*The Annals of Statistics*
2022, Vol. 50, No. 1, 197–224
https://doi.org/10.1214/21-AOS2101


# APPROXIMATE MESSAGE PASSING ALGORITHMS FOR ROTATIONALLY INVARIANT MATRICES

BY ZHOU FAN

*Department of Statistics and Data Science, Yale University, zhou.fan@yale.edu*

Approximate Message Passing (AMP) algorithms have seen widespread use across a variety of applications. However, the precise forms for their Onsager corrections and state evolutions depend on properties of the underlying random matrix ensemble, limiting the extent to which AMP algorithms derived for white noise may be applicable to data matrices that arise in practice.

In this work, we study more general AMP algorithms for random matrices $\mathbf{W}$ that satisfy orthogonal rotational invariance in law, where $\mathbf{W}$ may have a spectral distribution that is different from the semicircle and Marcenko–Pastur laws characteristic of white noise. The Onsager corrections and state evolutions in these algorithms are defined by the free cumulants or rectangular free cumulants of the spectral distribution of $\mathbf{W}$. Their forms were derived previously by Opper, Çakmak and Winther using nonrigorous dynamic functional theory techniques, and we provide rigorous proofs.

Our motivating application is a Bayes-AMP algorithm for Principal Components Analysis, when there is prior structure for the principal components (PCs) and possibly nonwhite noise. For sufficiently large signal strengths and any non-Gaussian prior distributions for the PCs, we show that this algorithm provably achieves higher estimation accuracy than the sample PCs.

**1. Introduction.** Approximate Message Passing (AMP) algorithms are a general family of iterative algorithms that have seen widespread use in a variety of applications. First developed for Bayesian linear regression and compressed sensing in [26–28, 35], they have since been applied to many high-dimensional problems arising in statistics and machine learning, including Lasso estimation and sparse linear regression [4, 39], generalized linear models and phase retrieval [52, 56, 60], robust linear regression [24], sparse or structured principal components analysis (PCA) [22, 23, 44, 53], group synchronization problems [51], deep learning [12, 13, 40] and optimization in spin glass models [1, 32, 42]. We refer to [30] for a recent review.

In their basic form as described in [3], given a data matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ and an initialization $\mathbf{u}_1 \in \mathbb{R}^m$, an AMP algorithm consists of the iterative updates

$$\begin{aligned} \mathbf{z}_t &= \mathbf{W}^\top \mathbf{u}_t - b_t \mathbf{v}_{t-1}, \\ \mathbf{v}_t &= v_t(\mathbf{z}_t), \\ \mathbf{y}_t &= \mathbf{W}\mathbf{v}_t - a_t \mathbf{u}_t, \\ \mathbf{u}_{t+1} &= u_{t+1}(\mathbf{y}_t). \end{aligned}$$

Here, $a_t, b_t \in \mathbb{R}$ are two sequences of debiasing coefficients, and $v_t : \mathbb{R} \to \mathbb{R}$ and $u_{t+1} : \mathbb{R} \to \mathbb{R}$ are two sequences of functions applied entrywise to $\mathbf{z}_t \in \mathbb{R}^n$ and $\mathbf{y}_t \in \mathbb{R}^m$. By appropriately designing these functions $v_t$ and $u_{t+1}$, possibly to also depend on additional "side information" such as response variables in regression problems, this basic iteration may be applied to perform optimization or Bayes posterior-mean estimation in the above applications.

Received October 2020; revised April 2021.
*MSC2020 subject classifications.* 62E20.
*Key words and phrases.* AMP, free probability theory, high-dimensional asymptotics, PCA.

A defining characteristic of the AMP algorithm is the subtraction of the two "memory" terms $b_t \mathbf{v}_{t-1}$ and $a_t \mathbf{u}_t$ in the definitions of $\mathbf{z}_t$ and $\mathbf{y}_t$, known as the *Onsager corrections*. This achieves the effect of removing a bias of $\mathbf{W}^\top \mathbf{u}_t$ and $\mathbf{W}\mathbf{v}_t$ in the directions of the preceding iterates, so that as $m, n \to \infty$, the empirical distributions of $\mathbf{y}_t$ and $\mathbf{z}_t$ converge to certain Gaussian limits

$$\mathbf{y}_t \to \mathcal{N}(0, \sigma_t^2) \quad \text{and} \quad \mathbf{z}_t \to \mathcal{N}(0, \omega_t^2). \tag{1.1}$$

This was proven rigorously in the Sherrington–Kirkpatrick model by Bolthausen in [11] and for general AMP algorithms of the above form by Bayati and Montanari in [3], and various extensions have been established in [2, 10, 19, 25, 34]. The description of the variances $\sigma_t^2$ and $\omega_t^2$ across iterations is known as the algorithm's *state evolution*. This ability to characterize the distributions of the iterates is a major appeal of the AMP approach, and has enabled a more precise theoretical understanding of many high-dimensional statistical estimators and the development of associated inference procedures that quantify statistical uncertainty [14, 45, 59–61].

A drawback of AMP algorithms, however, is that the correct forms of the debiasing coefficients $a_t, b_t$ and resulting variances $\sigma_t^2, \omega_t^2$ depend on the properties of the data matrix $\mathbf{W}$. When $\mathbf{W}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries, these quantities are given explicitly by

$$a_t = \langle v_t'(\mathbf{z}_t) \rangle, \qquad b_t = \gamma \langle u_t'(\mathbf{y}_{t-1}) \rangle, \qquad \sigma_t^2 = \langle v_t(\mathbf{z}_t)^2 \rangle, \qquad \omega_t^2 = \gamma \langle u_t(\mathbf{y}_{t-1})^2 \rangle,$$

where $\gamma = m/n$, $u_t'(\cdot), v_t'(\cdot), u_t(\cdot)^2, v_t(\cdot)^2$ denote the derivatives and squares of $u_t, v_t$ applied entrywise, and $\langle \cdot \rangle$ denotes the empirical average of coordinates. It has been shown in [2, 19] that these forms enjoy a certain amount of universality, being valid also for $\mathbf{W}$ having i.i.d. non-Gaussian entries. Extensions to $\mathbf{W}$ having independent entries with several blocks of differing variances were derived in [25, 34]. Unfortunately, these results do not apply to $\mathbf{W}$ with more complex correlation structure, which is common in data applications. A sizeable body of work has developed alternative algorithms or damping procedures to address this shortcoming [17, 18, 31, 36, 37, 41, 48–50, 54, 55, 57, 62], and the connections between several of these algorithms were discussed recently in [38]. However, many such algorithms are no longer characterized by a rigorous state evolution, and some have been empirically observed to exhibit slow convergence or divergent behavior.

1.1. *Contributions.* We develop a rigorous extension of general AMP procedures of the above form to rotationally invariant matrices. We then apply these general algorithms to a prototypical "structured PCA" problem of estimating a rank-one matrix in (possibly nonwhite) noise. In this PCA application, we develop a Bayes-AMP algorithm that provably achieves lower mean-squared-error than the rank-one estimate constructed from the sample principal components (PCs), for any sufficiently large signal strength and any prior distributions of the PCs that are not mean-zero Gaussian laws.

Let us first describe the general AMP algorithm in the simpler setting of symmetric square matrices. We study matrices $\mathbf{W} \in \mathbb{R}^{n \times n}$ that satisfy the equality in law

$$\mathbf{W} \overset{L}{=} \tilde{\mathbf{O}}^\top \mathbf{W} \tilde{\mathbf{O}}$$

for any deterministic orthogonal matrix $\tilde{\mathbf{O}} \in \mathbb{R}^{n \times n}$. Equivalently, such matrices admit the eigendecomposition $\mathbf{W} = \mathbf{O}^\top \mathbf{\Lambda} \mathbf{O}$ where the eigenvectors $\mathbf{O} \in \mathbb{R}^{n \times n}$ are independent of the eigenvalues $\mathbf{\Lambda}$ and are uniformly distributed on the orthogonal group. The AMP algorithm will take the form

$$\mathbf{z}_t = \mathbf{W}\mathbf{u}_t - b_{t1}\mathbf{u}_1 - b_{t2}\mathbf{u}_2 - \cdots - b_{tt}\mathbf{u}_t, \tag{1.2}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t), \tag{1.3}$$

where the coefficients $b_{ts}$ are defined so that each $\mathbf{z}_t$ has an empirical Gaussian limit as in (1.1). For greater generality and applicability, we will allow $u_{t+1} : \mathbb{R}^t \to \mathbb{R}$ to be a function of all previous iterates $\mathbf{z}_1, \dots, \mathbf{z}_t$, rather than only the preceding iterate $\mathbf{z}_t$. (Outside of the i.i.d. Gaussian setting, the full debiasing of $\mathbf{W}\mathbf{u}_t$ by $\mathbf{u}_1, \dots, \mathbf{u}_t$ is necessary even if $u_{t+1}(\cdot)$ depends only on $\mathbf{z}_t$.) The correct forms for $b_{t1}, \dots, b_{tt}$ and the corresponding state evolution

$$(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_t) \to \mathcal{N}(0, \mathbf{\Sigma}_t) \tag{1.4}$$

were derived previously by Opper, Çakmak and Winther using nonrigorous dynamic functional theory techniques [47]. These forms depend on the free cumulants of the eigenvalue distribution of $\mathbf{W}$, and we describe them in Section 4.1. Our work provides a rigorous proof of the validity of this state evolution.

In the rectangular setting, we study birotationally invariant matrices $\mathbf{W} \in \mathbb{R}^{m \times n}$ satisfying the equality in law

$$\mathbf{W} \overset{L}{=} \tilde{\mathbf{O}}^\top \mathbf{W} \tilde{\mathbf{Q}}$$

for any deterministic orthogonal matrices $\tilde{\mathbf{O}} \in \mathbb{R}^{m \times m}$ and $\tilde{\mathbf{Q}} \in \mathbb{R}^{n \times n}$. Equivalently, such matrices admit the singular value decomposition $\mathbf{W} = \mathbf{O}\mathbf{\Lambda}\mathbf{Q}^\top$ where the singular vectors $\mathbf{O} \in \mathbb{R}^{m \times m}$ and $\mathbf{Q} \in \mathbb{R}^{n \times n}$ are independent of the singular values $\mathbf{\Lambda}$ and are both uniformly distributed over the orthogonal groups. The analogous AMP algorithm takes the form

$$\mathbf{z}_t = \mathbf{W}^\top \mathbf{u}_t - b_{t1}\mathbf{v}_1 - b_{t2}\mathbf{v}_2 - \cdots - b_{t,t-1}\mathbf{v}_{t-1}, \tag{1.5}$$

$$\mathbf{v}_t = v_t(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_t), \tag{1.6}$$

$$\mathbf{y}_t = \mathbf{W}\mathbf{v}_t - a_{t1}\mathbf{u}_1 - a_{t2}\mathbf{u}_2 - \cdots - a_{tt}\mathbf{u}_t, \tag{1.7}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_t). \tag{1.8}$$

We describe in Section 5.1 the forms of the debiasing coefficients $a_{ts}, b_{ts}$ and the corresponding state evolutions

$$(\mathbf{y}_1, \dots, \mathbf{y}_t) \to \mathcal{N}(0, \mathbf{\Sigma}_t), \qquad (\mathbf{z}_1, \dots, \mathbf{z}_t) \to \mathcal{N}(0, \mathbf{\Omega}_t),$$

which are related to the rectangular free cumulants of the singular value distribution of $\mathbf{W}$ as introduced in [5, 6]. This algorithm has also been derived recently and independently in [16], using a dynamic functional theory approach similar to [47].

These classes of rotationally invariant matrices include, but are not restricted to, $\mathbf{W}$ having i.i.d. Gaussian entries. Importantly, the spectral distribution of $\mathbf{W}$ can be arbitrary, rather than following the behavior prescribed by the semicircle or Marcenko–Pastur law. Our primary motivation for studying such rotationally invariant models is that we expect the resulting AMP algorithms to be valid under a much larger universality class of matrices $\mathbf{W}$ than AMP algorithms derived in the i.i.d. Gaussian setting, and that this class may provide a more flexible model for data matrices arising in practice.

In the contexts of compressed sensing and generalized linear models, alternative "vector AMP" or "orthogonal AMP" approaches for rotationally-invariant matrices have been developed in [37, 54, 57, 62], and rigorous state evolutions for these algorithms were also derived. These derivations are based on analyses of denoising functions that satisfy the divergence-free conditions

$$\langle \partial_s v_t(\mathbf{z}_1, \dots, \mathbf{z}_t) \rangle = 0, \qquad \langle \partial_s u_{t+1}(\mathbf{y}_1, \dots, \mathbf{y}_t) \rangle = 0 \quad \text{for all } s \le t. \tag{1.9}$$

A similar idea was used in [15] to develop an algorithm for solving the TAP equations for Ising models with rotationally-invariant couplings. Analyses of certain "long-memory" Convolutional AMP algorithms for compressed sensing, related to our work, were recently

carried out in [63–65] by mapping these algorithms to a divergence-free form. Our proofs build on the insight in [54, 62] that Bolthausen's conditioning technique may be applied to rotationally-invariant models. However, we derive directly the forms of the Onsager corrections and state evolutions for AMP algorithms that do not restrict $v_t(\cdot)$ and $u_{t+1}(\cdot)$ to be divergence-free, in a general setting that extends beyond compressed sensing applications. We clarify the relation between certain long-memory algorithms and the AMP algorithms of [3] for Gaussian matrices, by relating their Onsager corrections and state evolutions to the free cumulants of the spectral distribution of $\mathbf{W}$.

1.2. *Organization of paper.* Section 2 establishes preliminary background and notation on Wasserstein convergence of empirical measures and free cumulants. Section 3 first discusses the specific application of structured PCA and the Bayes-AMP algorithms for this application that specialize the more general AMP algorithms to follow. Section 4 describes the general AMP algorithm and state evolution for symmetric square matrices, and Section 5 describes the analogous general algorithm for rectangular matrices. Section 6 provides a high-level overview of the proofs, which are contained in the supplementary Appendices [29].

## 2. Preliminaries on Wasserstein convergence and free probability.

2.1. *Notation and conventions.* For vectors $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^m$, we denote

$$\langle \mathbf{v} \rangle = \frac{1}{n}\sum_{i=1}^{n} v_i, \qquad \langle \mathbf{w} \rangle = \frac{1}{m}\sum_{i=1}^{m} w_i.$$

For a matrix $(\mathbf{v}_1, \dots, \mathbf{v}_k) \in \mathbb{R}^{n\times k}$ and a function $f : \mathbb{R}^k \to \mathbb{R}$, we write $f(\mathbf{v}_1, \dots, \mathbf{v}_k) \in \mathbb{R}^n$ as its rowwise evaluation.

For a weakly differentiable function $u : \mathbb{R}^k \to \mathbb{R}$, we denote by $\partial_s u$ (any version of) its $s$th partial derivative. For a matrix $(\mathbf{v}_1, \dots, \mathbf{v}_k) \in \mathbb{R}^{n\times k}$, we write $\Pi_{(\mathbf{v}_1,\dots,\mathbf{v}_k)} \in \mathbb{R}^{n\times n}$ for the orthogonal projection onto the linear span of $(\mathbf{v}_1, \dots, \mathbf{v}_k)$, and $\Pi_{(\mathbf{v}_1,\dots,\mathbf{v}_k)^\perp} = \mathrm{Id} - \Pi_{(\mathbf{v}_1,\dots,\mathbf{v}_k)}$ for the projection onto its orthogonal complement. Id is the identity matrix, and we write $\mathrm{Id}_{k\times k}$ to specify the dimension $k$. We will use the convention

$$\mathbf{M}^0 = \mathrm{Id}$$

for the zero-th power of any square matrix $\mathbf{M}$, even if some eigenvalues of $\mathbf{M}$ may be 0.

Products over the empty set are equal to 1, and sums over the empty set are equal to 0. $\|\cdot\|$ is the $\ell_2$ norm for vectors and $\ell_2 \to \ell_2$ operator norm for matrices. $\|\mathbf{v}\|_\infty = \max_i |v_i|$ is the vector $\ell_\infty$ norm, and $\|\mathbf{M}\|_F = (\sum_{i,j} m_{ij}^2)^{1/2}$ is the matrix Frobenius norm.

2.2. *Wasserstein convergence of empirical distributions.*

DEFINITION 2.1. For $p \geq 1$, a matrix $(\mathbf{v}_1, \dots, \mathbf{v}_k) = (v_{i,1}, \dots, v_{i,k})_{i=1}^n \in \mathbb{R}^{n\times k}$, and a probability distribution $\mathcal{L}$ over $\mathbb{R}^k$ or a random vector $(V_1, \dots, V_k) \sim \mathcal{L}$, we write

$$(\mathbf{v}_1, \dots, \mathbf{v}_k) \overset{W_p}{\to} \mathcal{L} \quad \text{or} \quad (\mathbf{v}_1, \dots, \mathbf{v}_k) \overset{W_p}{\to} (V_1, \dots, V_k)$$

for the convergence of the empirical distribution of rows of $(\mathbf{v}_1, \dots, \mathbf{v}_k)$ to $\mathcal{L}$ in the Wasserstein space of order $p$. This means, for any $C > 0$ and continuous function $f : \mathbb{R}^k \to \mathbb{R}$ satisfying

$$|f(v_1, \dots, v_k)| \leq C\big(1 + \|(v_1, \dots, v_k)\|^p\big), \tag{2.1}$$

as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} f(v_{i,1}, \dots, v_{i,k}) \to \mathbb{E}\big[f(V_1, \dots, V_k)\big]. \tag{2.2}$$

Implicit in this notation is the finite moment condition $\mathbb{E}_{(V_1,\dots,V_k)\sim\mathcal{L}}[\|(V_1, \dots, V_k)\|^p] < \infty$.

We write

$$(\mathbf{v}_1, \dots, \mathbf{v}_k) \overset{W}{\to} \mathcal{L} \quad \text{or} \quad (\mathbf{v}_1, \dots, \mathbf{v}_k) \overset{W}{\to} (V_1, \dots, V_k)$$

to mean that this convergence holds for every fixed $p \geq 1$, where $\mathcal{L}$ has finite moments of all orders.

We will use a certain calculus associated to these notation $\overset{W_p}{\to}$ and $\overset{W}{\to}$, which we review in Appendix E. By [66], Definition 6.7, to show that (2.2) holds for all continuous functions $f$ satisfying (2.1), it suffices to check that it holds for all bounded Lipschitz functions $f$ together with the function $f(v_1, \dots, v_k) = \|(v_1, \dots, v_k)\|^p$. See Chapter 6 of [66] for further background.

2.3. *Free cumulants*. We briefly review the notion of free cumulants, and refer readers to [46] for a more thorough and motivated introduction.

Let $X$ be a random variable with finite moments of all orders, and denote $m_k = \mathbb{E}[X^k]$. In what follows, the law of $X$ will be the empirical eigenvalue distribution of a symmetric matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$. Let $\mathrm{NC}(k)$ be the set of all noncrossing partitions of $\{1, \dots, k\}$. The free cumulants $\kappa_1, \kappa_2, \kappa_3, \dots$ of $X$ are defined recursively by the moment-cumulant relations

$$m_k = \sum_{\pi\in\mathrm{NC}(k)} \prod_{S\in\pi} \kappa_{|S|}, \tag{2.3}$$

where $|S|$ is the cardinality of the set $S \in \pi$. The first four free cumulants may be computed to be

$$\begin{aligned}
\kappa_1 &= m_1 = \mathbb{E}[X],\\
\kappa_2 &= m_2 - m_1^2 = \mathrm{Var}[X],\\
\kappa_3 &= m_3 - 3m_2m_1 + 2m_1^3,\\
\kappa_4 &= m_4 - 4m_3m_1 - 2m_2^2 + 10m_2m_1^2 - 5m_1^4,
\end{aligned}$$

where $\kappa_4$ is the first free cumulant that differs from the classical cumulants. The free cumulants linearize free additive convolution, describing the eigenvalue distribution of sums of freely independent symmetric square matrices. If $X$ has the Wigner semicircle law supported on $[-2, 2]$, then

$$\kappa_1 = 0, \qquad \kappa_2 = 1, \qquad \kappa_j = 0 \quad \text{for all } j \geq 3.$$

Defining the formal generating functions

$$M(z) = 1 + \sum_{k=1}^{\infty} m_k z^k, \qquad R(z) = \sum_{k=1}^{\infty} \kappa_k z^{k-1},$$

the relations (2.3) are equivalent to an identity of formal series (see [46], Section 2.5)

$$M(z) = 1 + zM(z) \cdot R\big(zM(z)\big).$$

Here, $R(z)$ is the R-transform of $X$. Comparing the coefficients of $z^k$ on both sides, each free cumulant $\kappa_k$ may be computed from $m_1, \dots, m_k$ and $\kappa_1, \dots, \kappa_{k-1}$ as

$$\kappa_k = m_k - [z^k] \sum_{j=1}^{k-1} \kappa_j \big(z + m_1 z^2 + m_2 z^3 + \cdots + m_{k-1} z^k\big)^j,$$

where $[z^k](q(z))$ denotes the coefficient of $z^k$ in the polynomial $q(z)$.

2.4. *Rectangular free cumulants.* For rectangular matrices $\mathbf{W} \in \mathbb{R}^{m \times n}$, we review the notion of rectangular free cumulants developed in [5]. This is an example of the operator-valued free cumulants described in [58], where freeness is with amalgamation over a 2-dimensional subalgebra corresponding to the $2 \times 2$ block structure of $\mathbb{R}^{(m+n)\times(m+n)}$.

We fix an aspect ratio parameter

$$\gamma = m/n > 0.$$

Let $X$ be a random variable with finite moments of all orders, and denote the even moments by $m_{2k} = \mathbb{E}[X^{2k}]$. The law of $X^2$ will be the empirical eigenvalue distribution of $\mathbf{W}\mathbf{W}^\top \in \mathbb{R}^{m\times m}$, so that $m_{2k}$ is the $k$th moment of this distribution. Define also an auxiliary sequence of even moments

$$\bar{m}_{2k} = \begin{cases} 1 & \text{if } k = 0, \\ \gamma \cdot m_{2k} & \text{if } k \geq 1. \end{cases} \tag{2.4}$$

Since the eigenvalues of $\mathbf{W}\mathbf{W}^\top$ and $\mathbf{W}^\top\mathbf{W}$ coincide up to the addition or removal of $|m-n|$ zeros, the value $\bar{m}_{2k}$ is the $k$th moment of the empirical eigenvalue distribution of $\mathbf{W}^\top\mathbf{W} \in \mathbb{R}^{n\times n}$.

Let $\mathrm{NC}'(2k)$ be the noncrossing partitions $\pi$ of $\{1, \dots, 2k\}$ where each set $S \in \pi$ has even cardinality. Then we may define two sequences of rectangular free cumulants $\kappa_2, \kappa_4, \kappa_6, \dots$ and $\bar{\kappa}_2, \bar{\kappa}_4, \bar{\kappa}_6, \dots$ by the moment-cumulant relations

$$m_{2k} = \sum_{\pi \in \mathrm{NC}'(2k)} \prod_{\substack{S\in\pi \\ \min S \text{ is odd}}} \kappa_{|S|} \cdot \prod_{\substack{S\in\pi \\ \min S \text{ is even}}} \bar{\kappa}_{|S|},$$

$$\bar{m}_{2k} = \sum_{\pi \in \mathrm{NC}'(2k)} \prod_{\substack{S\in\pi \\ \min S \text{ is odd}}} \bar{\kappa}_{|S|} \cdot \prod_{\substack{S\in\pi \\ \min S \text{ is even}}} \kappa_{|S|}.$$

See [5], equations (8)–(9). These cumulants have a simple relation given by

$$\bar{\kappa}_{2k} = \gamma \cdot \kappa_{2k} \quad \text{for all } k \geq 1, \tag{2.5}$$

so outside of the proofs, we will always refer to the first sequence $\{\kappa_{2k}\}_{k\geq 1}$ for simplicity.

Letting $e(\pi)$ be the number of sets $S \in \pi$ where the smallest element of $S$ is even, and letting $o(\pi)$ be the number where the smallest element is odd, applying (2.5) above implies

$$m_{2k} = \sum_{\pi \in \mathrm{NC}'(2k)} \gamma^{e(\pi)} \prod_{S\in\pi} \kappa_{|S|}, \qquad \bar{m}_{2k} = \sum_{\pi \in \mathrm{NC}'(2k)} \gamma^{o(\pi)} \prod_{S\in\pi} \kappa_{|S|}. \tag{2.6}$$

See also [5], Proposition 3.1. The first four rectangular free cumulants may be computed as

$$\begin{aligned}
\kappa_2 &= m_2 = \mathbb{E}\big[X^2\big], \\
\kappa_4 &= m_4 - (1+\gamma) m_2^2, \\
\kappa_6 &= m_6 - (3+3\gamma) m_4 m_2 + \big(2+3\gamma+2\gamma^2\big) m_2^3, \\
\kappa_8 &= m_8 - (4+4\gamma) m_6 m_2 - (2+2\gamma) m_4^2 + \big(10+16\gamma+10\gamma^2\big) m_4 m_2^2 \\
&\quad - \big(5+10\gamma+10\gamma^2+5\gamma^3\big) m_2^4.
\end{aligned}$$

The rectangular free cumulants linearize rectangular free additive convolution, describing the singular value distribution of sums of freely independent rectangular matrices. If $X^2$ has the Marcenko–Pastur law with aspect ratio $\gamma$, then

$$\kappa_2 = 1, \qquad \kappa_{2j} = 0 \quad \text{for all } j \geq 2.$$

The rectangular free cumulants may be computed from the following relation of generating functions: Let

$$M(z) = \sum_{k=1}^{\infty} m_{2k} z^k, \qquad R(z) = \sum_{k=1}^{\infty} \kappa_{2k} z^k.$$

Here, $R(z)$ is the rectangular R-transform of $X$. Then

$$M(z) = R\big(z\big(\gamma M(z) + 1\big)\big(M(z) + 1\big)\big); \tag{2.7}$$

see [5], Lemma 3.4. Thus, comparing the coefficients of $z^k$ on both sides, each value $\kappa_{2k}$ may be computed from $m_2, \ldots, m_{2k}$ and $\kappa_2, \ldots, \kappa_{2k-2}$ as

$$\kappa_{2k} = m_{2k} - \big[z^k\big] \sum_{j=1}^{k-1} \kappa_{2j} \big(z\big(\gamma M(z) + 1\big)\big(M(z) + 1\big)\big)^{2j},$$

where $[z^k](q(z))$ again denotes the coefficient of $z^k$ in the polynomial $q(z)$.

REMARK 2.2. The reasons for the appearance of the square/rectangular free cumulants in the forms of the Onsager corrections and state evolution for AMP are somewhat opaque in our work, as they will arise from a certain combinatorial unfolding of the moment-cumulant relations on the noncrossing partition lattice; we discuss this further in Section 6. Their emergence is conceptually clearer in the (nonrigorous, but illuminating) analysis of the limit characteristic function of the AMP iterates in [47], where they arise instead from the evaluation of a low-rank HCIZ integral over the Haar-orthogonal randomness in $\mathbf{W}$, and from the coefficients of the series expansion of the R-transform that describes this integral. See [20, 33] and [7] for this connection in the square and rectangular settings, respectively.

**3. Structured principal components analysis.** We study the problem of estimating a rank-one signal matrix in possibly nonwhite noise, where the singular vectors of the rank-one signal have some "prior" structure. For sufficiently large signal strengths, we describe a Bayes-AMP algorithm that provably achieves lower mean-squared-error than the rank-one estimate constructed from the sample principal components. This extends the types of AMP algorithms that were studied for i.i.d. Gaussian noise in [22, 23, 44, 53].

3.1. *Symmetric square matrices.* Suppose first that we observe a symmetric data matrix

$$\mathbf{X} = \frac{\alpha}{n} \mathbf{u}_* \mathbf{u}_*^\top + \mathbf{W} \in \mathbb{R}^{n \times n}$$

and seek to estimate $\mathbf{u}_* \in \mathbb{R}^n$. Writing the eigendecomposition $\mathbf{W} = \mathbf{O}^\top \mathbf{\Lambda} \mathbf{O}$ where $\mathbf{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$, we assume that $\mathbf{W}$ is rotationally-invariant in law and that as $n \to \infty$, the empirical distributions of $\boldsymbol{\lambda}$ and $\mathbf{u}_*$ satisfy

$$\boldsymbol{\lambda} \overset{W}{\to} \Lambda, \qquad \mathbf{u}_* \overset{W}{\to} U_* \tag{3.1}$$

for two limit laws $\Lambda$ and $U_*$. This notation $\overset{W}{\to}$ denotes Wasserstein convergence at all orders, as discussed in Section 2.2. To fix the scaling, we take $\|\mathbf{u}_*\| = \sqrt{n}$, so that

$$\mathbb{E}\big[U_*^2\big] = \lim_{n \to \infty} \frac{1}{n} \|\mathbf{u}_*\|^2 = 1.$$

Here, the law of $\Lambda$ is the limit spectral distribution of $\mathbf{W}$. The law of $U_*$ represents a prior distribution for the entries of $\mathbf{u}_*$, which may reflect assumptions of sparsity [22], nonnegativity [43], or a discrete support that encodes cluster or community membership [21].

We assume for simplicity that we have an initialization $\mathbf{u}_1 \in \mathbb{R}^n$ independent of $\mathbf{W}$, satisfying the joint empirical convergence

$$(\mathbf{u}_1, \mathbf{u}_*) \overset{W}{\to} (U_1, U_*), \qquad \mathbb{E}[U_1 U_*] > 0. \tag{3.2}$$

We then estimate $\mathbf{u}_*$ by the iterates $\mathbf{u}_t$ of an AMP algorithm

$$\mathbf{f}_t = \mathbf{X}\mathbf{u}_t - b_{t1}\mathbf{u}_1 - \cdots - b_{tt}\mathbf{u}_t, \tag{3.3}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{f}_t). \tag{3.4}$$

It will be shown that each iterate $\mathbf{f}_t$ behaves like $\mathbf{u}_*$ corrupted by entrywise Gaussian noise, so we take each function $u_{t+1}(\cdot)$ to be a scalar denoiser that estimates $\mathbf{u}_*$ from $\mathbf{f}_t$.

To describe the forms of the debiasing coefficients $b_{t1}, \dots, b_{tt}$, let us write $\lambda_1(\mathbf{X}) \geq \cdots \geq \lambda_n(\mathbf{X})$ as the eigenvalues of $\mathbf{X}$. For each $k \geq 1$, let

$$m_k = \frac{1}{n}\sum_{i=2}^{n} \lambda_i(\mathbf{X})^k \tag{3.5}$$

be the $k$th moment of the empirical eigenvalue distribution of $\mathbf{X}$ excluding its largest eigenvalue. Let $\{\kappa_k\}_{k\geq 1}$ be the free cumulants corresponding to this sequence of moments $\{m_k\}_{k\geq 1}$, as defined in Section 2.3. It is easy to check that under the assumption (3.1), as $n \to \infty$,

$$m_k \to m_k^\infty = \mathbb{E}[\Lambda^k], \qquad \kappa_k \to \kappa_k^\infty$$

for each fixed $k \geq 1$, where these limits are the moments and free cumulants of the limit spectral distribution $\Lambda$ of the noise $\mathbf{W}$. The debiasing coefficients in (3.3) are set as

$$b_{tt} = \kappa_1, \qquad b_{t,t-j} = \kappa_{j+1} \prod_{i=t-j+1}^{t} \langle u_i'(\mathbf{f}_{i-1}) \rangle \quad \text{for } j = 1, \dots, t-1. \tag{3.6}$$

The state evolution that describes the AMP iterations (3.3)–(3.4) is expressed in terms of a sequence of mean vectors $\boldsymbol{\mu}_T^\infty = (\mu_t^\infty)_{1\leq t\leq T}$ and covariance matrices $\boldsymbol{\Sigma}_T^\infty = (\sigma_{st}^\infty)_{1\leq s,t\leq T}$, defined recursively as follows: For $T = 1$, we set

$$\mu_1^\infty = \alpha \cdot \mathbb{E}[U_1 U_*], \qquad \sigma_{11}^\infty = \kappa_2^\infty \mathbb{E}[U_1^2].$$

Having defined $\boldsymbol{\mu}_T^\infty$ and $\boldsymbol{\Sigma}_T^\infty$, we denote

$$\begin{aligned} &U_t = u_t(F_{t-1}) \quad \text{for } t = 2, \dots, T+1, \\ &(F_1, \dots, F_T) = \boldsymbol{\mu}_T^\infty \cdot U_* + (Z_1, \dots, Z_T) \quad \text{and} \\ &(Z_1, \dots, Z_T) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_T^\infty) \quad \text{independent of } (U_1, U_*). \end{aligned} \tag{3.7}$$

We then define $\boldsymbol{\mu}_{T+1}^\infty$ and $\boldsymbol{\Sigma}_{T+1}^\infty$ to have the entries, for $1 \leq s, t \leq T+1$,

$$\begin{aligned} \mu_t^\infty &= \alpha \cdot \mathbb{E}[U_t U_*], \\ \sigma_{st}^\infty &= \sum_{j=0}^{s-1}\sum_{k=0}^{t-1} \kappa_{j+k+2}^\infty \Bigg( \prod_{i=s-j+1}^{s} \mathbb{E}[u_i'(F_{i-1})] \Bigg) \\ &\quad \times \Bigg( \prod_{i=t-k+1}^{t} \mathbb{E}[u_i'(F_{i-1})] \Bigg) \mathbb{E}[U_{s-j} U_{t-k}]. \end{aligned} \tag{3.8}$$

In the limit $n \to \infty$, the iterates of (3.3) will satisfy the second-order Wasserstein convergence

$$(\mathbf{f}_1, \dots, \mathbf{f}_T, \mathbf{u}_*) \overset{W_2}{\to} (F_1, \dots, F_T, U_*).$$

Thus, the rows of $(\mathbf{f}_1, \dots, \mathbf{f}_T)$ behave like Gaussian vectors with mean $\boldsymbol{\mu}_T^\infty \cdot U_*$ and covariance $\boldsymbol{\Sigma}_T^\infty$.

As one example of choosing the functions $u_{t+1}(\cdot)$, let us analyze this state evolution for the following "single-iterate posterior mean" denoisers: In the scalar Gaussian observation model,

$$F = \mu \cdot U_* + Z, \quad Z \sim \mathcal{N}(0, \sigma^2) \text{ independent of } U_*, \tag{3.9}$$

we denote the Bayes posterior-mean estimate of $U_*$ as

$$\eta(f \mid \mu, \sigma^2) = \mathbb{E}[U_* \mid F = f] = \frac{\mathbb{E}[U_* \exp(-(f - \mu \cdot U_*)^2/2\sigma^2)]}{\mathbb{E}[\exp(-(f - \mu \cdot U_*)^2/2\sigma^2)]}. \tag{3.10}$$

We denote the Bayes-optimal mean-squared-error of this estimate as

$$\operatorname{mmse}(\mu^2/\sigma^2) = \mathbb{E}\big[(U_* - \eta(F \mid \mu, \sigma^2))^2\big]. \tag{3.11}$$

The single-iterate posterior mean denoiser is the choice

$$u_{t+1}(f_t) = \eta(f_t \mid \mu_t^\infty, \sigma_{tt}^\infty), \tag{3.12}$$

where $\mu_t^\infty$ and $\sigma_{tt}^\infty$ are the above state evolution parameters that describe the univariate Gaussian law of $F_t$. These parameters may be replaced by consistent estimates in practice.

Let $R(x)$ be the R-transform of the limit spectral distribution $\Lambda$, as discussed in Section 2.3, and let $R'(x)$ be its derivative. For small $|x|$, these may be defined by the convergent series (see Proposition F.3)

$$R(x) = \sum_{k=1}^{\infty} \kappa_k^\infty x^{k-1}. \tag{3.13}$$

THEOREM 3.1. *Suppose* $\mathbf{W} = \mathbf{O}^\top \boldsymbol{\Lambda} \mathbf{O} \in \mathbb{R}^{n \times n}$ *where* $\mathbf{O}$ *is a Haar-uniform orthogonal matrix. Let* $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$, *where* $(\boldsymbol{\lambda}, \mathbf{u}_1, \mathbf{u}_*)$ *are independent of* $\mathbf{O}$, $\|\mathbf{u}_*\| = \sqrt{n}$, *and*

$$\boldsymbol{\lambda} \overset{W}{\to} \Lambda, \qquad (\mathbf{u}_1, \mathbf{u}_*) \overset{W}{\to} (U_1, U_*)$$

*almost surely as* $n \to \infty$. *Suppose* $\mathbb{E}[U_1^2] \le 1$, $\mathbb{E}[U_1 U_*] = \varepsilon > 0$, *and* $\|\boldsymbol{\lambda}\|_\infty \le C_0$ *almost surely for all large* $n$ *and some constants* $C_0, \varepsilon > 0$.

(a) *Let* $\alpha \ge 0$, *and let each function* $u_{t+1}(\cdot)$ *be continuously differentiable and Lipschitz on* $\mathbb{R}$. *Then for each fixed* $T \ge 1$, *almost surely as* $n \to \infty$,

$$(\mathbf{u}_1, \dots, \mathbf{u}_{T+1}, \mathbf{f}_1, \dots, \mathbf{f}_T, \mathbf{u}_*) \overset{W_2}{\to} (U_1, \dots, U_{T+1}, F_1, \dots, F_T, U_*),$$

*where the joint law of this limit is described by* (3.7).

(b) *Suppose each function* $u_{t+1}(\cdot)$ *is the posterior-mean denoiser in* (3.12), *and suppose this is Lipschitz on* $\mathbb{R}$. *Then there exist constants* $C, \alpha_0 > 0$ *depending only on* $C_0, \varepsilon$ *such that for all* $\alpha > \alpha_0$, *defining* $I_\Delta = [1 - C/\alpha^2, 1]$ *and* $I_\Sigma = [\kappa_2^\infty/2, 3\kappa_2^\infty/2]$, *there is a unique fixed point* $(\Delta_*, \Sigma_*) \in I_\Delta \times I_\Sigma$ *to the equations*

$$1 - \Delta_* = \operatorname{mmse}\left(\frac{\alpha^2 \Delta_*^2}{\Sigma_*}\right), \qquad \Sigma_* = \Delta_* R'\left(\frac{\alpha \Delta_* (1 - \Delta_*)}{\Sigma_*}\right). \tag{3.14}$$

*Furthermore*,

$$\lim_{T \to \infty} \left( \lim_{n \to \infty} \frac{1}{n} \mathbf{u}_T^\top \mathbf{u}_* \right) = \lim_{T \to \infty} \left( \lim_{n \to \infty} \frac{1}{n} \|\mathbf{u}_T\|^2 \right) = \Delta_*. \tag{3.15}$$

The proof of this result is provided in Appendix C. As discussed in Section 2.2, the notation $\xrightarrow{W_2}$ in part (a) guarantees that for any continuous function $f : \mathbb{R}^{2T+2} \to \mathbb{R}$ satisfying $\mathbb{E}[f(U_1, \dots, U_{T+1}, Z_1, \dots, Z_T, U_*)^2] < \infty$,

$$\langle f(\mathbf{u}_1, \dots, \mathbf{u}_{T+1}, \mathbf{z}_1, \dots, \mathbf{z}_T, \mathbf{u}_*)\rangle \to \mathbb{E}[f(U_1, \dots, U_{T+1}, Z_1, \dots, Z_T, U_*)],$$

where the left-hand side is the empirical average of this function $f$ evaluated across the $n$ rows.

REMARK 3.2. Theorem 3.1(b) implies that the asymptotic matrix mean-squared-error of the rank-one estimate $\mathbf{u}_T\mathbf{u}_T^\top$ for $\mathbf{u}_*\mathbf{u}_*^\top$, in the limit $T \to \infty$, is given by

$$\begin{aligned}\mathrm{MSE} &\equiv \lim_{T\to\infty}\left(\lim_{n\to\infty}\frac{1}{n^2}\|\mathbf{u}_T\mathbf{u}_T^\top - \mathbf{u}_*\mathbf{u}_*^\top\|_F^2\right)\\ &= \lim_{T\to\infty}\left(\lim_{n\to\infty}\frac{1}{n^2}(\|\mathbf{u}_T\|^2)^2 - \frac{2}{n^2}(\mathbf{u}_T^\top\mathbf{u}_*)^2 + \frac{1}{n^2}(\|\mathbf{u}_*\|^2)^2\right) = 1 - \Delta_*^2.\end{aligned}$$

Let us compare this with the matrix mean-squared-error of the best PCA estimate $c \cdot \hat{\mathbf{u}}_{\mathrm{PCA}}\hat{\mathbf{u}}_{\mathrm{PCA}}^\top$ optimized over $c > 0$, where $\hat{\mathbf{u}}_{\mathrm{PCA}}$ is the leading sample eigenvector of $\mathbf{X}$. Normalizing $\hat{\mathbf{u}}_{\mathrm{PCA}}$ such that $\|\hat{\mathbf{u}}_{\mathrm{PCA}}\| = \|\mathbf{u}_*\| = \sqrt{n}$, [8], Theorem 2.2(a), shows for sufficiently large $\alpha$ that

$$\lim_{n\to\infty}\left(\frac{1}{n}\hat{\mathbf{u}}_{\mathrm{PCA}}^\top\mathbf{u}_*\right)^2 = \Delta_{\mathrm{PCA}} \equiv \frac{-1}{\alpha^2 G'(G^{-1}(1/\alpha))}, \tag{3.16}$$

where $G(z) = \mathbb{E}[(z - \Lambda)^{-1}]$ is the Cauchy transform of $\Lambda$, and $G^{-1}(z)$ is the functional inverse of $G$ (which is well defined for small $|z|$). Then

$$\begin{aligned}\mathrm{MSE}_{\mathrm{PCA}} &\equiv \min_{c>0}\left(\lim_{n\to\infty}\frac{1}{n^2}\|c\cdot\hat{\mathbf{u}}_{\mathrm{PCA}}\hat{\mathbf{u}}_{\mathrm{PCA}}^\top - \mathbf{u}_*\mathbf{u}_*^\top\|_F^2\right)\\ &= \min_{c>0} c^2 - 2c\Delta_{\mathrm{PCA}} + 1 = 1 - \Delta_{\mathrm{PCA}}^2,\end{aligned}$$

with the minimum attained at the rescaling $c = \Delta_{\mathrm{PCA}} < 1$.

To see that $1 - \Delta_*^2 \le 1 - \Delta_{\mathrm{PCA}}^2$, observe that for any prior distribution $U_*$ satisfying our normalization $\mathbb{E}[U_*^2] = 1$, we have

$$\mathrm{mmse}(\mu^2/\sigma^2) \le \frac{1}{1 + \mu^2/\sigma^2}. \tag{3.17}$$

This is because under the scalar observation model (3.9), the right-hand side of (3.17) is the risk $\mathbb{E}[(\hat{U} - U_*)^2]$ of the linear estimator $\hat{U} = (\mu/(\sigma^2 + \mu^2))F$, which upper bounds the Bayes risk on the left-hand side of (3.17). Equality holds in (3.17) if and only if $\hat{U}$ is the Bayes estimator in this model, that is, if and only if the prior distribution is $U_* \sim \mathcal{N}(0, 1)$. Applying (3.17) to the first equation of (3.14) and rearranging, we obtain

$$\frac{\alpha\Delta_*(1 - \Delta_*)}{\Sigma_*} \le \frac{1}{\alpha}.$$

Now applying this to the second equation of (3.14), and using that $\kappa_2^\infty = \mathrm{Var}[\Lambda] > 0$ so that the function $xR'(x) = \kappa_2^\infty x + 2\kappa_3^\infty x^2 + 3\kappa_4^\infty x^3 + \cdots$ is increasing in a neighborhood of 0, we have for $\alpha > \alpha_0$ sufficiently large that

$$1 - \Delta_* = \frac{1}{\alpha}\cdot\frac{\alpha\Delta_*(1 - \Delta_*)}{\Sigma_*}R'\left(\frac{\alpha\Delta_*(1 - \Delta_*)}{\Sigma_*}\right) \le \frac{1}{\alpha^2}R'\left(\frac{1}{\alpha}\right).$$

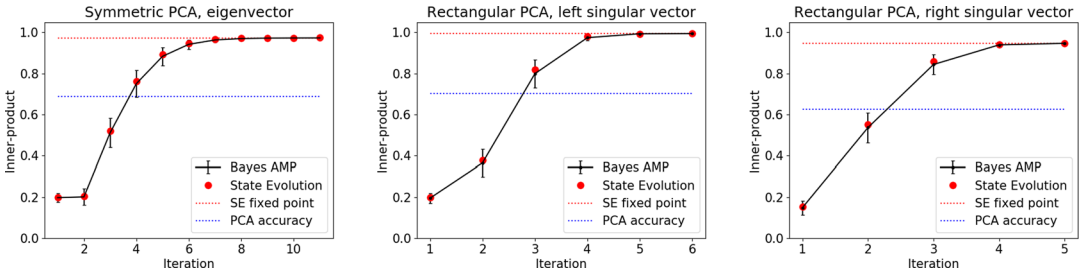


FIG. 1. *Simulations of the Bayes-AMP algorithms for PCA, with priors* $U_*, V_* \sim \text{Uniform}\{+1, -1\}$ *and the single-iterate posterior mean denoisers in* (3.12) *and* (3.27). *Shown are the mean and std. dev. of* $\langle \mathbf{u}_t \mathbf{u}_* \rangle$ *and* $\langle \mathbf{v}_t \mathbf{v}_* \rangle$ *across* 100 *simulations in black, their state evolution predictions computed from* (3.8), (3.24) *and* (3.26) *in red dots, and the fixed points* $\Delta_*$ *and* $\Gamma_*$ *of* (3.14) *and* (3.28) *in dashed red. For comparison,* $\Delta_{\text{PCA}}$ *and* $\Gamma_{\text{PCA}}$ *corresponding to the sample PCs are in dashed blue. Left*: $\mathbf{u}_1$ (*initialization*), $\mathbf{u}_2, \dots, \mathbf{u}_{11}$ *for symmetric square* $\mathbf{W}$ *with* $n = 2000$, $\alpha = 2.5$, *and eigenvalue distribution given by centering and scaling* Beta(1, 2) *to mean* 0 *and variance* 1. *Middle and right*: $\mathbf{u}_1$ (*initialization*), $\mathbf{u}_2, \dots, \mathbf{u}_6$ *and* $\mathbf{v}_1, \dots, \mathbf{v}_5$ *for rectangular* $\mathbf{W}$ *with* $m = 2000$, $n = 4000$, $\gamma = 0.5$, $\alpha = 1.5$, *and singular value distribution given by rescaling* Beta(1, 2) *to second-moment* 1.

Differentiating the R-transform identity $R(x) = G^{-1}(x) - 1/x$, this is equivalently written as

$$\Delta_* \geq 1 - \frac{1}{\alpha^2} R'\left(\frac{1}{\alpha}\right) = \frac{-1}{\alpha^2 G'(G^{-1}(1/\alpha))} = \Delta_{\text{PCA}},$$

so that

$$\text{MSE} = 1 - \Delta_*^2 \leq 1 - \Delta_{\text{PCA}}^2 = \text{MSE}_{\text{PCA}} \tag{3.18}$$

as desired. Equality holds here if and only if equality holds in (3.17), that is, when $U_* \sim \mathcal{N}(0, 1)$. Thus, for any signal strength $\alpha > \alpha_0$ sufficiently large and any distribution of $U_*$ other than $\mathcal{N}(0, 1)$, the above AMP algorithm achieves strictly better estimation accuracy than PCA.

An illustration of the algorithm and state evolution is presented in the left panel of Figure 1, with noise eigenvalues drawn from a centered and rescaled Beta(1, 2) distribution. We observe a close agreement with the state evolution predictions at sample size $n = 2000$, and a significant improvement in estimation accuracy over the naive principal components for this prior distribution $U_* \sim \text{Uniform}\{+1, -1\}$. Let us remark that although carrying out many iterations of this AMP algorithm would require estimating successively higher-order free cumulants of the spectral distribution of $\mathbf{W}$, for large signal strengths $\alpha$ the algorithm only needs a very small number of iterations to converge.

REMARK 3.3. In this algorithm, the Onsager corrections involving the free cumulants may be understood as iteratively constructing the series (3.13) for $R(\alpha\Delta_*(1 - \Delta_*)/\Sigma_*)$, whose derivative appears in the characterization of the fixed point in Theorem 3.1. This is somewhat analogous to the single-step-memory algorithm in [47] for solving the TAP equations in a related Ising model, which alternatively constructs a series for the inverse R-transform.

The convergence condition and final mean-squared-error of this algorithm are likely not Bayes-optimal. For example, we believe that the convergence of (3.3)–(3.4) requires convergence of the series (3.13) at $x = \alpha\Delta_*(1 - \Delta_*)/\Sigma_*$, which (depending on the spectral law of $\mathbf{W}$) may impose a stronger condition for the signal strength $\alpha$ than the spectral phase transition. One natural way to improve upon the algorithm is to consider more generally

$$u_{t+1}(\mathbf{f}_1, \dots, \mathbf{f}_t) = \eta\big(c_{t1}\mathbf{f}_1 + \cdots + c_{tt}\mathbf{f}_t \mid \mathbf{c}_t^\top \boldsymbol{\mu}_t^\infty, \mathbf{c}_t^\top \boldsymbol{\Sigma}_t^\infty \mathbf{c}_t\big)$$

for a vector $\mathbf{c}_t = (c_{t1}, \dots, c_{tt})$ in each iteration, or specialize this to $\mathbf{c}_t = (\mathbf{\Sigma}_t^\infty)^{-1}\boldsymbol{\mu}_t^\infty$ to obtain the posterior mean estimate of $U_*$ given all previous observations $(F_1, \dots, F_t)$. Our general results describe also the state evolution for these extensions, but analyses of their fixed points are more involved, and we will not pursue this in the current work.

These procedures differ from the "Vector AMP" or "memory-free" algorithms of [15, 54], whose forms may be derived from the Expectation Propagation framework of [41]. These latter algorithms operate directly on a resolvent of $\mathbf{W}$ and use divergence-free nonlinearities, corresponding to $\mathbf{\Phi}_t = \mathbf{\Psi}_t = 0$ in our notation to follow. Thus their state evolutions have simpler forms that depend on the first two moments of the resolvent but not (explicitly) on the free cumulants of $\mathbf{W}$. PCA differs from the applications in [15, 47, 54] in two important ways: First, the log-likelihood of $\mathbf{X}$ given $\mathbf{u}$ is not quadratic in $\mathbf{u}$ under general spectral laws of $\mathbf{W}$. Second, the noise matrix $\mathbf{W}$ is not directly observed in PCA, and its resolvent cannot be directly computed. Due to these differences, we believe that extending the algorithmic ideas of [15, 54] to PCA may be an interesting open question to study in future work.

3.2. *Rectangular matrices.* Consider now a rectangular data matrix

$$\mathbf{X} = \frac{\alpha}{m}\mathbf{u}_*\mathbf{v}_*^\top + \mathbf{W} \in \mathbb{R}^{m\times n},$$

and the task of estimating $\mathbf{u}_* \in \mathbb{R}^m$ and $\mathbf{v}_* \in \mathbb{R}^n$. Writing the singular value decomposition $\mathbf{W} = \mathbf{O}^\top\mathbf{\Lambda}\mathbf{Q}$ where $\mathbf{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$ and $\boldsymbol{\lambda} \in \mathbb{R}^{\min(m,n)}$, we assume that $\mathbf{W}$ is birotationally invariant in law and that

$$m/n = \gamma, \qquad \boldsymbol{\lambda} \overset{W}{\to} \Lambda, \qquad \mathbf{u}_* \overset{W}{\to} U_*, \qquad \mathbf{v}_* \overset{W}{\to} V_*$$

as $m, n \to \infty$, for some constant $\gamma \in (0, \infty)$ and some limit laws $\Lambda, U_*, V_*$. We fix the scalings $\|\mathbf{u}_*\| = \sqrt{m}$ and $\|\mathbf{v}_*\| = \sqrt{n}$, so that

$$\mathbb{E}[U_*^2] = \mathbb{E}[V_*^2] = 1.$$

Note that the rank-one signal component $(\alpha/m)\mathbf{u}_*\mathbf{v}_*^\top$ has singular value $\alpha/\sqrt{\gamma}$.

We again assume that we have an initialization $\mathbf{u}_1 \in \mathbb{R}^m$ independent of $\mathbf{W}$, for which

$$(\mathbf{u}_1, \mathbf{u}_*) \overset{W}{\to} (U_1, U_*), \qquad \mathbb{E}[U_1U_*] > 0.$$

We then estimate $\mathbf{u}_*$ and $\mathbf{v}_*$ by the iterates $\mathbf{u}_t$ and $\mathbf{v}_t$ of an AMP algorithm

$$\mathbf{g}_t = \mathbf{X}^\top\mathbf{u}_t - b_{t1}\mathbf{v}_1 - \cdots - b_{t,t-1}\mathbf{v}_{t-1}, \tag{3.19}$$

$$\mathbf{v}_t = v_t(\mathbf{g}_t), \tag{3.20}$$

$$\mathbf{f}_t = \mathbf{X}\mathbf{v}_t - a_{t1}\mathbf{u}_1 - \cdots - a_{tt}\mathbf{u}_t, \tag{3.21}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{f}_t), \tag{3.22}$$

where $u_{t+1}(\cdot)$ and $v_t(\cdot)$ are scalar denoisers that estimate $\mathbf{u}_*$ and $\mathbf{v}_*$ from $\mathbf{f}_t$ and $\mathbf{g}_t$.

To describe the forms of the debiasing coefficients $a_{ts}$ and $b_{ts}$, let us define $\boldsymbol{\lambda}_m \in \mathbb{R}^m$ to be $\boldsymbol{\lambda}$ if $m \le n$ or $\boldsymbol{\lambda}$ extended by $m - n$ additional 0's if $m > n$. We will work instead with the limit

$$\boldsymbol{\lambda}_m \overset{W}{\to} \Lambda_m,$$

which is a mixture of $\Lambda$ and a point mass at 0 if $\gamma = m/n > 1$. Denoting the singular values of $\mathbf{X}$ by $\lambda_1(\mathbf{X}) \ge \cdots \ge \lambda_{\min(m,n)}(\mathbf{X})$, for each $k \ge 1$ we set

$$m_{2k} = \frac{1}{m}\sum_{i=2}^{\min(m,n)} \lambda_i(\mathbf{X})^{2k}.$$

We then define $\{\kappa_{2k}\}_{k\geq 1}$ as the rectangular free cumulants associated to these even moments $\{m_{2k}\}_{k\geq 1}$ and aspect ratio $\gamma$, as defined in Section 2.4. It is easily checked that as $m, n \to \infty$,

$$m_{2k} \to m_{2k}^{\infty} = \mathbb{E}\big[\Lambda_m^{2k}\big], \qquad \kappa_{2k} \to \kappa_{2k}^{\infty},$$

where these limits are the even moments and rectangular free cumulants of $\Lambda_m$. Then the debiasing coefficients in (3.19)–(3.22) are set as

$$a_{t,t-j} = \kappa_{2(j+1)}\big\langle v_t'(\mathbf{g}_t)\big\rangle \prod_{i=t-j+1}^{t} \big\langle u_i'(\mathbf{f}_{i-1})\big\rangle\big\langle v_{i-1}'(\mathbf{g}_{i-1})\big\rangle \quad \text{for } j = 0, \dots, t-1,$$

$$b_{t,t-j} = \gamma\kappa_{2j}\big\langle u_t'(\mathbf{f}_{t-1})\big\rangle \prod_{i=t-j+1}^{t-1} \big\langle v_i'(\mathbf{g}_i)\big\rangle\big\langle u_i'(\mathbf{f}_{i-1})\big\rangle \quad \text{for } j = 1, \dots, t-1.$$

We use the convention that empty products equal 1, so the first coefficients here are simply

$$a_{tt} = \kappa_2\big\langle v_t'(\mathbf{g}_t)\big\rangle, \qquad b_{t,t-1} = \gamma\kappa_2\big\langle u_t'(\mathbf{f}_{t-1})\big\rangle.$$

The state evolution for this algorithm may be expressed in terms of two sequences of mean vectors $\boldsymbol{\mu}_T^{\infty} = (\mu_t^{\infty})_{1\leq t\leq T}$ and $\boldsymbol{\nu}_T^{\infty} = (\nu_t^{\infty})_{1\leq t\leq T}$ and covariance matrices $\boldsymbol{\Sigma}_T^{\infty} = (\sigma_{st}^{\infty})_{1\leq s,t\leq T}$ and $\boldsymbol{\Omega}_T^{\infty} = (\omega_{st}^{\infty})_{1\leq s,t\leq T}$, defined as follows: For $T = 1$, we set

$$\nu_1^{\infty} = \alpha \cdot \mathbb{E}[U_1 U_*], \qquad \omega_{11}^{\infty} = \gamma\kappa_2^{\infty} \cdot \mathbb{E}\big[U_1^2\big].$$

Having defined $\boldsymbol{\mu}_{T-1}^{\infty}$, $\boldsymbol{\Sigma}_{T-1}^{\infty}$, $\boldsymbol{\nu}_T^{\infty}$, and $\boldsymbol{\Omega}_T^{\infty}$, we denote

$$\begin{aligned}
&U_t = u_t(F_{t-1}) \quad \text{for } t = 2, \dots, T,\\
&(F_1, \dots, F_{T-1}) = \boldsymbol{\mu}_{T-1}^{\infty} \cdot U_* + (Y_1, \dots, Y_{T-1}),\\
&(Y_1, \dots, Y_{T-1}) \sim \mathcal{N}\big(0, \boldsymbol{\Sigma}_{T-1}^{\infty}\big) \text{ independent of } (U_1, U_*),\\
&V_t = v_t(G_t) \quad \text{for } t = 1, \dots, T,\\
&(G_1, \dots, G_T) = \boldsymbol{\nu}_T^{\infty} \cdot V_* + (Z_1, \dots, Z_T),\\
&(Z_1, \dots, Z_T) \sim \mathcal{N}\big(0, \boldsymbol{\Omega}_T^{\infty}\big) \text{ independent of } V_*.
\end{aligned} \tag{3.23}$$

We then define $\boldsymbol{\mu}_T^{\infty}$ and $\boldsymbol{\Sigma}_T^{\infty}$ with the entries, for $1 \leq s, t \leq T$,

$$\begin{aligned}
\mu_t^{\infty} &= (\alpha/\gamma) \cdot \mathbb{E}[V_t V_*],\\
\sigma_{st}^{\infty} &= \sum_{j=0}^{s-1}\sum_{k=0}^{t-1}\Bigg(\prod_{i=s-j+1}^{s} \mathbb{E}\big[v_i'(G_i)\big]\mathbb{E}\big[u_i'(F_{i-1})\big]\Bigg)\\
&\quad \times \Bigg(\prod_{i=t-k+1}^{t} \mathbb{E}\big[v_i'(G_i)\big]\mathbb{E}\big[u_i'(F_{i-1})\big]\Bigg)\\
&\quad \times \big(\kappa_{2(j+k+1)}^{\infty}\mathbb{E}[V_{s-j}V_{t-k}] + \kappa_{2(j+k+2)}^{\infty}\mathbb{E}\big[v_{s-j}'(G_{s-j})\big]\\
&\quad \times \mathbb{E}\big[v_{t-k}'(G_{t-k})\big]\mathbb{E}[U_{s-j}U_{t-k}]\big).
\end{aligned} \tag{3.24}$$

Now having defined $\boldsymbol{\mu}_T^{\infty}$ and $\boldsymbol{\Sigma}_T^{\infty}$, we extend (3.23) to

$$\begin{aligned}
&U_t = u_t(F_{t-1}) \quad \text{for } t = 2, \dots, T+1,\\
&(F_1, \dots, F_T) = \boldsymbol{\mu}_T^{\infty} \cdot U_* + (Y_1, \dots, Y_T),\\
&(Y_1, \dots, Y_T) \sim \mathcal{N}\big(0, \boldsymbol{\Sigma}_T^{\infty}\big) \text{ independent of } (U_1, U_*)
\end{aligned} \tag{3.25}$$

and define $\boldsymbol{\nu}_{T+1}^{\infty}$ and $\boldsymbol{\Omega}_{T+1}^{\infty}$ with the entries, for $1 \le s, t \le T+1$,

$$\begin{aligned}
\nu_t^{\infty} &= \alpha \cdot \mathbb{E}[U_t U_*], \\
\omega_{st}^{\infty} &= \gamma \sum_{j=0}^{s-1} \sum_{k=0}^{t-1} \left( \prod_{i=s-j+1}^{s} \mathbb{E}[u_i'(F_{i-1})] \mathbb{E}[v_{i-1}'(G_{i-1})] \right) \\
&\quad \times \left( \prod_{i=t-k+1}^{t} \mathbb{E}[u_i'(F_{i-1})] \mathbb{E}[v_{i-1}'(G_{i-1})] \right) \\
&\quad \times \left( \kappa_{2(j+k+1)}^{\infty} \mathbb{E}[U_{s-j} U_{t-k}] + \kappa_{2(j+k+2)}^{\infty} \mathbb{E}[u_{s-j}'(F_{s-j-1})] \right. \\
&\quad \left. \times \mathbb{E}[u_{t-k}'(F_{t-k-1})] \mathbb{E}[V_{s-j-1} V_{t-k-1}] \right).
\end{aligned} \tag{3.26}$$

We use the convention $V_0 = 0$, so that the second term of (3.26) is 0 for $j = s-1$ or $k = t-1$. In the limit $m, n \to \infty$, the iterates of (3.19)–(3.22) will satisfy

$$(\mathbf{f}_1, \dots, \mathbf{f}_T, \mathbf{u}_*) \overset{W_2}{\to} (F_1, \dots, F_T, U_*), \qquad (\mathbf{g}_1, \dots, \mathbf{g}_T, \mathbf{v}_*) \overset{W_2}{\to} (G_1, \dots, G_T, V_*).$$

As an example of choices for $v_t(\cdot)$ and $u_{t+1}(\cdot)$, let us again analyze the single-iterate posterior mean denoisers given by

$$v_t(g_t) = \eta(g_t \mid \nu_t, \omega_{tt}), \qquad u_{t+1}(f_t) = \eta(f_t \mid \mu_t, \sigma_{tt}), \tag{3.27}$$

where $\eta(\cdot)$ is as defined in (3.10), and $(\nu_t, \omega_{tt})$ and $(\mu_t, \sigma_{tt})$ are the state evolution parameters describing the univariate Gaussian laws of $G_t$ and $F_t$. We denote by $\mathrm{mmse}(\cdot)$ the scalar mean-squared-error function from (3.11), and by $R(x)$ the rectangular R-transform of $\Lambda_m$ with aspect ratio $\gamma$, as discussed in Section 2.4. This may be defined for small $|x|$ by the convergent series (see Proposition F.3)

$$R(x) = \sum_{k=1}^{\infty} \kappa_{2k}^{\infty} x^k,$$

where $\kappa_{2k}^{\infty}$ are the rectangular free cumulants of $\Lambda_m$ above. We denote $R'(x)$ as its derivative, and

$$S(x) = \left( \frac{R(x)}{x} \right)' = \frac{x R'(x) - R(x)}{x^2}.$$

THEOREM 3.4. *Suppose* $\mathbf{W} = \mathbf{O}^{\top} \boldsymbol{\Lambda} \mathbf{Q} \in \mathbb{R}^{m \times n}$ *where* $\mathbf{Q}$ *and* $\mathbf{O}$ *are Haar-uniform orthogonal matrices. Let* $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, *where* $(\boldsymbol{\lambda}, \mathbf{u}_1, \mathbf{u}_*, \mathbf{v}_*)$ *are independent of* $(\mathbf{O}, \mathbf{Q})$, $\|\mathbf{u}_*\| = \sqrt{n}$, $\|\mathbf{v}_*\| = \sqrt{m}$, *and*

$$\boldsymbol{\lambda} \overset{W}{\to} \Lambda, \qquad (\mathbf{u}_1, \mathbf{u}_*) \overset{W}{\to} (U_1, U_*), \qquad \mathbf{v}_* \overset{W}{\to} V_*, \qquad m/n = \gamma \in (0, \infty)$$

*as* $m, n \to \infty$. *Suppose* $\mathbb{E}[U_1^2] \le 1$, $\mathbb{E}[U_1 U_*] = \varepsilon > 0$, *and* $\|\boldsymbol{\lambda}\|_{\infty} \le C_0$, *almost surely for all large* $n$ *and some constants* $C_0, \varepsilon > 0$.

(i) *Let* $\alpha \ge 0$, *and let each function* $v_t(\cdot)$ *and* $u_{t+1}(\cdot)$ *be continuously differentiable and Lipschitz on* $\mathbb{R}$. *Then for each fixed* $T \ge 1$, *almost surely as* $m, n \to \infty$,

$$(\mathbf{u}_1, \dots, \mathbf{u}_{T+1}, \mathbf{f}_1, \dots, \mathbf{f}_T, \mathbf{u}_*) \overset{W}{\to} (U_1, \dots, U_{T+1}, F_1, \dots, F_T, U_*),$$

$$(\mathbf{v}_1, \dots, \mathbf{v}_T, \mathbf{g}_1, \dots, \mathbf{g}_T, \mathbf{v}_*) \overset{W}{\to} (V_1, \dots, V_T, G_1, \dots, G_T, V_*),$$

*where these limits are as defined in* (3.23) *and* (3.25).

(ii) *Suppose* $v_t(\cdot), u_{t+1}(\cdot)$ *are the posterior-mean denoisers in* (3.27) *and are Lipschitz on* $\mathbb{R}$. *There exist constants* $C, \alpha_0 > 0$ *depending only on* $C_0, \varepsilon, \gamma$ *such that for all* $\alpha > \alpha_0$, *setting*

$$I_\Delta = I_\Gamma = [1 - C/\alpha^2, 1], \qquad I_\Sigma = [\kappa_2^\infty/2, 3\kappa_2^\infty/2], \qquad I_\Omega = \gamma \cdot I_\Sigma,$$

*there is a unique fixed point* $(\Delta_*, \Sigma_*, \Gamma_*, \Omega_*, X_*) \in I_\Delta \times I_\Sigma \times I_\Gamma \times I_\Omega \times \mathbb{R}$ *to the equations*

$$\begin{aligned} X_* &= \frac{\alpha^2 \Delta_* \Gamma_* (1-\Delta_*)(1-\Gamma_*)}{\gamma \Sigma_* \Omega_*}, \qquad 1 - \Delta_* = \operatorname{mmse}\left(\frac{\alpha^2 \Gamma_*^2}{\gamma^2 \Sigma_*}\right), \\ 1 - \Gamma_* &= \operatorname{mmse}\left(\frac{\alpha^2 \Delta_*^2}{\Omega_*}\right), \\ \Sigma_* &= \Gamma_* R'(X_*) + \frac{\alpha^2 \Delta_*^3 (1-\Gamma_*)^2}{\Omega_*^2} S(X_*), \\ \Omega_* &= \gamma \Delta_* R'(X_*) + \frac{\alpha^2 \Gamma_*^3 (1-\Delta_*)^2}{\gamma \Sigma_*^2} S(X_*). \end{aligned} \tag{3.28}$$

*Furthermore*,

$$\begin{aligned} \lim_{T\to\infty}\left(\lim_{m,n\to\infty} \frac{1}{m}\mathbf{u}_T^\top \mathbf{u}_*\right) &= \lim_{T\to\infty}\left(\lim_{m,n\to\infty} \frac{1}{m}\|\mathbf{u}_T\|^2\right) = \Delta_*, \\ \lim_{T\to\infty}\left(\lim_{m,n\to\infty} \frac{1}{n}\mathbf{v}_T^\top \mathbf{v}_*\right) &= \lim_{T\to\infty}\left(\lim_{m,n\to\infty} \frac{1}{n}\|\mathbf{v}_T\|^2\right) = \Gamma_*. \end{aligned}$$

The proof of this result is provided in Appendix C.

REMARK 3.5. As in the symmetric square setting of Remark 3.2, the above fixed points imply that the asymptotic matrix mean-squared error is given by

$$\mathrm{MSE} \equiv \lim_{T\to\infty}\left(\lim_{m,n\to\infty} \frac{1}{mn}\left\|\mathbf{u}_T \mathbf{v}_T^\top - \mathbf{u}_* \mathbf{v}_*^\top\right\|_F^2\right) = 1 - \Delta_* \Gamma_*.$$

We may compare this with the asymptotic error of the PCA estimate: Assume without loss of generality that $\gamma = m/n \le 1$. Let $\hat{\mathbf{u}}_{\mathrm{PCA}}$ and $\hat{\mathbf{v}}_{\mathrm{PCA}}$ be the leading left and right singular vectors of $\mathbf{X}$, with the scalings $\|\hat{\mathbf{u}}_{\mathrm{PCA}}\| = \|\mathbf{u}_*\| = \sqrt{m}$ and $\|\hat{\mathbf{v}}_{\mathrm{PCA}}\| = \|\mathbf{v}_*\| = \sqrt{n}$. Recall that the singular value of the rank-one signal $(\alpha/m)\mathbf{u}_*\mathbf{v}_*^\top$ is $\alpha/\sqrt{\gamma}$, and set

$$x = \gamma/\alpha^2.$$

Then [9], Theorem 2.9, shows

$$\lim_{m,n\to\infty}\left(\frac{1}{m}\hat{\mathbf{u}}_{\mathrm{PCA}}^\top \mathbf{u}_*\right)^2 = \Delta_{\mathrm{PCA}} \equiv \frac{-2x\varphi(D^{-1}(x))}{D'(D^{-1}(x))}, \tag{3.29}$$

$$\lim_{m,n\to\infty}\left(\frac{1}{n}\hat{\mathbf{v}}_{\mathrm{PCA}}^\top \mathbf{v}_*\right)^2 = \Gamma_{\mathrm{PCA}} \equiv \frac{-2x\bar{\varphi}(D^{-1}(x))}{D'(D^{-1}(x))}, \tag{3.30}$$

where

$$\varphi(z) = \mathbb{E}\left[\frac{z}{z^2 - \Lambda_m^2}\right], \qquad \bar{\varphi}(z) = \gamma\varphi(z) + \frac{1-\gamma}{z}, \qquad D(z) = \varphi(z)\bar{\varphi}(z), \tag{3.31}$$

and $D^{-1}(z)$ is the functional inverse of $D$ for small $|z|$. Then the matrix mean-squared error for the best rescaling of the PCA estimate is

$$\begin{aligned}\mathrm{MSE}_{\mathrm{PCA}} &\equiv \min_{c>0}\Big(\lim_{m,n\to\infty}\frac{1}{mn}\big\|c\cdot\hat{\mathbf{u}}_{\mathrm{PCA}}\hat{\mathbf{v}}_{\mathrm{PCA}}^{\top}-\mathbf{u}_*\mathbf{v}_*^{\top}\big\|_F^2\Big)\\ &= \min_{c>0}\big(c^2-2c\sqrt{\Delta_{\mathrm{PCA}}\Gamma_{\mathrm{PCA}}}+1\big)=1-\Delta_{\mathrm{PCA}}\Gamma_{\mathrm{PCA}}\end{aligned}$$

with the minimum attained at $c=\sqrt{\Delta_{\mathrm{PCA}}\Gamma_{\mathrm{PCA}}}$. We verify in Appendix C.3 that for all $\alpha>\alpha_0$ sufficiently large, the fixed points of Theorem 3.4(b) satisfy

$$\mathrm{MSE}=1-\Delta_*\Gamma_*\le 1-\Delta_{\mathrm{PCA}}\Gamma_{\mathrm{PCA}}=\mathrm{MSE}_{\mathrm{PCA}}, \tag{3.32}$$

and that equality holds if and only if both $U_*\sim\mathcal{N}(0,1)$ and $V_*\sim\mathcal{N}(0,1)$. Thus, for sufficiently large signal strength and any non-Gaussian prior for either $U_*$ or $V_*$, the above AMP algorithm achieves strictly better estimation accuracy than PCA.

An illustration of this AMP algorithm and its state evolution is presented in the middle and right panels of Figure 1, with noise singular values drawn from a rescaled Beta(1, 2) distribution. Again, close agreement with the state evolution predictions is observed at these sample sizes $(m,n)=(2000,4000)$ and $\gamma=1/2$.

**4. General AMP algorithm for symmetric square matrices.** We now describe the general AMP algorithm for symmetric square matrices

$$\mathbf{W}=\mathbf{O}^{\top}\mathbf{\Lambda}\mathbf{O}\in\mathbb{R}^{n\times n},\quad \mathbf{\Lambda}=\mathrm{diag}(\boldsymbol{\lambda}) \tag{4.1}$$

and we state a formal theorem for its state evolution.

We consider an initialization $\mathbf{u}_1\in\mathbb{R}^n$, and also a possible matrix of side information

$$\mathbf{E}\in\mathbb{R}^{n\times k}$$

for a fixed dimension $k\ge 0$, both independent of $\mathbf{W}$. (We may take $k=0$ if there is no such side information.) Starting from this initialization $\mathbf{u}_1$, the AMP algorithm takes the form

$$\mathbf{z}_t=\mathbf{W}\mathbf{u}_t-b_{t1}\mathbf{u}_1-b_{t2}\mathbf{u}_2-\cdots-b_{tt}\mathbf{u}_t, \tag{4.2}$$

$$\mathbf{u}_{t+1}=u_{t+1}(\mathbf{z}_1,\dots,\mathbf{z}_t,\mathbf{E}). \tag{4.3}$$

Each function $u_{t+1}:\mathbb{R}^{t+k}\to\mathbb{R}$ is applied rowwise to $(\mathbf{z}_1,\dots,\mathbf{z}_t,\mathbf{E})\in\mathbb{R}^{n\times(t+k)}$. The debiasing coefficients $b_{t1},\dots,b_{tt}\in\mathbb{R}$ are defined to ensure the empirical convergence

$$(\mathbf{z}_1,\dots,\mathbf{z}_t)\overset{W}{\to}\mathcal{N}\big(0,\mathbf{\Sigma}_t^{\infty}\big)$$

as $n\to\infty$. The forms of $b_{t1},\dots,b_{tt}$ and $\mathbf{\Sigma}_t^{\infty}$ were first described in [47], and we review this in the next section.

4.1. *Debiasing coefficients and limit covariance.* Define the $t\times t$ matrices

$$\begin{aligned}\mathbf{\Delta}_t&=\begin{pmatrix}\langle\mathbf{u}_1^2\rangle & \langle\mathbf{u}_1\mathbf{u}_2\rangle & \cdots & \langle\mathbf{u}_1\mathbf{u}_t\rangle\\ \langle\mathbf{u}_2\mathbf{u}_1\rangle & \langle\mathbf{u}_2^2\rangle & \cdots & \langle\mathbf{u}_2\mathbf{u}_t\rangle\\ \vdots & \vdots & \ddots & \vdots\\ \langle\mathbf{u}_t\mathbf{u}_1\rangle & \langle\mathbf{u}_t\mathbf{u}_2\rangle & \cdots & \langle\mathbf{u}_t^2\rangle\end{pmatrix},\\ \mathbf{\Phi}_t&=\begin{pmatrix}0 & 0 & \cdots & 0 & 0\\ \langle\partial_1\mathbf{u}_2\rangle & 0 & \cdots & 0 & 0\\ \langle\partial_1\mathbf{u}_3\rangle & \langle\partial_2\mathbf{u}_3\rangle & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \langle\partial_1\mathbf{u}_t\rangle & \langle\partial_2\mathbf{u}_t\rangle & \cdots & \langle\partial_{t-1}\mathbf{u}_t\rangle & 0\end{pmatrix},\end{aligned} \tag{4.4}$$

where $\mathbf{u}_s\mathbf{u}_{s'} \in \mathbb{R}^n$, $\mathbf{u}_s^2 \in \mathbb{R}^n$ and $\partial_{s'}\mathbf{u}_s \in \mathbb{R}^n$ denote the entrywise product, square and partial derivative with respect to $z_{s'}$. For each $j \geq 0$, define

$$\boldsymbol{\Theta}_t^{(j)} = \sum_{i=0}^{j} \boldsymbol{\Phi}_t^i \boldsymbol{\Delta}_t (\boldsymbol{\Phi}_t^{j-i})^\top. \tag{4.5}$$

For example,

$$\boldsymbol{\Theta}_t^{(0)} = \boldsymbol{\Delta}_t, \boldsymbol{\Theta}_t^{(1)} = \boldsymbol{\Phi}_t\boldsymbol{\Delta}_t + \boldsymbol{\Delta}_t\boldsymbol{\Phi}_t^\top, \qquad \boldsymbol{\Theta}_t^{(2)} = \boldsymbol{\Phi}_t^2\boldsymbol{\Delta}_t + \boldsymbol{\Phi}_t\boldsymbol{\Delta}_t\boldsymbol{\Phi}_t^\top + \boldsymbol{\Delta}_t(\boldsymbol{\Phi}_t^2)^\top.$$

Let $\{\kappa_k\}_{k\geq 1}$ be the free cumulants of the empirical eigenvalue distribution of $\mathbf{W}$. These are the free cumulants as defined in Section 2.3 corresponding to the empirical moments

$$m_k = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^k, \tag{4.6}$$

where $(\lambda_1, \ldots, \lambda_n) = \boldsymbol{\lambda}$ are the eigenvalues of $\mathbf{W}$. Then define two matrices $\mathbf{B}_t$ and $\boldsymbol{\Sigma}_t$ by

$$\mathbf{B}_t = \left(\sum_{j=0}^{\infty} \kappa_{j+1}\boldsymbol{\Phi}_t^j\right)^\top, \qquad \boldsymbol{\Sigma}_t = \sum_{j=0}^{\infty} \kappa_{j+2}\boldsymbol{\Theta}_t^{(j)}. \tag{4.7}$$

Here, $\mathbf{B}_t$ may be interpreted as the R-transform applied to $\boldsymbol{\Phi}_t^\top$. Note that we write these as infinite series for convenience, but in fact the series are finite because $\boldsymbol{\Phi}_t^j = 0$ for all $j \geq t$, and hence also $\boldsymbol{\Theta}_t^{(j)} = 0$ for all $j \geq 2t-1$. So, for example,

$$\mathbf{B}_1 = \kappa_1 \mathrm{Id}_{1\times 1}, \qquad \mathbf{B}_2 = \kappa_1 \mathrm{Id}_{2\times 2} + \kappa_2\boldsymbol{\Phi}_2^\top,$$
$$\boldsymbol{\Sigma}_1 = \kappa_2\boldsymbol{\Theta}_1^{(0)}, \qquad \boldsymbol{\Sigma}_2 = \kappa_2\boldsymbol{\Theta}_2^{(0)} + \kappa_3\boldsymbol{\Theta}_2^{(1)} + \kappa_4\boldsymbol{\Theta}_2^{(2)}.$$

Each matrix $\mathbf{B}_t$ is upper triangular, which we may write entrywise as

$$\mathbf{B}_t = \begin{pmatrix} b_{11} & b_{21} & \cdots & b_{t1} \\ & b_{22} & \cdots & b_{t2} \\ & & \ddots & \vdots \\ & & & b_{tt} \end{pmatrix}.$$

The debiasing coefficients in (4.2) are defined to be the last column of $\mathbf{B}_t$. Note that the diagonal entries $b_{11}, b_{22}, \ldots, b_{tt}$ are all equal to $\kappa_1$, corresponding to the subtraction of $\kappa_1\mathbf{u}_t$ in (4.2) when the eigenvalue distribution of $\mathbf{W}$ has mean $\kappa_1$. If $\kappa_1 = 0$, then the debiasing for $\mathbf{W}\mathbf{u}_t$ depends only on the previous iterates $\mathbf{u}_1, \ldots, \mathbf{u}_{t-1}$.

Under the conditions to be imposed in Assumption 4.2, all of the matrices $\boldsymbol{\Delta}_t$, $\boldsymbol{\Phi}_t$, $\mathbf{B}_t$ and $\boldsymbol{\Sigma}_t$ will converge to deterministic $t \times t$ matrices in the $n \to \infty$ limit, which we denote as

$$(\boldsymbol{\Delta}_t^\infty, \boldsymbol{\Phi}_t^\infty, \mathbf{B}_t^\infty, \boldsymbol{\Sigma}_t^\infty) = \lim_{n\to\infty} (\boldsymbol{\Delta}_t, \boldsymbol{\Phi}_t, \mathbf{B}_t, \boldsymbol{\Sigma}_t).$$

This matrix $\boldsymbol{\Sigma}_t^\infty$ is the covariance defining the state evolution of the iterates $(\mathbf{z}_1, \ldots, \mathbf{z}_t)$. All of our results will hold equally if the debiasing coefficients in (4.2) are replaced by their limits $b_{ts}^\infty$, or by any consistent estimates of these limits.

We make two observations regarding this construction:

1. From the lower-triangular form of $\boldsymbol{\Phi}_t$, one may check that the upper left $(t-1)\times(t-1)$ submatrix of $(\boldsymbol{\Phi}_t^j)^\top$ is $(\boldsymbol{\Phi}_{t-1}^j)^\top$, and similarly the upper left $(t-1)\times(t-1)$ submatrix of $\boldsymbol{\Theta}_t^{(j)}$ is $\boldsymbol{\Theta}_{t-1}^{(j)}$. Thus, the upper left submatrices of $\mathbf{B}_t$ and $\boldsymbol{\Sigma}_t$ coincide with $\mathbf{B}_{t-1}$ and $\boldsymbol{\Sigma}_{t-1}$.
2. For each iteration $t \geq 1$, $\mathbf{B}_t$ depends on $\boldsymbol{\lambda}$ only via its first $t$ free cumulants $\kappa_1, \ldots, \kappa_t$, and $\boldsymbol{\Sigma}_t$ depends on $\boldsymbol{\lambda}$ only via its first $2t$ free cumulants $\kappa_1, \ldots, \kappa_{2t}$.

REMARK 4.1. In the Gaussian setting of $\mathbf{W} \sim \mathrm{GOE}(n)$, where $\mathbf{W}$ has independent $\mathcal{N}(0, 1/n)$ entries above the diagonal and $\mathcal{N}(0, 2/n)$ entries on the diagonal, the limit spectral distribution of $\mathbf{W}$ is the Wigner semicircle law. The limits of the free cumulants $\kappa_1, \kappa_2, \ldots$ in this case are

$$\kappa_1^\infty = 0, \qquad \kappa_2^\infty = 1, \qquad \kappa_j^\infty = 0 \quad \text{for all } j \geq 2.$$

This yields simply

$$\mathbf{B}_t^\infty = \big(\mathbf{\Phi}_t^\infty\big)^\top, \qquad \mathbf{\Sigma}_t^\infty = \mathbf{\Delta}_t^\infty.$$

If we further specialize to an algorithm where each $\mathbf{u}_t$ depends only on the previous iterate $\mathbf{z}_{t-1}$, then $\langle \partial_s \mathbf{u}_t \rangle = 0$ for $s \neq t-1$, and this yields the Gaussian AMP algorithm

$$\mathbf{z}_t = \mathbf{W}\mathbf{u}_t - \langle \partial_{t-1} \mathbf{u}_t \rangle \mathbf{u}_{t-1},$$
$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{z}_t, \mathbf{E})$$

as studied in [11] and [3], Section 4. Furthermore, the state evolution is such that each iterate $\mathbf{z}_t$ has the empirical limit $\mathcal{N}(0, \sigma_{tt}^\infty)$, where $\sigma_{tt}^\infty = \lim_{n\to\infty} \langle \mathbf{u}_t^2 \rangle$.

Note that outside of this Gaussian setting, we do not in general have the identity $\mathbf{\Sigma}_t^\infty = \mathbf{\Delta}_t^\infty$, that is, the empirical second moments of $\mathbf{z}_1, \ldots, \mathbf{z}_t$ do not coincide with those of $\mathbf{u}_1, \ldots, \mathbf{u}_t$ in the large-$n$ limit, even if $\mathbf{W}$ is scaled so that $\kappa_2 = 1$.

4.2. *Main result.* We impose the following assumptions on the model (4.1) and the AMP iterates (4.2)–(4.3). Note that here, we do not require the functions $u_{t+1}(\cdot)$ to be Lipschitz, but instead impose only the assumption (2.1) of polynomial growth.

ASSUMPTION 4.2.

(a) $\mathbf{O} \in \mathbb{R}^{n\times n}$ is a random and Haar-uniform orthogonal matrix.

(b) $\boldsymbol{\lambda} \in \mathbb{R}^n$ is independent of $\mathbf{O}$ and satisfies $\boldsymbol{\lambda} \overset{W}{\to} \Lambda$ almost surely as $n \to \infty$, for a random variable $\Lambda$ having finite moments of all orders.

(c) $\mathbf{u}_1 \in \mathbb{R}^n$ and $\mathbf{E} \in \mathbb{R}^{n\times k}$ are independent of $\mathbf{O}$ and satisfy $(\mathbf{u}_1, \mathbf{E}) \overset{W}{\to} (U_1, E)$ almost surely as $n \to \infty$, for a random vector $(U_1, E) \equiv (U_1, E_1, \ldots, E_k)$ having finite moments of all orders.

(d) Each function $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ satisfies (2.1) for some $C > 0$ and $p \geq 1$. Writing its argument as $(z, e)$ where $z \in \mathbb{R}^t$ and $e \in \mathbb{R}^k$, $u_{t+1}$ is weakly differentiable in $z$ and continuous in $e$. For each $s = 1, \ldots, t$, $\partial_s u_{t+1}$ also satisfies (2.1) for some $C > 0$ and $p \geq 1$, and $\partial_s u_{t+1}(z, e)$ is continuous at Lebesgue-a.e. $z \in \mathbb{R}^t$ for every $e \in \mathbb{R}^k$.

(e) $\mathrm{Var}[\Lambda] > 0$ and $\mathbb{E}[U_1^2] > 0$. Letting $(Z_1, \ldots, Z_t) \sim \mathcal{N}(0, \mathbf{\Sigma}_t^\infty)$ be independent of $(U_1, E)$, each function $u_{t+1}$ is such that there do not exist constants $\alpha_1, \ldots, \alpha_t, \beta_1, \ldots, \beta_t$ for which

$$u_{t+1}(Z_1, \ldots, Z_t, E) = \sum_{s=1}^t \alpha_s Z_s + \beta_1 U_1 + \sum_{s=2}^t \beta_s U_s(Z_1, \ldots, Z_{s-1}, E)$$

with probability 1 over $(U_1, E, Z_1, \ldots, Z_t)$.

We clarify that Theorem 4.3 below establishes the existence of the limit $\mathbf{\Sigma}_t^\infty$ provided that condition (e) holds for the functions $u_2, \ldots, u_t$, and this limit $\mathbf{\Sigma}_t^\infty$ then defines condition (e) for the next function $u_{t+1}$. This condition (e) is a nondegeneracy assumption that holds if each function $u_{t+1}$ has a nonlinear dependence on the preceding iterate $z_t$.

THEOREM 4.3. *Under Assumption 4.2, for each fixed $t \geq 1$, almost surely as $n \to \infty$: $\boldsymbol{\Sigma}_t \to \boldsymbol{\Sigma}_t^\infty$ for a deterministic nonsingular matrix $\boldsymbol{\Sigma}_t^\infty$, and*

$$(\mathbf{u}_1, \dots, \mathbf{u}_{t+1}, \mathbf{z}_1, \dots, \mathbf{z}_t, \mathbf{E}) \overset{W}{\to} (U_1, \dots, U_{t+1}, Z_1, \dots, Z_t, E),$$

*where $(Z_1, \dots, Z_t) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t^\infty)$, this vector $(Z_1, \dots, Z_t)$ is independent of $(U_1, E)$, and $U_s = u_s(Z_1, \dots, Z_{s-1}, E)$ for each $s = 2, \dots, t+1$.*

The proof of this result is provided in Appendix A. The limit $\boldsymbol{\Sigma}_t^\infty$ is given by replacing $\langle \mathbf{u}_s \mathbf{u}_{s'} \rangle$, $\langle \partial_{s'} \mathbf{u}_s \rangle$, and $\kappa_k$ in the definitions (4.4) and (4.7) with $\mathbb{E}[U_s U_{s'}]$, $\mathbb{E}[\partial_{s'} u_s(Z_1, \dots, Z_{s-1}, E)]$, and the free cumulants $\kappa_k^\infty$ of the limit spectral distribution $\Lambda$.

4.3. *Removing the nondegeneracy assumption.* The following corollary provides a version of Theorem 4.3 without the nondegeneracy condition of Assumption 4.2(e), under the stronger condition that each function $u_{t+1}$ is continuously-differentiable and Lipschitz. Note that the convergence established is only in $W_2$, rather than in $W_p$ for every order $p \geq 1$ as in Theorem 4.3.

The proof follows the idea of [10] by studying a perturbed AMP sequence and then taking the limit of this perturbation to 0. We provide this proof in Appendix D.

COROLLARY 4.4. *Suppose Assumption 4.2(a)–(c) holds, $\limsup_{n\to\infty} \|\boldsymbol{\lambda}\|_\infty < \infty$, each function $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ is continuously-differentiable, and*

$$\left|u_{t+1}(z, e) - u_{t+1}(z', e)\right| \leq C \|z - z'\|$$

*for a constant $C > 0$ and all $z, z' \in \mathbb{R}^t$ and $e \in \mathbb{R}^k$. Then for each fixed $t \geq 1$, almost surely as $n \to \infty$: $\boldsymbol{\Sigma}_t \to \boldsymbol{\Sigma}_t^\infty$ for a deterministic (possibly singular) matrix $\boldsymbol{\Sigma}_t^\infty$, and*

$$(\mathbf{u}_1, \dots, \mathbf{u}_{t+1}, \mathbf{z}_1, \dots, \mathbf{z}_t, \mathbf{E}) \overset{W_2}{\to} (U_1, \dots, U_{t+1}, Z_1, \dots, Z_t, E),$$

*where $(U_1, \dots, U_{t+1}, Z_1, \dots, Z_t, E)$ is as defined in Theorem 4.3.*

**5. AMP algorithm for rectangular matrices.** In this section, we describe the form of the general AMP algorithm for a rectangular matrix

$$\mathbf{W} = \mathbf{O}^\top \boldsymbol{\Lambda} \mathbf{Q} \in \mathbb{R}^{m \times n}, \quad \boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda}) \tag{5.1}$$

and state a formal theorem for its state evolution. We denote

$$\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{\min(m,n)}) \in \mathbb{R}^{\min(m,n)} \tag{5.2}$$

as the diagonal entries of $\boldsymbol{\Lambda}$, which are the singular values of $\mathbf{W}$.

We consider an initialization $\mathbf{u}_1 \in \mathbb{R}^m$, and two matrices of side information

$$\mathbf{E} \in \mathbb{R}^{m \times k} \quad \text{and} \quad \mathbf{F} \in \mathbb{R}^{n \times \ell}$$

for fixed dimensions $k, \ell \geq 0$, all independent of $\mathbf{W}$. (We may take $k, \ell = 0$ if there is no such side information.) Starting from this initialization, the AMP algorithm takes the form

$$\mathbf{z}_t = \mathbf{W}^\top \mathbf{u}_t - b_{t1}\mathbf{v}_1 - b_{t2}\mathbf{v}_2 - \cdots - b_{t,t-1}\mathbf{v}_{t-1}, \tag{5.3}$$

$$\mathbf{v}_t = v_t(\mathbf{z}_1, \dots, \mathbf{z}_t, \mathbf{F}), \tag{5.4}$$

$$\mathbf{y}_t = \mathbf{W}\mathbf{v}_t - a_{t1}\mathbf{u}_1 - a_{t2}\mathbf{u}_2 - \cdots - a_{tt}\mathbf{u}_t, \tag{5.5}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{y}_1, \dots, \mathbf{y}_t, \mathbf{E}), \tag{5.6}$$

for functions $v_t : \mathbb{R}^{t+\ell} \to \mathbb{R}$ and $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$. In the first iteration $t = 1$, (5.3) is simply $\mathbf{z}_1 = \mathbf{W}^\top \mathbf{u}_1$. The debiasing coefficients $a_{t1}, \dots, a_{tt}$ and $b_{t1}, \dots, b_{t,t-1}$ are defined to ensure that

$$(\mathbf{y}_1, \dots, \mathbf{y}_t) \overset{W}{\to} \mathcal{N}(0, \mathbf{\Sigma}_t^\infty) \quad \text{and} \quad (\mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W}{\to} \mathcal{N}(0, \mathbf{\Omega}_t^\infty)$$

as $m, n \to \infty$. We describe these debiasing coefficients and state evolution in the next section, in terms of the rectangular free cumulants of $\mathbf{W}$—these were also derived recently in [16].

5.1. *Debiasing coefficients and limit covariance.* Define the $t \times t$ matrices

$$\begin{aligned}
\mathbf{\Delta}_t &= \begin{pmatrix} \langle \mathbf{u}_1^2 \rangle & \langle \mathbf{u}_1 \mathbf{u}_2 \rangle & \cdots & \langle \mathbf{u}_1 \mathbf{u}_t \rangle \\ \langle \mathbf{u}_2 \mathbf{u}_1 \rangle & \langle \mathbf{u}_2^2 \rangle & \cdots & \langle \mathbf{u}_2 \mathbf{u}_t \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{u}_t \mathbf{u}_1 \rangle & \langle \mathbf{u}_t \mathbf{u}_2 \rangle & \cdots & \langle \mathbf{u}_t^2 \rangle \end{pmatrix}, \\
\mathbf{\Phi}_t &= \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ \langle \partial_1 \mathbf{u}_2 \rangle & 0 & \cdots & 0 & 0 \\ \langle \partial_1 \mathbf{u}_3 \rangle & \langle \partial_2 \mathbf{u}_3 \rangle & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \langle \partial_1 \mathbf{u}_t \rangle & \langle \partial_2 \mathbf{u}_t \rangle & \cdots & \langle \partial_{t-1} \mathbf{u}_t \rangle & 0 \end{pmatrix},
\end{aligned} \tag{5.7}$$

$$\begin{aligned}
\mathbf{\Gamma}_t &= \begin{pmatrix} \langle \mathbf{v}_1^2 \rangle & \langle \mathbf{v}_1 \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_1 \mathbf{v}_t \rangle \\ \langle \mathbf{v}_2 \mathbf{v}_1 \rangle & \langle \mathbf{v}_2^2 \rangle & \cdots & \langle \mathbf{v}_2 \mathbf{v}_t \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_t \mathbf{v}_1 \rangle & \langle \mathbf{v}_t \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_t^2 \rangle \end{pmatrix}, \\
\mathbf{\Psi}_t &= \begin{pmatrix} \langle \partial_1 \mathbf{v}_1 \rangle & 0 & \cdots & 0 \\ \langle \partial_1 \mathbf{v}_2 \rangle & \langle \partial_2 \mathbf{v}_2 \rangle & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \langle \partial_1 \mathbf{v}_t \rangle & \langle \partial_2 \mathbf{v}_t \rangle & \cdots & \langle \partial_t \mathbf{v}_t \rangle \end{pmatrix}.
\end{aligned} \tag{5.8}$$

For each $j \geq 0$, define

$$\mathbf{\Theta}_t^{(j)} = \sum_{i=0}^{j} (\mathbf{\Phi}_t \mathbf{\Psi}_t)^i \mathbf{\Delta}_t (\mathbf{\Psi}_t^\top \mathbf{\Phi}_t^\top)^{j-i} + \sum_{i=0}^{j-1} (\mathbf{\Phi}_t \mathbf{\Psi}_t)^i \mathbf{\Phi}_t \mathbf{\Gamma}_t \mathbf{\Phi}_t^\top (\mathbf{\Psi}_t^\top \mathbf{\Phi}_t^\top)^{j-1-i}, \tag{5.9}$$

$$\mathbf{\Xi}_t^{(j)} = \sum_{i=0}^{j} (\mathbf{\Psi}_t \mathbf{\Phi}_t)^i \mathbf{\Gamma}_t (\mathbf{\Phi}_t^\top \mathbf{\Psi}_t^\top)^{j-i} + \sum_{i=0}^{j-1} (\mathbf{\Psi}_t \mathbf{\Phi}_t)^i \mathbf{\Psi}_t \mathbf{\Delta}_t \mathbf{\Psi}_t^\top (\mathbf{\Phi}_t^\top \mathbf{\Psi}_t^\top)^{j-1-i}. \tag{5.10}$$

The second summations of (5.9) and (5.10) are not present for $j = 0$. So, for example,

$$\begin{aligned}
\mathbf{\Theta}_t^{(0)} &= \mathbf{\Delta}_t, \\
\mathbf{\Theta}_t^{(1)} &= \mathbf{\Phi}_t \mathbf{\Psi}_t \mathbf{\Delta}_t + \mathbf{\Phi}_t \mathbf{\Gamma}_t \mathbf{\Phi}_t^\top + \mathbf{\Delta}_t \mathbf{\Psi}_t^\top \mathbf{\Phi}_t^\top, \\
\mathbf{\Theta}_t^{(2)} &= \mathbf{\Phi}_t \mathbf{\Psi}_t \mathbf{\Phi}_t \mathbf{\Psi}_t \mathbf{\Delta}_t + \mathbf{\Phi}_t \mathbf{\Psi}_t \mathbf{\Phi}_t \mathbf{\Gamma}_t \mathbf{\Phi}_t^\top + \mathbf{\Phi}_t \mathbf{\Psi}_t \mathbf{\Delta}_t \mathbf{\Psi}_t^\top \mathbf{\Phi}_t^\top \\
&\quad + \mathbf{\Phi}_t \mathbf{\Gamma}_t \mathbf{\Phi}_t^\top \mathbf{\Psi}_t^\top \mathbf{\Phi}_t^\top + \mathbf{\Delta}_t \mathbf{\Psi}_t^\top \mathbf{\Phi}_t^\top \mathbf{\Psi}_t^\top \mathbf{\Phi}_t^\top, \\
\mathbf{\Xi}_t^{(0)} &= \mathbf{\Gamma}_t \\
\mathbf{\Xi}_t^{(1)} &= \mathbf{\Psi}_t \mathbf{\Phi}_t \mathbf{\Gamma}_t + \mathbf{\Psi}_t \mathbf{\Delta}_t \mathbf{\Psi}_t^\top + \mathbf{\Gamma}_t \mathbf{\Phi}_t^\top \mathbf{\Psi}_t^\top,
\end{aligned}$$

$$\boldsymbol{\Xi}_t^{(2)} = \boldsymbol{\Psi}_t\boldsymbol{\Phi}_t\boldsymbol{\Psi}_t\boldsymbol{\Phi}_t\boldsymbol{\Gamma}_t + \boldsymbol{\Psi}_t\boldsymbol{\Phi}_t\boldsymbol{\Psi}_t\boldsymbol{\Delta}_t\boldsymbol{\Psi}_t^\top + \boldsymbol{\Psi}_t\boldsymbol{\Phi}_t\boldsymbol{\Gamma}_t\boldsymbol{\Phi}_t^\top\boldsymbol{\Psi}_t^\top + \boldsymbol{\Psi}_t\boldsymbol{\Delta}_t\boldsymbol{\Psi}_t^\top\boldsymbol{\Phi}_t^\top\boldsymbol{\Psi}_t^\top + \boldsymbol{\Gamma}_t\boldsymbol{\Phi}_t^\top\boldsymbol{\Psi}_t^\top\boldsymbol{\Phi}_t^\top\boldsymbol{\Psi}_t^\top.$$

Let $\{\kappa_{2k}\}_{k\geq 1}$ be the rectangular free cumulants with aspect ratio $\gamma = m/n$ corresponding to the sequence of even moments

$$m_{2k} = \frac{1}{m}\sum_{i=1}^{\min(m,n)} \lambda_i^{2k}, \tag{5.11}$$

as defined in Section 2.4. Note that we always use the normalization $1/m$, so these are the moments of $\boldsymbol{\lambda}$ padded by $m-n$ additional 0's if $m > n$.

Define the $t \times t$ matrices

$$\mathbf{A}_t = \left(\sum_{j=0}^{\infty} \kappa_{2(j+1)}\boldsymbol{\Psi}_t(\boldsymbol{\Phi}_t\boldsymbol{\Psi}_t)^j\right)^\top, \qquad \mathbf{B}_t = \left(\gamma\sum_{j=0}^{\infty} \kappa_{2(j+1)}\boldsymbol{\Phi}_t(\boldsymbol{\Psi}_t\boldsymbol{\Phi}_t)^j\right)^\top, \tag{5.12}$$

$$\boldsymbol{\Sigma}_t = \sum_{j=0}^{\infty} \kappa_{2(j+1)}\boldsymbol{\Xi}_t^{(j)}, \qquad \boldsymbol{\Omega}_t = \gamma\sum_{j=0}^{\infty} \kappa_{2(j+1)}\boldsymbol{\Theta}_t^{(j)}. \tag{5.13}$$

These are in fact finite series, as it may be verified that

$$\begin{aligned}
\boldsymbol{\Psi}_t(\boldsymbol{\Phi}_t\boldsymbol{\Psi}_t)^j &= 0 \quad \text{for } j \geq t+1,\\
\boldsymbol{\Phi}_t(\boldsymbol{\Psi}_t\boldsymbol{\Phi}_t)^j &= 0 \quad \text{for } j \geq t,\\
\boldsymbol{\Xi}_t^{(j)} &= 0 \quad \text{for } j \geq 2t,\\
\boldsymbol{\Theta}_t^{(j)} &= 0 \quad \text{for } j \geq 2t-1.
\end{aligned}$$

So, for example,

$$\begin{aligned}
&\mathbf{A}_1 = \kappa_2\boldsymbol{\Psi}_1^\top, \qquad \mathbf{A}_2 = \kappa_2\boldsymbol{\Psi}_2^\top + \kappa_4(\boldsymbol{\Psi}_2\boldsymbol{\Phi}_2\boldsymbol{\Psi}_2)^\top, \quad \ldots\\
&\mathbf{B}_1 = 0, \qquad \mathbf{B}_2 = \gamma\kappa_2\boldsymbol{\Phi}_2^\top, \qquad \mathbf{B}_3 = \gamma\kappa_2\boldsymbol{\Phi}_3^\top + \gamma\kappa_4(\boldsymbol{\Phi}_3\boldsymbol{\Psi}_3\boldsymbol{\Phi}_3)^\top, \quad \ldots\\
&\boldsymbol{\Sigma}_1 = \kappa_2\boldsymbol{\Xi}_1^{(0)} + \kappa_4\boldsymbol{\Xi}_1^{(1)}, \qquad \boldsymbol{\Sigma}_2 = \kappa_2\boldsymbol{\Xi}_2^{(0)} + \kappa_4\boldsymbol{\Xi}_2^{(1)} + \kappa_6\boldsymbol{\Xi}_2^{(2)} + \kappa_8\boldsymbol{\Xi}_2^{(3)}, \ldots\\
&\boldsymbol{\Omega}_1 = \gamma\kappa_2\boldsymbol{\Theta}_1^{(0)}, \qquad \boldsymbol{\Omega}_2 = \gamma\kappa_2\boldsymbol{\Theta}_2^{(0)} + \gamma\kappa_4\boldsymbol{\Theta}_2^{(1)} + \gamma\kappa_6\boldsymbol{\Theta}_2^{(2)}, \quad \ldots.
\end{aligned}$$

The matrices $\mathbf{A}_t$ and $\mathbf{B}_t$ are upper triangular, with the forms

$$\mathbf{A}_t = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{t1}\\ & a_{22} & \cdots & a_{t2}\\ & & \ddots & \vdots\\ & & & a_{tt}\end{pmatrix}, \qquad \mathbf{B}_t = \begin{pmatrix} 0 & b_{21} & b_{31} & \cdots & b_{t1}\\ & 0 & b_{32} & \cdots & b_{t2}\\ & & \ddots & \ddots & \vdots\\ & & & 0 & b_{t,t-1}\\ & & & & 0\end{pmatrix}.$$

The debiasing coefficients $a_{t1}, \ldots, a_{tt}, b_{t1}, \ldots, b_{t,t-1}$ in (5.3) and (5.5) are defined as the last columns of $\mathbf{A}_t$ and $\mathbf{B}_t$. Under the conditions to be imposed in Assumption 5.2, these matrices all have deterministic $t \times t$ limits

$$(\boldsymbol{\Delta}_t^\infty, \boldsymbol{\Gamma}_t^\infty, \boldsymbol{\Phi}_t^\infty, \boldsymbol{\Psi}_t^\infty, \mathbf{A}_t^\infty, \mathbf{B}_t^\infty, \boldsymbol{\Sigma}_t^\infty, \boldsymbol{\Omega}_t^\infty) = \lim_{m,n\to\infty}(\boldsymbol{\Delta}_t, \boldsymbol{\Gamma}_t, \boldsymbol{\Phi}_t, \boldsymbol{\Psi}_t, \mathbf{A}_t, \mathbf{B}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\Omega}_t).$$

The matrices $\boldsymbol{\Sigma}_t^\infty$ and $\boldsymbol{\Omega}_t^\infty$ are the covariances in the state evolutions for $(\mathbf{y}_1, \ldots, \mathbf{y}_t)$ and $(\mathbf{z}_1, \ldots, \mathbf{z}_t)$. As in the symmetric square setting, the debiasing coefficients in (5.3) and (5.5) may be replaced by their limits $a_{ts}^\infty$ and $b_{ts}^\infty$, or by any consistent estimates of these limits.

We make the following observations about the above definitions:

1. The upper left $(t-1)\times(t-1)$ submatrices of $\mathbf{A}_t, \mathbf{B}_t, \mathbf{\Sigma}_t, \mathbf{\Omega}_t$ coincide with the matrices $\mathbf{A}_{t-1}, \mathbf{B}_{t-1}, \mathbf{\Sigma}_{t-1}, \mathbf{\Omega}_{t-1}$.

2. For each $t \geq 1$, $\mathbf{A}_t, \mathbf{B}_t, \mathbf{\Sigma}_t, \mathbf{\Omega}_t$ depend respectively only on the rectangular free cumulants of $\boldsymbol{\lambda}$ up to the orders $\kappa_{2t}, \kappa_{2t-2}, \kappa_{4t}, \kappa_{4t-2}$.

3. The matrices $\mathbf{A}_t, \mathbf{\Sigma}_t$ depend on $\mathbf{u}_1, \dots, \mathbf{u}_t, \mathbf{v}_1, \dots, \mathbf{v}_t$ and their derivatives. The matrices $\mathbf{B}_t, \mathbf{\Omega}_t$ depend on $\mathbf{u}_1, \dots, \mathbf{u}_t, \mathbf{v}_1, \dots, \mathbf{v}_{t-1}$ and their derivatives, but they do not depend on $\mathbf{v}_t$ or its derivatives. (Thus the debiasing coefficients and state evolution for $\mathbf{z}_t$ in (5.3) are well defined before defining $\mathbf{v}_t$ in (5.4).)

The first two statements are analogous to our observations in the symmetric square setting. The third statement holds from the definitions of $\mathbf{B}_t$ and $\mathbf{\Omega}_t$ in (5.12)–(5.13), because the last column of $\mathbf{\Phi}_t$ is 0, so $\mathbf{\Phi}_t\mathbf{\Psi}_t$ does not depend on the last row of $\mathbf{\Psi}_t$, and $\mathbf{\Phi}_t\mathbf{\Gamma}_t\mathbf{\Phi}_t^\top$ does not depend on the last row or column of $\mathbf{\Gamma}_t$.

REMARK 5.1. In the Gaussian setting where $\mathbf{W}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries, the limit spectral distribution of $\mathbf{W}\mathbf{W}^\top$ is the Marcenko–Pastur law, with limiting rectangular free cumulants

$$\kappa_2^\infty = 1, \qquad \kappa_{2j}^\infty = 0 \quad \text{for all } j \geq 2.$$

This yields simply

$$\mathbf{A}_t^\infty = \big(\mathbf{\Psi}_t^\infty\big)^\top, \qquad \mathbf{B}_t^\infty = \gamma\big(\mathbf{\Phi}_t^\infty\big)^\top, \qquad \mathbf{\Sigma}_t^\infty = \mathbf{\Gamma}_t^\infty, \qquad \mathbf{\Omega}_t = \gamma\,\mathbf{\Delta}_t^\infty.$$

If we further specialize to an algorithm where $v_t$ depends only on $z_t$ and $u_{t+1}$ depends only on $y_t$, then $\langle \partial_s \mathbf{u}_t\rangle = 0$ for all $s \neq t-1$ and $\langle \partial_s \mathbf{z}_t\rangle = 0$ for all $s \neq t$. This yields the Gaussian AMP algorithm

$$\begin{aligned}
\mathbf{z}_t &= \mathbf{W}^\top \mathbf{u}_t - \gamma\,\langle \partial_{t-1}\mathbf{u}_t\rangle \mathbf{v}_{t-1},\\
\mathbf{v}_t &= v_t(\mathbf{z}_t, \mathbf{F}),\\
\mathbf{y}_t &= \mathbf{W}\mathbf{v}_t - \langle \partial_t \mathbf{v}_t\rangle \mathbf{u}_t,\\
\mathbf{u}_{t+1} &= u_{t+1}(\mathbf{y}_t, \mathbf{E}),
\end{aligned}$$

as studied in [3], Section 3. Furthermore, the state evolution is such that $\mathbf{z}_t$ has the empirical limit $\mathcal{N}(0, \omega_{tt}^\infty)$ where $\omega_{tt}^\infty = \lim_{m,n\to\infty} \gamma\cdot\langle \mathbf{u}_t^2\rangle$, and $\mathbf{y}_t$ has the empirical limit $\mathcal{N}(0, \sigma_{tt}^\infty)$ where $\sigma_{tt}^\infty = \lim_{m,n\to\infty}\langle \mathbf{v}_t^2\rangle$.

Note that outside of this Gaussian setting, in general we do not have the identities $\mathbf{\Sigma}_t = \mathbf{\Gamma}_t$ and $\mathbf{\Omega}_t = \gamma\,\mathbf{\Delta}_t$ even when $\mathbf{W}$ is normalized such that $\kappa_2 = 1$.

5.2. *Main result.* We impose the following assumptions on the model (5.1)–(5.2) and the AMP iterates (5.3)–(5.6). Again, we do not require here $v_t(\cdot)$ and $u_{t+1}(\cdot)$ to be Lipschitz.

ASSUMPTION 5.2.

(a) $m, n \to \infty$ such that $m/n = \gamma \in (0, \infty)$ is a fixed constant.

(b) $\mathbf{O} \in \mathbb{R}^{m\times m}$ and $\mathbf{Q} \in \mathbb{R}^{n\times n}$ are independent random and Haar-uniform orthogonal matrices.

(c) $\boldsymbol{\lambda} \in \mathbb{R}^{\min(m,n)}$ is independent of $\mathbf{O}, \mathbf{Q}$ and satisfies $\boldsymbol{\lambda} \overset{W}{\to} \Lambda$ almost surely as $m, n \to \infty$, for a random variable $\Lambda$ having finite moments of all orders.

(d) $\mathbf{u}_1 \in \mathbb{R}^m$, $\mathbf{E} \in \mathbb{R}^{m\times k}$ and $\mathbf{F} \in \mathbb{R}^{n\times \ell}$ are independent of $\mathbf{O}, \mathbf{Q}$ and satisfy $(\mathbf{u}_1, \mathbf{E}) \overset{W}{\to} (U_1, E)$ and $\mathbf{F} \overset{W}{\to} F$ almost surely as $m, n \to \infty$, where $(U_1, E) \equiv (U_1, E_1, \dots, E_k)$ and $F \equiv (F_1, \dots, F_\ell)$ are random vectors having finite moments of all orders.

(e) Each function $v_t : \mathbb{R}^{t+\ell} \to \mathbb{R}$ and $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ satisfies (2.1) for some $C > 0$ and $p \geq 1$. Writing their arguments as $(z, f)$ and $(y, e)$ where $z, y \in \mathbb{R}^t$, $f \in \mathbb{R}^\ell$ and $e \in \mathbb{R}^k$, $v_t$ is weakly differentiable in $z$ and continuous in $f$, and $u_{t+1}$ is weakly differentiable in $y$ and continuous in $e$. For each $s = 1, \ldots, t$, $\partial_s v_t$ and $\partial_s u_{t+1}$ also satisfy (2.1) for some $C > 0$ and $p \geq 1$, where $\partial_s v_t(z, f)$ is continuous at Lebesgue-a.e. $z \in \mathbb{R}^t$ for every $f \in \mathbb{R}^\ell$, and $\partial_s u_{t+1}(y, e)$ is continuous at Lebesgue-a.e. $y \in \mathbb{R}^t$ for every $e \in \mathbb{R}^k$.

(f) $\mathrm{Var}[\Lambda] > 0$ and $\mathbb{E}[U_1^2] > 0$. Letting $(Z_1, \ldots, Z_t) \sim \mathcal{N}(0, \boldsymbol{\Omega}_t^\infty)$ be independent of $F$, there do not exist constants $\alpha_1, \ldots, \alpha_t, \beta_1, \ldots, \beta_{t-1}$ for which

$$v_t(Z_1, \ldots, Z_t, F) = \sum_{s=1}^{t} \alpha_s Z_s + \sum_{s=1}^{t-1} \beta_s v_s(Z_1, \ldots, Z_s, F)$$

with probability 1 over $(F, Z_1, \ldots, Z_t)$. Letting $(Y_1, \ldots, Y_t) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t^\infty)$ be independent of $(U_1, E)$, there do not exist constants $\alpha_1, \ldots, \alpha_t, \beta_1, \ldots, \beta_t$ for which

$$u_{t+1}(Y_1, \ldots, Y_t, E) = \sum_{s=1}^{t} \alpha_s Y_s + \beta_1 U_1 + \sum_{s=2}^{t} \beta_s u_s(Y_1, \ldots, Y_{s-1}, E)$$

with probability 1 over $(U_1, E, Y_1, \ldots, Y_t)$.

As in the symmetric square setting, we clarify that Theorem 5.3 below establishes the existence of $\boldsymbol{\Omega}_t^\infty$ when condition (f) holds for $u_1, \ldots, u_t$ and $v_1, \ldots, v_{t-1}$, and this limit $\boldsymbol{\Omega}_t^\infty$ then defines condition (f) for $v_t$. Similarly, the theorem establishes the existence of $\boldsymbol{\Sigma}_t^\infty$ when condition (f) holds for $u_1, \ldots, u_t$ and $v_1, \ldots, v_t$, and this limit $\boldsymbol{\Sigma}_t^\infty$ then defines the condition for $u_{t+1}$. This condition (f) is a nondegeneracy assumption that will hold as long as $u_{t+1}(\cdot)$ and $v_t(\cdot)$ depend nonlinearly on $y_t$ and $z_t$, respectively.

THEOREM 5.3. *Under Assumption 5.2, for each fixed $t \geq 1$, almost surely as $n \to \infty$: $\boldsymbol{\Sigma}_t \to \boldsymbol{\Sigma}_t^\infty$ and $\boldsymbol{\Omega}_t \to \boldsymbol{\Omega}_t^\infty$ for some deterministic nonsingular matrices $\boldsymbol{\Sigma}_t^\infty$ and $\boldsymbol{\Omega}_t^\infty$. Also,*

$$(\mathbf{u}_1, \ldots, \mathbf{u}_{t+1}, \mathbf{y}_1, \ldots, \mathbf{y}_t, \mathbf{E}) \xrightarrow{W} (U_1, \ldots, U_{t+1}, Y_1, \ldots, Y_t, E),$$

$$(\mathbf{v}_1, \ldots, \mathbf{v}_t, \mathbf{z}_1, \ldots, \mathbf{z}_t, \mathbf{F}) \xrightarrow{W} (V_1, \ldots, V_t, Z_1, \ldots, Z_t, F),$$

*where $(Y_1, \ldots, Y_t) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t^\infty)$ is independent of $(U_1, E)$; $(Z_1, \ldots, Z_t) \sim \mathcal{N}(0, \boldsymbol{\Omega}_t^\infty)$ is independent of $F$; $U_s = u_s(Z_1, \ldots, Z_{s-1}, E)$ for each $s = 2, \ldots, t+1$; and $V_s = v_s(Z_1, \ldots, Z_s, F)$ for each $s = 1, \ldots, t$.*

The limits $\boldsymbol{\Sigma}_t^\infty$ and $\boldsymbol{\Omega}_t^\infty$ are given by replacing $\langle \mathbf{u}_s \mathbf{u}_{s'} \rangle$, $\langle \mathbf{v}_s \mathbf{v}_{s'} \rangle$, $\langle \partial_{s'} \mathbf{u}_s \rangle$, $\langle \partial_{s'} \mathbf{v}_s \rangle$, and $\kappa_{2k}$ in the definitions (5.8)–(5.9) and (5.13) with $\mathbb{E}[U_s U_{s'}]$, $\mathbb{E}[V_s V_{s'}]$, $\mathbb{E}[\partial_{s'} u_s(Y_1, \ldots, Y_{s-1}, E)]$, $\mathbb{E}[\partial_{s'} v_s(Z_1, \ldots, Z_s, F)]$, and $\kappa_{2k}^\infty$.

The proof of this result is provided in Appendix B. As in Corollary 4.4, we may remove the nondegeneracy condition in Assumption 5.2(f) if $v_t$ and $u_{t+1}$ are continuously-differentiable and Lipschitz. This is stated in the following corollary. The proof follows the same argument as that of Corollary 4.4, and we omit this for brevity.

COROLLARY 5.4. *Suppose Assumption 5.2(a)–(d) holds, $\limsup_{n\to\infty} \|\boldsymbol{\lambda}\|_\infty < \infty$, each function $v_t : \mathbb{R}^{t+\ell} \to \mathbb{R}$ and $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ is continuously-differentiable, and*

$$|v_t(z, f) - v_t(z', f)| \leq C\|z - z'\|, \qquad |u_{t+1}(y, e) - u_{t+1}(y', e)| \leq C\|y - y'\|$$

*for a constant $C > 0$ and all $z, z', y, y' \in \mathbb{R}^t$, $e \in \mathbb{R}^k$, and $f \in \mathbb{R}^\ell$. Then for each fixed $t \geq 1$, almost surely as $n \to \infty$: $\boldsymbol{\Sigma}_t \to \boldsymbol{\Sigma}_t^\infty$ and $\boldsymbol{\Omega}_t \to \boldsymbol{\Omega}_t^\infty$ for some deterministic (possibly singular) matrices $\boldsymbol{\Sigma}_t^\infty$ and $\boldsymbol{\Omega}_t^\infty$, and*

$$(\mathbf{u}_1, \dots, \mathbf{u}_{t+1}, \mathbf{y}_1, \dots, \mathbf{y}_t, \mathbf{E}) \xrightarrow{W_2} (U_1, \dots, U_{t+1}, Y_1, \dots, Y_t, E),$$

$$(\mathbf{v}_1, \dots, \mathbf{v}_t, \mathbf{z}_1, \dots, \mathbf{z}_t, \mathbf{F}) \xrightarrow{W_2} (V_1, \dots, V_t, Z_1, \dots, Z_t, F),$$

*where these limits are as defined in Theorem 5.3.*

**6. Proof ideas.** We describe here the main ideas of the proofs. In the setting of a symmetric square matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, the basic strategy is to write $\mathbf{W} = \mathbf{O}^\top \boldsymbol{\Lambda} \mathbf{O}$, and to express the AMP iterations (1.2)–(1.3) in an expanded form as

$$\mathbf{r}_t = \mathbf{O}\mathbf{u}_t, \tag{6.1}$$

$$\mathbf{s}_t = \mathbf{O}^\top \boldsymbol{\Lambda} \mathbf{r}_t, \tag{6.2}$$

$$\mathbf{z}_t = \mathbf{s}_t - b_{t1}\mathbf{u}_1 - \cdots - b_{tt}\mathbf{u}_t, \tag{6.3}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{z}_1, \dots, \mathbf{z}_t). \tag{6.4}$$

All analyses are performed conditional on $\mathbf{u}_1$ and $\boldsymbol{\Lambda}$, so that the only randomness is in the Haar-orthogonal matrix $\mathbf{O}$. We apply Bolthausen's conditioning technique [11], analyzing sequentially each iterate $\mathbf{r}_1, \mathbf{s}_1, \mathbf{z}_1, \mathbf{u}_2, \mathbf{r}_2, \dots$ conditional on all preceding iterates. This requires understanding the law of $\mathbf{O}$ conditional on events of the form

$$\mathbf{O}\mathbf{X} = \mathbf{Y},$$

which was shown in [54, 62] to be

$$\mathbf{O}|_{\mathbf{O}\mathbf{X}=\mathbf{Y}} \overset{L}{=} \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{Y}^\top + \Pi_{\mathbf{X}^\perp} \tilde{\mathbf{O}} \Pi_{\mathbf{Y}^\perp}^\top. \tag{6.5}$$

Here, $\Pi_{\mathbf{X}^\perp}$ and $\Pi_{\mathbf{Y}^\perp}$ are matrices with orthonormal columns spanning the orthogonal complements of the column spans of $\mathbf{X}$ and $\mathbf{Y}$, and $\tilde{\mathbf{O}}$ is an independent Haar-orthogonal matrix. Applying (6.5) to the appearances of $\mathbf{O}$ in (6.1)–(6.2), we will exhibit decompositions

$$\mathbf{r}_t = \mathbf{r}_\parallel + \mathbf{r}_\perp, \qquad \mathbf{s}_t = \mathbf{s}_\parallel + \mathbf{s}_\perp.$$

The vectors $\mathbf{r}_\perp$ and $\mathbf{s}_\perp$ arise from the second term of (6.5) and have empirical distributions that are approximately Gaussian conditional on the preceding iterates. The vectors $\mathbf{r}_\parallel$ and $\mathbf{s}_\parallel$ arise from the first term of (6.5), are deterministic conditional on the preceding iterates and represent biases, respectively, in the directions of $(\mathbf{r}_1, \dots, \mathbf{r}_{t-1}, \boldsymbol{\Lambda}\mathbf{r}_{t-1}, \dots, \boldsymbol{\Lambda}\mathbf{r}_{t-1})$ and $(\mathbf{u}_1, \dots, \mathbf{u}_t, \mathbf{z}_1, \dots, \mathbf{z}_{t-1})$. The Onsager correction by $b_{t1}\mathbf{u}_1 + \cdots + b_{tt}\mathbf{u}_t$ in (6.3) is defined to exactly cancel the component of this bias $\mathbf{s}_\parallel$ in $(\mathbf{u}_1, \dots, \mathbf{u}_t)$, so that $(\mathbf{z}_1, \dots, \mathbf{z}_t)$ has an approximate joint Gaussian law. When the spectrum of $\mathbf{W}$ converges to Wigner's semicircle law, the forms of $\mathbf{r}_\parallel$ and $\mathbf{s}_\parallel$ and variances of $\mathbf{r}_\perp$ and $\mathbf{s}_\perp$ are more straightforward to track across iterations, and this produces a slightly different proof of the AMP analyses in [3, 11].

When the spectrum of $\mathbf{W}$ does not converge to the semicircle law, two difficulties arise in carrying out this conditional analysis. First, the forms of $\mathbf{r}_\parallel, \mathbf{r}_\perp, \mathbf{s}_\parallel, \mathbf{s}_\perp$ in iteration $T$ will depend on

$$n^{-1}\mathbf{u}_s^\top \mathbf{W}^k \mathbf{u}_t \equiv n^{-1}\mathbf{r}_s^\top \boldsymbol{\Lambda}^k \mathbf{r}_t \quad \text{for } k = 1, 2 \text{ and } s, t \leq T.$$

These values will in turn depend on

$$n^{-1}\mathbf{u}_s^\top \mathbf{W}^k \mathbf{u}_t \equiv n^{-1}\mathbf{r}_s^\top \boldsymbol{\Lambda}^k \mathbf{r}_t \quad \text{for } k = 1, \dots, 4 \text{ and } s, t \leq T - 1,$$

which will in turn depend on

$$n^{-1}\mathbf{u}_s^\top \mathbf{W}^k \mathbf{u}_t \equiv n^{-1}\mathbf{r}_s^\top \mathbf{\Lambda}^k \mathbf{r}_t \quad \text{for } k = 1, \dots, 6 \text{ and } s, t \le T - 2,$$

and so forth. The final dependence is on $n^{-1}\mathbf{u}_1^\top \mathbf{W}^k \mathbf{u}_1$ for $k = 1, \dots, 2T$, whose large-$n$ limits are given by the first $2T$ moments of the limit spectral distribution of $\mathbf{W}$, because the initialization $\mathbf{u}_1$ is independent of $\mathbf{W}$ which is rotationally invariant in law. The free cumulants of $\mathbf{W}$ that appear in the final forms of the Onsager correction and state evolution emerge by tracking these dependences. To provide an inductive argument that can describe these dependences for arbitrary iterations, our proof establishes a precise form of

$$\lim_{n\to\infty} n^{-1}\mathbf{u}_s^\top \mathbf{W}^k \mathbf{u}_t$$

for every fixed moment $k \ge 0$ and all fixed iterates $s, t \ge 1$. These forms depend on combinatorial coefficients that we call "partial moment coefficients," defined by summing over certain subsets of the noncrossing partition lattice, and which interpolate between the moments and free cumulants of the spectral distribution of $\mathbf{W}$. We define these coefficients in Appendix A.1.

A second technical difficulty which arises is that for the resulting conditioning events $\mathbf{OX} = \mathbf{Y}$, the form of the matrix $\mathbf{X}^\top\mathbf{X}$ in (6.5) becomes complicated, depending on series of matrices with these partial moment coefficients, and $(\mathbf{X}^\top\mathbf{X})^{-1}$ does not admit a tractable description. Instead, we handle matrix-vector products $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{v}$ arising in the computation by "guessing" the form $\mathbf{w}$ for this product, and then verifying that $(\mathbf{X}^\top\mathbf{X})\mathbf{w} = \mathbf{v}$. This type of verification is contained in Lemma A.3, and relies on combinatorial identities for these partial moment coefficients.

The proof ideas in the rectangular setting are similar: We write $\mathbf{W} = \mathbf{O}^\top\mathbf{\Lambda}\mathbf{Q}$ and express (1.5)–(1.8) in an expanded form analogous to (6.1)–(6.4) above. A key component of the proof is then to identify the large-$(m, n)$ limits of the four quantities

$$m^{-1}\mathbf{u}_s^\top(\mathbf{W}\mathbf{W}^\top)^k\mathbf{u}_t, \qquad m^{-1}\mathbf{v}_s^\top\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top)^k\mathbf{u}_t,$$
$$n^{-1}\mathbf{u}_s^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^k\mathbf{v}_t, \qquad n^{-1}\mathbf{v}_s^\top(\mathbf{W}^\top\mathbf{W})^k\mathbf{v}_t$$

for all fixed moments $k \ge 0$ and iterates $s, t \ge 1$. These will depend on certain partial moment coefficients that interpolate between the moments and rectangular free cumulants of the limit singular value distribution of $\mathbf{W}$, and which are defined by summing over subsets of the lattice of noncrossing partitions of sets with even cardinality. These coefficients are defined in Appendix B.1, and the corresponding identities involving $(\mathbf{X}^\top\mathbf{X})^{-1}$ are contained in Lemma B.3.

For the analyses of the Bayes-AMP algorithms for PCA in Section 3, part (a) of Theorems 3.1 and 3.4 are straightforward consequences of the results for the general AMP algorithms. Part (b) of these theorems require an analysis of the state evolutions for the single-iterate posterior mean denoisers, which we carry out in Appendix C.2. This analysis applies a contractive mapping argument to show that for sufficiently large signal strengths, the matrices $\mathbf{\Delta}_t$, $\mathbf{\Sigma}_t$, $\mathbf{\Gamma}_t$ and $\mathbf{\Omega}_t$ all converge as $t \to \infty$ in a space of "infinite matrices" equipped with a weighted $\ell_\infty$ metric.

**Acknowledgments.** I am grateful to my advisor Andrea Montanari, who first introduced me to the beautiful worlds of both free probability and AMP. I would like to thank Keigo Takeuchi and Galen Reeves for helpful discussions and pointers to related literature, and Yufan Li for pointing out an error in a previous version of the manuscript

**Funding.** This research is supported in part by NSF Grant DMS-1916198.

## SUPPLEMENTARY MATERIAL

**Supplementary appendices** (DOI: 10.1214/21-AOS2101SUPP; .pdf). The supplementary appendices contain the proofs of the theoretical results.

## REFERENCES


[1] ALAOUI, A. E., MONTANARI, A. and SELLKE, M. (2020). Optimization of mean-field spin glasses. Preprint. Available at arXiv:2001.00904.

[2] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[3] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[4] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612

[5] BENAYCH-GEORGES, F. (2009). Rectangular random matrices, related convolution. *Probab. Theory Related Fields* **144** 471–515. MR2496440 https://doi.org/10.1007/s00440-008-0152-z

[6] BENAYCH-GEORGES, F. (2009). Rectangular random matrices, entropy, and Fisher's information. *J. Operator Theory* **62** 371–419. MR2552088

[7] BENAYCH-GEORGES, F. (2011). Rectangular $R$-transform as the limit of rectangular spherical integrals. *J. Theoret. Probab.* **24** 969–987. MR2851240 https://doi.org/10.1007/s10959-011-0362-7

[8] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201 https://doi.org/10.1016/j.aim.2011.02.007

[9] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Anal.* **111** 120–135. MR2944410 https://doi.org/10.1016/j.jmva.2012.04.019

[10] BERTHIER, R., MONTANARI, A. and NGUYEN, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Inf. Inference* **9** 33–79. MR4079177 https://doi.org/10.1093/imaiai/iay021

[11] BOLTHAUSEN, E. (2014). An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Comm. Math. Phys.* **325** 333–366. MR3147441 https://doi.org/10.1007/s00220-013-1862-3

[12] BORGERDING, M. and SCHNITER, P. (2016). Onsager-corrected deep learning for sparse linear inverse problems. In 2016 *IEEE Global Conference on Signal and Information Processing* (*GlobalSIP*) 227–231. IEEE, New York.

[13] BORGERDING, M., SCHNITER, P. and RANGAN, S. (2017). AMP-inspired deep networks for sparse linear inverse problems. *IEEE Trans. Signal Process.* **65** 4293–4308. MR3684065 https://doi.org/10.1109/TSP.2017.2708040

[14] BU, Z., KLUSOWSKI, J., RUSH, C. and SU, W. (2019). Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. In *Advances in Neural Information Processing Systems* 9366–9376.

[15] ÇAKMAK, B. and OPPER, M. (2019). Memory-free dynamics for the Thouless–Anderson–Palmer equations of Ising models with arbitrary rotation-invariant ensembles of random coupling matrices. *Phys. Rev. E* **99** 062140. MR3984544

[16] ÇAKMAK, B. and OPPER, M. (2020). A dynamical mean-field theory for learning in restricted Boltzmann machines. *J. Stat. Mech. Theory Exp.* **10** 103303. MR4197533 https://doi.org/10.1088/1742-5468/abb8c9

[17] CAKMAK, B., WINTHER, O. and FLEURY, B. H. (2014). S-AMP: Approximate message passing for general matrix ensembles. In 2014 *IEEE Information Theory Workshop* (*ITW* 2014) 192–196. IEEE, New York.

[18] CALTAGIRONE, F., ZDEBOROVÁ, L. and KRZAKALA, F. (2014). On convergence of approximate message passing. In 2014 *IEEE International Symposium on Information Theory* 1812–1816. IEEE, New York.

[19] CHEN, W.-K. and LAM, W.-K. (2021). Universality of approximate message passing algorithms. *Electron. J. Probab.* **26** Paper No. 36. MR4235487

[20] COLLINS, B. (2003). Moments and cumulants of polynomial random variables on unitary groups, the Itzykson–Zuber integral, and free probability. *Int. Math. Res. Not.* **17** 953–982. MR1959915 https://doi.org/10.1155/S107379280320917X

[21] DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference* **6** 125–170. MR3671474 https://doi.org/10.1093/imaiai/iaw017

[22] DESHPANDE, Y. and MONTANARI, A. (2014). Information-theoretically optimal sparse PCA. In 2014 *IEEE International Symposium on Information Theory* 2197–2201. IEEE, New York.

[23] DIA, M., MACRIS, N., KRZAKALA, F., LESIEUR, T. and ZDEBOROVÁ, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems* 424–432.

[24] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z

[25] DONOHO, D. L., JAVANMARD, A. and MONTANARI, A. (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Inf. Theory* **59** 7434–7464. MR3124654 https://doi.org/10.1109/TIT.2013.2274513

[26] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[27] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2010a). Message passing algorithms for compressed sensing: I. Motivation and construction. In 2010 *IEEE Information Theory Workshop on Information Theory* (*ITW* 2010, *Cairo*) 1–5. IEEE.

[28] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2010b). Message passing algorithms for compressed sensing: II. Analysis and validation. In 2010 *IEEE Information Theory Workshop on Information Theory* (*ITW* 2010, *Cairo*) 1–5. IEEE.

[29] FAN, Z. (2022). Supplement to "Approximate Message Passing algorithms for rotationally invariant matrices." https://doi.org/10.1214/21-AOS2101SUPP

[30] FENG, O. Y., VENKATARAMANAN, R., RUSH, C. and SAMWORTH, R. J. (2021). A unifying tutorial on Approximate Message Passing. Preprint. Available at arXiv:2105.02180.

[31] FLETCHER, A., SAHRAEE-ARDAKAN, M., RANGAN, S. and SCHNITER, P. (2016). Expectation consistent approximate inference: Generalizations and convergence. In 2016 *IEEE International Symposium on Information Theory* (*ISIT*) 190–194. IEEE.

[32] GAMARNIK, D. and JAGANNATH, A. (2021). The overlap gap property and approximate message passing algorithms for $p$-spin models. *Ann. Probab.* **49** 180–205. MR4203336 https://doi.org/10.1214/20-AOP1448

[33] GUIONNET, A. and MAÏDA, M. (2005). A Fourier view on the $R$-transform and related asymptotics of spherical integrals. *J. Funct. Anal.* **222** 435–490. MR2132396 https://doi.org/10.1016/j.jfa.2004.09.015

[34] JAVANMARD, A. and MONTANARI, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf. Inference* **2** 115–144. MR3311445 https://doi.org/10.1093/imaiai/iat004

[35] KABASHIMA, Y. (2003). A CDMA multiuser detection algorithm on the basis of belief propagation. *J. Phys. A*: *Math. Gen.* **36** 11111.

[36] KABASHIMA, Y. and VEHKAPERÄ, M. (2014). Signal recovery using expectation consistent approximation for linear observations. In 2014 *IEEE International Symposium on Information Theory* 226–230. IEEE, New York.

[37] MA, J. and PING, L. (2017). Orthogonal AMP. *IEEE Access* **5** 2020–2033.

[38] MAILLARD, A., FOINI, L., LAGE CASTELLANOS, A., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2019). High-temperature expansions and message passing algorithms. *J. Stat. Mech. Theory Exp.* **11** 113301. MR4059712 https://doi.org/10.1088/1742-5468/ab4bbb

[39] MALEKI, A., ANITORI, L., YANG, Z. and BARANIUK, R. G. (2013). Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Trans. Inf. Theory* **59** 4290–4308. MR3071330 https://doi.org/10.1109/TIT.2013.2252232

[40] METZLER, C., MOUSAVI, A. and BARANIUK, R. (2017). Learned D-AMP: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems* 1772–1783.

[41] MINKA, T. P. (2001). A family of algorithms for approximate Bayesian inference. Ph.D. thesis, Massachusetts Institute of Technology.

[42] MONTANARI, A. (2019). Optimization of the Sherrington–Kirkpatrick Hamiltonian. In 2019 *IEEE* 60*th Annual Symposium on Foundations of Computer Science* (*FOCS*) 1417–1433. IEEE, New York.

[43] MONTANARI, A. and RICHARD, E. (2016). Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Trans. Inf. Theory* **62** 1458–1484. MR3472260 https://doi.org/10.1109/TIT.2015.2457942

[44] MONTANARI, A. and VENKATARAMANAN, R. (2021). Estimation of low-rank matrices via approximate message passing. *Ann. Statist.* **49** 321–345. MR4206680 https://doi.org/10.1214/20-AOS1958

[45] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2018). Consistent parameter estimation for LASSO and approximate message passing. *Ann. Statist.* **46** 119–148. MR3766948 https://doi.org/10.1214/17-AOS1544

[46] NOVAK, J. (2014). Three lectures on free probability. *Random matrix theory, interacting particle systems, and integrable systems* **65** 13.

[47] OPPER, M., ÇAKMAK, B. and WINTHER, O. (2016). A theory of solving TAP equations for Ising models with general invariant random matrices. *J. Phys. A* **49** 114002. MR3462332 https://doi.org/10.1088/1751-8113/49/11/114002

[48] OPPER, M. and WINTHER, O. (2001). Adaptive and self-averaging Thouless–Anderson–Palmer mean-field theory for probabilistic modeling. *Phys. Rev. E* **64** 056131.

[49] OPPER, M. and WINTHER, O. (2001). Tractable approximations for probabilistic models: The adaptive Thouless–Anderson–Palmer mean field approach. *Phys. Rev. Lett.* **86** 3695–3699. https://doi.org/10.1103/PhysRevLett.86.3695

[50] OPPER, M. and WINTHER, O. (2005). Expectation consistent approximate inference. *J. Mach. Learn. Res.* **6** 2177–2204. MR2249885

[51] PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2018). Message-passing algorithms for synchronization problems over compact groups. *Comm. Pure Appl. Math.* **71** 2275–2322. MR3862091 https://doi.org/10.1002/cpa.21750

[52] RANGAN, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In 2011 *IEEE International Symposium on Information Theory Proceedings* 2168–2172. IEEE, New York.

[53] RANGAN, S. and FLETCHER, A. K. (2012). Iterative estimation of constrained rank-one matrices in noise. In 2012 *IEEE International Symposium on Information Theory Proceedings* 1246–1250. IEEE.

[54] RANGAN, S., SCHNITER, P. and FLETCHER, A. K. (2019). Vector approximate message passing. *IEEE Trans. Inf. Theory* **65** 6664–6684. MR4009222 https://doi.org/10.1109/TIT.2019.2916359

[55] RANGAN, S., SCHNITER, P., FLETCHER, A. K. and SARKAR, S. (2019). On the convergence of approximate message passing with arbitrary matrices. *IEEE Trans. Inf. Theory* **65** 5339–5351. MR4009237 https://doi.org/10.1109/TIT.2019.2913109

[56] SCHNITER, P. and RANGAN, S. (2015). Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Signal Process.* **63** 1043–1055. MR3311635 https://doi.org/10.1109/TSP.2014.2386294

[57] SCHNITER, P., RANGAN, S. and FLETCHER, A. K. (2016). Vector approximate message passing for the generalized linear model. In 2016 50*th Asilomar Conference on Signals, Systems and Computers* 1525–1529. IEEE.

[58] SPEICHER, R. (1998). *Combinatorial Theory of the Free Product with Amalgamation and Operator-Valued Free Probability Theory* **627**. AMS, Providence.

[59] SU, W., BOGDAN, M. and CANDÈS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150. MR3718164 https://doi.org/10.1214/16-AOS1521

[60] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

[61] SUR, P., CHEN, Y. and CANDÈS, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a *rescaled* chi-square. *Probab. Theory Related Fields* **175** 487–558. MR4009715 https://doi.org/10.1007/s00440-018-00896-9

[62] TAKEUCHI, K. (2017). Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 501–505. IEEE, New York.

[63] TAKEUCHI, K. (2019). A unified framework of state evolution for message-passing algorithms. In 2019 *IEEE International Symposium on Information Theory* (*ISIT*) 151–155. IEEE, New York.

[64] TAKEUCHI, K. (2020). Convolutional approximate message-passing. *IEEE Signal Process. Lett.* **27** 416–420.

[65] TAKEUCHI, K. (2020). Bayes-optimal convolutional AMP. Preprint. Available at arXiv:2003.12245.

[66] VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9