialization. Here, n = 10,000 and $v^?$ $\overset{\text{i.i.d.}}{}$ Unif($\frac{1}{}$) ($_i$ n). We generate 20 independent copies of M according to Eq. **1** and report the averaged results, with the width of the shaded region reflecting (twice) the SD. Plots (*A*) and (*B*) correspond to = 1:15 and = 1:2, respectively.

Main Contributions and Technical Challenges. In the present paper, we provide a nonasymptotic analysis that allows one to predict how AMP evolves over time from random initialization, even when the signal strength is exceedingly close to the information-theoretic limit. Our theory is able to track the correlation of the AMP iterates and the truth $v^?$. In particular, we demonstrate in Theorem 1 that the signal component in the AMP iterates increases exponentially fast at the initial stage, taking no more than O $\frac{\log n}{1}$ iterations to grow from $\Theta(\frac{1}{\sqrt{n}})$ to $O(\sqrt{\ 2\ \ 1})$ (the latter of which coincides with the correlation of spectral initialization and the truth). Furthermore, once the signal component surpasses $O(\sqrt{\ 2\ \ 1})$ in magnitude, the finite-sample AMP dynamics are very well predicted by the asymptotic state evolution recursion derived previously for any fixed t and n ! 1 (even though we are working with the finite-sample regime). Our paper characterizes the performance of AMP when initialized randomly, justifying and advocating the use of random initialization. Put another way, a carefully designed warm start is not necessary at all for this problem.

Built upon the analysis recipe recently developed by Li and Wei (50), the development of our theory requires ideas far beyond this framework in order to track AMP from random initialization. Before continuing, we take a moment to single out the key technical hurdles that need to be overcome.

- Prior theory based on state evolution analysis falls short of offering "fine-grained" understanding about the AMP iterates when they have vanishingly small correlation with the truth. More precisely, past theory fails to measure the progress of AMP during the initial stage when its signal component is of strength $o(1)$ (in fact, as small as $\Theta(\frac{1}{\sqrt{n}})$ when initialized), but instead treats the signal strength as 0 in the large-n limit.
- Another technical challenge results from the complicated statistical dependency across iterations, which is particularly difficult to cope with when the algorithm starts with random initialization and when the number of iterations grows with the dimension n. While prior literature tackles this issue for other nonconvex optimization methods by resorting to either delicate leave-one-out decoupling arguments (see, e.g. ref. 57) or global landscape analysis (see, e.g. ref. 58), these approaches remain unavailable when analyzing AMP.

Notation. Finally, let us introduce a set of notation that shall be useful throughout. We use $'(\ )$ (resp. $'_n(\ )$) to denote the probability density function (p.d.f.) of a standard Gaussian random variable (resp. a Gaussian random vector $\mathrm{N}(0, I_n)$). For any matrix M, we let kMk and kMk$_{\mathrm{F}}$ denote the spectral norm and the Frobenius norm of M, respectively. For any vector x 2 $[x_i]_{1in}$ = R^n, we denote by jxj$_{(i)}$ (resp. $x_{(i)}$) the absolute value (resp. value) of the i-th largest entry of x in magnitude. We write $\mathrm{S}^{d\ 1}$ = fx 2 R^d j kxk$_2$ = 1g as the unit sphere in ^{d}. Moreover, for any two vectors x, y R^n, we write x y for their Kronecker product, namely, x y = $(x_1y_1, :::, x_ny_n)^>$ R^n. When a function is applied to a vector, it should be understood as being applied in a component-wise fashion; for instance, for any vector x = $[x\ ]_{\ i\ 1in}$, we let x + 1 := $[x_i + 1]_{1in}$.

In addition, given two functions $f(n)$ and $g(n)$, we write $f(n)$. $g(n)$ or $f(n)$ = $O(g(n))$ to indicate that j$f(n)$j c $g_1(n)$ for some universal constant c $\geq$ 0 independent of n, and similarly, $f(n)$ & $g(n)$ means that $f(n)$ c_2j$g(n)$j for some universal constant c_2 > 0. We write $f(n)$ = $\Theta(g(n))$ if $f(n)$ = $O(g(n))$ up to logarithm factors. We also adopt the notation $f(n)$ $g(n)$ to indicate that both $f(n)$. $g(n)$ and $f(n)$ & $g(n)$ hold simultaneously. Moreover, when we write $f(n)$ $g(n)$ or $f(n)$ = $o(g(n))$, it means $f(n)$=$g(n)$! 0 as n ! 1 ; we also write $f(n)$ $g(n)$ if $g(n)$=$f(n)$! 0 as n ! 1 . We use c, C to denote universal constants that do not depend on n, whose values might change from line to line.

2. Main Results

In this section, we provide precise statements of our main theoretical guarantees for randomly initialized AMP. For notational convenience, let us introduce

$$_{t+1} := v^{?>}{}_t(x_t), \qquad \textbf{[5]}$$

which captures the projection of the t-th iterate (after denoising) onto the direction of the truth $v^\star$. In some sense, this quantity captures the size of the signal component carried by the t-th iterate. With this notation in place, we single out a key threshold

as follows:

$$\& := \min\ t : |\mu_t| \ge \frac{1}{2}\sqrt{1+\lambda^2}, \qquad [6]$$

which reflects the time taken for the AMP iterate to carry a significant signal component (note that a random initial guess obeys $|v^{\star\top} x_1| \asymp \frac{1}{\sqrt{n}}$, meaning that the initial signal component is exceedingly small). Additionally, we define the state evolution recursion starting from the &-th iteration as follows for any $t \ge \&$

$$\tau_\& = \lambda^{-1/2}|\&| \quad \text{and} \quad \tau_{t+1}^2 = \int \tanh^2(\tau_t z + \lambda)\,\varphi(z)\,dz, \qquad [7]$$

Notably, the asymptotic state evolution recursion (which is concerned with a 1-dimensional sequence in this case) is known to faithfully track the dynamics of AMP for any fixed t in the limit when $n \to \infty$, although its utility in the finite-sample regime was poorly understood in theory.

Equipped with the above definitions, our main results are summarized in the following theorem.

Theorem 1. *Consider the Z_2 synchronization problem with*

$$\lambda \ge n^{1/9}\log n. \qquad \lambda > 0.2.$$

*Suppose we run AMP (cf. Eqs. **2** and **3**) with random initialization Eq. **4**. Consider any t obeying $1 \le t \le \frac{cn(\lambda-1)^5}{\log^2 n}$, where $c > 0$ is some universal constant. Then, with probability at least $1 - O(n^{-10})$, the following results hold:*

- *(Decomposition and error bound). The AMP iterates admit the decomposition*

$$x_t = \mu_t v^\star + \sum_{k=1}^{t} \xi_{tk} u_k + w_t, \qquad [8a]$$

*where μ_t is defined in Eq. **5**, the u_k's are i.i.d. Gaussian vectors obeying $u_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n} I_n)$, and*

$$\|\xi_t\|_2 := (\xi_{t1}, \xi_{t2}, \ldots, \xi_{tt})_2 = \|\xi_t(x_t)\|_2 = 1, \qquad [8b]$$

$$\|w_t\|_2 \le \sqrt{\frac{t\log n}{n(\lambda-1)^2}} + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}; \qquad [8c]$$

- *(Crossing time). The threshold & defined in Eq. **6** satisfies*

$$\& = O\left(\frac{\log n}{\lambda-1}\right); \qquad [9]$$

- *(Nonasymptotic state evolution). For any t obeying $\& \le t \le \frac{cn(\lambda-1)^5}{\log^2 n}$, we have*

$$\mu_t \ge \left(1 + O\left(t\,\frac{\log^3 n}{(\lambda-1)^5 n}\right)\right)\tau_t, \qquad [10]$$

*where $\{\tau_t\}$ stand for the asymptotic state evolution parameters defined in Eq. **7**.*

***Remark 1* (Range of λ):** Theorem 1 only focuses on the regime where λ is larger than but close to 1. In fact, $\lambda = 1$ represents the phase transition point for Z_2 synchronization (54), in the sense that i) when $\lambda < 1$, no estimator performs better than the 0 estimator asymptotically, and ii) when λ is strictly larger than 1, it is possible to achieve nontrivial correlation with $v^\star$. We focus on the feasible regime by considering a more refined yet highly challenging case with $\lambda - 1 \gtrsim n^{-1/9}\log n$ (so that λ can be very close to 1). While it is possible to improve the exponent 1/9, it is beyond the scope of this paper. The upper bound $\lambda > 1.2$ is not crucial at all as the problem becomes easier as λ increases. In fact, our result continues to hold when $\lambda > 1.2$, which can be justified via a more refined characterization of the residual term w_t as well as ξ_t. This paper imposes this assumption $\lambda > 1.2$ merely to streamline our presentation and analysis.

Remark 2: We remark that while the iterates x_t are random quantities that depend on the randomnesses in W and $v^\star$, the decomposition Eq. **8a** is purely deterministic. For definitions and properties of $\{u_k\}_{k\le t+1}$ and $\{f_k^t(\cdot)\}_{k\le t+1}$, we refer the readers to *SI Appendix*, section A.2.2. In order to ensure that each u_k yields a homogeneous Gaussian distribution $\mathcal{N}(0, \frac{1}{n} I_n)$, we have included in f_k^t additional terms that involve extra randomnesses $\{g_j\}_{j\le k}$. These terms are properly subtracted and reflected in the residual w_{t+1}. As a result, the right-hand side of expression Eq. **8a** is a function of W and $v^\star$ and therefore measurable with respect to W and $v^\star$:

In the sequel, we provide some interpretations of Theorem 1 and discussions about its implications. It is assumed below that $\lambda > 1$.

Gaussian Approximation. The first result Eq. **8a** in Theorem 1 asserts that each AMP iterate is composed of three components: i) a signal component $\mu_t v^\star$ that aligns with the true signal $v^\star$, ii) a noise component $\sum_{k=1}^{t} \xi_{tk} u_k$ that is a linear combination of i.i.d. Gaussian vectors, and iii) a residual component w_{t+1}. While this decomposition resembles that of ref. 50, we justify its validity even in the absence of carefully designed spectral initialization. A few remarks are in order.

- Regarding the noise component, Theorem 1 implies that the 1-Wasserstein distance between its distribution (denoted by $\mathcal{P}_{\sum_{k=1}^{t} \xi_{tk} u_k}$) and a Gaussian distribution $\mathcal{N}(0, \frac{1}{n} I_n)$ is at most

$$W_1\left(\mathcal{P}_{\sum_{k=1}^{t} \xi_{tk} u_k}, \mathcal{N}\left(0, \frac{1}{n} I_n\right)\right) \le \sqrt{\frac{t}{n(\lambda-1)^2}} + \sqrt{\frac{\log n}{n}}. \qquad [11]$$

  For t not too large, the noise component well approximates a Gaussian vector $\mathcal{N}(0, \frac{1}{n} I_n)$.
- Regarding the signal component $\mu_t v^\star$, it is self-evident that μ_t governs how effective AMP is in recovering the true signal. Importantly, once $|\mu_t|$ exceeds the threshold $\frac{1}{2}\sqrt{1+\lambda^2}$, it follows a nonasymptotic state evolution that closely resembles the asymptotic counterpart τ_t (Eq. **10**), a result that is made possible thanks to the nonasymptotic nature of our analysis.

To summarize, up to a small error term at most $O\left(\sqrt{\frac{t}{n(\lambda-1)^2}} + \sqrt{\frac{1}{n(\lambda-1)^3}}\right)$, the AMP iterate is approximately

$$x_t \approx \mu_t v^\star + \mathcal{N}\left(0, \frac{1}{n} I_n\right), \qquad t < O\left(\frac{n(\lambda-1)^5}{\log^2 n}\right),$$

even when initialized randomly. An asymptotic version of this observation has been made in ref. 45, although the result therein required both informative initialization and a fixed t that does not grow with n.

Dynamics after Random Initialization. The most challenging element of Theorem 1 lies in analyzing the initial stage after random initialization. As shall be made clear from our analysis, we can understand the AMP trajectory by dividing it into three phases.

- Phase #1: escaping from random initialization. When initialized randomly with $x_1 \sim \mathcal{N}(0, \frac{1}{n} I_n)$, AMP starts with an extremely small signal component about the order of $\mathcal{O}(\frac{1}{\sqrt{n}})$, for which the canonical state evolution becomes vacuous. To overcome this technical hurdle, we develop fine-grained characterizations regarding how $|v^t|$ evolves in this phase (before $|v^t|$ surpasses $\sqrt{1/n}$), that is,

$$|v^{t+1}| \approx |v^t| + g_{t+1}, \quad \text{with } g_{t+1} \sim \mathcal{N}\Big(0, \frac{1}{n}\Big); \tag{12}$$

  see *SI Appendix*, section B.4 for details. This approximate noisy recursion tells us that while the signal component might be initially buried under the noise term, it takes at most $O(\frac{\log n}{1})$ iterations for the signal component to rise above the noise size and reach the order of $\sqrt{1/n}$ (*SI Appendix*, section A.2.2).
- Phase #2: exponential growth. Once the signal component exceeds $\sqrt{1/n}$ in size, the AMP iterate correlates nontrivially with the true signal. Interestingly, the signal strength $|v^t|$ starts to grow exponentially until reaching the order of $\sqrt{2}$ 1: As we shall justify in *SI Appendix*, section A.2.2, $|v^{t+1}|$ obeys

$$|v^{t+1}| \geq \sqrt{1 + \frac{1}{3}\Big(\frac{1}{\sqrt{2}} - o(1)\Big)}\,|v^t|, \tag{13}$$

  in this phase, which accounts for at most $O(\frac{\log n}{1})$ iterations.
- Phase #3: local refinement. Upon reaching the order of $\sqrt{2}$ 1, $|v^t|$ enters a local refinement phase, during which randomly initialized AMP behaves similarly as AMP with spectral or other informative initialization. In this phase, the asymptotic state evolution Eq. **7** also starts to be effective when predicting the evolution of $|v^t|$ (Eq. **10**). As we shall solidify in *SI Appendix*, section A.2.4, the signal strength $|v^t|$ satisfies

$$\big||v^{t+1}| - \sqrt{2}\big| \leq \Big(1 - \frac{1}{2}\Big)^{t-\tau} + O\Big(e^{-\frac{t+\frac{1}{2}}{n(1-\frac{1}{2})^5}}\Big) \tag{14}$$

  where $\sqrt{2}$ (determined by) denotes the limit of $\sqrt{t}$ as $t \to \infty$ (cf. Eq. **7**) and is unique solution of

$$\sqrt{2} = \sqrt{2}\,\mathbb{E}\tanh(\sqrt{}(\sqrt{} + G)), \quad \text{with } G \sim \mathcal{N}(0,1). \tag{15}$$

Bayes Optimality. As was shown previously [see e.g., (46, Lemma A.7)], we can construct an AMP-based estimator whose risk coincides with that of the Bayes optimal estimator $x^{\text{bayes}} := \mathbb{E}[vv^\top \mid M]$. More precisely, taking the AMP-based estimator as

$$u_t := \frac{1}{\sqrt{n(t + \frac{1}{2})}} \tanh(\cdot x_t), \tag{16}$$

its asymptotic risk satisfies [*SI Appendix*, section C and (54)]:

$$\lim_{t\to\infty}\lim_{n\to\infty} \mathbb{E}\big\|vv^\top - u_t u_t^\top\big\|_F^2 = \lim_{n\to\infty} \mathbb{E}\big\|vv^\top - x^{\text{bayes}}(x^{\text{bayes}})^\top\big\|_F^2 = 1 - \frac{?^4}{\cdot\,4} \tag{17}$$

where $?$ is the fixed point of the limiting state evolution (cf. Eq. **15**). This together with the nonasymptotic results in Theorem 1 leads to a more refined risk characterization, as we shall prove in *SI Appendix*, section C.

Corollary 1. *With probability at least* $1 - O(n^{-10})$, *there exists some* $t = O(\frac{\log n}{1})$ *such that*

$$\big\|vv^\top - u_t u_t^\top\big\|_F \leq 1 - \frac{?^2}{4} + O\Bigg(\sqrt{\frac{\log^4 n}{n(\cdot - 1)^6}}\Bigg). \tag{18}$$

In words, it only takes the AMP algorithm at most $O(\log n)$ number of iterations to achieve—up to a discrepancy of $O\big(\sqrt{\frac{1}{n(\cdot - 1)^6}}\big)$—the Bayes optimal risk.

Roadmap for the Proof of Theorem 1. To provide some intuition underlying Theorem 1, we briefly give an outline of the proof; details can be found in *SI Appendix*.

- First, focusing on the initial stage obeying $1 \leq t \leq \min\{\&, c\log n\}$ for some constant $c > 0$, we develop an upper bound on $\|v^t v^\top\|_2$ in *SI Appendix*, section A.2.1 as:

$$\|v^t v^\top\|_2 \lesssim \sqrt{t}\sqrt{\frac{\log n}{n}}; \tag{19}$$

  here, $\&$ is a threshold defined in Eq. **6**. This step, which is accomplished by means of an inductive argument, helps us justify the validity of the decomposition Eq. **8a** with small residual terms before the crossing time $\&$.
- Second, with the above decomposition Eq. **8a** in place, we can readily investigate (using the derived Gaussian approximation) how the signal strength $|v^t|$ evolves during the execution of AMP (*SI Appendix*, section A.2.2). Crucially, recalling that $\&$ reflects the first time t that satisfies $|v^t| \geq \sqrt{2} - 1$ (cf. Eq. **6**), we can use the dynamics of $|v^t|$ demonstrate that

$$\& \lesssim \frac{\log n}{1}; \tag{20}$$

  in words, in spite of random (and hence uninformative) initialization, it takes AMP at most $O(\frac{\log n}{1})$ iterations to find an informative estimate.
- Third, with the above control of $\&$ in place, we go on to develop a more complete upper bound on $\|vv^\top - u_t u_t^\top\|_2$ that covers the iterations after $\&$, that is,

$$\|v^t v^\top\|_2 \lesssim \sqrt{\frac{t\,\mathbb{1}(t > \&)\log n}{n(\cdot - 1)^2}} + \sqrt{\frac{\min\{t, \&\}^3 \log n}{n}}, \tag{21}$$

for any $t < \frac{cn(\ \ 1)^5}{\log^2 n}$. In order words, when the number of iterations grows larger that an order of $\frac{\log^3 n}{\ \ 1}$, the size of the residual scales as

$$\|\ _t\|_2 \lesssim \sqrt{\frac{t \log n}{n(\ \ 1)^2}}.$$

This is the main content of *SI Appendix*, section A.2.3, accomplished again via an inductive argument.

- Finally, after the iteration number exceeds the threshold &, we demonstrate in *SI Appendix*, section A.2.4 that the asymptotic state evolution (the one characterizing large-system limits) becomes fairly accurate in the finite-sample/finite-time regime. In particular, a connection is established between the nonasymptotic state evolution and its asymptotic analog, namely,

$$\frac{|\ ^2_{t+1} - \ ^{?2}_{t+1}|}{1 - \ ^{t}(\ \ 1)} \le \frac{|\ ^2_t - \ ^{?2}_t|}{1 - \ _t^{?2}} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\ \ 1})\log n}{n(\ \ 1)^3}}\right) \quad \text{for some } c > 0,$$

which plays a critical role in characterizing the finite-sample convergence behavior of AMP.

Comparisons to Li and Wei (50). While Li and Wei (50) provided a general decomposition for the AMP iterates $\{x_t\}$, the theory therein is far from sufficient when studying AMP from random initialization. A key reason is that during the initial stage of AMP, the signal component is vanishingly small and asymptotically vanishing compared to the magnitude of the residual. A direct application of ref. 50 leads to a vacuous upper bound on $\|\ _t\|_2$ and does not reveal the effectiveness of random initialization. In contrast, the current paper focuses on showing that the signal component will undergo a rapid growth phase and reach a level comparable to the noise. A crucial step of our analysis is to prove that $(x_t)\ \ _t v^? + \ _t\ \ _{t\ 1}$ at the initial stage, by demonstrating that $\{\ _t(x_t)\}$ are almost orthogonal to each other (see *SI Appendix*, section B.4 for more details). Based on this approximation, we then argue that it takes only $O(\log n)$ iterations for the signal strength to reach a nontrivial level. Once the signal strength has reached this level, we then proceed to uncover a new stage in which the signal strength starts to grow exponentially fast. Establishing all these phenomena requires fine-grained analyses about how AMP behaves in different stages, which was not achievable by existing analysis in ref. 50.

3. Discussions

In this paper, we have pinned down the finite-sample convergence behavior of AMP when initialized randomly, focusing on the prototypical $\mathbb{Z}_2$ synchronization problem. This algorithm has been shown to enjoy fast global convergence, as it takes no more than $O(\frac{\log n}{\ \ 1})$ iterations to arrive at a point whose risk is $O(\sqrt{\frac{\log^4 n}{n(\ \ 1)^6}})$ close to Bayes optimal. Our theory offers rigorous evidence supporting the effectiveness of randomly initialized AMP in low-rank matrix estimation. While the present paper concentrates on a specific choice of denoising functions tailored to $\mathbb{Z}_2$ synchronization, we expect our analysis framework to be generalizable to a broader family of separable and Lipschitz-continuous denoising functions.

Moving forward, there is no shortage of research directions worth exploring. One natural extension is concerned with other structural prior about $v^?$; for instance, it would be interesting to see how randomly initialized AMP performs when $v^?$ is known to satisfy general cone constraints (see e.g., refs. 59 and 60). Another direction of interest is to go beyond the spiked Gaussian Wigner model. A recent work along this line (61) studied the role of random initialization for power iteration in the problem of tensor decomposition, which leverages upon the AMP-type analysis for analyzing tensor power methods. Can we further extend these to understand (randomly initialized) AMP toward solving more challenging problems like low-rank matrix completion and tensor completion? Moreover, while AMP serves as a versatile machinery for understanding various statistical procedures in high dimensions, there are several alternative analysis frameworks like the convex Gaussian min-max theorem (CGMT) (62–64) and the leave-one-out analysis (2, 65, 66) that also prove effective and enjoy their own benefits. Is there any effective way to combine them so as to exploit all of their advantages at once? Finally, moving beyond $\mathbb{Z}_2$ synchronization, we believe that our nonasymptotic framework and the analysis ideas for understanding random initialization can both be extended to accommodate other important settings such as sparse linear regression and generalized linear models (GLMs). Take generalized approximate message passing (GAMP) for instance (27, 67), which can often be viewed as AMP applied to asymmetric matrix models. More specifically, given an asymmetric design matrix X, GAMP maintains two sequences of updates as follows

$$\begin{aligned} s_t &= XF_t(\ _t) \\ &\ \ F^0 G_{t\ 1}(s_{t\ 1}), \quad \ _{t+1} = X^\top G_t(s_t) \\ &\ \ \ _t G^0 F_t(\ _t), \end{aligned}$$

thus resembling the update rule considered in the current paper. One can then employ similar analysis ideas as in ref. 50, while in the meantime keeping track of two sets of orthogonal bases and two sequences of Gaussian random vectors. Once we are equipped with the nonasymptotic decomposition for each sequence, the role of random initialization can be understood via similar yet more complicated arguments as the ones provided in the current paper, given that these two sequences are intertwined and rely heavily on each other. We leave these questions for future investigation.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. This work was partially supported by NSF grants DMS 2147546/2015447, the NSF CAREER award DMS-2143215, and the Google Research Scholar Award.

1. A. Singer, Angular synchronization by eigenvectors and semidefinite programming. Appl. Comput. Harmonic Anal. 30, 20–36 (2011).
2. E. Abbe, J. Fan, K. Wang, Y. Zhong, Entrywise eigenvector analysis of random matrices with low expected rank. Ann. Stat. 48, 1452 (2020).
3. R. Keshavan, A. Montanari, S. Oh, Matrix completion from noisy entries. Adv. Neural Inf. Process. Syst. 22 (2009).
4. A. Lemon, A. Man-Cho So, Y. Ye, Low-rank semidefinite programming: Theory and applications. Found. Trends Opt. 2, 1–156 (2016).

5. E. J. Candes, Y. Plan, Matrix completion with noise. Proc. IEEE 98, 925–936 (2010).
6. Y. Chi, Y. M. Lu, Y. Chen, Nonconvex optimization meets low-rank matrix factorization: An overview. IEEE Trans. Sig. Process. 67, 5239–5269 (2019).
7. A. Javanmard, A. Montanari, Localization from incomplete noisy distance measurements. Found. Comput. Math. 13, 297–345 (2013).
8. S. Athey, M. Bayati, N. Doudchenko, G. Imbens, K. Khosravi, Matrix completion methods for causal panel data models. J. Am. Stat. Assoc. 116, 1716–1730 (2021).
9. S. Péché, The largest eigenvalue of small rank perturbations of Hermitian random matrices. Probability Theory Related Fields 134, 127–173 (2006).
10. J. Baik, G. B. Arous, S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann. Prob. 33, 1643–1697 (2005).
11. D. Féral, S. Péché, The largest eigenvalue of rank one deformation of large Wigner matrices. Commun. Math. Phys. 272, 185–228 (2007).
12. M. Capitaine, C. Donati-Martin, D. Féral, The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. Ann. Probab. 37, 1–47 (2009).
13. C. Cheng, Y. Wei, Y. Chen, Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. IEEE Trans. Inf. Theory 67, 7380–7419 (2021).
14. Y. Chen, Y. Chi, J. Fan, C. Ma, Spectral methods for data science: A statistical perspective. Found. Trends Mach. Learn. 14, 566–806 (2021).
15. T. Tony Cai, A. Zhang, Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. Ann. Stat. 46, 60–89 (2018).
16. Y. Yan, Y. Chen, J. Fan, Inference for heteroskedastic PCA with missing data. arXiv [Preprint] (2021). https://arxiv.org/abs/2107.12365
17. A. Perry, A. S. Wein, A. S. Bandeira, A. Moitra, Message-passing algorithms for synchronization problems over compact groups. Commun. Pure Appl. Math. 71, 2275–2322 (2018).
18. Y. Deshpande, A. Montanari, E. Richard, Cone-constrained principal component analysis. Adv. Neural Inf. Process. Syst. 27 (2014).
19. T. Lesieur, F. Krzakala, L. Zdeborová, Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. J. Stat. Mech. Theory Exp. 2017, 073403 (2017).
20. I. M. Johnstone, A. Y. Lu, On consistency and sparsity for principal components analysis in high dimensions. J. Am. Stat. Assoc. 104, 682–693 (2009).
21. Q. Berthet, P. Rigollet, Optimal detection of sparse principal components in high dimension. Ann. Stat. 41, 1780–1815 (2013).
22. O. Y. Feng, R. Venkataramanan, C. Rush, R. J. Samworth, A unifying tutorial on approximate message passing. Found. Trends Mach. Learn. 15, 335–536 (2022).
23. D. L. Donoho, A. Maleki, A. Montanari, Message-passing algorithms for compressed sensing. Proc. Natl. Acad. Sci. U.S.A. 106, 18914–18919 (2009).
24. D. L. Donoho, A. Maleki, A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction." in 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo, IEEE, 2010), pp. 1–5.
25. M. Bayati, A. Montanari, The dynamics of message passing on dense graphs, with applications to compressed sensing. IEEE Trans. Inf. Theory 57, 764–785 (2011).
26. Z. Fan, Approximate message passing algorithms for rotationally invariant matrices. Ann. Stat. 50, 197–224 (2022).
27. S. Rangan, Generalized approximate message passing for estimation with random linear mixing" in 2011 IEEE International Symposium on Information Theory Proceedings (IEEE) (2011), pp. 2168–2172.
28. M. Celentano, A. Montanari, Fundamental barriers to high-dimensional regression with convex penalties. Ann. Stat. 50, 170–196 (2022).
29. M. Bayati, A. Montanari, The LASSO risk for Gaussian matrices. IEEE Trans. Inf. Theory 58, 1997–2017 (2011).
30. Y. Li, Y. Wei, Minimum ell$_1$-norm interpolators: Precise asymptotics and multiple descent. arXiv [Preprint] (2021). https://arxiv.org/abs/2110.09502
31. P. Sur, Y. Chen, E. J. Candès, The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. Prob. Theory Related Fields 175, 487–558 (2019).
32. Y. Zhang, M. Mondelli, R. Venkataramanan, Precise asymptotics for spectral methods in mixed generalized linear models. arXiv [Preprint] (2022). http://arxiv.org/abs/2211.11368.
33. D. Donoho, A. Montanari, High dimensional robust m-estimation: Asymptotic variance via approximate message passing. Probab. Theory Relat. Fields 166, 935–969 (2016).
34. J. Ma, J. Xu, A. Maleki, Optimization-based AMP for phase retrieval: The impact of initialization and ell$_2$-regularization. arXiv [Preprint] (2018). https://arxiv.org/abs/1801.01170.
35. M. Lelarge, Léo. Miolane,, Fundamental limits of symmetric low-rank matrix estimation. Probab. Theory Relat. Fields 173, 859–929 (2019).
36. A. Javanmard, A. Montanari, F. Ricci-Tersenghi, Phase transitions in semidefinite relaxations. Proc. Natl. Acad. Sci. U.S.A. 113, E2218–E2223 (2016).
37. B. Zhiqi, J. M. Klusowski, C. Rush, W. J. Su, Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. IEEE Trans. Inf. Theory 67, 506–537 (2020).
38. P. Sur, E. J. Candès, A modern maximum-likelihood theory for high-dimensional logistic regression. Proc. Natl. Acad. Sci. U.S.A. 116, 14516–14525 (2019)
39. A. K. Fletcher, S. Rangan, Scalable inference for neuronal connectivity from calcium imaging. Adv. Neural Inf. Process Syst. 27 (2014).
40. C. Jeon, R. Ghods, A. Maleki, C. Studer, "Optimality of large mimo detection via approximate message passing" in 2015 IEEE International Symposium on Information Theory (ISIT) (IEEE) (2015), pp. 1227–1231.
41. C. Rush, A. Greig, R. Venkataramanan, Capacity-achieving sparse superposition codes via approximate message passing decoding. IEEE Trans. Inf. Theory 63, 1476–1500 (2017).
42. P. Pandit, M. Sahraee, S. Rangan, A. K. Fletcher, "Asymptotics of map inference in deep networks" in 2019 IEEE International Symposium on Information Theory (ISIT) (IEEE) (2019), pp. 842–846.
43. J. Barbier, F. Krzakala, Approximate message-passing decoder and capacity achieving sparse superposition codes. IEEE Trans. Inf. Theory 63, 4894–4927 (2017).
44. C. Ma, K. Wang, Y. Chi, Y. Chen, Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. Found. Comut. Math. 20, 451–632 (2020).
45. A. Montanari, R. Venkataramanan, Estimation of low-rank matrices via approximate message passing. Ann. Stat. 49, 321–345 (2021).
46. M. Celentano, Z. Fan, S. Mei, Local convexity of the TAP free energy and AMP convergence for Z2-synchronization. arXiv [Preprint] (2021). https://arxiv.org/abs/2106.11428.
47. X. Zhong, T. Wang, Z. Fan, Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. arXiv [Preprint] (2021). https://arxiv.org/abs/2110.02318.
48. C. Rush, R. Venkataramanan, Finite sample analysis of approximate message passing algorithms. IEEE Trans. Inf. Theory 64, 7264–7286 (2018).
49. C. Cademartori, C. Rush, A non-asymptotic analysis of generalized approximate message passing algorithms with right rotationally invariant designs. arXiv [Preprint] (2023). https://arxiv.org/abs/2302.00088.
50. G. Li, Y. Wei, A non-asymptotic framework for approximate message passing in spiked models. arXiv [Preprint] (2022). https://arxiv.org/abs/2208.03313.
51. Y. Chen, E. J. Candès, The projected power method: An efficient algorithm for joint alignment from pairwise differences. Commun. Pure Appl. Anal. 71, 1648–1714 (2018).
52. Y. Zhong, N. Boumal, Near-optimal bounds for phase synchronization. SIAM J. Optim. 28, 989–1016 (2018).
53. C. Gao, A. Y. Zhang, SDP achieves exact minimax optimality in phase synchronization. IEEE Trans. Inf. Theory (2022).
54. Y. Deshpande, E. Abbe, A. Montanari, Asymptotic mutual information for the balanced binary stochastic block model. Inf. Inference: J. IMA 6, 125–170 (2017).
55. Z. Fan, S. Mei, A. Montanari, TAP free energy, spin glasses and variational inference. Ann. Probab. 49, 1–45 (2021).
56. M. Mondelli, R. Venkataramanan, PCA initialization for approximate message passing in rotationally invariant models. Adv. Neural Inf. Process Syst. 34, 29616–29629 (2021).
57. Y. Chen, Y. Chi, J. Fan, C. Ma, Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. Math. Program. 176, 5–37 (2019).
58. R. Ge, C. Jin, Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis" in International Conference on Machine Learning (2017), pp. 1233–1242.
59. A. S. Bandeira, D. Kunisky, A. S. Wein, Computational hardness of certifying bounds on constrained PCA problems. arXiv [Preprint] (2019). http://arxiv.org/abs/1902.07324.
60. Y. Wei, M. J. Wainwright, A. Guntuboyina, The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. Ann. Stat. 47, 994–1024 (2019).
61. W. Yuchen, K. Zhou, Lower bounds for the convergence of tensor power iteration on random overcomplete models. arXiv [Preprint] (2022). http://arxiv.org/abs/2211.03827.
62. M. Celentano, A. Montanari, Y. Wei, The Lasso with general Gaussian designs with applications to hypothesis testing. arXiv [Preprint] (2020). http://arxiv.org/abs/2007.13716.
63. L. Miolane, A. Montanari, The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. Ann. Stat. 49, 2313–2335 (2021).
64. C. Thrampoulidis, E. Abbasi, B. Hassibi, Precise error analysis of regularized m-estimators in high dimensions. IEEE Trans. Inf. Theory 64, 5592–5628 (2018).
65. N. El Karoui, On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. Probab. Theory Relat. Fields 170, 95–175 (2018).
66. Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. SIAM J. Optim. 30, 3098–3121 (2020).
67. M. Mondelli, R. Venkataramanan, "Approximate message passing with spectral initialization for generalized linear models" in International Conference on Artificial Intelligence and Statistics PMLR 2021 (2021), pp. 397–405.