# AMP Algorithms for Rotationally-Invariant Models: A Unified Approach via Orthogonal Decomposition

Songbin Liu  Junjie Ma*

**Abstract**

This paper aims to provide a unified framework for constructing approximate message passing (AMP) algorithms for rotationally-invariant models. Starting with a general iterative algorithm template, we show how to derive the correct Onsager terms in a principled way via reduction to long-memory orthogonal AMP (OAMP) algorithms. We use this method to re-derive an AMP algorithm recently proposed by Fan and Opper et al. Unlike the works of Opper et al. and Fan, free cumulants of the spectral law emerge in our derivations of AMP from a recursive centering operation, potentially of independent interest. To showcase the versatility of our approach, we devise two novel variants of AMP and demonstrate its applications to estimation in spiked models.

## Contents



*The authors are with the Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China. Emails:{liusongbin,majunjie}@lsec.cc.ac.cn

# 1 Introduction

## 1.1 AMP Algorithm

Approximate message passing (AMP) [4, 6, 12, 21, 23] has been applied to a broad range of high-dimensional estimation problems. Assume that $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ is a symmetric random matrix sampled from the Gaussian orthogonal ensemble (GOE), namely, $\boldsymbol{W} = \frac{1}{\sqrt{2N}}\left(\boldsymbol{G}+\boldsymbol{G}^{\mathsf{T}}\right)$, where $\boldsymbol{G}$ consists of independent identically distributed (i.i.d.) Gaussian entries. Starting from an initial guess $\boldsymbol{u}_0$, the AMP algorithm proceeds as follows:

$$\boldsymbol{r}_t = \boldsymbol{W}\boldsymbol{u}_t - \langle \partial_{t-1}\boldsymbol{u}_t\rangle \cdot \boldsymbol{u}_{t-1}, \tag{1a}$$

$$\boldsymbol{u}_{t+1} = \eta_{t+1}(\boldsymbol{r}_t), \tag{1b}$$

where $\eta_t : \mathbb{R} \mapsto \mathbb{R}$ is Lipschitz continuous and applies to elements of the input vector $\boldsymbol{r}_t$ separately and $\langle \partial_{t-1}\boldsymbol{u}_t\rangle := \sum_{i-1}^{N} \eta_t'\left(r_{t-1}[i]\right)/N$ is the divergence of the nonlinearity $\eta_t$. A distinguishing characteristic of the AMP algorithm is that the empirical law of the iterates $\boldsymbol{r}_t$ converges to a Gaussian distribution in the high-dimensional limit:

$$\boldsymbol{r}_t \to \mathcal{N}(0,\tau_t^2), \quad \forall t \geq 1.$$

Moreover, the variance of the limiting Gaussian distribution can be tracked by a simple recursion known as *state evolution*:

$$\tau_t^2 = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\eta_{t-1}^2\left(\tau_{t-1}Z\right)\right].$$

This property enables precise characterization of the performance of AMP, and has shed important insight to various high-dimensional estimation and optimization problems [1, 5, 7, 13, 14, 19, 21, 37, 38, 47, 49, 55, 58].

A major limitation of the Gaussian AMP algorithm (referred to as *Gaussian AMP* hereafter) is that its theoretical guarantee relies on the fact that $\boldsymbol{W}$ is GOE. Extending AMP beyond this setup is the subject of extensive studies in recent years [8, 9, 15–18, 20, 22, 27–30, 33, 42, 46, 48, 48, 50–54, 56, 59]. In particular, building on the earlier work of Opper et al. [42], Fan [20] proposed the following AMP algorithm for *rotationally-invariant* models:

$$\boldsymbol{r}_t = \boldsymbol{W}\boldsymbol{u}_t - \left(\mathsf{b}_{t,1}\boldsymbol{u}_1 + \mathsf{b}_{t,2}\boldsymbol{u}_2 + \cdots + \mathsf{b}_{t,t}\boldsymbol{u}_t\right), \tag{2a}$$

$$\boldsymbol{u}_{t+1} = \eta_{t+1}(\boldsymbol{r}_t). \tag{2b}$$

where the coefficients $(\mathsf{b}_{t,i})_{1\leq t,1\leq i\leq t}$ depend on the *free cumulants* of the spectral law of $\boldsymbol{W}$. This AMP algorithm, which we refer to as rotationally-invariant AMP (RI-AMP), was first proposed in [42] using the non-rigorous dynamical functional theory in the context of Ising models; see [34, 42] and the references therein for more information. Similar to the original AMP, the iterate $\boldsymbol{r}_t$ in RI-AMP is asymptotically Gaussian distributed. (Note that unlike the Gaussian AMP in (1), the Onsager term in the above RI-AMP algorithm involves all iterates $(\boldsymbol{u}_i)_{i\leq t}$, even when the nonlinearity $\eta_t$ only depends on $\boldsymbol{r}_t$.) Moreover, the variance of the limiting Gaussian distribution can be tracked by a state evolution, which was rigorously proved in [20] using a conditioning technique pioneered in [4, 6] for Gaussian models and generalized in [46, 50] for rotationally-invariant models.

The above RI-AMP algorithm has been generalized to various other setups, including the rectangular matrix setting [20], generalized linear models (GLM) [54], and multi-layer GLM [57].

## 1.2 Orthogonal AMP Algorithm

The orthogonal AMP (OAMP) [29] or vector AMP (VAMP) [46] algorithms, which can be derived based on expectation propagation [36, 41], represent another line of work that generalizes AMP to rotationally-invariant models. This algorithm has been applied in various high-dimensional estimation problems [11, 31, 32, 43]. OAMP relies on the use of *trace-free* matrix denoisers and *divergence-free* iterate denoisers:

$$\boldsymbol{x}_t = \left( f_t(\boldsymbol{W}) - \frac{\mathrm{tr}\left(f_t(\boldsymbol{W})\right)}{N} \cdot \boldsymbol{I}_N \right) \bar{\boldsymbol{x}}_t, \tag{3a}$$

$$\bar{\boldsymbol{x}}_{t+1} = g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t) - \sum_{i=1}^{t} \langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t) \rangle \cdot \boldsymbol{x}_i, \tag{3b}$$

where $f_t : \mathbb{R} \mapsto \mathbb{R}$ is applied to the matrix $\boldsymbol{W}$ in the following sense: let $\boldsymbol{W} = \boldsymbol{O}\mathrm{diag}\left(\lambda_1, \ldots, \lambda_N\right)\boldsymbol{O}^{\mathsf{T}}$ be the eigenvalue decomposition of $\boldsymbol{W}$, then $f_t(\boldsymbol{W}) := \boldsymbol{O}\mathrm{diag}\left(f_t(\lambda_1), f_t(\lambda_2), \ldots, f_t(\lambda_N)\right)\boldsymbol{O}^{\mathsf{T}}$; and $\partial_s g_t$ denotes the partial derivative with respect to the $s$th argument of $g_t$. A precise definition of the OAMP algorithm can be found in Section 3. The original OAMP algorithm [29] assumes $g_t$ to be a univariate function of $\boldsymbol{r}_t$, and the extension to the multivariate case is due to Takeuchi [51]. Note that the long-memory OAMP algorithm originally proposed in [51] is more general than (3). The above presentation of the OAMP algorithm follows Dudeja et al. [18] and is general enough for the purpose of the present paper. Similar to RI-AMP, the iterate $\boldsymbol{x}_t$ in OAMP converges to Gaussian $\boldsymbol{x}_t \to \mathcal{N}(0, \tau_t^2)$, $\forall t \geq 1$, thanks to the use of trace-free matrix denoiser and divergence-free iterate denoisers.

The OAMP iteration in (3) has been used to construct various AMP algorithms, either explicitly or implicitly (e.g., [28, 52]). In a different direction, Takeuchi [51] showed that the original Gaussian AMP algorithm can be mapped to some OAMP algorithm, in the context of compressed sensing. The idea of reducing a general iterative algorithm to certain OAMP algorithms has been explored in Dudeja et al. [18] for proving universality (with respect to the sensing matrix) of the performance of convex regularized least squares estimators, and in Dudeja et al. [17] for analyzing the statistically optimal performance achievable within a broad class of iterative algorithms for spiked models.

## 1.3 Contributions

Compared with the Gaussian AMP counterpart, the derivations of the rotationally-invariant AMP in [20, 42] are more complicated. In this paper, we aim to provide a unified, and arguably more elementary, approach to constructing AMP algorithms (more specifically, their Onsager terms) for rotationally-invariant models. Our approach can be used to re-derive existing AMP algorithms and devise new variants quite easily. The main contributions of this paper include:

- We introduce a unified framework for constructing AMP algorithms for rotationally-invariant models. Our approach is through reduction to long-memory OAMP algorithms, based on *orthogonal decomposition* of the iterates and recursive unfolding the algorithm. The orthogonal decomposition idea was first introduced in [18] but for a different purpose. We use this approach to re-derive the RI-AMP algorithm. Our results provide an alternative and more interpretable state evolution of RI-AMP.

- In our derivation of RI-AMP, the de-biasing coefficients are naturally represented as the normalized traces of certain polynomials of $\boldsymbol{W}$. We prove that these coefficients are related to the free cumulants of the spectral law of $\boldsymbol{W}$. Our proof relies on a recursive characterization of the free cumulants, which appears to be novel and may be of independent interest.

- We demonstrate the versatility of our approach by devising two variants of RI-AMP. The first variant, which we refer to as RI-AMP-DF, employs a different form of the Onsager term which only cancels out the essential non-Gaussian terms. We show that RI-AMP-DF is equivalent to RI-AMP with a change of variables. The second variant (called RI-AMP-MP), which applies a nonlinear matrix processing on $\boldsymbol{W}$, is inspired by Barbier et al. [2]. We further apply RI-AMP-DF to a signal estimation problem in spiked models. Our approach provides a generalization of the BAMP algorithm proposed in [2] to handle general non-polynomial matrix processing functions.

### 1.4 Organization and Notations

**Organization.** This paper is organized as follows. We start with some preliminary results on free cumulants in Section 2. In Section 3, we review some existing results on the orthogonal AMP algorithm. Section 4 contains the main results of this paper. Section 5 provides some generalizations and applications of our framework. Section 6 includes some numerical experiments. The appendices contain omitted proofs and detailed calculations.

**Notation.** The sets $\mathbb{R}, \mathbb{C}$ represent the set real numbers and complex numbers. $[N]$ represents the set $\{1, 2, \ldots, N\}$ and $\mathbb{O}(N)$ denotes the set of $N \times N$ orthogonal matrices. We use the bold-face font for vectors and matrices. $\|\boldsymbol{u}\|$ denotes the $\ell_2$ norm of the vector $\boldsymbol{u}$. $\boldsymbol{A} \otimes \boldsymbol{B}$ and $\boldsymbol{A} \circ \boldsymbol{B}$ denote Kronecker product and Hadamard product of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. $\mathrm{diag}(\boldsymbol{u})$ represents diagonal matrix with diagonal entries given by the entries of $\boldsymbol{u}$. $\mathrm{diag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_t)$ denotes a block diagonal matrix with the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_t$ placed on the diagonal blocks. $\mathrm{tr}(\boldsymbol{M}), \|\boldsymbol{M}\|_{\mathrm{op}}, \|\boldsymbol{M}\|$ represent the trace, operator norm, and Frobenius norm of the matrix $\boldsymbol{M}$ respectively. $\boldsymbol{I}_k$ denotes the $k \times k$ identity matrix. We use $\mathbb{E}[\cdot], \mathrm{Var}[\cdot], \mathrm{Cov}[\cdot]$ to denote expectations, variances, and covariances of random variables. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We use $\mathsf{Unif}(\mathbb{O}(N))$ to denote the Haar measure on the orthogonal group $\mathbb{O}(N)$. The probability measure $\delta_x$ on $\mathbb{R}$ denotes the point mass at $x \in \mathbb{R}$. We use $\xrightarrow{\mathbb{P}}$ to denote convergence in probability. For a sequence of real-valued random variables $(Y_N)_{N \geq 1}$, we say that $\operatorname{plim} Y_N = y$ if $Y_N \xrightarrow{\mathbb{P}} y$.

## 2 Preliminaries

The Onsager term of the rotationally-invariant AMP (RI-AMP) algorithm in (2) involves the *free cumulants* [40] of the spectral measure $\mu$. In this section, we will first review the definition of free cumulants and then introduce a recursive characterization of free cumulants. This recursive characterization, which appears to be novel, will be used in our derivation of RI-AMP.

### 2.1 Free Cumulant

**Non-crossing partition.** A *partition* of the set $[k] := \{1, 2, \ldots, k\}$ is a collection of nonempty disjoint sets $B_1, B_2, \ldots, B_k$, called *blocks*, whose union is $[k]$. A partition is *non-crossing* if there are no four distinct elements $1 \leq a < b < c < d \leq k$ such that $a, c$ are in the same block while $b, d$ are in another block. The collection of all non-crossing partition of $[k]$ is denoted as $\mathrm{NC}(k)$. An example of a non-crossing partition of [5] is shown in Fig. 1.

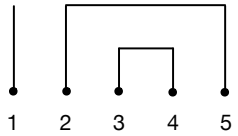



Figure 1: A non-crossing partition of $\{1, 2, 3, 4, 5\}$: $(\{1\}, \{2, 5\}, \{3, 4\})$.

**Free cumulant.** Let $m_k := \mathbb{E}[X^k]$ be the $k$-th moment of a random variable $X$. The free cumulants $(\kappa_k)_{k \geq 1}$ of $X$ are defined implicitly in terms of the moments $(m_k)_{m \geq 1}$ through the moment-cumulant formula [40]:

$$m_k = \sum_{\pi \in \mathrm{NC}(k)} \kappa_\pi, \tag{4}$$

where $\kappa_\pi := \prod_{B \in \pi} \kappa_{|B|}$ is the product of free cumulants corresponding to the cardinality of every block $B \in \pi$. For example, $\mathrm{NC}(3)$ includes the following five non-crossing partitions: $(\{1, 2, 3\})$, $(\{1, 2\}, \{3\})$, $(\{1, 3\}, \{2\})$, $(\{1\}, \{2, 3\})$, $(\{1\}, \{2\}, \{3\})$. (In fact, all partitions of $[3] = \{1, 2, 3\}$ are non-crossing.) Then,

according to (4),

$$m_3 = \kappa_3 + \kappa_2 \cdot \kappa_1 + \kappa_2 \cdot \kappa_1 + \kappa_1 \cdot \kappa_2 + \kappa_1^3.$$

Note that the recursive formula (4) uniquely determines the sequence of free cumulants $(\kappa_k)_{k\geq 1}$ from the sequence of moments $(m_k)_{k\geq 1}$ [40, Lecture 10].

## 2.2 A Recursive Characterization of Free Cumulants

Proposition 1 below introduces a recursive characterization of free cumulants. This recursion naturally appears in our derivation of the RI-AMP algorithm to be detailed in Section 4.

**Proposition 1.** *Assume that the moments $(m_n)_{n\geq 1}$ of a random variable $\Lambda$ exist for all orders. Let $(\kappa_n)_{n\geq 1}$ be the free cumulants of $\Lambda$. Define a sequence of random variables $(Q_n)_{n\geq 0}$ recursively as follows:*

$$Q_n = \Lambda Q_{n-1} - \sum_{i=1}^{n} \mathbb{E}[\Lambda Q_{i-1}] \cdot Q_{n-i}, \quad \forall n \geq 1, \tag{5}$$

*where $Q_0 := 1$. Then, we have*

$$\mathbb{E}[\Lambda Q_{n-1}] = \kappa_n, \quad \forall n \geq 1. \tag{6}$$

*Proof.* See Appendix A. □

*Remark* 1 (Connection with the partial moments in [20]). The random variables $(Q_n)_{n\geq 0}$ defined in Proposition 1 are closely related to the "partial moments" introduced by Fan [20], which are doubly-indexed sequence of coefficients that interpolate moments and free cumulants. Let $(m_n)_{n\geq 1}$ and $(\kappa_n)_{n\geq 0}$ (with the convention $\kappa_0 = 1$) be the moments and free cumulants of $\Lambda \sim \mu$. The partial moments $(c_{k,j})_{k,j\geq 0}$ are defined via [20, Appendix A.1]:

$$c_{k,j} = \sum_{m=0}^{j+1} c_{k-1,m} \cdot \kappa_{j+1-m}, \quad \forall k \geq 1, j \geq 0, \tag{7}$$

with initialization $c_{0,0} = 1$ and $c_{0,j} = 0, \forall j \geq 1$. It can be shown that

$$c_{k,j} = \mathbb{E}\left[\Lambda^k Q_j\right], \quad \forall k, j \geq 0. \tag{8}$$

The proof of (8) can be found in Appendix A.4.

Proposition 1 suggests a way of computing the sequence of free cumulants $(\kappa_n)_{n\geq 1}$ from the sequence of moments $(m_n)_{n\geq 1}$. Note that $Q_n$ is a degree-$n$ polynomial of $\Lambda$ (see (5)). The polynomial coefficients can be computed recursively, as demonstrated in the following corollary.

**Corollary 1.** *Let $(\alpha_{n,i})_{1\leq n, 0\leq i\leq n}$ be the coefficients for the polynomial representations of $(Q_n)_{n\geq 1}$:*

$$Q_n = \sum_{i=0}^{n} \alpha_{n,i} \Lambda^i, \quad \forall n \geq 1. \tag{9}$$

*Then, $(\alpha_{n,i})_{1\leq n, 0\leq i\leq n}$ satisfy the following recursion*

$$\alpha_{n,i} = \alpha_{n-1,i-1} - \sum_{j=1}^{n-i} \kappa_j \cdot \alpha_{n-j,i}, \quad \forall n \geq 1,\ 0 \leq i \leq n-1, \tag{10}$$

*with $\alpha_{i,i} = 1$, and $\alpha_{i,-1} = 0$, $\forall i \geq 0$.*

The proof of Corollary 1 can be found in Appendix A.1. The free cumulant $\kappa_n$ can be calculated via

$$\kappa_n = \mathbb{E}[\Lambda Q_{n-1}] \overset{(9)}{=} \sum_{i=0}^{n-1} \alpha_{n-1,i} \cdot \mathbb{E}\left[\Lambda^{i+1}\right] = \sum_{i=0}^{n-1} \alpha_{n-1,i} \cdot m_{i+1}, \quad \forall n \geq 1. \tag{11}$$

The sequence of free cumulants $(\kappa_n)_{n\geq 1}$ can be calculated by alternatively updating (10) and (11) in a recursive fashion. For the reader's convenience, we describe the whole procedure in Algorithm 1.

**Algorithm 1** Calculating free cumulants from moments

**Input:** $(m_n)_{n\geq 1}$
**Initilization:** $\alpha_{n,n}=1$, $\alpha_{n,-1}=0$, $\forall n\geq 0$. $\kappa_1=m_1$.
**for** $n=1,2,\ldots,$ **do**
  **for** $i=0,1,\ldots,n-1$ **do**
    $\alpha_{n,i}=\alpha_{n-1,i-1}-\sum_{j=1}^{n-i}\kappa_j\cdot\alpha_{n-j,i}$
  **end for**
  $\kappa_n=\sum_{j=0}^{n-1}\alpha_{n-1,j}\cdot m_{j+1}$
**end for**
**Output:** $(\kappa_n)_{n\geq 1}$

*Remark* 2 (Alternative calculations of free cumulants). A more common approach to compute the free cumulant $\kappa_n$ of a measure from its moments $m_1,\ldots,m_n$ is the following recursive formula:

$$\kappa_n = m_n - \sum_{j=1}^{n-1} \prod_{\substack{k_1,\ldots,k_j\geq 0\\ k_1+\cdots+k_j=n-j}} \kappa_j m_{k_1}\cdots m_{k_j}. \tag{12}$$

For a derivation of (12) from the moment-cumulant formula (4), refer to [35, Proposition 17]. The proposed method is computationally favorable than a naive implementation of (12).

## 2.3 Monte Carlo Estimator of Free Cumulants

The Onager term of the RI-AMP algorithm involves the free cumulants of the limiting spectral law $\Lambda\sim\mu$. In applications, it is often the case that $\mu$ is not known and only a sample of $\boldsymbol{W}$ is available. In this case, we may simply replace the moments $(m_n)_{n\geq 1}$ in Algorithm 1 or (12) by their estimates $(\widehat{m}_n)_{n\geq 1}$. Assuming that the empirical eigenvalue distribution of $\boldsymbol{W}$ converges weakly to a compactly supported distribution, we have $\frac{1}{N}\boldsymbol{g}^\mathsf{T}\boldsymbol{W}^n\boldsymbol{g}\xrightarrow{a.s.}\mathbb{E}[\Lambda^n]=m_n$ where $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)$. Motivated by this fact, we consider the following estimator of $m_n$:

$$\widehat{m}_n := \frac{1}{N}\boldsymbol{g}^\mathsf{T}\boldsymbol{W}^n\boldsymbol{g},\quad \boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}). \tag{13}$$

This estimator does not require eigenvalue decomposition of $\boldsymbol{W}$ and is thus computationally favorable for large-scale problems. Similar idea was proposed by Liu et al. [28] in the context of linear models.

**Algorithm 2** Monte Carlo estimator of free cumulants

**Input:** $\boldsymbol{W}$, $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)$, $\boldsymbol{z}_0:=\boldsymbol{g}$, $\widehat{\kappa}_0=1$, $\boldsymbol{h}:=\boldsymbol{W}\boldsymbol{g}$
**for** $n=1,2,\ldots,$ **do**
  $\widehat{\kappa}_n=\frac{1}{N}\boldsymbol{h}^\mathsf{T}\boldsymbol{z}_{n-1}$
  $\boldsymbol{z}_n=\boldsymbol{W}\boldsymbol{z}_{n-1}-\sum_{i=1}^{n}\widehat{\kappa}_i\cdot\boldsymbol{z}_{n-i}$
**end for**
**Output:** $(\widehat{\kappa}_n)_{n\geq 1}$

Here we propose an alternative Monte Carlo estimator of the free cumulants based on Proposition 1. Let $(\boldsymbol{Q}_n)_{n\geq 1}$ be a sequence of polynomials of $\boldsymbol{W}$ defined via (cf. (5))

$$\boldsymbol{Q}_n = \boldsymbol{W}\boldsymbol{Q}_{n-1}-\sum_{i=1}^{n}\kappa_i\cdot\boldsymbol{Q}_{n-i},\quad \forall n\geq 1 \tag{14}$$

with $\boldsymbol{Q}_0=\boldsymbol{I}_N$. Similar to (13) we can estimate $\kappa_k$ as follows

$$\widehat{\kappa}_n = \frac{1}{N}\boldsymbol{g}^\mathsf{T}\boldsymbol{W}\boldsymbol{Q}_{n-1}\boldsymbol{g},\quad \boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0}_N,\boldsymbol{I}_N).$$

Note that we do not need to compute the matrices $(\boldsymbol{Q}_n)_{n\geq 1}$ explicitly. The implementation details of this idea can be found in Algorithm 2.

# 3 Orthogonal AMP Algorithm

Our approach to derive the RI-AMP algorithm is by reducing it (in a recursive fashion) to certain long-memory orthogonal AMP (OAMP) algorithm [51] which has very simple state evolution characterization. In this section, we collect some existing results about the OAMP algorithm for later use.

**Definition 1** (Orthogonal AMP algorithm)**.** Starting from an initialization $\bar{\boldsymbol{x}}_1 \in \mathbb{R}^N$, an orthogonal AMP algorithm proceeds as follows:

$$\boldsymbol{x}_t = \left( f_t(\boldsymbol{W}) - \frac{\operatorname{tr}\left(f_t(\boldsymbol{W})\right)}{N} \cdot \boldsymbol{I}_N \right) \bar{\boldsymbol{x}}_t, \tag{15a}$$

$$\bar{\boldsymbol{x}}_{t+1} = g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) - \sum_{i=1}^{t} \langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) \rangle \cdot \boldsymbol{x}_i, \tag{15b}$$

where

- $f_t : \mathbb{R} \mapsto \mathbb{R}$ is applied to $\boldsymbol{W}$ in the following sense: let $\boldsymbol{W} = \boldsymbol{O}\operatorname{diag}(\lambda_1, \ldots, \lambda_N)\boldsymbol{O}^{\mathsf{T}}$ be the eigenvalue decomposition of $\boldsymbol{W}$, then $f_t(\boldsymbol{W}) := \boldsymbol{O}\operatorname{diag}\left(f_t(\lambda_1), f_t(\lambda_2), \ldots, f_t(\lambda_N)\right)\boldsymbol{O}^{\mathsf{T}}$.
- $g_{t+1} : \mathbb{R}^t \times \mathbb{R}^k \mapsto \mathbb{R}$ acts separately on the $N$ rows of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) \in \mathbb{R}^{N \times t} \times \mathbb{R}^{N \times k}$, and $\langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) \rangle$ denotes the empirical average of the partial derivative of $g_{t+1}$ w.r.t. the $i$-th argument, i.e.,

$$\langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}[n]) \rangle := \frac{1}{N} \sum_{n=1}^{N} \frac{\partial g_{t+1}(\boldsymbol{x}_1[n], \ldots, \boldsymbol{x}_t[n]; \boldsymbol{a}[n])}{\partial \boldsymbol{x}_i[n]}.$$

  Here, $\boldsymbol{a} \in \mathbb{R}^{N \times k}$ represents available side information, and $\boldsymbol{x}_i[n] \in \mathbb{R}$ and $\boldsymbol{a}[n] \in \mathbb{R}^k$ denote the $n$-th component of $\boldsymbol{x}_i$ and $n$-th row of $\boldsymbol{a}$ respectively.

Throughout this paper, we will use the following convergence of high-dimensional vectors. The reader is referred to [4, 20, 21] for more information about this notion of convergence.

**Definition 2** (Convergence of high-dimensional vectors)**.** Let $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell) \in \mathbb{R}^{N \times \ell}$ be a collection of random vectors. We say its empirical distribution converges to random variables $(\mathsf{Z}_1, \ldots, \mathsf{Z}_\ell)$ as $N \to \infty$, which we denote as

$$(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell) \xrightarrow{W_2} (\mathsf{Z}_1, \ldots, \mathsf{Z}_\ell),$$

if for any test function $\psi : \mathbb{R}^\ell \mapsto \mathbb{R}$ satisfying

$$|\psi(\boldsymbol{a}) - \psi(\boldsymbol{b})| \le L\|\boldsymbol{a} - \boldsymbol{b}\|_2 (1 + \|\boldsymbol{a}\|_2 + \|\boldsymbol{b}\|_2), \quad \forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k, \tag{16}$$

the following holds as $N \to \infty$:

$$\frac{1}{N} \sum_{n=1}^{N} \psi\left(\boldsymbol{z}_1[n], \ldots, \boldsymbol{z}_\ell[n]\right) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\psi\left(\mathsf{Z}_1, \ldots, \mathsf{Z}_\ell\right)\right]. \tag{17}$$

The following assumptions are needed for the high-dimensional asymptotic characterization of OAMP.

**Assumption 1** (Assumptions for OAMP)**.**

(1) Let $\boldsymbol{W} = \boldsymbol{O}\operatorname{diag}(\boldsymbol{\lambda})\boldsymbol{O}^{\mathsf{T}}$ be the eigenvalue decomposition of $\boldsymbol{W}$. We assume $\boldsymbol{O} \sim \mathsf{Unif}(\mathbb{O}(N))$ and $\boldsymbol{\lambda} \in \mathbb{R}^N$ is deterministic. Moreover, the empirical distribution of $\boldsymbol{\lambda}$ converges weakly to a compactly support probability measure $\mu$. Additionally, the operator norm of $\boldsymbol{W}$ is upper bounded by an $N$-independent constant $C$.

(2) The side information $\boldsymbol{a} \in \mathbb{R}^{N \times k}$ and the initialization $\bar{\boldsymbol{x}}_1$ are independent of $\boldsymbol{O}$. Moreover, $(\bar{\boldsymbol{x}}_1, \boldsymbol{a}) \xrightarrow{W_2} (\bar{\mathsf{X}}_1, \mathsf{A})$, where $\mathbb{E}[\bar{\mathsf{X}}_1^2] < \infty$ and $\mathbb{E}[\|\mathsf{A}\|^2] < \infty$.

(3) For all $t \geq 1$, the matrix denoiser $f_t : \mathbb{R} \mapsto \mathbb{R}$ is continuous, and does not depend on $N$.

(4) For all $t \geq 1$, the function $g_{t+1} : \mathbb{R}^t \times \mathbb{R}^k \mapsto \mathbb{R}$ is continuously-differentiable and Lipschitz, and does not depend on $N$.

The high-dimensional asymptotic performance of OAMP admits a state evolution description. This was first conjectured in [29] for the OAMP algorithm with univariate denoisers. A rigorous proof was provided by [46, 50] based on a generalization of the conditioning technique in [4, 6] to Haar random orthogonal matrices. (The proof in [46] was for the vector AMP (VAMP) algorithm. In this paper, the acronyms OAMP and VAMP are used interchangeably.) Generalizations of OAMP to the case of multivariate denoisers are introduced in [18, 28, 51]. The following theorem provides a state evolution characterization for the version of OAMP algorithm as defined in Definition 1. Its proof, which is based on a simple modification of [18], is deferred to Appendix B.

**Theorem 1** (State evolution of OAMP [18])**.** *Let $(\boldsymbol{x}_t)_{t\geq 1}$ be generated via the OAMP algorithm. Suppose Assumption 1 holds. Then, the following holds as $N \to \infty$:*

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) \xrightarrow{W_2} (\mathsf{X}_1, \ldots, \mathsf{X}_t; \mathsf{A}), \quad \forall t \geq 1, \tag{18a}$$

*where $(\mathsf{X}_1, \ldots, \mathsf{X}_t) \sim \mathcal{N}(\mathbf{0}_t, \boldsymbol{\Omega}_t)$ is independent of $(\bar{\mathsf{X}}_1, \mathsf{A})$, and*

$$\boldsymbol{\Omega}_t = \begin{bmatrix} \mathrm{Cov}_\mu[f_1, f_1] & \cdots & \mathrm{Cov}_\mu[f_1, f_t] \\ \mathrm{Cov}_\mu[f_2, f_1] & \cdots & \mathrm{Cov}_\mu[f_2, f_t] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}_\mu[f_t, f_1] & \cdots & \mathrm{Cov}_\mu[f_t, f_t] \end{bmatrix} \circ \begin{bmatrix} \mathbb{E}[\bar{\mathsf{X}}_1\bar{\mathsf{X}}_1] & \cdots & \mathbb{E}[\bar{\mathsf{X}}_1\bar{\mathsf{X}}_t] \\ \mathbb{E}[\bar{\mathsf{X}}_2\bar{\mathsf{X}}_1] & \cdots & \mathbb{E}[\bar{\mathsf{X}}_2\bar{\mathsf{X}}_t] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\bar{\mathsf{X}}_t\bar{\mathsf{X}}_1] & \cdots & \mathbb{E}[\bar{\mathsf{X}}_t\bar{\mathsf{X}}_t] \end{bmatrix}, \tag{18b}$$

*where $\mathrm{Cov}_\mu[f_i, f_j]$ denotes the covariance between the random variables $f_i(\Lambda)$ and $f_j(\Lambda)$ (with $\Lambda \sim \mu$), and $\circ$ denotes Hadamard product and*

$$\bar{\mathsf{X}}_t := g_t(\mathsf{X}_1, \ldots, \mathsf{X}_{t-1}; \mathsf{A}) - \sum_{i=1}^{t-1} \mathbb{E}[\partial_i g_t(\mathsf{X}_1, \ldots, \mathsf{X}_{t-1}; \mathsf{A})] \cdot \mathsf{Z}_i, \quad \forall t \geq 2. \tag{18c}$$

*In the above equation, $\mu$ denotes the limiting eigenvalue distribution of $\boldsymbol{W}$.*

*Remark* 3 (Interpretation of OAMP's state evolution). The state evolution equation (18b) makes precise the following approximation:

$$\frac{1}{N}\boldsymbol{x}_s^\mathsf{T}\boldsymbol{x}_t \overset{(15)}{=} \frac{1}{N}\bar{\boldsymbol{x}}_s^\mathsf{T}\left(f_s(\boldsymbol{W}) - \frac{\mathrm{tr}\,(f_s(\boldsymbol{W}))}{N}\cdot\boldsymbol{I}_N\right)\left(f_t(\boldsymbol{W}) - \frac{\mathrm{tr}\,(f_t(\boldsymbol{W}))}{N}\cdot\boldsymbol{I}_N\right)\bar{\boldsymbol{x}}_t, \quad \forall s,t \geq 1 \tag{19a}$$

$$\simeq \frac{1}{N}\mathsf{tr}\left(f_s(\boldsymbol{W}) - \frac{\mathrm{tr}\,(f_s(\boldsymbol{W}))}{N}\cdot\boldsymbol{I}_N\right)\left(f_t(\boldsymbol{W}) - \frac{\mathrm{tr}\,(f_t(\boldsymbol{W}))}{N}\cdot\boldsymbol{I}_N\right)\cdot\frac{1}{N}\bar{\boldsymbol{x}}_s^\mathsf{T}\bar{\boldsymbol{x}}_t. \tag{19b}$$

Standard random matrix results show that the above approximation is exact (under mild conditions) in the large-$N$ limit, if that $\bar{\boldsymbol{x}}_t$ and $\bar{\boldsymbol{x}}_s$ were independent of the noise matrix $\boldsymbol{W}$. Therefore, for a heuristic interpretation of the state evolution, we may treat the iterates as if they are independent of $\boldsymbol{W}$. This property significantly simplifies the dynamics of OAMP algorithms and renders the state evolution easily interpretable.

A distinguishing feature of OAMP algorithm is the pairwise asymptotic orthogonality between the iterates $\boldsymbol{x}_s$ and $\bar{\boldsymbol{x}}_t$, $\forall s,t \geq 1$. This orthogonality property justifies the name "orthogonal AMP".

**Proposition 2** (Pairwise orthogonality)**.** *The following holds as $N \to \infty$,*

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\boldsymbol{x}_s^\mathsf{T}\bar{\boldsymbol{x}}_t = 0, \quad \forall s,t \geq 1. \tag{20}$$

*Proof.* The case $t = 1$ is straightforward. We consider the case $t > 1$. From Theorem 1, the joint empirical distributions of $\boldsymbol{x}_s$ and $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}; \boldsymbol{a})$ converges in the sense of Definition 2. Consequently,

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} \frac{1}{N}\boldsymbol{x}_s^{\mathsf{T}}\bar{\boldsymbol{x}}_t &\overset{\text{(a)}}{=} \operatorname*{plim}_{N\to\infty} \frac{1}{N}\boldsymbol{x}_s^{\mathsf{T}}\left(g_t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}; \boldsymbol{a}) - \sum_{i=1}^{t-1}\langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}; \boldsymbol{a})\rangle \cdot \boldsymbol{x}_i\right) \\ &\overset{\text{(b)}}{=} \mathbb{E}\left[\mathsf{X}_s g_t(\mathsf{X}_1, \ldots, \mathsf{X}_{t-1}; \mathsf{A})\right] - \sum_{i=1}^{t-1}\mathbb{E}\left[\partial_i g_t(\mathsf{X}_1, \ldots, \mathsf{X}_{t-1}; \mathsf{A})\right] \cdot \mathbb{E}[\mathsf{X}_s \mathsf{X}_i] \\ &\overset{\text{(c)}}{=} 0,\end{aligned}$$

where step (a) is from the definition of $\bar{\boldsymbol{x}}_t$ in (15b); step (b) is due to the convergence result in Theorem 1 and the fact that $\partial_i g_t$ is continuous and bounded (from the assumption that $g_t$ is Lipschitz and continuously-differentiable); step (c) is from the multivariate version of Stein's lemma. □

*Remark* 4. When $s \geq t$, $\boldsymbol{x}_s$ is not an immediate input of $\bar{\boldsymbol{x}}_t := g_t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}; \boldsymbol{a}) - \sum_{i=1}^{t-1}\langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}; \boldsymbol{a})\rangle \cdot \boldsymbol{x}_i$. Note that the orthogonality property (20) still holds in such cases.

*Remark* 5 (Orthogonal decomposition). We can decompose $g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a})$ as

$$g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) = \sum_{i=1}^{t}\langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a})\rangle \cdot \boldsymbol{x}_i + \bar{\boldsymbol{x}}_{t+1}, \tag{21}$$

where $\bar{\boldsymbol{x}}_{t+1} = g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) - \sum_{i=1}^{t}\langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a})\rangle \cdot \boldsymbol{x}_i$. Proposition 2 guarantees that $\bar{\boldsymbol{x}}_{t+1}$ is asymptotically orthogonal to $\boldsymbol{x}_i$, $\forall i \in [t]$. Namely, (21) is a decomposition of $g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a})$ into a component that lies in $\operatorname{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ and a component perpendicular to it asymptotically.

OAMP algorithms have simple and interpretable state evolution, thanks to the use of trace-free and divergence-free denoisers. Moreover, with the introduction of general long-memory processing, the OAMP framework is very flexible and allows for convenient derivation of AMP algorithms (as we show in subsequent sections). These attributes make OAMP a unified and simple approach for deriving AMP algorithms tailored to rotationally-invariant models.

# 4 Main Results

This section presents the main results of this paper. We provide a unified way of deriving the AMP algorithms for rotationally-invariant models.

## 4.1 Rotationally-Invariant AMP Algorithm

The starting point of our derivation of RI-AMP algorithm is the first order method (FOM) defined below. We remark that FOM has the same form as RI-AMP, but the de-biasing coefficients in FOM are free parameters and not necessarily set as in RI-AMP.

**Definition 3** (First order method (FOM))**.** Starting from an initialization $\boldsymbol{u}_1 \in \mathbb{R}^N$, a first order method (FOM) proceeds as

$$\boldsymbol{r}_t = \boldsymbol{W}\boldsymbol{u}_t - \left(\mathsf{b}_{t,1}\boldsymbol{u}_1 + \mathsf{b}_{t,2}\boldsymbol{u}_2 + \cdots + \mathsf{b}_{t,t}\boldsymbol{u}_t\right), \quad \forall t \in \mathbb{N}, \tag{22a}$$

$$\boldsymbol{u}_{t+1} = \eta_{t+1}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t), \tag{22b}$$

where $\eta_{t+1} : \mathbb{R}^t \mapsto \mathbb{R}$ acts separately on the $N$ rows of $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t) \in \mathbb{R}^{N\times t}$.

For convenience, let $\mathbf{B}_t \in \mathbb{R}^{t\times t}$ be the matrix collecting the de-biasing coefficients $(\mathsf{b}_{t,i})_{1\leq t, 1\leq i\leq t}$ as

$$\mathbf{B}_t := \begin{bmatrix} \mathsf{b}_{1,1} & & & \\ \mathsf{b}_{2,1} & \mathsf{b}_{2,2} & & \\ \vdots & \vdots & \ddots & \\ \mathsf{b}_{t,1} & \mathsf{b}_{t,2} & \cdots & \mathsf{b}_{t,t} \end{bmatrix}. \tag{23}$$

The rotationally-invariant AMP algorithm is a specific way of setting the de-biasing matrix $\mathbf{B}_t$ in order to make the iterates $(\boldsymbol{r}_t)_{t\geq 1}$ asymptotically Gaussian. We can now define the RI-AMP algorithm formally.

**Definition 4** (RI-AMP algorithm [20])**.** The rotationally-invariant AMP (RI-AMP) algorithm is a first order method (FOM) with the de-biasing coefficients matrix $\mathbf{B}_t$ set to

$$\mathbf{B}_t = \sum_{i=1}^{t} \kappa_i (\widehat{\boldsymbol{\Phi}}_t)^{i-1}, \tag{24}$$

where $(\kappa_i)_{i\geq 1}$ are the free cumulants of the limiting spectral law $\mu$, and the matrix $\widehat{\boldsymbol{\Phi}}_t \in \mathbb{R}^{t\times t}$ collects the empirical partial derivatives of $(\eta_i)_{i\in[t]}$ (i.e., "divergences"):

$$\widehat{\boldsymbol{\Phi}}_t := \begin{bmatrix} 0 & & & & \\ \langle \partial_1 \boldsymbol{u}_2 \rangle & 0 & & & \\ \langle \partial_1 \boldsymbol{u}_3 \rangle & \langle \partial_2 \boldsymbol{u}_3 \rangle & 0 & & \\ \vdots & \vdots & \ddots & & \\ \langle \partial_1 \boldsymbol{u}_t \rangle & \langle \partial_2 \boldsymbol{u}_t \rangle & \cdots & \langle \partial_{t-1} \boldsymbol{u}_t \rangle & 0 \end{bmatrix}. \tag{25}$$

*Remark* 6 (R-transform). Recall that the free cumulants $(\kappa_i)_{i\geq 1}$ are the coefficients in the power series expansion of the R-transform [40]. Consequently, the de-biasing matrix formula (24) can be expressed in terms of the R-transform, as employed in the work of Opper et al. [42]. The R-transform emerges in Opper et al. [42] as a result of evaluating certain high-dimensional integrals asymptotically. In the present work, free cumulants arise naturally within a recursive centering operation.

*Remark* 7 (Reduction to standard Gaussian AMP). For the special iid Gaussian model, namely when $\boldsymbol{W}$ is drawn from the Gaussian orthogonal Ensemble (GOE), the RI-AMP algorithm simplifies significantly. In this case, the free cumulants vanish except for $\kappa_2$:

$$(\kappa_1, \kappa_2, \kappa_3, \kappa_4, \ldots) = (0, 1, 0, 0, \ldots).$$

Hence, the de-biasing matrix formula (24) reduces to $\mathbf{B}_t = \widehat{\boldsymbol{\Phi}}_t$, and RI-AMP reduces to the standard (long-memory) Gaussian AMP.

Following [20], we make the following assumptions in order to analyze the asymptotic performance of RI-AMP.

**Assumption 2** (Assumptions for RI-AMP)**.**

(1) Let $\boldsymbol{W} = \boldsymbol{O}\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{O}^{\mathsf{T}}$ be the eigenvalue decomposition of $\boldsymbol{W}$. We assume $\boldsymbol{O} \sim \mathsf{Unif}(\mathbb{O}(N))$ and $\boldsymbol{\lambda} \in \mathbb{R}^N$ is deterministic. Moreover, the empirical distribution of $\boldsymbol{\lambda}$ converges weakly to a compactly support probability measure $\mu$. Additionally, the operator norm of $\boldsymbol{W}$ is upper bounded by an $N$-independent constant $C$.

(2) The initialization $\boldsymbol{u}_1 \in \mathbb{R}^N$ is independent of $\boldsymbol{O}$. Moreover, $\boldsymbol{u}_1 \xrightarrow{W_2} U_1$, where the random variable $U_1$ has moments with all orders.

(3) For all $t \geq 1$, the function $\eta_t : \mathbb{R}^t \mapsto \mathbb{R}$ is Lipschitz continuous.

## 4.2 Overview of Our Approach

Before presenting the technical details, we first give a brief overview of our approach to AMP algorithms for rotationally-invariant models. Recall that the RI-AMP algorithm has the following form:

$$\boldsymbol{r}_t = \boldsymbol{W}\boldsymbol{u}_t - (\mathsf{b}_{t,1}\boldsymbol{u}_1 + \mathsf{b}_{t,2}\boldsymbol{u}_2 + \cdots + \mathsf{b}_{t,t}\boldsymbol{u}_t), \tag{26a}$$

$$\boldsymbol{u}_{t+1} = \eta_{t+1}(\boldsymbol{r}_t). \tag{26b}$$

For ease of discussion, here we assumed that $\eta_t(\cdot)$ is a single-iterate function that only depends on $\boldsymbol{r}_t$. The question we aim to address is:

- *How to derive the correct de-biasing coefficients* $(\mathsf{b}_{t,i})_{1\le t,1\le i\le t}$ *such that* $(\boldsymbol{r}_t)_{t\ge 1}$ *are asymptotically Gaussian distributed?*

Our approach to this question is based on an orthogonal decomposition idea first introduced by Dudeja et al. [18]. Specifically, we decompose the iterate $\boldsymbol{u}_{t+1} := \eta_{t+1}(\boldsymbol{r}_t)$, for each $t \ge 1$, into a component that is parallel to the input vector $\boldsymbol{r}_t$ and a component orthogonal to it (denoted as $\bar{\boldsymbol{u}}_{t+1}$). For the more general scenario where $\eta_{t+1}(\cdot)$ depends on all past iterates $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t)$, this orthogonal decomposition idea can be generalized naturally: we decompose $\eta_{t+1}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t)$ into a component that lies in $\mathrm{span}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t)$ and a component orthogonal to it. (Notice that, if the input vectors $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t)$ are jointly Gaussian, the residual orthogonal component $\bar{\boldsymbol{u}}_{t+1}$ is nothing but the divergence-free estimate defined in Section 3.) Recursively unfolding the algorithm using this orthogonal decomposition, it is possible to represent the iterates $(\boldsymbol{r}_t)_{t\ge 1}$ as linear combinations of the orthogonal components $(\bar{\boldsymbol{u}}_t)_{t\ge 1}$. This representation is already close to an OAMP algorithm defined in Section 3. It turns out that, by properly choosing the de-biasing coefficients, it is always possible to make the matrices appeared in this linear representation of $(\bar{\boldsymbol{u}}_t)_{t\ge 1}$ asymptotically trace-free, thereby casting the algorithm into an OAMP algorithm. The asymptotic Gaussian distribution then follows from the general property of OAMP.

We illustrate the idea by considering the first few iterations. In our discussions, we will encounter a centering operation that makes a matrix asymptotically trace-free, and we shall denote this centering operation by $\mathcal{T}$. Specifically, letting $f : \mathbb{R} \mapsto \mathbb{R}$ be a matrix denoising function which acts on input matrices as in (15), we denote

$$\mathcal{T}(\boldsymbol{A}) := \boldsymbol{A} - \mathbb{E}_{\Lambda\sim\mu}[f(\Lambda)] \cdot \boldsymbol{I}_N, \quad \text{where } \boldsymbol{A} := f(\boldsymbol{W}). \tag{27}$$

We now consider the first iteration of the algorithm. From (26), we have $\boldsymbol{r}_1 = (\boldsymbol{W} - \mathsf{b}_{1,1}\boldsymbol{I}_N)\,\boldsymbol{u}_1$. By setting $\mathsf{b}_{1,1} = m_1 := \mathbb{E}[\Lambda]$, we have

$$\boldsymbol{r}_1 = \mathcal{T}(\boldsymbol{W})\boldsymbol{u}_1 := Q_1(\boldsymbol{W})\boldsymbol{u}_1. \tag{28}$$

This guarantees $\boldsymbol{r}_1$ to be asymptotically Gaussian by Theorem 1 in Section 3. Next, we show how to choose $\mathsf{b}_{2,1}$ and $\mathsf{b}_{2,2}$ to make $\boldsymbol{r}_2$ asymptotically Gaussian. Our approach is based on an orthogonal decomposition of $\boldsymbol{u}_1$:

$$\boldsymbol{u}_2 := \mathsf{d}_1 \cdot \boldsymbol{r}_1 + \bar{\boldsymbol{u}}_2, \tag{29}$$

where the first term is parallel to $\boldsymbol{r}_1$ and the second component orthogonal to $\boldsymbol{r}_1$ (in certain asymptotic sense). When $\boldsymbol{r}_1$ is asymptotically Gaussian, by appealing to Stein's lemma, the choice $\mathsf{d}_1 := \langle \partial_1 \boldsymbol{u}_2 \rangle$ satisfies the asymptotic orthogonality condition. and $\bar{\boldsymbol{u}}_2 := \eta_1(\boldsymbol{r}_1) - \mathsf{d}_1 \cdot \boldsymbol{r}_1$ is a divergence-free function of $\boldsymbol{r}_1$. For notational simplicity, we assume $\langle \partial_1 \boldsymbol{u}_2 \rangle = 1$ in the following discussions. Now, substituting (29) into (26) and denoting $\boldsymbol{v}_1 := \boldsymbol{u}_1$ yields

$$\begin{aligned}
\boldsymbol{r}_2 &= \boldsymbol{W}\boldsymbol{u}_2 - (\mathsf{b}_{2,1}\boldsymbol{u}_1 + \mathsf{b}_{2,2}\boldsymbol{u}_2) \\
&\overset{(29)}{=} \boldsymbol{W}(\boldsymbol{r}_1 + \bar{\boldsymbol{u}}_2) - \mathsf{b}_{2,1}\bar{\boldsymbol{u}}_1 - \mathsf{b}_{2,2}(\boldsymbol{r}_1 + \bar{\boldsymbol{u}}_2) \\
&\overset{(28)}{=} (\boldsymbol{W}Q_1(\boldsymbol{W}) - \mathsf{b}_{2,2}Q_1(\boldsymbol{W}) - \mathsf{b}_{2,1}\boldsymbol{I})\bar{\boldsymbol{u}}_1 + (\boldsymbol{W} - \mathsf{b}_{2,2}\boldsymbol{I})\bar{\boldsymbol{u}}_2 \\
&= \underbrace{\mathcal{T}\big(\boldsymbol{W}Q_1(\boldsymbol{W}) - \mathsf{b}_{2,2}Q_1(\boldsymbol{W}) - \mathsf{b}_{2,1}\boldsymbol{I}\big)}_{Q_2(\boldsymbol{W})}\bar{\boldsymbol{u}}_1 + \underbrace{\mathcal{T}(\boldsymbol{W} - \mathsf{b}_{2,2}\boldsymbol{I})}_{Q_1(\boldsymbol{W})}\bar{\boldsymbol{u}}_2,
\end{aligned}$$

where the last step holds when $\mathsf{b}_{2,1}$ and $\mathsf{b}_{2,2}$ are chosen to center the respective matrices:

$$\begin{aligned}
\mathsf{b}_{2,2} &= \lim_{N\to\infty} \frac{1}{N}\mathrm{tr}(\boldsymbol{W}) = m_1, \\
\mathsf{b}_{2,1} &= \lim_{N\to\infty} \frac{1}{N}\mathrm{tr}(\boldsymbol{W}Q_1(\boldsymbol{W}) - \mathsf{b}_{2,2}Q_1(\boldsymbol{W})) = \lim_{N\to\infty} \frac{1}{N}\mathrm{tr}(\boldsymbol{W}Q_1(\boldsymbol{W})) \overset{(28)}{=} m_2 - m_1^2.
\end{aligned} \tag{31}$$

Again, by Theorem 1 in Section 3, vectors in the form of $P(\boldsymbol{W})\bar{\boldsymbol{u}}$ are asymptotically Gaussian with $P(\cdot)$ a polynomial and $P(\boldsymbol{W})$ *trace-free*, and $\bar{\boldsymbol{u}}$ *divergence-free*. Hence, both $Q_2(\boldsymbol{W})\bar{\boldsymbol{u}}_1$ and $Q_1(\boldsymbol{W})\bar{\boldsymbol{u}}_2$ are asymptotically Gaussian, and so is $\boldsymbol{r}_2 = Q_2(\boldsymbol{W})\bar{\boldsymbol{u}}_1 + Q_1(\boldsymbol{W})\bar{\boldsymbol{u}}_2$.

We can continue this process. In each step, we decompose the estimate $\eta_t(\boldsymbol{r}_t)$ as a linear term plus a divergence-free term. To illustrate the main idea and simplify discussions, we assume that $\langle \partial_t(\boldsymbol{u}_{t+1}) \rangle = 1$

$\forall t \geq 1$ in this section. It is not difficult to show that, under proper choices of the de-biasing coefficients, we could represent $(\boldsymbol{r}_t)_{t\geq 1}$ as:

$$\begin{aligned}
\boldsymbol{r}_1 &= Q_1(\boldsymbol{W})\bar{\boldsymbol{u}}_1, \\
\boldsymbol{r}_2 &= Q_2(\boldsymbol{W})\bar{\boldsymbol{u}}_1 + Q_1(\boldsymbol{W})\bar{\boldsymbol{u}}_2, \\
\boldsymbol{r}_3 &= Q_3(\boldsymbol{W})\bar{\boldsymbol{u}}_1 + Q_2(\boldsymbol{W})\bar{\boldsymbol{u}}_2 + Q_1(\boldsymbol{W})\bar{\boldsymbol{u}}_3, \\
&\vdots \\
\boldsymbol{r}_t &= Q_t(\boldsymbol{W})\bar{\boldsymbol{u}}_1 + Q_{t-1}(\boldsymbol{W})\bar{\boldsymbol{u}}_2 + Q_{t-2}(\boldsymbol{W})\bar{\boldsymbol{u}}_3 + \cdots + Q_1(\boldsymbol{W})\bar{\boldsymbol{u}}_t,
\end{aligned} \tag{32}$$

where $\bar{\boldsymbol{u}}_1 := \boldsymbol{u}_1$. In the above display, $(Q_t)_{t\geq 1}$ is a sequence of polynomials that have zero-mean w.r.t. $\mu$ (i.e., $Q_t(\boldsymbol{W})$ is trace-free) and they satisfy certain recursive relationship.

In our approach, the way of choosing the de-biasing coefficients is conceptually very simple: *we choose* $\mathsf{B}_t$ *to center the matrices in the representations of* $(\boldsymbol{r}_t)_{t\geq 1}$ *as linear combinations of* $(\bar{\boldsymbol{u}}_t)_{t\geq 1}$. We shall prove in the next subsections that such choice of $\mathsf{B}_t$ is unique, and is precisely the one used in RI-AMP.

## 4.3 Derivation of Rotationally-Invariant AMP

In this section, we present our approach of deriving the RI-AMP algorithm. The basic idea is to start with a FOM and figure out the correct de-biasing matrix such that the FOM can be recursively reduced to certain OAMP algorithm.

Similar to the strategy used in Section 4.2, we introduce a set of intermediate variables $(\bar{\boldsymbol{u}}_t)_{t\geq 1}$. We reformulate the iterates $(\boldsymbol{r}_t)_{t\geq 1}$ in a FOM as linear combinations of $(\bar{\boldsymbol{u}}_t)_{t\geq 1}$, which we record in Lemma 1 below. We emphasize that the results in this lemma are purely *algebraic*. In particular, the quantities $(\mathsf{d}_{t,i})$ are understood as free parameters and need not be the divergences of the de-noisers.

**Lemma 1** (Reformulation of FOM)**.** *Let* $(\boldsymbol{u}_t)_{t\geq 1}$ *and* $(\boldsymbol{r}_t)_{t\geq 1}$ *be generated as in* (22)*. Let* $(\mathsf{d}_{t,i})_{1\leq t,1\leq i\leq t}$ *be an arbitrary sequence. Define*

$$\bar{\boldsymbol{u}}_t := \boldsymbol{u}_t - \left(\mathsf{d}_{t-1,1}\cdot\boldsymbol{r}_1 + \cdots + \mathsf{d}_{t-1,t-1}\cdot\boldsymbol{r}_{t-1}\right), \quad \forall t > 1. \tag{33}$$

*Moreover,* $\bar{\boldsymbol{u}}_1 := \boldsymbol{u}_1$*. Then,* $(\boldsymbol{r}_t)_{t\geq 1}$ *can be represented as*

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = \begin{bmatrix} P_{1,1}(\boldsymbol{W}) & & & \\ P_{2,1}(\boldsymbol{W}) & P_{2,2}(\boldsymbol{W}) & & \\ \vdots & \vdots & \ddots & \\ P_{t,1}(\boldsymbol{W}) & P_{t,1}(\boldsymbol{W}) & \cdots & P_{t,t}(\boldsymbol{W}) \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}, \tag{34}$$

*where* $(P_{t,i})_{1\leq t,1\leq i\leq t}$ *is a sequence of polynomials. Let* $\mathsf{P}_t : \mathbb{R} \mapsto \mathbb{R}^{t\times t}$ *be a collection of these polynomials:*

$$\mathsf{P}_t(\lambda) := \begin{bmatrix} P_{1,1}(\lambda) & & & \\ P_{2,1}(\lambda) & P_{2,2}(\lambda) & & \\ \vdots & \vdots & \ddots & \\ P_{t,1}(\lambda) & P_{t,1}(\lambda) & \cdots & P_{t,t}(\lambda) \end{bmatrix}, \quad \forall \lambda \in \mathbb{R}. \tag{35}$$

*Then,* $\mathsf{P}_t(\lambda)$ *admits the following explicit expression:*

$$\mathsf{P}_t(\lambda) = (\mathsf{I}_t - \lambda\mathsf{D}_t + \mathsf{B}_t\mathsf{D}_t)^{-1}(\lambda\mathsf{I}_t - \mathsf{B}_t), \quad \forall \lambda \in \mathbb{R}. \tag{36}$$

*In the above equation,* $\mathsf{B}_t \in \mathbb{R}^{t\times t}$ *denotes the de-biasing matrix* (23) *and* $\mathsf{D}_t \in \mathbb{R}^{t\times t}$ *is defined as*

$$\mathsf{D}_t := \begin{bmatrix} 0 & & & & \\ \mathsf{d}_{1,1} & 0 & & & \\ \mathsf{d}_{2,1} & \mathsf{d}_{2,2} & 0 & & \\ \vdots & \vdots & \ddots & & \\ \mathsf{d}_{t-1,1} & \mathsf{d}_{t-1,2} & \cdots & \mathsf{d}_{t-1,t-1} & 0 \end{bmatrix}. \tag{37}$$

*Proof.* See Appendix C.1. □

In Lemma 1, we have rewritten the iterate $\boldsymbol{r}_t$ as a linear combination of the variables $(\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_t)$, which we defined in (33). (The reason why we introduce these new variables will be clear in our later discussions.) Towards reducing the FOM to a OAMP algorithm and as the second step of our derivation, we impose a trace-free constraint on the matrices $(P_{t,i})_{1\le t, 1\le i\le t}$ in (33), in order to cast the FOM into certain long-memory OAMP algorithm. The following lemma shows that this trace-free constraint naturally yields the form of the de-biasing matrix $\mathbf{B}_t$ used in RI-AMP.

**Lemma 2** (Choice of de-biasing matrix $\mathbf{B}_t$)**.** *Let $\mathbf{B} \in \mathbb{R}^{t\times t}$ and $\mathbf{D}_t \in \mathbb{R}^{t\times t}$ be two deterministic matrices defined as in (23) and (37). Define $\mathbf{P}_t : \mathbb{R} \mapsto \mathbb{R}^{t\times t}$ as*

$$\mathbf{P}_t(\lambda) := (\mathbf{I}_t - \lambda \mathbf{D}_t + \mathbf{B}_t \mathbf{D}_t)^{-1}(\lambda \mathbf{I}_t - \mathbf{B}_t), \quad \forall \lambda \in \mathbb{R}. \tag{38}$$

*Then, the following hold true.*

*(1) For any given $\mathbf{D}_t$, the following equation has a unique solution in $\mathbf{B}_t$:*

$$\mathbb{E}\left[\mathbf{P}_t(\Lambda)\right] = \mathbf{0}_{t\times t}, \quad \Lambda \sim \mu. \tag{39}$$

*(2) The matrix $\mathbf{B}_t$ that solves (39) can be represented as*

$$\mathbf{B}_t = \sum_{i=1}^{t} \alpha_i \mathbf{D}_t^{i-1}, \tag{40}$$

*where we adopted the convention $\mathbf{D}_t^0 := \boldsymbol{I}$. Let $(Q_n)_{n\ge 0}$ be a sequence of polynomials recursively defined by ($Q_0(\lambda) := 1, \forall \lambda \in \mathbb{R}$):*

$$Q_n(\lambda) = \lambda Q_{n-1}(\lambda) - \sum_{i=1}^{n} \mathbb{E}\left[\Lambda Q_{i-1}(\Lambda)\right] \cdot Q_{n-i}(\lambda), \quad \forall \lambda \in \mathbb{R}, n \ge 1. \tag{41a}$$

*Then, the coefficients $(\alpha_n)_{n\ge 1}$ in (40) are given by*

$$\alpha_n := \mathbb{E}\left[\Lambda Q_{n-1}(\Lambda)\right], \quad \forall n \ge 1. \tag{41b}$$

*(3) The matrix defined in (40) can be alternatively represented as*

$$\mathbf{B}_t = \sum_{i=1}^{t} \kappa_i \mathbf{D}_t^{i-1}, \tag{42}$$

*where $(\kappa_i)_{i\ge 1}$ denotes the free cumulants of $\mu$.*

*(4) With $\mathbf{B}_t$ given by (40), $\mathbf{P}_t(\lambda)$ can be represented as*

$$\mathbf{P}_t(\lambda) = \sum_{i=1}^{t} Q_i(\lambda) \mathbf{D}_t^{i-1}. \tag{43}$$

*Proof.* See Appendix C.2. □

*Remark* 8 (On the appearance of free cumulants). Lemma 2 shows that, to make the matrices in the linear representation (34) asymptotically trace-free (in the sense of (39)), the de-biasing matrix $\mathbf{B}_t$ is uniquely given by a polynomial of $\mathbf{D}_t$, whose coefficients are free cumulants of the limiting spectral law $\mu$. We remark that, in our derivation of RI-AMP, the appearance of the recursion (41) is natural, while its connection with free cumulants (as established in Proposition 1) may not be immediately clear. In Section 5.1, we will introduce a variant of RI-AMP where the Onsager term is a linear combination of all past divergence-free estimates (which may be interpreted as "essential" non-Gaussian terms). In this variant of RI-AMP, the de-biasing coefficients satisfy a recurrence different from (41) and has no direct relationship with the free cumulants.

The above lemma shows that choosing de-biasing matrix $\mathbf{B}_t = \sum_{i=1}^{t} \kappa_i \mathbf{D}_t^{i-1}$ ensures the matrices $(\boldsymbol{P}_{i,j}(\boldsymbol{W}))_{1\le i,1\le j\le i}$ to be asymptotically trace-free. To fully specify the FOM, it remains to determine the matrix $\mathbf{D}_t$ which appears in the definition of $(\bar{\boldsymbol{u}}_t)_{t\ge 1}$. In RI-AMP, this matrix is set to be $\mathbf{D}_t = \widehat{\boldsymbol{\Phi}}_t$, where $\widehat{\boldsymbol{\Phi}}_t$ collects the empirical divergences of the de-noisers; see (25). Lemma 3 below summarizes our new formulation of RI-AMP in terms of the intermediate variables $(\bar{\boldsymbol{u}}_t)_{t\ge 1}$.

**Lemma 3** (Choice of $\mathbf{D}_t$ and final RI-AMP algorithm)**.** *Let $(\boldsymbol{r}_t)_{t\ge 1}$ be the iterates generated by the RI-AMP algorithm. With slight abuse of notations, define $(\bar{\boldsymbol{u}}_t)_{t\ge 1}$ as in (33), but with $\mathbf{D}_t = \widehat{\boldsymbol{\Phi}}_t$. Namely, $\bar{\boldsymbol{u}}_1 := \boldsymbol{u}_1$ and*

$$\bar{\boldsymbol{u}}_t := \boldsymbol{u}_t - \left(\langle \partial_1 \boldsymbol{u}_t\rangle \cdot \boldsymbol{r}_1 + \cdots + \langle \partial_{t-1} \boldsymbol{u}_t\rangle \cdot \boldsymbol{r}_{t-1}\right), \quad \forall t > 1. \tag{44}$$

*Then, $(\boldsymbol{r}_t)_{t\ge 1}$ can be represented as*

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = \begin{bmatrix} \widehat{P}_{1,1}(\boldsymbol{W}) & & & \\ \widehat{P}_{2,1}(\boldsymbol{W}) & \widehat{P}_{2,2}(\boldsymbol{W}) & & \\ \vdots & \vdots & \ddots & \\ \widehat{P}_{t,1}(\boldsymbol{W}) & \widehat{P}_{t,1}(\boldsymbol{W}) & \cdots & \widehat{P}_{t,t}(\boldsymbol{W}) \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}, \tag{45}$$

*where $(\widehat{P}_{t,i})_{1\le t,1\le i\le t}$ are a sequence of polynomials. Let $\widehat{\mathbf{P}}_t(\lambda) \in \mathbb{R}^{t\times t}$ be the collection of these polynomials (cf. (35)). We have*

$$\widehat{\mathbf{P}}_t(\lambda) = \sum_{i=1}^{t} Q_i(\lambda) \widehat{\boldsymbol{\Phi}}_t^{i-1}, \quad \forall \lambda \in \mathbb{R}, \tag{46}$$

*where $(Q_i(\lambda))_{i\ge 1}$ are defined recursively*

$$Q_n(\lambda) = \lambda Q_{n-1}(\lambda) - \sum_{i=1}^{n} \mathbb{E}[\Lambda Q_{i-1}(\Lambda)] \cdot Q_{n-i}(\lambda), \quad \forall \lambda \in \mathbb{R}, n \ge 1, \tag{47}$$

*under the initialization $Q_0(\lambda) = 1, \forall \lambda \in \mathbb{R}$.*

*Proof.* Note that RI-AMP is an instance of FOM with $\mathbf{B}_t = \sum_{i=1}^{t} \kappa_i \widehat{\boldsymbol{\Phi}}_t^{i-1}$. Then, the claimed result is a consequence of Lemma 1 and Lemma 2. However, we emphasize one subtle point here. Note that we have assumed $\mathbf{B}_t$ and $\mathbf{D}_t$ to be deterministic in Lemma 2, based on which $\mathbb{E}_{\Lambda\sim\mu}[\mathbf{P}_t(\Lambda)]$ was calculated. On the other hand, when $\mathbf{D}_t = \widehat{\boldsymbol{\Phi}}_t$, which contains the empirical divergences of the de-noisers, the matrix $\mathbf{D}_t$ is no longer deterministic and correlated with $\boldsymbol{W}$. Nevertheless, Lemma 3 still holds. To see this, note that by Lemma 1, (45) holds with

$$\widehat{\mathbf{P}}_t(\lambda) = (\mathbf{I}_t - \lambda \widehat{\boldsymbol{\Phi}}_t + \mathbf{B}_t \widehat{\boldsymbol{\Phi}}_t)^{-1} (\lambda \mathbf{I}_t - \mathbf{B}_t) \tag{48a}$$

$$= \sum_{i=1}^{t} \left(\lambda \widehat{\boldsymbol{\Phi}}_t - \mathbf{B}_t \widehat{\boldsymbol{\Phi}}_t\right)^{i-1} (\lambda \mathbf{I}_t - \mathbf{B}_t), \tag{48b}$$

where the second step is due to the following identity for a strictly lower triangular matrix $\mathbf{A} \in \mathbb{R}^{t\times t}$: $(\mathbf{I}_t - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{t-1}$. Now, $\mathbf{B}_t = \sum_{i=1}^{t} \kappa_i \widehat{\boldsymbol{\Phi}}_t^{i-1}$, we have

$$\mathbf{B}_t = \sum_{i=1}^{t} \kappa_i \widehat{\boldsymbol{\Phi}}_t^{i-1} = \sum_{i=1}^{t} \mathbb{E}[\Lambda Q_{i-1}(\Lambda)] \cdot \widehat{\boldsymbol{\Phi}}_t^{i-1}, \tag{49}$$

where $(Q_i(\lambda))_{i\ge 1}$ are defined recursively in (47). The second equality in the above equation is due to the recursive characterization of free cumulants in Lemma 1. From (48) and (49), $\widehat{\mathbf{P}}_t(\lambda)$ is a polynomial of $\widehat{\boldsymbol{\Phi}}_t$. Applying similar manipulations as the proof of Lemma 2-(2) yields the desired expression of $\widehat{\boldsymbol{\Phi}}_t$ (70b). □

*Remark* 9 (RI-AMP with univariate denoisers). When the de-noiser $\eta_{t+1}(\cdot)$ only depends on $\boldsymbol{r}_t$, the above representation of RI-AMP can be further simplified. In particular, if $\langle \partial_t \boldsymbol{u}_{t+1} \rangle = 1, \forall t \geq 1$, the matrix $\widehat{\mathbf{P}}_t(\lambda)$ becomes (cf. (70b))

$$\widehat{\mathbf{P}}_t(\lambda) = \sum_{i=1}^{t} Q_i(\lambda) \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 0 & 1 & 0 & & \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}^{i-1} = \begin{bmatrix} Q_1(\lambda) & & & & \\ Q_2(\lambda) & Q_1(\lambda) & & & \\ Q_3(\lambda) & Q_2(\lambda) & Q_1(\lambda) & & \\ \vdots & \vdots & \ddots & & \\ Q_t(\lambda) & Q_{t-1}(\lambda) & \cdots & Q_2(\lambda) & Q_1(\lambda) \end{bmatrix}.$$

Hence, the representation of $(\boldsymbol{r}_t)_{t\geq 1}$ in (45) recovers (32) which we introduced in Section 4.2.

Notice that the representation of the iterates $(\boldsymbol{r}_t)_{t\geq 1}$ of RI-AMP in (45) is *deterministic*. At this point, we have rewritten the RI-AMP algorithm as a OAMP algorithm. The asymptotic Gaussian distribution of $(\boldsymbol{r}_t)_{t\geq 1}$ then follows from the state evolution results of OAMP. Theorem 2 below summarizes the state evolution of RI-AMP, which reproduces the state evolution result of RI-AMP [20], but presented differently.

**Theorem 2** (State evolution of RI-AMP). *Let $(\boldsymbol{r}_t)_{t\geq 1}$ be generated by the RI-AMP algorithm. Suppose that Assumption 2 holds. Then, the following holds as $N \to \infty$:*

$$(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t) \xrightarrow{W_2} (\mathsf{R}_1, \ldots, \mathsf{R}_t) \sim \mathcal{N}(\mathbf{0}_t, \boldsymbol{\Sigma}_t), \quad \forall t \geq 1, \tag{51a}$$

*where*

$$\boldsymbol{\Sigma}_t = \mathbb{E}_{\Lambda\sim\mu}\left[\mathbf{P}_t(\Lambda)\, \bar{\boldsymbol{\Delta}}_t\, \mathbf{P}_t(\Lambda)^{\mathsf{T}}\right], \tag{51b}$$

*with*

$$\mathbf{P}_t(\Lambda) := \sum_{i=1}^{t} Q_i(\Lambda) \boldsymbol{\Phi}_t^{i-1}. \tag{51c}$$

*In the above equations,*

- *The sequence of polynomials $(Q_i)_{i\in\mathbb{N}}$ are defined in (41).*
- *$\boldsymbol{\Phi}_t$ is defined as in (25) but with the empirical divergence $\langle \partial_i \boldsymbol{u}_j \rangle := \langle \partial_i \eta_j(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{j-1}) \rangle$ replaced by $\mathbb{E}\left[\partial_i \eta_j(\mathsf{R}_1, \ldots, \mathsf{R}_{j-1})\right]$, for all $j \in [t]$ and $i \in [j-1]$.*
- *$\bar{\boldsymbol{\Delta}}_t$ is defined as*

$$\bar{\boldsymbol{\Delta}}_t := \begin{bmatrix} \mathbb{E}\left[\bar{\mathsf{U}}_1\bar{\mathsf{U}}_1\right] & \mathbb{E}\left[\bar{\mathsf{U}}_1\bar{\mathsf{U}}_2\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{U}}_1\bar{\mathsf{U}}_t\right] \\ \mathbb{E}\left[\bar{\mathsf{U}}_2\bar{\mathsf{U}}_1\right] & \mathbb{E}\left[\bar{\mathsf{U}}_2^2\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{U}}_2\bar{\mathsf{U}}_t\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\bar{\mathsf{U}}_t\bar{\mathsf{U}}_1\right] & \mathbb{E}\left[\bar{\mathsf{U}}_t\bar{\mathsf{U}}_2\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{U}}_t\bar{\mathsf{U}}_t\right] \end{bmatrix}, \tag{51d}$$

  *where $\bar{\mathsf{U}}_j := \eta_j(\mathsf{R}_1, \ldots, \mathsf{R}_{j-1}) - \sum_{i=1}^{j-1} \mathbb{E}\left[\partial_i \eta_j(\mathsf{R}_1, \ldots, \mathsf{R}_{j-1})\right] \cdot \mathsf{R}_i$, $\forall j \geq 2$.*

*Proof.* See Appendix C.3. □

The state evolution of RI-AMP in Theorem 2 is presented in a way that is most natural from the perspective of OAMP, and is in a different format as that given in [20, Section 4]. The following proposition shows that these two formulations are equivalent.

**Proposition 3** (Consistency with the state evolution in [20]). *The covariance $\boldsymbol{\Sigma}_t$ (51b) can be equivalently written as*

$$\boldsymbol{\Sigma}_t = \sum_{j=0}^{\infty} \sum_{i=0}^{j} \kappa_{j+2} \boldsymbol{\Phi}_t^i \boldsymbol{\Delta}_t \left((\boldsymbol{\Phi}_t)^{j-i}\right)^{\mathsf{T}}, \tag{52a}$$

*where*

$$\boldsymbol{\Delta}_t := \begin{bmatrix} \mathbb{E}\left[\mathsf{U}_1\mathsf{U}_1\right] & \mathbb{E}\left[\mathsf{U}_1\mathsf{U}_2\right] & \cdots & \mathbb{E}\left[\mathsf{U}_1\mathsf{U}_t\right] \\ \mathbb{E}\left[\mathsf{U}_2\mathsf{U}_1\right] & \mathbb{E}\left[\mathsf{U}_2^2\right] & \cdots & \mathbb{E}\left[\mathsf{U}_2\mathsf{U}_t\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\mathsf{U}_t\mathsf{U}_1\right] & \mathbb{E}\left[\mathsf{U}_t\mathsf{U}_2\right] & \cdots & \mathbb{E}\left[\mathsf{U}_t\mathsf{U}_t\right] \end{bmatrix} \tag{52b}$$

*with* $\mathsf{U}_j := \eta_j(\mathsf{R}_1, \ldots, \mathsf{R}_{j-1})$*, and* $(\mathsf{R}_1, \ldots, \mathsf{R}_t)$ *are the state evolution random variables described in* (51a)*.*

*Proof.* See Appendix C.4. □

# 5 Generalizations and Applications

In the preceding section, we demonstrated how our approach can be employed to derive an existing AMP algorithm. In this section, we showcase the versatility of our approach by devising two novel AMP variants. The first variant employs a different form of Onsager term. The second variant allows an additional matrix denoising step, inspired by the recent line of work [2, 3, 17]. It is possible to produce numerous other variants of AMP, but we do not pursue it further in this paper. Finally, we discuss the applications in spiked models.

## 5.1 Generalization of RI-AMP: Different Form of Onsager Terms

In the RI-AMP algorithm, the Onsager term $\mathsf{b}_{t,1}\boldsymbol{u}_1 + \cdots + \mathsf{b}_{t,t}\boldsymbol{u}_t$ is employed to cancel out the non-Gaussian component within $\boldsymbol{W}\boldsymbol{u}_t$. Our derivation of RI-AMP in Section 4 indicates that it is possible to achieve the same goal by using an Onsager term that is a linear combination of $(\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_t)$. Motivated by this insight, we introduce a new variant of RI-AMP, coined RI-AMP-DF.

**Definition 5** (A variant of RI-AMP: RI-AMP-DF)**.** Starting from an initialization $\boldsymbol{u}_1 = \bar{\boldsymbol{u}}_1 \in \mathbb{R}^N$, RI-AMP-DF proceeds as

$$\boldsymbol{r}_t = \boldsymbol{W}\boldsymbol{u}_t - \left(\mathsf{c}_{t,1}\bar{\boldsymbol{u}}_1 + \mathsf{c}_{t,2}\bar{\boldsymbol{u}}_2 + \cdots + \mathsf{c}_{t,t}\bar{\boldsymbol{u}}_t\right), \quad \forall t \geq 1, \tag{53a}$$

$$\boldsymbol{u}_{t+1} = \eta_{t+1}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t), \tag{53b}$$

$$\bar{\boldsymbol{u}}_{t+1} = \eta_{t+1}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t) - \left(\langle \partial_1 \boldsymbol{u}_{t+1} \rangle \cdot \boldsymbol{r}_1 + \cdots + \langle \partial_t \boldsymbol{u}_{t+1} \rangle \cdot \boldsymbol{r}_t\right), \tag{53c}$$

where $\eta_{t+1} : \mathbb{R}^t \mapsto \mathbb{R}$ acts separately on the $N$ rows of $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t) \in \mathbb{R}^{N \times t}$. The de-biasing coefficients are set to

$$\mathbf{C}_t := \begin{bmatrix} \mathsf{c}_{1,1} & & & \\ \mathsf{c}_{2,1} & \mathsf{c}_{2,2} & & \\ \vdots & \vdots & \ddots & \\ \mathsf{c}_{t,1} & \mathsf{c}_{t,2} & \cdots & \mathsf{c}_{t,t} \end{bmatrix} = \sum_{i=1}^{t} \gamma_i (\widehat{\boldsymbol{\Phi}}_t)^{i-1}, \tag{54}$$

where $\widehat{\boldsymbol{\Phi}}_t$ is defined in (25) and the sequence $(\gamma_n)_{n \in \mathbb{N}}$ are defined as

$$\gamma_n = \mathbb{E}\left[\mathsf{\Lambda} H_{n-1}(\mathsf{\Lambda})\right], \quad \forall n \geq 1, \quad \mathsf{\Lambda} \sim \mu, \tag{55}$$

with $(H_i(\lambda))_{i \geq 1}$ a sequence of polynomials satisfying the recurrence ($H_0(\lambda) := 1$):

$$H_n(\lambda) = \lambda H_{n-1}(\lambda) - \mathbb{E}\left[\mathsf{\Lambda} H_{n-1}(\mathsf{\Lambda})\right], \quad \forall n \geq 1, \quad \lambda \in \mathbb{R}. \tag{56}$$

The de-biasing matrix $\mathbf{C}_t$ (54) is derived using the approach detailed in Section 4.3. Note that both the de-biasing matrix $\mathbf{B}_t$ (24) for RI-AMP and the de-biasing matrix $\mathbf{C}_t$ for RI-AMP-DF are polynomials of the divergence matrix $\widehat{\boldsymbol{\Phi}}_t$. Moreover, the coefficients in both polynomial representations admit recursive characterization; see (41) and (56).

As before, we can reduce RI-AMP-DF to a certain OAMP algorithm. The following theorem summarizes the reduction result.

**Theorem 3** (Reduction of RI-AMP-DF to OAMP and state evolution)**.** *Let* $(\boldsymbol{r}_t)_{t \geq 1}$ *and* $(\bar{\boldsymbol{u}}_t)_{t \geq 1}$ *be generated via the RI-AMP-DF algorithm. Then, the following statements hold.*

*(1) For all $t \geq 1$,*

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = \begin{bmatrix} \widehat{G}_{1,1}(\boldsymbol{W}) & & & \\ \widehat{G}_{2,1}(\boldsymbol{W}) & \widehat{G}_{2,2}(\boldsymbol{W}) & & \\ \vdots & \vdots & \ddots & \\ \widehat{G}_{t,1}(\boldsymbol{W}) & \widehat{G}_{t,2}(\boldsymbol{W}) & \cdots & \widehat{G}_{t,t}(\boldsymbol{W}) \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}, \tag{57a}$$

*Let $\widehat{\mathbf{G}}_t(\lambda)$ be the matrix that collects the polynomials $(\widehat{G}_{i,j})_{1\leq i\leq t, 1\leq j\leq i}$ (cf. (35)). Then, $\widehat{\mathbf{G}}_t(\lambda)$ admits the following representation:*

$$\widehat{\mathbf{G}}_t(\lambda) = \sum_{n=1}^{t} H_n(\lambda) \widehat{\boldsymbol{\Phi}}_t^{n-1}, \quad \forall \lambda \in \mathbb{R}, \tag{58}$$

*where $(H_n(\lambda))_{n\geq 1}$ are defined in (56).*

*(2) Suppose that Assumption 2 holds. As $N \to \infty$,*

$$(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t) \xrightarrow{W_2} (\mathsf{R}_1, \ldots, \mathsf{R}_t) \sim \mathcal{N}(\mathbf{0}_t, \boldsymbol{\Delta}_t), \quad \forall t \geq 1, \tag{59a}$$

*where*

$$\boldsymbol{\Delta}_t = \mathbb{E}_{\Lambda\sim\mu}\left[\mathbf{G}_t(\Lambda)\, \bar{\boldsymbol{\Delta}}_t\, \mathbf{G}_t(\Lambda)^{\mathsf{T}}\right] \tag{59b}$$

*with*

$$\mathbf{G}_t(\Lambda) := \sum_{n=1}^{t} H_n(\Lambda) \boldsymbol{\Phi}_t^{n-1}. \tag{59c}$$

*In the above equation, $\bar{\boldsymbol{\Delta}}_t$ and $\boldsymbol{\Phi}_t$ are defined similarly as in Theorem 2.*

*Proof.* See Appendix D.1. □

*Remark* 10 (RI-AMP-DF can implement any GFOM). The Onsager terms of RI-AMP-DF and the original RI-AMP algorithm are slightly different. However, these two algorithms are essentially equivalent up to a proper change of variables. In fact, both RI-AMP-DF and RI-AMP are equivalent to a broader class algorithms, namely, the generalized first order methods (GFOM) introduced in [10, 39]. More details about this reduction can be found in Appendix D.2.

## 5.2 Generalization of RI-AMP: Matrix Processing

Motivated by [2], we introduce another variant of RI-AMP that applies a nonlinear processing function on the matrix $\boldsymbol{W}$ in each iteration. In the following, we call this algorithm RI-AMP-MP. It is shown in [2] that adding the matrix processing operation could lead to better signal estimation for spiked models.

**Definition 6** (A variant of RI-AMP: RI-AMP-MP)**.** Starting from an initialization $\boldsymbol{u}_1 \in \mathbb{R}^N$, RI-AMP-MP proceeds as follows

$$\boldsymbol{r}_t = f_t(\boldsymbol{W})\boldsymbol{u}_t - \left(\mathsf{e}_{t,1}\boldsymbol{u}_1 + \mathsf{e}_{t,2}\boldsymbol{u}_2 + \cdots + \mathsf{e}_{t,t}\boldsymbol{u}_t\right), \tag{60a}$$

$$\boldsymbol{u}_{t+1} = \eta_t(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t), \tag{60b}$$

where $f_t : \mathbb{R} \mapsto \mathbb{R}$ is continuous and applied to the eigenvalues of $\boldsymbol{W}$ with the eigenvectors unchanged. Let $\mathbf{E}_t \in \mathbb{R}^{t\times t}$ be a matrix collecting the de-biasing coefficients:

$$\mathbf{E}_t := \begin{bmatrix} \mathsf{e}_{1,1} & & & \\ \mathsf{e}_{2,1} & \mathsf{e}_{2,2} & & \\ \vdots & \vdots & \ddots & \\ \mathsf{e}_{t,1} & \mathsf{e}_{t,2} & \cdots & \mathsf{e}_{t,t} \end{bmatrix}. \tag{61}$$

The de-biasing matrix $\mathbf{E}_t$ will be set according to Lemma 4 below.

We make two remarks about the RI-AMP-MP algorithm.

(1) If the matrix processing function $(f_t)_{t\geq 1}$ are identical across different iterations, then RI-AMP-MP effectively reduces to an RI-AMP with the original rotationally-invariant matrix replaced by $\widehat{\boldsymbol{W}} := f(\boldsymbol{W})$. Employing an iteration-dependent matrix processing matrix complicates the derivations of the algorithm.

(2) In [2], the matrix processing functions $(f_t)_{t\geq 1}$ are assumed to be polynomials, and the Onsager terms and the associated state evolution are derived by mapping the corresponding algorithm to certain RI-AMP algorithm together with re-indexing. Here, we consider general continuous matrix processing functions.

The following lemma summarizes the reduction of RI-AMP-MP to OAMP. Its proof is similar to that of Lemma 1 and Lemma 2, and thus omitted.

**Lemma 4** (Reduction of RI-AMP-MP to OAMP)**.** *Let $(\boldsymbol{r}_t)_{t\geq 1}$ be generated via the RI-AMP-MP algorithm and let $(\bar{\boldsymbol{u}}_t)_{t\geq 1}$ be defined as in (44). Then, the following holds for all $t \geq 1$:*

*(1) For any fixed $\mathsf{E}_t$, we have*

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = \begin{bmatrix} \widehat{J}_{1,1}(\boldsymbol{W}) & & & \\ \widehat{J}_{2,1}(\boldsymbol{W}) & \widehat{J}_{2,2}(\boldsymbol{W}) & & \\ \vdots & \vdots & \ddots & \\ \widehat{J}_{t,1}(\boldsymbol{W}) & \widehat{J}_{t,2}(\boldsymbol{W}) & \cdots & \widehat{J}_{t,t}(\boldsymbol{W}) \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}, \tag{62a}$$

*where $(\widehat{J}_{i,j})_{1\leq i,1\leq j\leq i}$ is a sequence of functions. Let $\widehat{\mathbf{J}}_t : \mathbb{R} \mapsto \mathbb{R}^{t\times t}$ be a matrix representation of these functions (similar to (35)). Then, $\widehat{\mathbf{J}}_t(\lambda)$ can be written into the following explicit form:*

$$\widehat{\mathbf{J}}(\lambda) := \left(\mathbf{I}_t - \mathrm{diag}\{f_1(\lambda), \ldots, f_t(\lambda)\}\widehat{\boldsymbol{\Phi}}_t + \mathsf{E}_t\widehat{\boldsymbol{\Phi}}_t\right)^{-1} \left(\mathrm{diag}\{f_1(\lambda), \ldots, f_t(\lambda)\} - \mathsf{E}_t\right), \quad \forall \lambda \in \mathbb{R}. \tag{62b}$$

*(2) For any $\widehat{\boldsymbol{\Phi}}_t \in \mathbb{R}^{t\times t}$, the following equation has a unique solution in $\mathsf{E}_t$:*

$$\mathbb{E}\left[\widehat{\mathbf{J}}(\mathsf{\Lambda})\right] = \mathbf{0}_{t\times t}, \tag{63}$$

*where the expectation is taken w.r.t. $\mathsf{\Lambda} \sim \mu$, with $\mathsf{\Lambda} \sim \mu$ independent of $\widehat{\boldsymbol{\Phi}}_t$.*

Some comments are in order:

(1) It seems that the de-biasing matrix $\mathsf{E}_t$ in RI-AMP-MP, determined by (63), cannot be expressed as a simple polynomial of $\widehat{\boldsymbol{\Phi}}_t$. While we can solve (63) for $\mathsf{E}_t$ recursively (namely, row-by-row), as in the proof of Lemma 2-(1), the non-commutativity of $\mathrm{diag}\{f_1(\lambda), \ldots, f_t(\lambda)\}$ and $\widehat{\boldsymbol{\Phi}}_t$ precludes a simple polynomial form.

(2) Similar to RI-AMP-DF, we may use an Onsager term that is a linear combination of the divergence-free estimates. In this case, the de-biasing matrix can be expressed explicitly, but still cannot be be written as a simple polynomial of $\widehat{\boldsymbol{\Phi}}_t$.

(3) The state evolution equations of RI-AMP-MP can be readily derived from the representation (62). We omit the details here.

### 5.3 Application: Spiked Matrix Model

Until now, we have focused on the derivation of general AMP iterations. In this section, we illustrate the application of our approach to one concrete signal estimation problem, namely, spiked models. AMP algorithms have been instrumental to the theoretical underpinnings of spiked models [24–26, 37, 44, 45].

We shall show that our approach can conveniently recover and generalize existing AMP algorithms in the context of spiked models.

Consider a *spiked matrix model* where the observation matrix $\boldsymbol{Y} \in \mathbb{R}^{N \times N}$ reads:

$$\boldsymbol{Y} = \frac{\theta}{N} \boldsymbol{x}_\star \boldsymbol{x}_\star^\mathsf{T} + \boldsymbol{W}. \tag{64}$$

In the above equation, $\boldsymbol{x}_\star \in \mathbb{R}^N$ is the signal vector to be estimated, $\theta > 0$ is a parameter that dictates the SNR, and $\boldsymbol{W}$ is a noise matrix which we assume to be symmetric and rotationally-invariant. We first recall the definitions of the OAMP algorithm introduced in [17] for spiked models. Note that compared with the OAMP algorithms in (15), the only difference is that the matrix $\boldsymbol{W}$ in (15) is replaced by the observation matrix $\boldsymbol{Y}$.

**Definition 7** (OAMP algorithm for spiked models [17])**.** Starting from an initialization $\bar{\boldsymbol{x}}_1 \in \mathbb{R}^N$, an orthogonal AMP algorithm for the spiked model (64) proceeds as follows:

$$\boldsymbol{x}_t = \left( f_t(\boldsymbol{Y}) - \frac{\mathrm{tr}\left(f_t(\boldsymbol{Y})\right)}{N} \cdot \boldsymbol{I}_N \right) \bar{\boldsymbol{x}}_t, \tag{65a}$$

$$\bar{\boldsymbol{x}}_{t+1} = g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) - \sum_{i=1}^{t} \langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) \rangle \cdot \boldsymbol{x}_i, \tag{65b}$$

where $f_t : \mathbb{R} \mapsto \mathbb{R}$ and $g_{t+1} : \mathbb{R}^t \times \mathbb{R}^k \mapsto \mathbb{R}$ are defined similarly as in Definition 1.

The above definition of OAMP is slightly different from that in [17] in the sense that the empirical estimates $\mathrm{tr}\left(f_t(\boldsymbol{Y})\right)/N$ and $\langle \partial_i g_{t+1}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t; \boldsymbol{a}) \rangle$ are involved in (65), whereas the OAMP in [17] uses their corresponding large-$N$ limits. Despite this difference, the state evolution result in [17] still applies and we record the results in the following lemma. (This can be shown following the same arguments as in the proof of Theorem 1, and we omit the details.) We also refer the reader to [17, Section 4.1] for a heuristic derivation of the state evolution equations.

**Lemma 5** (State evolution of OAMP for spiked models [17])**.** *Let $(\boldsymbol{x}_t)_{t \in \mathbb{N}}$ be generated via (65). Suppose Assumption 1 holds. Assume additionally that $(\boldsymbol{x}_\star; \boldsymbol{a}) \xrightarrow{W_2} (\mathsf{X}_\star; \mathsf{A}) \sim \pi$ where $\mathbb{E}[\mathsf{X}_\star^2] = 1$ and $\mathbb{E}[\|\mathsf{A}\|^2] < \infty$. Then, the following holds for all $t \geq 1$:*

$$(\boldsymbol{x}_\star, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a}) \xrightarrow{W_2} (\mathsf{X}_\star, \mathsf{X}_1, \ldots, \mathsf{X}_t; \mathsf{A}), \tag{66a}$$

*where $(\mathsf{X}_\star, \mathsf{A}) \sim \pi$ and*

$$[\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_t]^\mathsf{T} = \boldsymbol{\beta}_t \mathsf{X}_\star + [\mathsf{Z}_1, \mathsf{Z}_2, \ldots, \mathsf{Z}_t]^\mathsf{T}, \tag{66b}$$

*where $[\mathsf{Z}_1, \ldots, \mathsf{Z}_t] \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_{t,1} + \boldsymbol{\Omega}_{t,2})$ is independent of $\mathsf{X}_\star$, and $(\boldsymbol{\beta}_t, \boldsymbol{\Omega}_{t,1}, \boldsymbol{\Omega}_{t,2})$ are given by*

$$\boldsymbol{\beta}_t = \begin{bmatrix} \mathbb{E}\left[\bar{f}_1(\Lambda_\nu)\right] \\ \vdots \\ \mathbb{E}\left[\bar{f}_t(\Lambda_\nu)\right] \end{bmatrix} \circ \boldsymbol{\alpha}_t, \tag{66c}$$

$$\boldsymbol{\Omega}_{t,1} = \begin{bmatrix} \mathrm{Cov}_\nu\left[\bar{f}_1, \bar{f}_1\right] & \cdots & \mathrm{Cov}_\nu\left[\bar{f}_1, \bar{f}_t\right] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}_\nu\left[\bar{f}_t, \bar{f}_1\right] & \cdots & \mathrm{Cov}_\nu\left[\bar{f}_t, \bar{f}_t\right] \end{bmatrix} \circ (\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\mathsf{T}), \tag{66d}$$

$$\boldsymbol{\Omega}_{t,2} = \begin{bmatrix} \mathrm{Cov}_\mu\left[\bar{f}_1, \bar{f}_1\right] & \cdots & \mathrm{Cov}_\mu\left[\bar{f}_1, \bar{f}_t\right] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}_\mu\left[\bar{f}_t, \bar{f}_1\right] & \cdots & \mathrm{Cov}_\mu\left[\bar{f}_t, \bar{f}_t\right] \end{bmatrix} \circ (\bar{\boldsymbol{\Delta}}_t - \boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\mathsf{T}), \tag{66e}$$

*where $\bar{f}_i(\lambda) := f_i(\lambda) - \mathbb{E}_{\Lambda \sim \mu}[\Lambda], \forall \lambda \in \mathbb{R}, i \in [t]$ and $\mathrm{Cov}_\mu\left[\bar{f}_i, \bar{f}_j\right]$ is defined as in Theorem 1, and $(\boldsymbol{\alpha}_t, \bar{\boldsymbol{\Delta}}_t)$ are defined as*

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \mathbb{E}\left[\mathsf{X}_\star \bar{\mathsf{X}}_1\right] \\ \vdots \\ \mathbb{E}\left[\mathsf{X}_\star \bar{\mathsf{X}}_t\right] \end{bmatrix} \quad \text{and} \quad \bar{\boldsymbol{\Delta}}_t = \begin{bmatrix} \mathbb{E}\left[\bar{\mathsf{X}}_1 \bar{\mathsf{X}}_1\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{X}}_1 \bar{\mathsf{X}}_t\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[\bar{\mathsf{X}}_t \bar{\mathsf{X}}_1\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{X}}_t \bar{\mathsf{X}}_t\right] \end{bmatrix}. \tag{66f}$$

*In the above equations, $\Lambda_\nu \sim \nu$ where $\nu$ is a probability measure whose Stieltjes transform (defined as $m_\nu(z) := \int_{\mathbb{R}} \frac{\nu(\mathrm{d}\lambda)}{z-\lambda}, \forall z \in \mathbb{C}\backslash\mathbb{R}$) is given by:*

$$m_\nu(z) = \frac{m_\mu(z)}{1-\theta m_\mu(z)}, \quad \forall z \in \mathbb{C}\backslash\mathbb{R}. \tag{66g}$$

*Further,*

$$\bar{\mathsf{X}}_t := g_t(\mathsf{X}_1, \ldots, \mathsf{X}_{t-1}; \mathsf{A}) - \sum_{i=1}^{t-1} \mathbb{E}\left[\partial_i g_t(\mathsf{X}_1, \ldots, \mathsf{X}_{t-1}; \mathsf{A})\right] \cdot \mathsf{X}_i, \quad \forall t \geq 2. \tag{66h}$$

*Remark* 11 (On the measure $\nu$). The state evolution equations in the above theorem involve a probability measure $\nu$. As shown in [17], the measure $\nu$ is the large-$N$ limit of the following empirical measure:

$$\nu_N := \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{x}_\star^{\mathsf{T}} \boldsymbol{u}_i(\boldsymbol{Y})\right)^2 \cdot \delta_{\lambda_i(\boldsymbol{Y})}, \tag{67}$$

where $(\lambda_i(\boldsymbol{Y}))_{1\leq i\leq N}$ and $(\boldsymbol{u}_i(\boldsymbol{Y}))_{1\leq i\leq N}$ denote the eigenvalues and eigenvectors of $\boldsymbol{Y}$ respectively. For more explicit characterization of $\nu$ by Stieltjes inversion see [17, Appendix A.1].

We can now apply our approach to derive AMP algorithms by reduction to certain OAMP algorithms. Here, we consider an application of the RI-AMP-MP algorithm (see Definition 6) to signal estimation in spiked models. We consider a variant of RI-AMP-MP where the matrix processing function does not change across iterations.

**Definition 8** (RI-AMP-MP algorithm for spiked models)**.** Starting from an initialization $\boldsymbol{u}_1 \in \mathbb{R}^N$, RI-AMP-MP generates a sequence of iterates via

$$\boldsymbol{r}_t = f(\boldsymbol{Y})\boldsymbol{u}_t - \left(\mathsf{e}_{t,1}\boldsymbol{u}_1 + \mathsf{e}_{t,2}\boldsymbol{u}_2 + \cdots + \mathsf{e}_{t,t}\boldsymbol{u}_t\right), \tag{68a}$$

$$\boldsymbol{u}_{t+1} = \eta_t(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t), \tag{68b}$$

where $f : \mathbb{R} \mapsto \mathbb{R}$ is a continuous function that does not change across iterations, and the de-biasing matrix $\mathbf{E}_t \in \mathbb{R}^{t\times t}$ is given by

$$\mathbf{E}_t = \sum_{i=1}^{t} \widetilde{\alpha}_i (\widehat{\boldsymbol{\Phi}}_t)^{i-1}, \tag{69}$$

where $\widehat{\boldsymbol{\Phi}}_t$ is defined as in (25) and $(\widetilde{\alpha}_i)_{i\in\mathbb{N}}$ are the free cumulants of the random variable $f(\Lambda)$, $\Lambda \sim \mu$.

*Remark* 12 (Connection with [2]). The above algorithm is essentially the "Bayes-optimal AMP" (BAMP) algorithm proposed in [2]. It generalizes BAMP in the sense that the matrix processing function $f(\cdot)$ can be non-polynomial.

Again, we can use the same technique to reduce the above RI-AMP-MP algorithm to an OAMP algorithm, and then apply Lemma 5 to derive a state evolution result equation for RI-AMP-MP. Our results are summarized in the following theorem.

**Theorem 4** (State evolution of RI-AMP-MP for spiked models)**.** *Let $(\boldsymbol{r}_t)_{t\geq 1}$ be generated via (68) and let $(\bar{\boldsymbol{u}}_t)_{t\geq 1}$ be defined as in (44). Then, the following statements hold.*

*(1) For all $t \geq 1$:*

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = \begin{bmatrix} \widehat{J}_{1,1}(\boldsymbol{Y}) & & & \\ \widehat{J}_{2,1}(\boldsymbol{Y}) & \widehat{J}_{2,2}(\boldsymbol{Y}) & & \\ \vdots & \vdots & \ddots & \\ \widehat{J}_{t,1}(\boldsymbol{Y}) & \widehat{J}_{t,2}(\boldsymbol{Y}) & \cdots & \widehat{J}_{t,t}(\boldsymbol{Y}) \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}. \tag{70a}$$

*Let* $\widehat{\mathbf{J}}_t : \mathbb{R} \mapsto \mathbb{R}^{t\times t}$ *be a matrix representation of the functions* $(\widehat{J}_{i,j})_{1\le i\le t, 1\le j\le i}$. *Then,* $\widehat{\mathbf{J}}_t(\lambda)$ *can be written into the following explicit form:*

$$\widehat{\mathbf{J}}_t(\lambda) = \sum_{i=1}^{t} K_i(\lambda)\widehat{\boldsymbol{\Phi}}_t^{i-1}, \quad \forall \lambda \in \mathbb{R}, \tag{70b}$$

*where the sequence of functions* $(K_n)_{n\ge 1}$ *is specified by the recurrence (with* $K_0(\lambda) := 1, \forall \lambda \in \mathbb{R}$*)*

$$K_n(\lambda) = f(\lambda)K_{n-1}(\lambda) - \sum_{i=1}^{n} \mathbb{E}_{\Lambda\sim\mu}\left[f(\Lambda)K_{i-1}(\Lambda)\right]\cdot K_{n-i}(\lambda), \quad \forall \lambda \in \mathbb{R},\ n \ge 1. \tag{70c}$$

*(2) Suppose that Assumption 2 holds. Further, assume that* $\boldsymbol{x}_\star \xrightarrow{W_2} \mathsf{X}_\star$ *where* $\mathbb{E}[\mathsf{X}_\star^2] = 1$. *As* $N \to \infty$*:*

$$(\boldsymbol{x}_\star, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_t) \xrightarrow{W_2} (\mathsf{X}_\star, \mathsf{R}_1, \ldots, \mathsf{R}_t), \quad \forall t \ge 1, \tag{71a}$$

*where*

$$[\mathsf{R}_1, \mathsf{R}_2, \ldots, \mathsf{R}_t]^{\mathsf{T}} = \boldsymbol{\beta}_t \mathsf{X}_\star + [\mathsf{Z}_1, \mathsf{Z}_2, \ldots, \mathsf{Z}_t]^{\mathsf{T}}, \tag{71b}$$

*where* $[\mathsf{Z}_1, \ldots, \mathsf{Z}_t] \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{t,1} + \boldsymbol{\Sigma}_{t,2})$ *is independent of* $\mathsf{X}_\star$*, and* $(\boldsymbol{\beta}_t, \boldsymbol{\Sigma}_{t,1}, \boldsymbol{\Sigma}_{t,2})$ *are given by*

$$\boldsymbol{\beta}_t = \mathbb{E}_{\Lambda_\nu\sim\nu}\left[\mathbf{J}_t(\Lambda_\nu)\right]\boldsymbol{\alpha}_t, \tag{71c}$$

$$\boldsymbol{\Sigma}_{t,1} = \mathbb{E}_{\Lambda_\nu\sim\nu}\left[\mathbf{J}_t(\Lambda_\nu)\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^{\mathsf{T}}\mathbf{J}_t(\Lambda_\nu)^{\mathsf{T}}\right] - \boldsymbol{\beta}_t\boldsymbol{\beta}_t^{\mathsf{T}}, \tag{71d}$$

$$\boldsymbol{\Sigma}_{t,2} = \mathbb{E}_{\Lambda\sim\mu}\left[\mathbf{J}_t(\Lambda)(\bar{\boldsymbol{\Delta}}_t - \boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^{\mathsf{T}})\mathbf{J}_t(\Lambda)^{\mathsf{T}}\right], \tag{71e}$$

*where* $\mathbf{J}_t(\lambda) := \sum_{i=1}^{t} K_i(\lambda)\boldsymbol{\Phi}_t^{i-1}$, $\forall \lambda \in \mathbb{R}$*, the probability measure* $\nu$ *is defined in Lemma 5, and* $(\boldsymbol{\alpha}_t, \bar{\boldsymbol{\Delta}}_t)$ *are defined as*

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \mathbb{E}\left[\mathsf{X}_\star\bar{\mathsf{X}}_1\right] \\ \vdots \\ \mathbb{E}\left[\mathsf{X}_\star\bar{\mathsf{X}}_t\right] \end{bmatrix} \quad \text{and} \quad \bar{\boldsymbol{\Delta}}_t = \begin{bmatrix} \mathbb{E}\left[\bar{\mathsf{X}}_1\bar{\mathsf{X}}_1\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{X}}_1\bar{\mathsf{X}}_t\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[\bar{\mathsf{X}}_t\bar{\mathsf{X}}_1\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{X}}_t\bar{\mathsf{X}}_t\right] \end{bmatrix}. \tag{71f}$$

*Proof.* Claim (1) can be proved analogously to Lemma 3, with $\boldsymbol{W}$ replaced by $\boldsymbol{Y}$. Note that the reduction of RI-AMP to OAMP as presented in Lemma 3 is a deterministic result, and hence still holds with $\boldsymbol{W}$ replaced by $\boldsymbol{Y}$. The proof of Claim (2) is similar to that of Theorem 2, the only difference being that we now appeal to Lemma 5 (state evolution of OAMP for spiked models). □

# 6 Numerical Results

We conduct a few numerical experiments to demonstrate the performance of the RI-AMP-MP algorithm (Definition 6) and show the accuracy of its theoretical state evolution prediction. Note that RI-AMP-MP generalizes the BAMP algorithm proposed in [2] in the sense that the matrix denoising function $f(\cdot)$ can be non-polynomial. For polynomial $f(\cdot)$, the de-biasing coefficients in BAMP algorithm and the RI-AMP-MP algorithm are slightly different (the former is based on reduction to certain RI-AMP [20] with proper time re-indexing), but equivalent asymptotically.

Fig. 2 shows the empirical and theoretical mean square error (MSE) performances of RI-AMP-MP. We set $f(\cdot)$ using the function proposed in [2, 3]. The theoretical state evolution prediction is based on the formula given in Theorem 4. In sub-figure (a), the eigenvalue distribution $\mu$ corresponds to the *trace ensemble* [2] with pure quartic potential. Both BAMP and RI-AMP-MP use the single-variate MMSE denoiser. For this setup, the function $f(\cdot)$ is a polynomial and RI-AMP-MP is asymptotically equivalent to BAMP. Indeed, as shown in Fig. 2-(a), the performances of BAMP and RI-AMP-MP are very close and are close to state evolution prediction.

Fig. 2-(b) considers a setup where $f(\cdot)$ is non-polynomial and BAMP is not directly applicable. (In principle, it is possible to approximate $f(\cdot)$ using polynomials within BAMP, but the Onager term and the

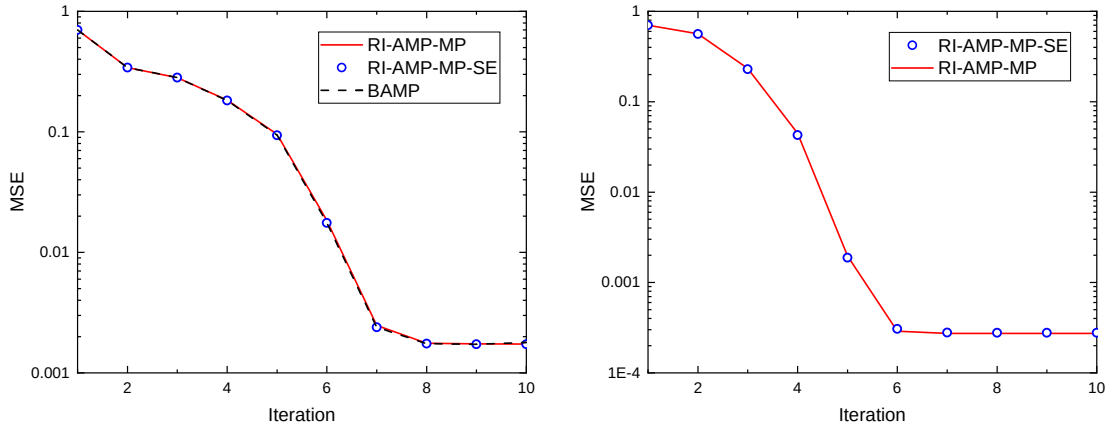
state evolution become somewhat cumbersome as the degree increases.) For experiment purposes, we set $\mu$ to the Marcenko–Pastur distribution to

$$\mu(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(a_+ - \lambda)(\lambda - a_-)}}{\alpha\lambda}, \quad a_+ := (1+\sqrt{\alpha})^2,\ a_- := (1-\sqrt{\alpha})^2, \tag{72}$$

where $\alpha$ is a parameter (which is set to $\alpha = 0.2$ in our experiement). Following [3], we set the matrix denoising function $f(\cdot)$ as

$$f(\lambda) = \frac{\theta}{\alpha}\left(1 + \frac{\alpha - 1}{\lambda}\right) - \frac{\theta^2}{\alpha\lambda}. \tag{73}$$

In the above equation, $\theta > 0$ denotes the SNR parameter for the spiked model (64). The signal denoisers $\eta_t(\cdot)$ are uses the strategy in [20, Remark 3.3], namely, optimal linear combining followed by the single-variate MMSE function. Fig. 2-(b) confirms that our theoretical state evolution prediction is still quite accurate for this setup.




(a) Pure quartic distribution.

(b) Marcenko–Pastur distribution.

Figure 2: MSE performance of the RI-AMP-MP algorithm and the BAMP algorithm [2]. In both experiments, we initialize the algorithms by $\boldsymbol{u}_1 = \sqrt{\omega}\boldsymbol{x}_\star + \sqrt{(1-\omega)}\boldsymbol{n}$ where $\boldsymbol{n}$ is standard Gaussian independent of $\boldsymbol{x}_\star$ and $\omega = 0.3$. The empirical results are average over 50 independent runs. $N = 5000$. The SNR parameter is $\theta = 3$ for figure (a) and $\theta = 1.3$ for figure (b).


## Acknowledgement

We are grateful to Rishabh Dudeja for many insightful conversations and comments that substantially impacted this work. We also thank Zhou Fan for helpful discussions. The second author would like to thank Arian Maleki for discussions that ignited his interest on this problem and Jizu Huang for discussions about the proof of Lemma 2.


## References


[1] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[2] Jean Barbier, Francesco Camilli, Marco Mondelli, and Manuel Sáenz. Fundamental limits in structured principal component analysis and how to reach them. *Proceedings of the National Academy of Sciences*, 120(30):e2302028120, 2023.

[3] Jean Barbier, Francesco Camilli, Marco Mondelli, and Yizhou Xu. Information limits and Thouless-Anderson-Palmer equations for spiked matrix models with structured noise. *arXiv preprint arXiv:2405.20993*, 2024.

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

[5] Mohsen Bayati and Andrea Montanari. The LASSO risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[6] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[7] Zhiqi Bu, Jason M Klusowski, Cynthia Rush, and Weijie J Su. Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Transactions on Information Theory*, 67 (1):506–537, 2020.

[8] Collin Cademartori and Cynthia Rush. A non-asymptotic analysis of generalized vector approximate message passing algorithms with rotationally invariant designs. *IEEE Transactions on Information Theory*, 2024.

[9] Burak Cakmak, Ole Winther, and Bernard H Fleury. S-amp: Approximate message passing for general matrix ensembles. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pages 192–196. IEEE, 2014.

[10] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR, 2020.

[11] Yiyao Cheng, Lei Liu, and Li Ping. Orthogonal AMP for massive access in channels with spatial and temporal correlations. *IEEE Journal on Selected Areas in Communications*, 39(3):726–740, 2020.

[12] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[13] David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

[14] David L Donoho, Adel Javanmard, and Andrea Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE transactions on information theory*, 59(11):7434–7464, 2013.

[15] Rishabh Dudeja and Milad Bakhshizadeh. Universality of linearized message passing for phase retrieval with structured sensing matrices. *IEEE Transactions on Information Theory*, 68(11):7545–7574, 2022.

[16] Rishabh Dudeja, Yue M. Lu, and Subhabrata Sen. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.

[17] Rishabh Dudeja, Songbin Liu, and Junjie Ma. Optimality of approximate message passing algorithms for spiked matrix models with rotationally invariant noise. *arXiv preprint arXiv:2405.18081*, 2024.

[18] Rishabh Dudeja, Subhabrata Sen, and Yue M Lu. Spectral universality in regularized linear regression with nearly deterministic sensing matrices. *IEEE Transactions on Information Theory*, 2024.

[19] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *The Annals of Probability*, 49(6):2922–2960, 2021.

[20] Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224, 2022.

[21] Oliver Y Feng, Ramji Venkataramanan, Cynthia Rush, Richard J Samworth, et al. A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15(4):335–536, 2022.

[22] Alyson K Fletcher, Sundeep Rangan, and Philip Schniter. Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1884–1888. IEEE, 2018.

[23] Yoshiyuki Kabashima. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111, 2003.

[24] Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in Bayes-optimal matrix factorization. *IEEE Transactions on information theory*, 62(7):4228–4265, 2016.

[25] Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022.

[26] Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization. *Proceedings of the National Academy of Sciences*, 120(31):e2302930120, 2023.

[27] Yufan Li and Pragya Sur. Spectrum-aware adjustment: A new debiasing framework with applications to principal components regression. *arXiv preprint arXiv:2309.07810*, 2023.

[28] Lei Liu, Shunqi Huang, and Brian M Kurkoski. Memory AMP. *IEEE Transactions on Information Theory*, 68(12):8015–8039, 2022.

[29] Junjie Ma and Li Ping. Orthogonal AMP. *IEEE Access*, 5:2020–2033, 2017.

[30] Junjie Ma, Xiaojun Yuan, and Li Ping. Turbo compressed sensing with partial DFT sensing matrix. *IEEE Signal Processing Letters*, 22(2):158–161, 2014.

[31] Junjie Ma, Rishabh Dudeja, Ji Xu, Arian Maleki, and Xiaodong Wang. Spectral method for phase retrieval: an expectation propagation perspective. *IEEE Transactions on Information Theory*, 67(2): 1332–1355, 2021.

[32] Junjie Ma, Ji Xu, and Arian Maleki. Towards designing optimal sensing matrices for generalized linear inverse problems. *IEEE Transactions on Information Theory*, 2023.

[33] Antoine Maillard, Laura Foini, Alejandro Lage Castellanos, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. High-temperature expansions and message passing algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(11):113301, 2019.

[34] Kazushi Mimura and Masato Okada. Generating functional analysis for iterative CDMA multiuser detectors. *IEEE transactions on information theory*, 60(6):3645–3670, 2014.

[35] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[36] Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

[37] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1), 2021.

[38] Andrea Montanari and Alexander S Wein. Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. *Probability Theory and Related Fields*, pages 1–53, 2024.

[39] Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: a proof via orthogonalization. *Information and Inference: A Journal of the IMA*, 13(4):iaae027, 2024.

[40] Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

[41] Manfred Opper, Ole Winther, and Michael J Jordan. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6(12), 2005.

[42] Manfred Opper, Burak Cakmak, and Ole Winther. A theory of solving TAP equations for Ising models with general invariant random matrices. *Journal of Physics A: Mathematical and Theoretical*, 49(11): 114002, 2016.

[43] Parthe Pandit, Mojtaba Sahraee-Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*, 1(1):336–347, 2020.

[44] Jason T Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing—part i: Derivation. *IEEE Transactions on Signal Processing*, 62(22):5839–5853, 2014.

[45] Sundeep Rangan and Alyson K Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 1246–1250. IEEE, 2012.

[46] Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

[47] Galen Reeves and Henry D Pfister. The replica-symmetric prediction for random linear estimation with gaussian matrices is exact. *IEEE Transactions on Information Theory*, 65(4):2252–2283, 2019.

[48] Riccardo Rossetti, Bobak Nazer, and Galen Reeves. Linear operator approximate message passing (opamp). *arXiv preprint arXiv:2405.08225*, 2024.

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

[50] Keigo Takeuchi. Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Transactions on Information Theory*, 66(1):368–386, 2019.

[51] Keigo Takeuchi. A unified framework of state evolution for message-passing algorithms. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 151–155. IEEE, 2019.

[52] Keigo Takeuchi. Bayes-optimal convolutional AMP. *IEEE Transactions on Information Theory*, 67(7): 4405–4428, 2021.

[53] Keigo Takeuchi. Orthogonal approximate message-passing for spatially coupled linear models. *IEEE Transactions on Information Theory*, 2023.

[54] Ramji Venkataramanan, Kevin Kögler, and Marco Mondelli. Estimation in rotationally invariant generalized linear models via approximate message passing. In *International Conference on Machine Learning*, pages 22120–22144. PMLR, 2022.

[55] Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791–2823, 2020.

[56] Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *The Annals of Applied Probability*, 34(4):3943–3994, 2024.

[57] Yizhou Xu, TianQi Hou, ShanSuo Liang, and Marco Mondelli. Approximate message passing for multilayer estimation in rotationally invariant models. In *2023 IEEE Information Theory Workshop (ITW)*, pages 294–298. IEEE, 2023.

[58] Arvind Yedla, Yung-Yih Jian, Phong S Nguyen, and Henry D Pfister. A simple proof of maxwell saturation for coupled scalar recursions. *IEEE Transactions on Information Theory*, 60(11):6943–6965, 2014.

[59] Xinyi Zhong, Tianhao Wang, and Zhou Fan. Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. *Information and Inference: A Journal of the IMA*, 13(3):iaae024, 2024.

# Appendix A Proofs for Preliminaries

## A.1 Derivations of the Polynomial Coefficients (Proof of Corollary 1)

Recall that $(Q_n)_{n\geq 1}$ satisfies the recursion (see (5)):

$$Q_n = \Lambda Q_{n-1} - \sum_{i=1}^{n} \mathbb{E}[\Lambda Q_{i-1}] \cdot Q_{n-i}, \quad \forall n \geq 1 \tag{74a}$$

$$\overset{(6)}{=} \Lambda Q_{n-1} - \sum_{i=1}^{n} \kappa_i \cdot Q_{n-i}. \tag{74b}$$

The polynomial representation of $Q_n$ is denoted as (see (9))

$$Q_n = \sum_{i=0}^{n} \alpha_{n,i} \Lambda^i, \quad \forall n \geq 1. \tag{75}$$

Combining these two equations yields

$$\sum_{i=0}^{n} \alpha_{n,i}\Lambda^i = \sum_{i=0}^{n-1} \alpha_{n-1,i}\Lambda^{i+1} - \sum_{j=1}^{n}\sum_{i=0}^{n-j} \kappa_j \alpha_{n-j,i}\Lambda^i \tag{76a}$$

$$= \sum_{i=1}^{n} \alpha_{n-1,i-1}\Lambda^i - \sum_{j=1}^{n}\sum_{i=0}^{n-j} \kappa_j \alpha_{n-j,i}\Lambda^i \tag{76b}$$

$$\overset{(a)}{=} \left(\alpha_{n-1,n-1}\Lambda^n + \sum_{i=1}^{n-1} \alpha_{n-1,i-1}\Lambda^i\right) - \sum_{i=0}^{n-1}\sum_{j=1}^{n-i} \kappa_j \alpha_{n-j,i}\Lambda^i \tag{76c}$$

$$= \alpha_{n-1,n-1}\Lambda^n + \sum_{i=1}^{n-1} \alpha_{n-1,i-1}\Lambda^i - \sum_{j=1}^{n} \kappa_j \alpha_{n-j,0} - \sum_{i=1}^{n-1}\left(\sum_{j=1}^{n-i} \kappa_j \alpha_{n-j,i}\right)\Lambda^i, \quad \forall n \geq 2 \tag{76d}$$

$$= \alpha_{n-1,n-1}\Lambda^n + \sum_{i=1}^{n-1}\left(\alpha_{n-1,i-1} - \sum_{j=1}^{n-i} \kappa_j \alpha_{n-j,i}\right)\Lambda^i - \sum_{j=1}^{n} \kappa_j \alpha_{n-j,0}, \quad \forall n \geq 2. \tag{76e}$$

Comparing the coefficients of the two sides shows the following ($\forall n \geq 2, \forall 1 \leq i \leq n-1$):

$$\alpha_{n,n} = \alpha_{n-1,n-1}, \tag{77a}$$

$$\alpha_{n,0} = -\sum_{j=1}^{n} \kappa_j \alpha_{n-j,0}, \tag{77b}$$

$$\alpha_{n,i} = \alpha_{n-1,i-1} - \sum_{j=1}^{n-i} \kappa_j \alpha_{n-j,i}. \tag{77c}$$

For $n = 1$, it is easy to verify that $\alpha_{1,0} = -\kappa_1$ and $\alpha_{1,1} = 1$. Overall, the above recursion can be written into a unified formula (10) together with the initialization $\alpha_{i,i} = 1, \alpha_{i,-1} = 0, \forall i \geq 0$.

## A.2 A Reformulation of the Moment-Cumulant Formula

We will introduce a recursive characterization of free cumulants in Proposition 1 which is is useful for our derivation of RI-AMP. Before that, we present a useful reformulation of the right-hand side of the moment-cumulant formula (4).

**Lemma 6.** *Let $q_0 := 1$. The following holds for any sequence $(q_\ell)_{\ell \geq 1}$:*

$$\sum_{\pi \in \mathrm{NC}(\ell)} \prod_{B \in \pi} q_{|B|} = \sum_{(s_1,\ldots,s_\ell) \in \mathcal{S}(\ell)} \prod_{i=1}^{\ell} q_{s_i}, \quad \forall \ell \geq 1, \tag{78}$$

*where $\mathcal{S}(\ell)$ denotes the collection of $\ell$-tuple $(s_1, \ldots, s_\ell)$ satisfying*

*(1)* $s_m \geq 0, \ \forall m \in [\ell]$.

*(2)* $\sum_{j=1}^{m} s_j \leq m, \ \forall m \in [\ell]$.

*(3)* $\sum_{j=1}^{\ell} s_j = \ell$.

*Proof.* First, following similar reasoning as [40, pp. 373-375], we can prove that there is a bijection between $\mathrm{NC}(\ell)$ and $\mathcal{S}(\ell)$. For completeness, we include the proof here. We shall provide the explicit forms of the maps for both directions.

**Map from NC($\ell$) to $\mathcal{S}(\ell)$:** Consider $\pi := \{B_1, \ldots, B_N\} \in \mathrm{NC}(\ell)$, where $B_i$ denotes the $i$-th block of $\pi$. Define a map $\phi_\ell : \pi \mapsto (s_1, \ldots, s_\ell)$:

$$s_i := \begin{cases} |B_j|, & \text{if } i \text{ is the largest element of some block } B_j, \text{ where } j \in [\ell], \\ 0, & \text{otherwise.} \end{cases} \tag{79}$$

We claim that $\phi_\ell(\pi) := (s_1, \ldots, s_\ell) \in \mathcal{S}(\ell)$ for any non-crossing partition $\pi \in \mathrm{NC}(\ell)$. To see this, we verify that such $(s_1, \ldots, s_\ell)$ satisfies the following three conditions: (1) $s_m \geq 0, \ \forall m \in [\ell]$; (2) $\sum_{j=1}^{m} s_j \leq m, \ \forall m \in [\ell]$; (3) $\sum_{j=1}^{\ell} s_j = \ell$. Conditions (1) and (3) directly follow from the definition (79). Condition (2) is a consequence of the definition (79) as well as the fact that $\pi$ is non-crossing.

**Map from $\mathcal{S}(\ell)$ to NC($\ell$):** Let $(s_1, \ldots, s_\ell) \in \mathcal{S}(\ell)$. We start from $s_\ell$ and count backwards towards $s_1$ until we meet some nonzero $s_k$. We claim that one of the following scenarios must happen: $s_k = 1$ or $(s_{k-j}, \ldots, s_k) = (\underbrace{0, \ldots, 0}_{j-1 \text{ zeros}}, j)$ for some $1 \leq j \leq k-1$. Otherwise, the defining properties of $\mathcal{S}(\ell)$ would be violated. In the latter case, we form a block consisting of $\{s_{k-j}, \ldots, s_k\}$. Then, we remove the elements in this block and repeat the whole procedure (starting from the largest elements) until no elements in $[\ell]$ are left. The blocks produced by this procedure constitute a partition of $[\ell]$. Finally, this construction guarantees that the partition is non-crossing.

**Proof of Lemma 6.** Let $\pi \in \mathrm{NC}(\ell)$ be an arbitrary non-crossing partition, and denote $\phi_\ell(\pi) = (s_1(\pi), \ldots, s_n(\pi)) \in \mathcal{S}(\ell)$. By the definition of the map $\phi_\ell$ and $q_0 := 1$, we have

$$q_\pi := \prod_{B \in \pi} q_{|B|} = \prod_{j=1}^{\ell} q_{s_j(\pi)}.$$

Because of the bijection between $\mathrm{NC}(\ell)$ and $\mathcal{S}(\ell)$, we have

$$\sum_{\pi \in \mathrm{NC}(\ell)} \prod_{B \in \pi} q_{|B|} = \sum_{(s_1,\ldots,s_\ell) \in \mathcal{S}(\ell)} \prod_{j=1}^{\ell} q_{s_j}, \tag{80}$$

which completes the proof. □

*Remark* 13 (Connections of $\mathcal{S}(n)$, Dyck paths and non-crossing partitions). The object $\mathcal{S}(\ell)$ is also closely related to (and bijective to) *Dyck path* [40]. A Dyck path of length $2\ell$ is a lattice path in $\mathbb{Z}^2$ from $(0, 0)$ to $(k, k)$, consisting of $k$ east steps and $k$ north steps, which lives below the diagonal line $y = x$ (but may touch it).[1] Let $\mathrm{Dyck}(\ell)$ be the collection of Dyck paths of length $2\ell$. An example of a Dyck path of length 10 is shown in Fig. 3.

[1] Notice that the definition of Dyck path here is a flipped version of the usual definition, but they are equivalent.

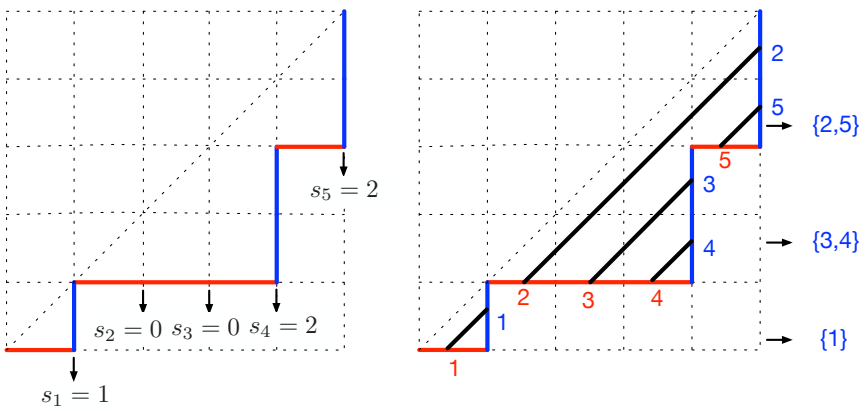



Figure 3: **Left:** Map from $(s_1, \ldots, s_5) = (1, 0, 0, 2, 2) \in \mathcal{S}(5)$ to a Dyck path of length 10 marked in solid lines. **Right:** Map from the Dyck path to a non-crossing partition of $[5]$: $(\{1\}, \{2, 5\}, \{3, 4\})$. (The non-crossing partition is depicted in Fig. 3.) The 5 horizontal/vertical steps of the Dyck path are marked in red/blue colors.

The three objects $\mathcal{S}(n)$, Dyck paths and non-crossing partitions are bijective to each other. For the connection between $\mathcal{S}(n)$ and Dyck paths, we note that each $s_i$ ($\forall i \in [\ell]$) can be interpreted as the vertical increment of a Dyck path at the vertical axis $x = i$; see illustration in the first sub-figure of Fig. 3. Therefore, if $s_i > 0$, the Dyck path has a vertical step (marked in blue color) of length $s_i$ at the axis $x = i$; while if $s_i = 0$, the Dyck path has no vertical step at the axis $x = i$. The connection between Dyck paths and non-crossing partitions are well-known [40], and is illustrated in the second sub-figure of Fig. 3. To map a Dyck path to a non-crossing partition, we label the $\ell$ horizontal steps of the Dyck paths in increasing order from left to right. The $\ell$ vertical steps of the Dyck paths are labeled in a way as shown in Fig. 3. We group the labels in the same vertical axis into one block. These blocks form a non-crossing partition of $[\ell]$. From this map, it is clear that the cardinalities of the blocks of the non-crossing partition $(|B_1|, \ldots, |B_k|)$, which appear in the cumulant-moment formula (4), correspond to the lengths of the vertical segments of a Dyck path.

### A.3 A Recursive Characterization of Free Cumulants (Proof of Proposition 1)

*Proof.* Let $(Q_n)_{n \geq 1}$ be defined as in (5). Denote

$$q_n := \mathbb{E}[\Lambda Q_{n-1}], \quad \forall n \geq 1. \tag{81}$$

We first prove that the following holds for all $\ell \geq 1$, $n \geq 0$:

$$\mathcal{H}_{\ell,n}: \quad \mathbb{E}[\Lambda^\ell Q_n] = \sum_{(s_1,\ldots,s_\ell) \in \mathcal{S}(\ell,n)} \prod_{j=1}^{\ell} q_{s_j}, \tag{82}$$

where $q_0 := 1$ and $\mathcal{S}(\ell, n)$ denotes the collection of $\ell$-tuple $(s_1, \ldots, s_\ell)$ satisfying

(1) $s_m \geq 0$, $\forall m \in [\ell]$.

(2) $\sum_{j=1}^{m} s_j \leq m + n$, $\forall m \in [\ell]$.

(3) $\sum_{j=1}^{\ell} s_j = \ell + n$.

Next, we shall prove (82) by induction on $\ell$:

1. Base case: $\{\mathcal{H}_{1,n}, \forall n \geq 0\}$.
2. Induction step: for all $\ell \geq 1$, $\{\mathcal{H}_{\ell,n}, \forall n \geq 0\} \implies \{\mathcal{H}_{\ell+1,n}, \forall n \geq 0\}$.

**Base case:** When $\ell = 1$, (82) reduces to

$$\mathcal{H}_{1,n}: \quad \mathbb{E}[\Lambda Q_n] = q_{1+n}, \tag{83}$$

which follows from the definition in (81).

**Induction step:** We shall prove the following holds for all $\ell \geq 1$ and $n \geq 1$:

$$\mathcal{H}_{\ell,n} \quad \Longrightarrow \quad \mathcal{H}_{\ell+1,n-1}. \tag{84}$$

This would imply $\{\mathcal{H}_{\ell,n}, \forall n \geq 1\} \Longrightarrow \{\mathcal{H}_{\ell+1,n'}, \forall n' \geq 0\}$ and hence the desired induction step $\{\mathcal{H}_{\ell,n}, \forall n \geq 0\} \Longrightarrow \{\mathcal{H}_{\ell+1,n'}, \forall n' \geq 0\}$. To prove (84), we unfold $Q_{n+1}$:

$$\begin{aligned}\mathbb{E}[\Lambda^\ell Q_{n+1}] &= \mathbb{E}\left[\Lambda^\ell\left(\Lambda Q_n - \sum_{i=1}^{n+1} \mathbb{E}[\Lambda^\ell Q_{i-1}] \cdot Q_{n+1-i}\right)\right] \\ &= \mathbb{E}[\Lambda^{\ell+1} Q_n] - \sum_{i=1}^{n+1} q_i \cdot \mathbb{E}\left[\Lambda^\ell Q_{n+1-i}\right].\end{aligned}$$

Re-arranging terms and noting $q_0 = 1$:

$$\mathbb{E}[\Lambda^{\ell+1} Q_n] = \sum_{i=0}^{n+1} q_i \cdot \mathbb{E}\left[\Lambda^\ell Q_{n+1-i}\right] \tag{85a}$$

$$\overset{(a)}{=} \sum_{i=0}^{n+1} q_i \cdot \left(\sum_{(s_1,\ldots,s_\ell)\in\mathcal{S}(\ell,n+1-i)} q_{s_1}\cdots q_{s_\ell}\right) \tag{85b}$$

$$\overset{(b)}{=} \sum_{(i,s_1,\ldots,s_\ell)\in\mathcal{S}(\ell+1,n)} q_i \cdot q_{s_1}\cdots p_{s_\ell} \tag{85c}$$

where step (a) is a consequence of the induction hypothesis $\{\mathcal{H}_{\ell,n}, \forall n \geq 0\}$ and step (b) is from the following recursion (which can be verified from the definition of $\mathcal{S}(\ell+1, n)$)

$$\mathcal{S}(\ell+1, n) = \big\{(i, s_1, \ldots, s_\ell) : 0 \leq i \leq n+1,\, (s_1, \ldots, s_\ell) \in \mathcal{S}(\ell, n+1-i)\big\}.$$

This proves (84) and completes the proof of (82).

Finally, setting $n = 0$ in (82) yields

$$m_\ell \overset{(a)}{=} \mathbb{E}[\Lambda^\ell Q_0] = \sum_{(s_1,\ldots,s_\ell)\in\mathcal{S}(\ell,0)} \prod_{j=1}^{\ell} q_{s_j} \tag{86a}$$

$$\overset{(b)}{=} \sum_{(s_1,\ldots,s_\ell)\in\mathcal{S}(\ell)} \prod_{j=1}^{\ell} q_{s_j} \tag{86b}$$

$$\overset{(c)}{=} \sum_{\pi\in\mathrm{NC}(\ell)} \prod_{B\in\pi} q_{|B|}, \tag{86c}$$

where step (a) follows from $Q_0 := 1$; step (b) is from the fact that $\mathcal{S}(\ell, 0)$ in (86) is equal to $\mathcal{S}(\ell)$ defined in Lemma 78; and step (c) is due to Lemma 78. Note that (86) is precisely the moment-cumulant formula that defines free cumulants. We can invert the above formula [40, Lecture 10] and conclude that the sequence $(p_\ell)_{\ell\geq 1}$ is equal to the sequence of free cumulants $(\kappa_\ell)_{\ell\geq 1}$. □

## A.4 Connection with the Partial Moments in [20]

**Proof of (8):** We consider the following cases:

- $t = 0, j = 0$: we have $\mathbb{E}[Q_0] = 1 = c_{0,0}$.

- $t = 0, j \geq 1$: we have $\mathbb{E}[Q_j] = 0 = c_{0,j}$.
- $t = 1$: From the definition (7), we have

$$c_{1,j} = \sum_{m=0}^{j+1} c_{0,m}\, \kappa_{j+1-m} \overset{\text{(a)}}{=} \kappa_{j+1},$$

where step (a) is due to the definition of $c_{0,m}$ for the above two corner cases. On the other hand, Proposition 1 shows that $\mathbb{E}[\Lambda Q_j] = \kappa_{j+1}, \forall j \geq 0$. Hence,

$$\mathbb{E}[\Lambda Q_j] = c_{1,j}, \quad \forall j \geq 0.$$

- $t > 1$: we shall prove via induction. Suppose that $c_{k,j} = \mathbb{E}[\Lambda^k Q_j]$ for all $0 \leq k \leq t-1$. Then, the following holds $\forall j \geq 0$:

$$\mathbb{E}[\Lambda^t Q_j] \overset{\text{(a)}}{=} \sum_{i=0}^{j+1} \kappa_i \cdot \mathbb{E}\left[\Lambda^{t-1} Q_{j+1-i}\right] \tag{87a}$$

$$\overset{\text{(b)}}{=} \sum_{i=0}^{j+1} \kappa_i \cdot c_{t-1,j+1-i} \tag{87b}$$

$$\overset{\text{(c)}}{=} \sum_{m=0}^{j+1} \kappa_{j+1-m} \cdot c_{t-1,m} \tag{87c}$$

$$\overset{\text{(d)}}{=} c_{t,j}, \tag{87d}$$

where

(a) This step is due to (85). Note that from the definition of $q_i$ in (81), $q_i := \mathbb{E}[\Lambda Q_{i-1}]$. On the other hand, Proposition 1 shows $\mathbb{E}[\Lambda Q_{i-1}] = \kappa_i$. Substituting $q_i = \kappa_i$ into (85) proves this step.

(b) This step follows from the induction hypothesis.

(c) This is a change of variable.

(d) This is from the definition of $c_{t,j}$; see (7).

The proof is now complete.

# Appendix B State Evolution of OAMP (Proof of Theorem 1)

Consider the following minor variant of the OAMP algorithm in (15):

$$\boldsymbol{w}_t = \left(f_t(\boldsymbol{W}) - \mathbb{E}\left[f_t(\Lambda)\right] \cdot \boldsymbol{I}_N\right) \bar{\boldsymbol{u}}_t, \tag{88a}$$

$$\bar{\boldsymbol{u}}_{t+1} = g_{t+1}(\boldsymbol{w}_{\leq t}; \boldsymbol{a}) - \sum_{i=1}^{t} \mathbb{E}\left[\partial_i g_{t+1}(\mathsf{X}_{\leq t}; \mathsf{A})\right] \cdot \boldsymbol{w}_i, \tag{88b}$$

where the initialization $\bar{\boldsymbol{u}}_1 = \bar{\boldsymbol{x}}_1$, and the expectations in (88b) are taken w.r.t. the random variables $(\mathsf{X}_1, \ldots, \mathsf{X}_t; \mathsf{A})$ defined recursively in (18b)-(18c). Note that (88) is an instance of the vector AMP algorithm as defined in [18, Section 4.1.2]. In the above equation, we have introduced the shorthand

$$\boldsymbol{w}_{\leq t} := (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t).$$

(We will keep this notation throughout this section.) By [18, Theorem 2], the following convergence holds

$$(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_t; \boldsymbol{a}) \xrightarrow{W_2} (\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_t; \mathsf{A}), \tag{89}$$

where the state evolution random variables $(\mathsf{X}_1, \ldots, \mathsf{X}_t; \mathsf{A})$ are defined in (18b)-(18c). Compared with the original OAMP algorithm (15), the iteration (88) replaces $\frac{1}{N}\mathrm{tr}\left(f_t(\boldsymbol{W})\right)$ and $(\langle \partial_i g_{t+1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t; \boldsymbol{a})\rangle)_{i\in[t]}$ by $\mathbb{E}\left[f_t(\Lambda)\right]$ and $(\mathbb{E}\left[\partial_i g_{t+1}(\mathsf{X}_1, \ldots, \mathsf{X}_t; \mathsf{A})\right])_{i\in[t]}$, respectively. Similar to the arguments used in the proof of [17, Theorem 1], we can show that the difference between these two algorithms is asymptotically negligible.

Following [17], we say $\boldsymbol{u} \overset{N\to\infty}{\simeq} \boldsymbol{v}$ for two random vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N$, if

$$\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{N} \xrightarrow{\mathbb{P}} 0 \quad \text{as} \quad N \to \infty.$$

Given the convergence result in (89), it suffices to prove that the the iterates generated by the original OAMP algorithm and that generated by the auxiliary iteration (88) satisfies

$$\boldsymbol{x}_t \overset{N\to\infty}{\simeq} \boldsymbol{w}_t, \quad \forall t \geq 1. \tag{90}$$

We show this by induction. Suppose that this is true up to $t$. Recall the definitions in (3) and (88):

$$\boldsymbol{x}_{t+1} = \left(f_{t+1}(\boldsymbol{W}) - \frac{\mathrm{tr}\left(f_{t+1}(\boldsymbol{W})\right)}{N} \cdot \boldsymbol{I}_N\right) \bar{\boldsymbol{x}}_{t+1}, \tag{91a}$$

$$\boldsymbol{w}_{t+1} = \left(f_{t+1}(\boldsymbol{W}) - \mathbb{E}[f_{t+1}(\Lambda)] \cdot \boldsymbol{I}_N\right) \bar{\boldsymbol{u}}_{t+1}. \tag{91b}$$

Hence,

$$\begin{aligned}
\frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{w}_{t+1}\|^2}{N} &\leq \frac{1}{N}\left\|\left(f_{t+1}(\boldsymbol{W}) - \frac{\mathrm{tr}\left(f_{t+1}(\boldsymbol{W})\right)}{N} \cdot \boldsymbol{I}_N\right)\left(\bar{\boldsymbol{x}}_{t+1}-\bar{\boldsymbol{u}}_{t+1}\right)\right\|^2 \\
&\quad + \left(\mathbb{E}[f_{t+1}(\Lambda)] - \frac{\mathrm{tr}\left(f_{t+1}(\boldsymbol{W})\right)}{N}\right)^2 \cdot \frac{1}{N}\left\|\bar{\boldsymbol{u}}_{t+1}\right\|^2 \qquad (92a) \\
&\leq \left\|f_{t+1}(\boldsymbol{W}) - \frac{\mathrm{tr}\left(f_{t+1}(\boldsymbol{W})\right)}{N} \cdot \boldsymbol{I}_N\right\|_{\mathrm{op}} \cdot \frac{1}{N}\left\|\bar{\boldsymbol{x}}_{t+1}-\bar{\boldsymbol{u}}_{t+1}\right\|^2 \\
&\quad + \left(\mathbb{E}[f_{t+1}(\Lambda)] - \frac{\mathrm{tr}\left(f_{t+1}(\boldsymbol{W})\right)}{N}\right)^2 \cdot \frac{1}{N}\left\|\bar{\boldsymbol{u}}_{t+1}\right\|^2. \qquad (92b)
\end{aligned}$$

To control the above term, we bound the operator norm of the matrix $f_{t+1}(\boldsymbol{W}) - \frac{\mathrm{tr}(f_{t+1}(\boldsymbol{W}))}{N} \cdot \boldsymbol{I}_N$:

$$\begin{aligned}
\left\|f_{t+1}(\boldsymbol{W}) - \frac{\mathrm{tr}\left(f_{t+1}(\boldsymbol{W})\right)}{N} \cdot \boldsymbol{I}_N\right\|_{\mathrm{op}} &= \max_{1\leq i\leq N}\left|f_{t+1}(\lambda_i) - \frac{1}{N}\sum_{j=1}^N f_{t+1}(\lambda_j)\right| \qquad (93a) \\
&\leq \max_{1\leq i,i'\leq N} \left|f_{t+1}(\lambda_i) - f_{t+1}(\lambda_{i'})\right| \qquad (93b) \\
&\leq 2\max_{1\leq i\leq N} \left|f_{t+1}(\lambda_i)\right| \leq C', \qquad (93c)
\end{aligned}$$

for some $N$-independent constant $C'$. The last step is due to the fact that $\max_{1\leq i\leq N}\ |\lambda_i|$ is bounded by an $N$-independent constant and $f_{t+1}$ is continuous and independent of $N$; see Assumption 1. Further, the empirical eigenvalue distribution of $\boldsymbol{W}$ converges to a compactly support measure, hence

$$\lim_{N\to\infty} \frac{\mathrm{tr}\left(f_{t+1}(\boldsymbol{W})\right)}{N} = \mathbb{E}[f_{t+1}(\Lambda)]. \tag{94}$$

In light of (92)-(94), to prove (90), it remains to show

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\left\|\bar{\boldsymbol{x}}_{t+1}-\bar{\boldsymbol{u}}_{t+1}\right\|^2 = 0, \tag{95a}$$

$$\operatorname*{plim\,sup}_{N\to\infty} \frac{1}{N}\left\|\bar{\boldsymbol{u}}_{t+1}\right\|^2 < \infty. \tag{95b}$$

Note that $\bar{\boldsymbol{u}}_{t+1}$ is generated via the auxiliary iteration (88), and (95b) follows from the convergence result (89). Finally, to prove (95a), recall the definitions of $\bar{\boldsymbol{x}}_{t+1}$ and $\bar{\boldsymbol{u}}_{t+1}$ in (3) and (88):

$$\begin{aligned}
&\frac{1}{N}\left\|\bar{\boldsymbol{x}}_{t+1}-\bar{\boldsymbol{u}}_{t+1}\right\|^2 \\
&=\frac{1}{N}\left\|g_{t+1}(\boldsymbol{x}_{\le t};\boldsymbol{a})-\sum_{i=1}^{t}\langle\partial_i g_t(\boldsymbol{x}_{\le t};\boldsymbol{a})\rangle\cdot\boldsymbol{x}_i-\left(g_{t+1}(\boldsymbol{w}_{\le t};\boldsymbol{a})-\sum_{i=1}^{t}\mathbb{E}\left[\partial_i g_{t+1}(\mathsf{X}_{\le t};\mathsf{A})\right]\cdot\boldsymbol{w}_i\right)\right\|^2 \\
&\le\frac{1}{N}\left\|g_{t+1}(\boldsymbol{x}_{\le t};\boldsymbol{a})-g_{t+1}(\boldsymbol{w}_{\le t};\boldsymbol{a})\right\|^2+\sum_{i=1}^{t}\frac{1}{N}\left\|\langle\partial_i g_t(\boldsymbol{x}_{\le t})\rangle\cdot\boldsymbol{x}_i-\mathbb{E}\left[\partial_i g_{t+1}(\mathsf{X}_{\le t};\mathsf{A})\right]\cdot\boldsymbol{w}_i\right\|^2.
\end{aligned}$$

The above term converges to zero in probability due to the following facts:

- Induction hypothesis: $\boldsymbol{x}_s \overset{N\to\infty}{\simeq} \boldsymbol{w}_s, \quad \forall s=1,\dots,t.$
- The assumption that $g_{t+1}$ is Lipschitz continuous.
- The convergence $\langle\partial_i g_t(\boldsymbol{x}_{\le t};\boldsymbol{a})\rangle \xrightarrow{\mathbb{P}} \mathbb{E}\left[\partial_i g_{t+1}(\mathsf{X}_{\le t};\mathsf{A})\right]$, $\forall i=1,\dots,t$. This is a consequence of: (1) the induction hypothesis: $\boldsymbol{x}_s \overset{N\to\infty}{\simeq} \boldsymbol{w}_s, \quad \forall s=1,\dots,t$; (2) the assumption that $g_{t+1}$ is Lipschitz and continuously differentiable (and hence $\partial_i g_{t+1}$ is continuously bounded).

The proof is now complete.

# Appendix C Proofs for Main Results

## C.1 Reformulation of First-order Method (Proof of Lemma 1)

*Proof.* We stack the iterates of the RI-AMP algorithm as

$$\begin{aligned}
\begin{bmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\\ \vdots\\ \boldsymbol{r}_t\end{bmatrix} &= \begin{bmatrix}\boldsymbol{W}\boldsymbol{u}_1\\ \boldsymbol{W}\boldsymbol{u}_2\\ \vdots\\ \boldsymbol{W}\boldsymbol{u}_t\end{bmatrix} - \begin{bmatrix}\mathsf{b}_{1,1}\boldsymbol{I}_N & & & \\ \mathsf{b}_{2,1}\boldsymbol{I}_N & \mathsf{b}_{2,2}\boldsymbol{I}_N & & \\ \vdots & \vdots & \ddots & \\ \mathsf{b}_{t,1}\boldsymbol{I}_N & \mathsf{b}_{t,2}\boldsymbol{I}_N & \cdots & \mathsf{b}_{t,t}\boldsymbol{I}_N\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\\ \vdots\\ \boldsymbol{u}_t\end{bmatrix} && \text{(97a)}\\
&= (\mathsf{I}_t\otimes\boldsymbol{W}-\mathsf{B}_t\otimes\boldsymbol{I}_N)\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\\ \vdots\\ \boldsymbol{u}_t\end{bmatrix}. && \text{(97b)}
\end{aligned}$$

Using the definition of $\bar{\boldsymbol{u}}_t$ in (33), we have

$$\begin{aligned}
\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\\ \vdots\\ \boldsymbol{u}_t\end{bmatrix} &= \begin{bmatrix}\bar{\boldsymbol{u}}_1\\ \bar{\boldsymbol{u}}_2\\ \vdots\\ \bar{\boldsymbol{u}}_t\end{bmatrix} + \begin{bmatrix}\boldsymbol{0}_N & & & & \\ \mathsf{d}_{1,1}\boldsymbol{I}_N & \boldsymbol{0}_N & & & \\ \mathsf{d}_{2,1}\boldsymbol{I}_N & \mathsf{d}_{2,2}\boldsymbol{I}_N & \boldsymbol{0}_N & & \\ \vdots & \vdots & \vdots & \ddots & \\ \mathsf{d}_{t-1,1}\boldsymbol{I}_N & \mathsf{d}_{t-1,2}\boldsymbol{I}_N & \cdots & \mathsf{d}_{t-1,t-1}\boldsymbol{I}_N & \boldsymbol{0}_N\end{bmatrix}\begin{bmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\\ \vdots\\ \boldsymbol{r}_t\end{bmatrix} && \text{(98a)}\\
&= \begin{bmatrix}\bar{\boldsymbol{u}}_1\\ \bar{\boldsymbol{u}}_2\\ \vdots\\ \bar{\boldsymbol{u}}_t\end{bmatrix} + (\mathsf{D}_t\otimes\boldsymbol{I}_N)\begin{bmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\\ \vdots\\ \boldsymbol{r}_t\end{bmatrix}. && \text{(98b)}
\end{aligned}$$

Substituting (98) into (97) yields

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = (\mathbf{I}_t \otimes \boldsymbol{W} - \mathbf{B}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix} + (\mathbf{D}_t \otimes \boldsymbol{W} - \mathbf{B}_t\mathbf{D}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix}. \tag{99a}$$

Re-arranging the equation, we have

$$[\boldsymbol{I}_{tN} - (\mathbf{D}_t \otimes \boldsymbol{W} - \mathbf{B}_t\mathbf{D}_t \otimes \boldsymbol{I}_N)] \begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = (\mathbf{I}_t \otimes \boldsymbol{W} - \mathbf{B}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}. \tag{100}$$

Note that both $\mathbf{B}_t$ and $\mathbf{D}_t$ are lower triangular matrices. Moreover, $\mathbf{D}_t$ is strictly lower triangular. We can then verify that $\boldsymbol{I}_{tN} - (\mathbf{D}_t \otimes \boldsymbol{W} - \mathbf{B}_t\mathbf{D}_t \otimes \boldsymbol{I}_N)$ is lower triangular with diagonal elements all equal to one, and thus invertible. Solving the above equation yields

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = [\boldsymbol{I}_{tN} - (\mathbf{D}_t \otimes \boldsymbol{W} - \mathbf{B}_t\mathbf{D}_t \otimes \boldsymbol{I}_N)]^{-1} (\mathbf{I}_t \otimes \boldsymbol{W} - \mathbf{B}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}. \tag{101}$$

At this point, we have written $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t)$ as a linear combination of $(\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_t)$. It remains to verify the equivalence between (101) and (34).

Let $\boldsymbol{W} = \boldsymbol{O}\boldsymbol{\Lambda}\boldsymbol{O}^\mathsf{T}$ be the eigenvalue decomposition of $\boldsymbol{W}$ with $\boldsymbol{\Lambda} := \{\lambda_1, \ldots, \lambda_N\}$. Using this decomposition, we have

$$\begin{aligned}
& [\boldsymbol{I}_{tN} - (\mathbf{D}_t \otimes \boldsymbol{W} - \mathbf{B}_t\mathbf{D}_t \otimes \boldsymbol{I}_N)]^{-1} (\mathbf{I}_t \otimes \boldsymbol{W} - \mathbf{B}_t \otimes \boldsymbol{I}_N) \\
= \; & (\boldsymbol{I}_t \otimes \boldsymbol{O}) [\boldsymbol{I}_{tN} - \mathbf{D}_t \otimes \boldsymbol{\Lambda} + \mathbf{B}_t\mathbf{D}_t \otimes \boldsymbol{I}_N]^{-1} (\mathbf{I}_t \otimes \boldsymbol{\Lambda} - \mathbf{B}_t \otimes \boldsymbol{I}_N) (\mathbf{I}_t \otimes \boldsymbol{O})^\mathsf{T} \\
\overset{(a)}{=} \; & (\mathbf{I}_t \otimes \boldsymbol{O}) \left[\boldsymbol{\Pi} (\boldsymbol{I}_{tN} - \boldsymbol{\Lambda} \otimes \mathbf{D}_t + \boldsymbol{I}_N \otimes \mathbf{B}_t\mathbf{D}_t) \boldsymbol{\Pi}^\mathsf{T}\right]^{-1} \boldsymbol{\Pi} (\boldsymbol{\Lambda} \otimes \mathbf{I}_t - \boldsymbol{I}_N \otimes \mathbf{B}_t) \boldsymbol{\Pi}^\mathsf{T} (\mathbf{I}_t \otimes \boldsymbol{O})^\mathsf{T} \\
= \; & (\mathbf{I}_t \otimes \boldsymbol{O}) \boldsymbol{\Pi} [\boldsymbol{I}_{tN} - \boldsymbol{\Lambda} \otimes \mathbf{D}_t + \boldsymbol{I}_N \otimes \mathbf{B}_t\mathbf{D}_t]^{-1} (\boldsymbol{\Lambda} \otimes \mathbf{I}_t - \boldsymbol{I}_N \otimes \mathbf{B}_t) \boldsymbol{\Pi}^\mathsf{T} (\mathbf{I}_t \otimes \boldsymbol{O})^\mathsf{T} \\
= \; & (\mathbf{I}_t \otimes \boldsymbol{O}) \boldsymbol{\Pi} \operatorname{diag} \left\{ (\mathbf{I}_t - \lambda_1 \mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)^{-1} (\lambda_1 \mathbf{I}_t - \mathbf{B}_t), \ldots, (\mathbf{I}_t - \lambda_N \mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)^{-1} (\lambda_N \mathbf{I}_t - \mathbf{B}_t) \right\} \boldsymbol{\Pi}^\mathsf{T} (\mathbf{I}_t \otimes \boldsymbol{O})^\mathsf{T} \\
\overset{(b)}{=} \; & (\mathbf{I}_t \otimes \boldsymbol{O}) \boldsymbol{\Pi} \operatorname{diag} \{\mathbf{P}(\lambda_1), \ldots, \mathbf{P}(\lambda_N)\} \boldsymbol{\Pi}^\mathsf{T} (\mathbf{I}_t \otimes \boldsymbol{O})^\mathsf{T} \\
\overset{(c)}{=} \; & (\mathbf{I}_t \otimes \boldsymbol{O}) \begin{bmatrix} P_{1,1}(\boldsymbol{\Lambda}) & & & \\ P_{2,1}(\boldsymbol{\Lambda}) & P_{2,2}(\boldsymbol{\Lambda}) & & \\ \vdots & \vdots & \ddots & \\ P_{t,1}(\boldsymbol{\Lambda}) & P_{t,2}(\boldsymbol{\Lambda}) & \cdots & P_{t,t}(\boldsymbol{\Lambda}) \end{bmatrix} (\mathbf{I}_t \otimes \boldsymbol{O})^\mathsf{T} \\
= \; & \begin{bmatrix} P_{1,1}(\boldsymbol{W}) & & & \\ P_{2,1}(\boldsymbol{W}) & P_{2,2}(\boldsymbol{W}) & & \\ \vdots & \vdots & \ddots & \\ P_{t,1}(\boldsymbol{W}) & P_{t,2}(\boldsymbol{W}) & \cdots & P_{t,t}(\boldsymbol{W}) \end{bmatrix},
\end{aligned}$$

where

- In step (a), $\boldsymbol{\Pi} \in \mathbb{R}^{tN \times tN}$ denotes the unique permutation matrix that reverses the order of the Kronecker product:

$$\boldsymbol{M}_t \otimes \boldsymbol{M}_N = \boldsymbol{\Pi}(\boldsymbol{M}_N \otimes \boldsymbol{M}_t)\boldsymbol{\Pi}^\mathsf{T}, \quad \forall \boldsymbol{M}_t \in \mathbb{R}^{t \times t}, \; \boldsymbol{M}_N \in \mathbb{R}^{N \times N}. \tag{103a}$$

  Specifically, let $\sigma : [tN] \mapsto [tN]$ be a permutation with the following mapping rule:

$$\sigma\left((i_1 - 1)N + i_2\right) = (i_2 - 1)t + i_1, \quad \forall i_1 \in [t], \; i_2 \in [N]. \tag{103b}$$

Then, the operation $\boldsymbol{A} \mapsto \boldsymbol{\Pi}\boldsymbol{A}\boldsymbol{\Pi}^{\mathsf{T}}$ ($\forall \boldsymbol{A} \in \mathbb{R}^{tN \times tN}$) is a re-ordering of the entries of $\boldsymbol{A}$ according to

$$\boldsymbol{A}_{m,n} \mapsto \boldsymbol{A}_{\sigma(m),\sigma(n)}, \quad \forall m, n \in [tN]. \tag{103c}$$

- In step (b), we introduced the notation:

$$\mathbf{P}(\lambda_i) := (\mathbf{I}_t - \lambda_i \mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)^{-1}(\lambda_i \mathbf{I}_t - \mathbf{B}_t) \in \mathbb{R}^{t \times t}, \quad \forall i \in [N].$$

  Note that since $\mathbf{B}_t$ and $\mathbf{D}_t$ are lower triangular matrices, $(\mathbf{P}(\lambda_i))_{i\in[N]}$ are also lower triangular.

- Step (c) is consequence of the definition of the permutation matrix $\boldsymbol{\Pi}$, cf. (103). Here, $(P_{i,j})_{1\le i\le t, 1\le j\le i}$ are understood as a sequence of scalar functions, which are defined as follows:

$$\begin{bmatrix} P_{1,1}(\lambda) & 0 & \cdots & 0 \\ P_{2,1}(\lambda) & P_{2,2}(\lambda) & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ P_{t,1}(\lambda) & P_{t,1}(\lambda) & \cdots & P_{t,t}(\lambda) \end{bmatrix} = (\mathbf{I}_t - \lambda \mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)^{-1}(\lambda \mathbf{I}_t - \mathbf{B}_t), \quad \forall \lambda \in \mathbb{R}. \tag{104a}$$

  Note that the lower triangular structure of the matrix on the LHS of (104a) is a consequence of the lower triangular structure of the matrix on the RHS, which is further due to the lower triangular structures of $\mathbf{B}_t$ and $\mathbf{D}_t$.

This completes the proof. □

## C.2 Choice of De-biasing Matrix (Proof of Lemma 2)

*Proof.* **Proof of (1).** For $t = 1$, we have $\mathbf{B}_1 = \mathsf{b}_{1,1}$ and $\mathbf{D}_1 = 0$. Then, $\mathbf{P}_1(\lambda) = \lambda - \mathsf{b}_{1,1}$. Clearly, $\mathsf{b}_{1,1} = \mathbb{E}[\Lambda]$ is the unique solution to $\mathbb{E}[\mathbf{P}_1(\Lambda)] = \mathbf{0}_{t\times t}$. In what follows, we assume $t \ge 2$. We partition $\mathbf{B}_t$ (cf. (23)) and $\mathbf{D}_t$ (cf. (37)) as follows

$$\mathbf{B}_t = \begin{bmatrix} \mathbf{B}_{t-1} & \mathbf{0}_{t-1\times 1} \\ \mathbf{b}_t & \mathsf{b}_{t,t} \end{bmatrix}, \quad \mathbf{D}_t = \begin{bmatrix} \mathbf{D}_{t-1} & \mathbf{0}_{t-1\times 1} \\ \mathbf{d}_t & 0 \end{bmatrix}, \tag{105}$$

where $\mathbf{b}_t, \mathbf{d}_t \in \mathbb{R}^{1\times(t-1)}$. Then,

$$\begin{aligned}
\mathbf{P}_t(\lambda) &= (\mathbf{I}_t - \lambda\mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)^{-1}(\lambda\mathbf{I}_t - \mathbf{B}_t) \\
&= \begin{bmatrix} \mathbf{I}_{t-1} - (\lambda\mathbf{I}_{t-1} - \mathbf{B}_{t-1})\mathbf{D}_{t-1} & \mathbf{0}_{t-1\times 1} \\ -\mathbf{b}_t\mathbf{D}_{t-1} + (\lambda - \mathsf{b}_{t,t}\mathbf{d}_t) & 1 \end{bmatrix}^{-1} \begin{bmatrix} \lambda\mathbf{I}_{t-1} - \mathbf{B}_{t-1} & \mathbf{0}_{t-1\times 1} \\ -\mathbf{b}_t & \lambda - \mathsf{b}_{t,t} \end{bmatrix} \\
&\overset{(a)}{=} \begin{bmatrix} [\mathbf{I}_{t-1} - (\lambda\mathbf{I}_{t-1} - \mathbf{B}_{t-1}\mathbf{D}_{t-1})]^{-1} & \mathbf{0}_{t-1\times 1} \\ (-\mathbf{b}_t\mathbf{D}_{t-1} + (\lambda - \mathsf{b}_{t,t})\mathbf{d}_t)\,[\mathbf{I}_{t-1} - (\lambda\mathbf{I}_{t-1} - \mathbf{B}_{t-1}\mathbf{D}_{t-1})]^{-1} & 1 \end{bmatrix} \begin{bmatrix} \lambda\mathbf{I}_{t-1} - \mathbf{B}_{t-1} & \mathbf{0}_{t-1\times 1} \\ -\mathbf{b}_t & \lambda - \mathsf{b}_{t,t} \end{bmatrix} \\
&\overset{(b)}{=} \begin{bmatrix} \mathbf{P}_{t-1}(\lambda) & \mathbf{0}_{t-1\times 1} \\ (-\mathbf{b}_t\mathbf{D}_{t-1} + (\lambda - \mathsf{b}_{t,t})\mathbf{d}_t)\,\mathbf{P}_{t-1}(\lambda) - \mathbf{b}_t & \lambda - \mathsf{b}_{t,t} \end{bmatrix}
\end{aligned}$$

Hence, enforcing the condition $\mathbb{E}[\mathbf{P}_t(\Lambda)] = \mathbf{0}_{t\times t}$ gives us

$$\mathbb{E}[\mathbf{P}_{t-1}(\Lambda)] = \mathbf{0}_{t-1\times t-1}, \tag{107a}$$

$$\mathbb{E}[(-\mathbf{b}_t\mathbf{D}_{t-1} + (\Lambda - \mathsf{b}_{t,t})\mathbf{d}_t)\,\mathbf{P}_{t-1}(\Lambda) - \mathbf{b}_t] = \mathbf{0}_{1\times t-1}, \tag{107b}$$

$$\mathbb{E}[\Lambda - \mathsf{b}_{t,t}] = 0. \tag{107c}$$

Notice that $\mathbf{P}_{t-1}(\lambda)$ only depends on the sub-matrices $\mathbf{B}_{t-1}$ and $\mathbf{D}_{t-1}$ (cf. (105)) but not on the last rows of $\mathbf{B}_t$ and $\boldsymbol{D}_t$. Suppose that the equation $\mathbb{E}[\mathbf{P}_{t-1}(\Lambda)] = \mathbf{0}_{t-1\times t-1}$ uniquely determines the sub-matrix $\mathbf{B}_{t-1}$ (as a function of $\mathbf{D}_{t-1}$). The last row of $\mathbf{B}_{t-1}$, namely $(\mathbf{b}_t, \mathsf{b}_{t,t})$ are uniquely determined as

$$\begin{aligned}
\mathsf{b}_{t,t} &= m_1, \\
\mathbf{b}_t &= \mathbf{d}_t\mathbb{E}\left[(\Lambda - m_1)\,\mathbf{P}_{t-1}(\Lambda)\,(\mathbf{I}_{t-1} + \mathbf{D}_{t-1}\mathbf{P}_{t-1}(\Lambda))^{-1}\right] \\
&= \mathbf{d}_t\mathbb{E}\left[\Lambda\mathbf{P}_{t-1}(\Lambda)\,(\mathbf{I}_{t-1} + \mathbf{D}_{t-1}\mathbf{P}_{t-1}(\Lambda))^{-1}\right],
\end{aligned}$$

where the last step follows from $\mathbb{E}\left[\mathbf{P}_{t-1}(\Lambda)\right] = \mathbf{0}_{t-1\times t-1}$. Hence, we have shown that, if $\mathbb{E}\left[\mathbf{P}_{t-1}(\Lambda)\right] = \mathbf{0}_{t-1\times t-1}$ uniquely determines the sub-matrix $\mathbf{B}_{t-1}$, then $\mathbb{E}\left[\mathbf{P}_{t}(\Lambda)\right] = \mathbf{0}_{t\times t}$ uniquely determines the matrix $\mathbf{B}_t$. Hence, the claim holds by induction.

**Proof of (2) and (4).** Claim (1) shows that the solution (39) is unique. It remains to verify that there always exists a solution $\mathbf{B}_t$ which is polynomial in $\mathbf{D}_t$. Moreover, the coefficients of the polynomial $(\alpha_i)_{i\geq 1}$ only depend on the law $\mu$, but not on $\mathbf{D}_t$. In other words, there exists a single polynomial that solves (39) for all $\mathbf{D}_t$. Assume that

$$\mathbf{B}_t = \sum_{i=1}^{t} \alpha_i \mathbf{D}_t^{i-1}, \tag{109}$$

where the sequence $(\alpha_i)_{i\geq 1}$ are yet to be determined. (Note that $\mathbf{D}_t$ is strictly lower triangular and hence $\mathbf{D}_t^i = \mathbf{0}_{t\times t}, \forall i \geq t$.) The matrix $\mathbf{P}_t$ can be written as (cf. (38)):

$$\begin{aligned}\mathbf{P}_t(\lambda) &= (\mathbf{I}_t - \lambda\mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)^{-1}(\lambda\mathbf{I}_t - \mathbf{B}_t)\\ &= \sum_{i=1}^{t} \left(\lambda\mathbf{D}_t - \mathbf{B}_t\mathbf{D}_t\right)^{i-1}(\lambda\mathbf{I}_t - \mathbf{B}_t),\end{aligned}$$

where the second step is due to the following identity for a strictly lower triangular matrix $\mathbf{A} \in \mathbb{R}^{t\times t}$: $(\mathbf{I}_t - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{t-1}$. Clearly, assuming $\mathbf{B}_t$ is a polynomial of $\mathbf{D}_t$, $\mathbf{P}_t$ is also a polynomial of $\mathbf{D}_t$, which we denote as:

$$\mathbf{P}_t(\lambda) = \sum_{i=1}^{t} Q_i(\lambda)\mathbf{D}_t^{i-1}, \tag{111}$$

where the coefficients $(Q_i(\lambda))_{i\in\mathbb{N}}$ depend on $\lambda$. We next provide a recursive characterization of $(Q_i(\lambda))_{i\in\mathbb{N}}$ in terms of $(a_i)_{i\in\mathbb{N}}$. From the definition of $\mathbf{P}_t(\lambda)$ (cf. (38)), the following identity holds

$$(\mathbf{I}_t - \lambda\mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)\mathbf{P}_t(\lambda) - \lambda\mathbf{I}_t + \mathbf{B}_t = \mathbf{0}_{t\times t}. \tag{112}$$

Using the polynomial representations of $\mathbf{B}_t$ and $\mathbf{P}_t(\lambda)$ in (109) and (111), we write the LHS of the above equation as

$$\begin{aligned}
&(\mathbf{I}_t - \lambda\mathbf{D}_t + \mathbf{B}_t\mathbf{D}_t)\mathbf{P}_t(\lambda) - \lambda\mathbf{I}_t + \mathbf{B}_t && \text{(113a)}\\
&= \left(\mathbf{I}_t - \lambda\mathbf{D}_t + \sum_{i=1}^{\infty}\alpha_i\mathbf{D}_t^i\right)\sum_{i=1}^{\infty} Q_i(\lambda)\mathbf{D}_t^{i-1} - \lambda\mathbf{I}_t + \sum_{i=1}^{\infty}\alpha_i\mathbf{D}_t^{i-1} && \text{(113b)}\\
&= \sum_{i=1}^{\infty} Q_i(\lambda)\mathbf{D}_t^{i-1} - \sum_{i=1}^{\infty}\lambda Q_i(\lambda)\mathbf{D}_t^{i} + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\alpha_i Q_j(\lambda)\mathbf{D}_t^{i+j-1} - \lambda\mathbf{I}_t + \sum_{i=1}^{\infty}\alpha_i\mathbf{D}_t^{i-1} && \text{(113c)}\\
&= \sum_{i=1}^{\infty} Q_i(\lambda)\mathbf{D}_t^{i-1} - \sum_{i=1}^{\infty}\lambda Q_i(\lambda)\mathbf{D}_t^{i} + \sum_{i=1}^{\infty}\sum_{k=1+i}^{\infty}\alpha_i Q_{k-i}(\lambda)\mathbf{D}_t^{k-1} - \lambda\mathbf{I}_t + \sum_{i=1}^{\infty}\alpha_i\mathbf{D}_t^{i-1} && \text{(113d)}\\
&\overset{\text{(a)}}{=} \sum_{i=1}^{\infty} Q_i(\lambda)\mathbf{D}_t^{i-1} - \sum_{i=2}^{\infty}\lambda Q_{i-1}(\lambda)\mathbf{D}_t^{i-1} + \sum_{k=2}^{\infty}\left(\sum_{i=1}^{k-1}\alpha_i Q_{k-i}(\lambda)\right)\mathbf{D}_t^{k-1} - \lambda\mathbf{I}_t + \sum_{i=1}^{\infty}\alpha_i\mathbf{D}_t^{i-1} && \text{(113e)}\\
&= (Q_1(\lambda) - \lambda + \alpha_1)\mathbf{I}_t + \sum_{n=2}^{\infty}\left(Q_n(\lambda) - \lambda Q_{n-1}(\lambda) + \sum_{i=1}^{n-1}\alpha_i Q_{n-i}(\lambda) + \alpha_n\right)\mathbf{D}_t^{n-1}, && \text{(113f)}
\end{aligned}$$

where step (a) is due to a swap of the summation order. Note that we have represented the finite order polynomials of $\mathbf{D}_t$ as infinite power series (since $\mathbf{D}_t^i = \mathbf{0}_{t\times t}, \forall i \geq t$) in the above equation. This representation is somewhat more convenient for calculations. Since (113) is identically zero (from (112)), we must have that

$$Q_1(\lambda) = \lambda - \alpha_1, \tag{114a}$$

$$Q_n(\lambda) = \lambda Q_{n-1}(\lambda) - \sum_{i=1}^{n-1}\alpha_i Q_{n-i}(\lambda) - \alpha_n, \quad \forall n \geq 2, \tag{114b}$$

which provides a recursive definition of $(Q_i(\lambda))_{i\geq 1}$ in terms of $(a_i)_{i\geq 1}$. Recall that $\mathbf{P}_t(\lambda) = \sum_{i=1}^t Q_i(\lambda)\mathbf{D}_t^{i-1}$. Therefore, if $\mathbb{E}[Q_i(\Lambda)] = 0, \forall i \in [t]$, then $\mathbb{E}[\mathbf{P}_t(\Lambda)] = \mathbf{0}_{t\times t}$, $\forall \mathbf{D}_t$. Taking expectations over $\Lambda \sim \mu$ in (114) and setting $\mathbb{E}[Q_i(\Lambda)] = 0, \forall i \in [t]$ leads to the following choice of $(\alpha_i)_{i\geq 1}$:

$$\alpha_1 = \mathbb{E}[\Lambda], \tag{115a}$$
$$\alpha_n = \mathbb{E}[\Lambda Q_{n-1}(\Lambda)], \quad \forall n \geq 2. \tag{115b}$$

Substituting the above equation into (114) leads to the following recursive characterization of $(Q_i(\lambda))_{i\geq 1}$:

$$Q_n(\lambda) = \lambda Q_{n-1}(\lambda) - \sum_{i=1}^{n-1} \alpha_i Q_{n-i}(\lambda) - \alpha_n \tag{116a}$$
$$= \lambda Q_{n-1}(\lambda) - \sum_{i=1}^{n-1} \mathbb{E}[\Lambda Q_{i-1}(\Lambda)] \cdot Q_{n-i}(\lambda) - \mathbb{E}[\Lambda Q_{n-1}(\Lambda)] \tag{116b}$$
$$= \lambda Q_{n-1}(\lambda) - \sum_{i=1}^{n} \mathbb{E}[\Lambda Q_{i-1}(\Lambda)] \cdot Q_{n-i}(\lambda), \tag{116c}$$

where we defined $Q_0(\lambda) := 1$ in the last step. This proves the claim.

**Proof of (3).** Item (2) shows that the matrix $\mathbf{B}_t$ that solves (39) is a polynomial in $\mathbf{D}_t$ with coefficients characterized by (41). The claimed result is then a consequence of this recursion together with Proposition 5. □

## C.3 State Evolution of RI-AMP (Proof of Theorem 2)

We first address some subtle points in reducing RI-AMP to OAMP. We then apply the general state evolution result of OAMP to calculate the claimed update equation of the covariance matrix $\mathbf{\Sigma}_t$.

**Reduction to OAMP:** From Lemma 3, the iterates $\boldsymbol{r}_t$ in RI-AMP can be written as (cf. (45))

$$\boldsymbol{r}_t = \widehat{P}_{t,1}(\boldsymbol{W})\bar{\boldsymbol{u}}_1 + \cdots + \widehat{P}_{t,t}(\boldsymbol{W})\bar{\boldsymbol{u}}_t, \tag{117a}$$
$$\boldsymbol{u}_{t+1} = \eta_{t+1}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t), \tag{117b}$$
$$\bar{\boldsymbol{u}}_{t+1} = \eta_{t+1}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t) - \left(\langle \partial_1 \boldsymbol{u}_{t+1}\rangle \cdot \boldsymbol{r}_1 + \cdots + \langle \partial_t \boldsymbol{u}_{t+1}\rangle \cdot \boldsymbol{r}_t\right), \tag{117c}$$

where the sequence of functions $(\widehat{P}_{t,i}(\lambda))_{1\leq t, 1\leq i\leq t}$ are given by the last row of $\widehat{\mathbf{P}}_t(\lambda) \in \mathbb{R}^{t\times t}$:

$$\widehat{\mathbf{P}}_t(\lambda) = \sum_{i=1}^{t} Q_i(\lambda)\widehat{\mathbf{\Phi}}_t^{i-1}. \tag{118}$$

The above iterations can be written into an OAMP algorithm as defined in Definition 1 by introducing intermediate variables $\boldsymbol{z}_{t,i} = \widehat{P}_{t,i}(\boldsymbol{W})\bar{\boldsymbol{u}}_i$ and properly re-indexing the iterates. The only subtle point here is that the functions $(\widehat{P}_{t,i}(\lambda))_{1\leq t,1\leq i\leq t}$, which depend on the empirical divergences, are random and satisfy the required trace-free condition in OAMP only in certain asymptotical sense. Nevertheless, similar to the proof in Appendix B, we can use a simple approximation argument to show that this difference is asymptotically negligible. More specifically, we note that if we replace the functions $(\widehat{P}_{t,i}(\lambda))_{1\leq t,1\leq i\leq t}$ by $(P_{t,i}(\lambda))_{1\leq t,1\leq i\leq t}$, which are given by the last row of the following matrix, then the state evolution in Appendix B applies:

$$\mathbf{P}_t(\lambda) = \sum_{i=1}^{t} Q_i(\lambda)\mathbf{\Phi}_t^{i-1}, \tag{119}$$

where the empirical divergences $\widehat{\mathbf{\Phi}}_t$ are replaced by their limits $\mathbf{\Phi}_t := \operatorname*{plim}_{N\to\infty} \widehat{\mathbf{\Phi}}_t$. (Note that the above limit holds by an inductive argument: we assume $\widehat{\mathbf{\Phi}}_t \xrightarrow{\mathbb{P}} \mathbf{\Phi}_t$, then prove the state evolution using our following arguments, and then show $\widehat{\mathbf{\Phi}}_{t+1} \xrightarrow{\mathbb{P}} \mathbf{\Phi}_{t+1}$.) This matrix satisfies the trace-free condition: $\mathbb{E}_{\Lambda\sim\mu}[\mathbf{P}_t(\Lambda)] = \mathbf{0}_{t\times t}$.

Then, we bound the approximation error between these two versions of OAMP, based on the same arguments used in Appendix B. We do not repeat the full argument, but it suffices to show $\|P_{t,i}(\boldsymbol{W}) - \widehat{P}_{t,i}(\boldsymbol{W})\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0$ for all $i \in [t]$. Note that $\widehat{P}_{t,i}(\lambda)$ and $P_{t,i}(\lambda)$ are linear combinations of $(Q_i(\lambda))_{i\in[t]}$, which we denote as

$$\widehat{P}_{t,i}(\lambda) := \widehat{\alpha}_{t,1}Q_1(\lambda) + \widehat{\alpha}_{t,2}Q_2(\lambda) + \cdots + \widehat{\alpha}_{t,t}Q_t(\lambda), \tag{120a}$$
$$P_{t,i}(\lambda) := \alpha_{t,1}Q_1(\lambda) + \alpha_{t,2}Q_2(\lambda) + \cdots + \alpha_{t,t}Q_t(\lambda). \tag{120b}$$

Moreover, from the fact that $\widehat{\boldsymbol{\Phi}}_t \xrightarrow{\mathbb{P}} \boldsymbol{\Phi}_t$, we have $\widehat{\alpha}_{t,i} \xrightarrow{\mathbb{P}} \alpha_{t,i}, \forall i = 1, \ldots, t$. Then,

$$\|P_{t,i}(\boldsymbol{W}) - \widehat{P}_{t,i}(\boldsymbol{W})\|_{\mathrm{op}} = \left\|\sum_{i=1}^{t}(\alpha_{t,i} - \widehat{\alpha}_{t,i})Q_i(\boldsymbol{W})\right\|_{\mathrm{op}} \leq \sum_{i=1}^{t}|\alpha_{t,i} - \widehat{\alpha}_{t,i}| \cdot \|Q_i(\boldsymbol{W})\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0, \tag{121a}$$

where the last step is due to the assumption that $\|\boldsymbol{W}\|_{\mathrm{op}}$ is bounded by an $N$-independent constant (see Assumption 2) and $(Q_i)_{i\in[t]}$ are continuous functions (in the current case, polynomials).

**Covariance matrix in state evolution:** From the above arguments, we can apply Theorem 1 to show that the empirical distributions of $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t)$ converges to a joint Gaussian distribution with zero mean and covariance

$$\boldsymbol{\Sigma}_t \overset{\text{(a)}}{=} \mathbb{E}\left[\mathsf{P}_t(\Lambda)\begin{bmatrix}\bar{\mathsf{U}}_1\\ \vdots \\ \bar{\mathsf{U}}_t\end{bmatrix}\left[\bar{\mathsf{U}}_1, \ldots, \bar{\mathsf{U}}_t\right]\mathsf{P}_t(\Lambda)^{\mathsf{T}}\right], \quad \Lambda \sim \mu, \tag{122a}$$
$$\overset{\text{(b)}}{=} \mathbb{E}\left[\mathsf{P}_t(\Lambda)\bar{\boldsymbol{\Delta}}_t\mathsf{P}_t(\Lambda)^{\mathsf{T}}\right] \tag{122b}$$
$$\tag{122c}$$

where step (a) follows Theorem 1 with the empirical divergences replaced by the population-level divergences, namely (cf. (70b)), $\mathsf{P}_t(\lambda) = \sum_{i=1}^{t} Q_i(\lambda)\boldsymbol{\Phi}_t^{i-1}$; step (b) is from the fact that the state evolution random variables $(\bar{\mathsf{U}}_1, \ldots, \bar{\mathsf{U}}_t)$ are independent of $\Lambda$.

## C.4 Equivalence of State Evolution (Proof of Proposition 3)

The following lemma will be used in our proof of Proposition 3.

**Lemma 7.** *Let $(Q_n)_{n\geq 1}$ and be defined as in (5). Then, the following holds*

$$\mathbb{E}\left[Q_I Q_J\right] = \sum_{m=1}^{I}\sum_{n=1}^{J}\mathbb{E}\left[Q_{I-m}Q_{J-n}\right] \cdot \kappa_{n+m}, \quad \forall I, J \geq 1. \tag{123}$$

*Proof.* From Proposition 1, (5) can be rewritten as

$$Q_n = \Lambda Q_{n-1} - \sum_{i=1}^{n}\kappa_i \cdot Q_{n-i}, \quad \forall n \geq 1, \tag{124}$$

with $Q_0 := 1$. Using the above identity, we can relate $Q_I Q_J$ and $Q_{I-1}Q_{J+1}$ as follows:

$$Q_I Q_J \overset{\text{(a)}}{=} \left(\Lambda Q_{I-1} - \sum_{i=1}^{I} \kappa_i Q_{I-i}\right) Q_J \tag{125a}$$

$$= \Lambda Q_{I-1}Q_J - \sum_{i=1}^{I} \kappa_i Q_{I-i}Q_J \tag{125b}$$

$$= \Lambda Q_{I-1}Q_J - \sum_{j=1}^{J+1} \kappa_j Q_{I-1}Q_{J+1-j} + \sum_{j=1}^{J+1} \kappa_j Q_{I-1}Q_{J+1-j} - \sum_{i=1}^{I} \kappa_i Q_{I-i}Q_J \tag{125c}$$

$$= Q_{I-1}\Big(\Lambda Q_J - \sum_{j=1}^{J+1} \kappa_j Q_{J+1-j}\Big) + Q_{I-1}\Big(\sum_{j=1}^{J+1} \kappa_j Q_{J+1-j}\Big) - \sum_{i=1}^{I} \kappa_i Q_{I-i}Q_J \tag{125d}$$

$$\overset{\text{(b)}}{=} Q_{I-1}Q_{J+1} + Q_{I-1}\Big(\sum_{j=1}^{J+1} \kappa_j Q_{J+1-j}\Big) - \sum_{i=1}^{I} \kappa_i Q_{I-i}Q_J, \tag{125e}$$

where both step (a) and step (b) used the identity (124). We can apply the same manipulations to relate $Q_{I-1}Q_{J+1}$ and $Q_{I-2}Q_{J+2}$. Continuing for $\ell$ steps eventually us the following identity between $Q_{I-1}Q_{J+1}$ and $Q_{I-\ell}Q_{J+\ell}$:

$$Q_I Q_J = Q_{I-\ell}Q_{J+\ell} + \sum_{k=1}^{\ell}\sum_{j=1}^{J+k} \kappa_j Q_{I-k}Q_{J+k-j} - \sum_{k=1}^{\ell}\sum_{i=1}^{I-k} \kappa_i Q_{I-k+1-i}Q_{J+k-1}. \tag{126}$$

Setting $\ell = I$ in the above identity yields

$$Q_I Q_J = Q_0 Q_{J+I} + \sum_{k=1}^{I}\sum_{j=1}^{J+k} \kappa_j Q_{I-k}Q_{J+k-j} - \sum_{k=1}^{I}\sum_{i=1}^{I-k} \kappa_i Q_{I-k+1-i}Q_{J+k-1} \tag{127a}$$

$$= Q_{J+I} + \sum_{k=1}^{I}\sum_{j=1}^{J+k} \kappa_j Q_{I-k}Q_{J+k-j} - \sum_{k=1}^{I}\sum_{i=1}^{I-k} \kappa_i Q_{I-k+1-i}Q_{J+k-1} \quad (Q_0 = 1) \tag{127b}$$

$$= Q_{J+I} + \underbrace{\sum_{k=1}^{I}\sum_{j=k+1}^{J+k} \kappa_j Q_{I-k}Q_{J+k-j}}_{\textsf{Term I}} + \underbrace{\sum_{k=1}^{I}\sum_{j=1}^{k} \kappa_j Q_{I-k}Q_{J+k-j} - \sum_{k=1}^{I}\sum_{i=1}^{I-k} \kappa_i Q_{I-k+1-i}Q_{J+k-1}}_{\textsf{Term II}}, \tag{127c}$$

where in the last step we split the sum over $j$ into two terms. By a change of variable, we rewrite Term I as

$$\textsf{Term I} = \sum_{k=1}^{I}\sum_{j'=1}^{J} \kappa_{k+j'} Q_{I-k}Q_{J-j'}. \tag{128}$$

Note that Term II involves $(Q_i)_{J+1\le i\le I+J-1}$, which are higher order terms that do not appear in the desired result. It turns out that Term II vanishes:

$$\textsf{Term II} = \sum_{k=1}^{I}\sum_{j=1}^{k} \kappa_j Q_{I-k}Q_{J+k-j} - \sum_{k=1}^{I}\sum_{i=1}^{I-k} \kappa_i Q_{I-k+1-i}Q_{J+k-1} \tag{129a}$$

$$\overset{\text{(a)}}{=} \sum_{k'=1}^{I}\sum_{i=1}^{I-k'} \kappa_i Q_{I-(k'+i-1)}Q_{J+k'-1} - \sum_{k=1}^{I}\sum_{i=1}^{I-k} \kappa_i Q_{I-k+1-i}Q_{J+k-1}, \tag{129b}$$

$$= 0, \tag{129c}$$

where step (a) is due to a change of variable $(k, j) \mapsto (k', i)$ via the map

$$\begin{bmatrix} k' \\ i \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} k \\ j \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Note that the map is one-to-one from $\{(k,j): 1 \le k \le I, 1 \le j \le k\}$ to $\{(k',i): 1 \le k' \le I, 1 \le i \le I-k'\}$, and hence the reformulation of the nested sums holds.

Summarizing (127)-(129) yields

$$Q_I Q_J = Q_{J+I} + \sum_{k=1}^{I} \sum_{j=1}^{J} \kappa_{k+j} Q_{I-k} Q_{J-j}. \tag{130}$$

The claimed identity then follows from taking expectation in the above identity and recalling that the random variables $(Q_i)_{i\ge 1}$ have zero mean. □

**Proof of Proposition 3.** Recall the following definitions that will be used in our proof:

$$\boldsymbol{\Sigma}_t := \begin{bmatrix} \mathbb{E}\left[\mathsf{R}_1^2\right] & \mathbb{E}\left[\mathsf{R}_1\mathsf{R}_2\right] & \cdots & \mathbb{E}\left[\mathsf{R}_1\mathsf{R}_t\right] \\ \mathbb{E}\left[\mathsf{R}_2\mathsf{R}_1\right] & \mathbb{E}\left[\mathsf{R}_2^2\right] & \cdots & \mathbb{E}\left[\mathsf{R}_2\mathsf{R}_t\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\mathsf{R}_t\mathsf{R}_1\right] & \mathbb{E}\left[\mathsf{R}_t\mathsf{R}_2\right] & \cdots & \mathbb{E}\left[\mathsf{R}_t\mathsf{R}_t\right] \end{bmatrix}, \tag{131a}$$

$$\boldsymbol{\Delta}_t := \begin{bmatrix} \mathbb{E}\left[\mathsf{U}_1^2\right] & \mathbb{E}\left[\mathsf{U}_1\mathsf{U}_2\right] & \cdots & \mathbb{E}\left[\mathsf{U}_1\mathsf{U}_t\right] \\ \mathbb{E}\left[\mathsf{U}_2\mathsf{U}_1\right] & \mathbb{E}\left[\mathsf{U}_2^2\right] & \cdots & \mathbb{E}\left[\mathsf{U}_2\mathsf{U}_t\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\mathsf{U}_t\mathsf{U}_1\right] & \mathbb{E}\left[\mathsf{U}_t\mathsf{U}_2\right] & \cdots & \mathbb{E}\left[\mathsf{U}_t\mathsf{U}_t\right] \end{bmatrix}, \tag{131b}$$

$$\bar{\boldsymbol{\Delta}}_t := \begin{bmatrix} \mathbb{E}\left[\bar{\mathsf{U}}_1^2\right] & \mathbb{E}\left[\bar{\mathsf{U}}_1\bar{\mathsf{U}}_2\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{U}}_1\bar{\mathsf{U}}_t\right] \\ \mathbb{E}\left[\bar{\mathsf{U}}_2\bar{\mathsf{U}}_1\right] & \mathbb{E}\left[\bar{\mathsf{U}}_2^2\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{U}}_2\bar{\mathsf{U}}_t\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\bar{\mathsf{U}}_t\bar{\mathsf{U}}_1\right] & \mathbb{E}\left[\bar{\mathsf{U}}_t\bar{\mathsf{U}}_2\right] & \cdots & \mathbb{E}\left[\bar{\mathsf{U}}_t\bar{\mathsf{U}}_t\right] \end{bmatrix}, \tag{131c}$$

$$\boldsymbol{\Phi}_t := \begin{bmatrix} 0 & & & & \\ \mathbb{E}\left[\partial_1\mathsf{U}_2\right] & 0 & & & \\ \mathbb{E}\left[\partial_1\mathsf{U}_3\right] & \mathbb{E}\left[\partial_2\mathsf{U}_3\right] & 0 & & \\ \vdots & \vdots & \ddots & & \\ \mathbb{E}\left[\partial_1\mathsf{U}_t\right] & \mathbb{E}\left[\partial_2\mathsf{U}_t\right] & \cdots & \mathbb{E}\left[\partial_{t-1}\mathsf{U}_t\right] & 0 \end{bmatrix}, \tag{131d}$$

where $(\mathsf{R}_1, \ldots, \mathsf{R}_{j-1})$ are jointly Gaussian with zero mean, and we denoted

$$\mathbb{E}\left[\partial_i \mathsf{U}_j\right] := \mathbb{E}\left[\partial_i \eta_j(\mathsf{R}_1, \ldots, \mathsf{R}_{j-1})\right], \quad \forall j > 1, i \in [j-1]. \tag{131e}$$

Recall that the covariance $\boldsymbol{\Sigma}_t$ in (51a) is expressed using $\bar{\boldsymbol{\Delta}}_t$, while the covariance in (52) is expressed using $\boldsymbol{\Delta}_t$, where $\bar{\boldsymbol{\Delta}}_t$ and $\boldsymbol{\Delta}_t$ denote the covariance of the state evolution random variables $(\bar{\mathsf{U}}_j)_{j\in[t]}$ and $(\mathsf{U}_j)_{j\in[t]}$ respectively (see (51d) and (52b)):

To related these two expressions, recall the definition

$$\mathsf{U}_j = \bar{\mathsf{U}}_j + \sum_{i=1}^{j-1} \mathbb{E}\left[\partial_i \mathsf{U}_j\right] \cdot \mathsf{R}_i, \quad \forall j > 1, \tag{132}$$

which can be written into a matrix form

$$\begin{bmatrix} \mathsf{U}_1 \\ \mathsf{U}_2 \\ \vdots \\ \mathsf{U}_t \end{bmatrix} = \begin{bmatrix} \bar{\mathsf{U}}_1 \\ \bar{\mathsf{U}}_2 \\ \vdots \\ \bar{\mathsf{U}}_t \end{bmatrix} + \begin{bmatrix} 0 & & & & \\ \mathbb{E}\left[\partial_1\mathsf{U}_2\right] & 0 & & & \\ \mathbb{E}\left[\partial_1\mathsf{U}_3\right] & \mathbb{E}\left[\partial_2\mathsf{U}_3\right] & 0 & & \\ \vdots & \vdots & \ddots & & \\ \mathbb{E}\left[\partial_1\mathsf{U}_t\right] & \mathbb{E}\left[\partial_2\mathsf{U}_t\right] & \cdots & \mathbb{E}\left[\partial_{t-1}\mathsf{U}_t\right] & 0 \end{bmatrix} \begin{bmatrix} \mathsf{R}_1 \\ \mathsf{R}_2 \\ \vdots \\ \mathsf{R}_t \end{bmatrix} \tag{133a}$$

$$:= \begin{bmatrix} \bar{\mathsf{U}}_1 \\ \bar{\mathsf{U}}_2 \\ \vdots \\ \bar{\mathsf{U}}_t \end{bmatrix} + \boldsymbol{\Phi}_t \begin{bmatrix} \mathsf{R}_1 \\ \mathsf{R}_2 \\ \vdots \\ \mathsf{R}_t \end{bmatrix}. \tag{133b}$$

From Proposition 2, we have the following orthogonality property:

$$\mathbb{E}\left[\bar{\mathsf{U}}_j \mathsf{R}_i\right] = 0, \quad \forall i, j \geq 1. \tag{134}$$

Following (131), (133) and (134), the covariance matrices $\boldsymbol{\Delta}_t$ and $\bar{\boldsymbol{\Delta}}_t$ satisfy the following relation:

$$\boldsymbol{\Delta}_t = \bar{\boldsymbol{\Delta}}_t + \boldsymbol{\Phi}_t \boldsymbol{\Sigma}_t \boldsymbol{\Phi}_t^{\mathsf{T}}. \tag{135}$$

Using (135), we can rewrite (51b) as

$$\begin{aligned}
\boldsymbol{\Sigma}_t &= \mathbb{E}\left[\mathbf{P}_t(\Lambda) \bar{\boldsymbol{\Delta}}_t \mathbf{P}_t(\Lambda)^{\mathsf{T}}\right] && \text{(136a)}\\
&= \mathbb{E}\left[\left(\sum_{i=1}^{t} Q_i(\Lambda) \boldsymbol{\Phi}_t^{i-1}\right) \bar{\boldsymbol{\Delta}}_t \left(\sum_{i=1}^{t} Q_i(\Lambda) \boldsymbol{\Phi}_t^{i-1}\right)^{\mathsf{T}}\right] && \text{(136b)}\\
&= \sum_{i=1}^{t}\sum_{j=1}^{t} \mathbb{E}_{\Lambda\sim\mu}\left[Q_i(\Lambda) Q_j(\Lambda)\right] \cdot \boldsymbol{\Phi}_t^{i-1} \bar{\boldsymbol{\Delta}}_t \left(\boldsymbol{\Phi}_t^{j-1}\right)^{\mathsf{T}} && \text{(136c)}\\
&:= \sum_{i=1}^{t}\sum_{j=1}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i-1} \bar{\boldsymbol{\Delta}}_t \left(\boldsymbol{\Phi}_t^{j-1}\right)^{\mathsf{T}} && \text{(136d)}\\
&= \sum_{i=1}^{t}\sum_{j=1}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i-1} \left(\boldsymbol{\Delta}_t - \boldsymbol{\Phi}_t \boldsymbol{\Sigma}_t \boldsymbol{\Phi}_t^{\mathsf{T}}\right) \left(\boldsymbol{\Phi}_t^{j-1}\right)^{\mathsf{T}} && \text{(136e)}\\
&= \sum_{i=1}^{t}\sum_{j=1}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i-1} \boldsymbol{\Delta}_t \left(\boldsymbol{\Phi}_t^{j-1}\right)^{\mathsf{T}} - \sum_{i=1}^{t}\sum_{j=1}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i} \boldsymbol{\Sigma}_t \left(\boldsymbol{\Phi}_t^{j}\right)^{\mathsf{T}}, && \text{(136f)}
\end{aligned}$$

where for convenience we denoted

$$\Omega_{i,j} := \mathbb{E}_{\Lambda\sim\mu}\left[Q_i(\Lambda) Q_j(\Lambda)\right], \quad \forall i \geq 0, j \geq 0. \tag{136g}$$

Since $Q_0 = 1$, we can write $\boldsymbol{\Sigma}_t = \mathbb{E}_{\Lambda\sim\mu}\left[Q_0(\Lambda) Q_0(\Lambda)\right] \cdot \boldsymbol{\Phi}_t^0 \boldsymbol{\Sigma}_t \left(\boldsymbol{\Phi}_t^0\right)^{\mathsf{T}}$. Moreover, $\mathbb{E}[Q_i(\Lambda)] = 0$, $\forall i \geq 1$. Hence,

$$\boldsymbol{\Sigma}_t + \sum_{i=1}^{t}\sum_{j=1}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i} \boldsymbol{\Sigma}_t \left(\boldsymbol{\Phi}_t^{j}\right)^{\mathsf{T}} = \sum_{i=0}^{t}\sum_{j=0}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i} \boldsymbol{\Sigma}_t \left(\boldsymbol{\Phi}_t^{j}\right)^{\mathsf{T}}. \tag{137}$$

Therefore, we can rewrite (136) as follows

$$\sum_{i=0}^{t}\sum_{j=0}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i} \boldsymbol{\Sigma}_t \left(\boldsymbol{\Phi}_t^{j}\right)^{\mathsf{T}} = \sum_{i=1}^{t}\sum_{j=1}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^{i-1} \boldsymbol{\Delta}_t \left(\boldsymbol{\Phi}_t^{j-1}\right)^{\mathsf{T}}. \tag{138}$$

We treat (138) as an equation of $\boldsymbol{\Sigma}_t$, with $(\boldsymbol{\Phi}_t, \boldsymbol{\Delta}_t)$ being fixed matrices. We remark that this equation has a unique solution. To see this, we use the identity $\mathrm{vec}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}^{\mathsf{T}}) = (\boldsymbol{B} \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{X})$. Then, a vectorization of the LHS of (138) reads

$$\left(\boldsymbol{I}_{t^2} + \boldsymbol{I}_t \otimes \boldsymbol{\Phi}_t + \boldsymbol{\Phi}_t \otimes \boldsymbol{I}_t + \boldsymbol{I}_t \otimes \boldsymbol{\Phi}_t^2 + \cdots\right) \mathrm{vec}(\boldsymbol{\Sigma}_t). \tag{139}$$

Note that $\boldsymbol{\Phi}_t$ is strictly lower triangular. Hence, the matrix $\left(\boldsymbol{I}_{t^2} + \boldsymbol{I}_t \otimes \boldsymbol{\Phi}_t + \boldsymbol{\Phi}_t \otimes \boldsymbol{I}_t + \boldsymbol{I}_t \otimes \boldsymbol{\Phi}_t^2 + \cdots\right)$ is lower triangular with diagonal elements all equal to one. Therefore, $\left(\boldsymbol{I}_{t^2} + \boldsymbol{I}_t \otimes \boldsymbol{\Phi}_t + \boldsymbol{\Phi}_t \otimes \boldsymbol{I}_t + \boldsymbol{I}_t \otimes \boldsymbol{\Phi}_t^2 + \cdots\right)$ is invertible and hence (138) has a unique solution.

Next, we verify that one solution (and hence the only one) to (138) is

$$\boldsymbol{\Sigma}_t = \sum_{j=0}^{\infty}\sum_{i=0}^{j} \kappa_{j+2} \boldsymbol{\Phi}_t^{i} \boldsymbol{\Delta}_t \left((\boldsymbol{\Phi}_t)^{j-i}\right)^{\mathsf{T}}, \tag{140}$$

which is precisely (52). This would conclude the equivalence between (51b) and (52) which we aim to prove. We first make a change-of-variable $(i,j) \mapsto (m,n)$ in (140) via

$$\begin{bmatrix} i \\ j \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} m \\ n \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix}. \tag{141}$$

This map is bijective from $\{(i,j) : 0 \le j < \infty, 0 \le i \le j\}$ to $\{(m,n) : 1 \le m < \infty, 1 \le n < \infty\}$. We then write (140) as

$$\boldsymbol{\Sigma}_t = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \kappa_{m+n} \boldsymbol{\Phi}_t^{m-1} \boldsymbol{\Delta}_t \left( (\boldsymbol{\Phi}_t)^{n-1} \right)^{\mathsf{T}} = \sum_{m=1}^{t} \sum_{n=1}^{t} \kappa_{m+n} \boldsymbol{\Phi}_t^{m-1} \boldsymbol{\Delta}_t \left( (\boldsymbol{\Phi}_t)^{n-1} \right)^{\mathsf{T}}, \tag{142}$$

where the second step is due to the fact that $\boldsymbol{\Phi}_t$ strictly lower triangular. Towards proving that (142) is a solution to (138), we substitute (140) into the LHS of (138):

$$\begin{aligned}
\sum_{i=0}^{t} \sum_{j=0}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^i \boldsymbol{\Sigma}_t \left( \boldsymbol{\Phi}_t^j \right)^{\mathsf{T}} &= \sum_{i=0}^{t} \sum_{j=0}^{t} \Omega_{i,j} \boldsymbol{\Phi}_t^i \left( \sum_{m=1}^{t} \sum_{n=1}^{t} \kappa_{m+n} \boldsymbol{\Phi}_t^{m-1} \boldsymbol{\Delta}_t \left( (\boldsymbol{\Phi}_t)^{n-1} \right)^{\mathsf{T}} \right) \left( \boldsymbol{\Phi}_t^j \right)^{\mathsf{T}} && \text{(143a)} \\
&= \sum_{i=0}^{t} \sum_{j=0}^{t} \sum_{m=1}^{t} \sum_{n=1}^{t} \Omega_{i,j} \kappa_{m+n} \boldsymbol{\Phi}_t^{i+m-1} \boldsymbol{\Delta}_t \left( (\boldsymbol{\Phi}_t)^{n+j-1} \right)^{\mathsf{T}} && \text{(143b)} \\
&\overset{\text{(a)}}{=} \sum_{(I,J) \in [t] \times [t]} \left( \sum_{(m',n') \in [I] \times [J]} \Omega_{I-m',J-n'} \cdot \kappa_{m'+n'} \right) \cdot \boldsymbol{\Phi}_t^{I-1} \boldsymbol{\Delta}_t \left( \boldsymbol{\Phi}_t^{J-1} \right)^{\mathsf{T}} && \text{(143c)} \\
&\overset{\text{(b)}}{=} \sum_{(I,J) \in [t] \times [t]} \Omega_{I,J} \cdot \boldsymbol{\Phi}_t^{I-1} \boldsymbol{\Delta}_t \left( \boldsymbol{\Phi}_t^{J-1} \right)^{\mathsf{T}}, && \text{(143d)}
\end{aligned}$$

where step (a) is due to a change of variable $(m,i) \mapsto (m',I)$ and $(n,j) \mapsto (n',J)$ via the map:

$$\begin{aligned}
m' &= m, \quad n' = n, \\
I &= m + i, \quad J = n + j,
\end{aligned}$$

and step (b) follows from Lemma 7 (and the definition $\Omega_{i,j} = \mathbb{E}\left[Q_i Q_j\right]$). We now recognize that (143) is identical to the RHS of (138). This concludes our proof of Proposition 3.

# Appendix D RI-AMP-DF Algorithm

## D.1 Reduction of RI-AMP-DF to OAMP (Proof of Theorem 3)

**Proof of item (1):** The de-biasing matrix (54) in RI-AMP-DF can be derived following the calculations in Appendix C.1. We collect the iterates of RI-AMP-DF into the following form (cf. (53)):

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{W}\boldsymbol{u}_1 \\ \boldsymbol{W}\boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{W}\boldsymbol{u}_t \end{bmatrix} - \begin{bmatrix} \mathsf{c}_{1,1}\boldsymbol{I}_N & & & \\ \mathsf{c}_{2,1}\boldsymbol{I}_N & \mathsf{b}_{2,2}\boldsymbol{I}_N & & \\ \vdots & \vdots & \ddots & \\ \mathsf{c}_{t,1}\boldsymbol{I}_N & \mathsf{b}_{t,2}\boldsymbol{I}_N & \cdots & \mathsf{b}_{t,t}\boldsymbol{I}_N \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix} \tag{144a}$$

$$= (\mathsf{I}_t \otimes \boldsymbol{W}) \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_t \end{bmatrix} - (\mathsf{C}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix} \tag{144b}$$

$$\overset{\text{(a)}}{=} (\mathsf{I}_t \otimes \boldsymbol{W}) \left( \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix} + (\widehat{\boldsymbol{\Phi}}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} \right) - (\mathsf{C}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix} \tag{144c}$$

$$= (\widehat{\boldsymbol{\Phi}}_t \otimes \boldsymbol{W}) \begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} + (\mathsf{I}_t \otimes \boldsymbol{W} - \mathsf{C}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}, \tag{144d}$$

where step (a) is due to the relationship between $\boldsymbol{u}_t$ and $\bar{\boldsymbol{u}}_t$ (which follows the same calculation as (98) but with $\mathsf{D}_t$ replaced by $\widehat{\boldsymbol{\Phi}}_t$). Then, following the procedure in Appendix C.1, it can be shown that

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_t \end{bmatrix} = \left( \boldsymbol{I}_{tN} - \widehat{\boldsymbol{\Phi}}_t \otimes \boldsymbol{W} \right)^{-1} (\mathsf{I}_t \otimes \boldsymbol{W} - \mathsf{C}_t \otimes \boldsymbol{I}_N) \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix} \tag{145a}$$

$$:= \begin{bmatrix} \widehat{G}_{1,1}(\boldsymbol{W}) & & & \\ \widehat{G}_{2,1}(\boldsymbol{W}) & \widehat{G}_{2,2}(\boldsymbol{W}) & & \\ \vdots & \vdots & \ddots & \\ \widehat{G}_{t,1}(\boldsymbol{W}) & \widehat{G}_{t,2}(\boldsymbol{W}) & \cdots & \widehat{G}_{t,t}(\boldsymbol{W}) \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{u}}_1 \\ \bar{\boldsymbol{u}}_2 \\ \vdots \\ \bar{\boldsymbol{u}}_t \end{bmatrix}, \tag{145b}$$

where $(\widehat{G}_{i,j})_{i\in[t],j\in[t]}$ is a sequence of polynomials defined as

$$\widehat{\mathsf{G}}_t(\lambda) := \begin{bmatrix} \widehat{G}_{1,1}(\lambda) & 0 & \cdots & 0 \\ \widehat{G}_{2,1}(\lambda) & \widehat{G}_{2,2}(\lambda) & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ \widehat{G}_{t,1}(\lambda) & \widehat{G}_{t,1}(\lambda) & \cdots & \widehat{G}_{t,t}(\lambda) \end{bmatrix} = (\mathsf{I}_t - \lambda\widehat{\boldsymbol{\Phi}}_t)^{-1}(\lambda\mathsf{I}_t - \mathsf{C}_t), \quad \forall \lambda \in \mathbb{R}. \tag{145c}$$

Following Lemma 2, we set the de-biasing matrix $\mathsf{C}_t$ as the solution to the equation

$$\mathbb{E}\left[\widehat{\mathsf{G}}_t(\Lambda)\right] = \mathbf{0}_{t\times t}, \tag{146}$$

where $\widehat{\mathsf{G}}_t(\lambda)$ is defined in (145c) and the expectation in (146) is taken w.r.t. $\Lambda \sim \mu$ independent of $\widehat{\boldsymbol{\Phi}}_t$.

Enforcing the above condition yields

$$\mathbf{C}_t = \left(\mathbb{E}\left[(\mathbf{I}_t - \Lambda\widehat{\mathbf{\Phi}}_t)^{-1}\right]\right)^{-1} \mathbb{E}\left[\Lambda(\mathbf{I}_t - \Lambda\widehat{\mathbf{\Phi}}_t)^{-1}\right] \tag{147a}$$

$$\overset{\text{(a)}}{=} \left(\sum_{i=1}^{t} m_{i-1}\widehat{\mathbf{\Phi}}_t^{i-1}\right)^{-1} \left(\sum_{i=1}^{t} m_i\widehat{\mathbf{\Phi}}_t^{i-1}\right), \tag{147b}$$

where $m_i := \mathbb{E}[\Lambda^i]$, $\forall i = 0, 1, \ldots$, and step (a) is due to the expansion

$$(\mathbf{I}_t - \Lambda\widehat{\mathbf{\Phi}}_t)^{-1} = \sum_{i=1}^{t} \Lambda^{i-1}\widehat{\mathbf{\Phi}}_t^{i-1}.$$

(Recall that $\widehat{\mathbf{\Phi}}_t$ is strictly lower triangular.) From (147), $\mathbf{C}_t$ is clearly a polynomial of $\widehat{\mathbf{\Phi}}_t$. Further, from (145c), $\widehat{\mathbf{G}}_t(\lambda)$ is also a polynomial of $\widehat{\mathbf{\Phi}}_t$. In the following, we apply the trick in the proof of Lemma 2-(2) to derive a recursive formula for the coefficients in this polynomial representations of $\mathbf{C}_t$ and $\widehat{\mathbf{G}}_t(\lambda)$.

Let us denote

$$\mathbf{C}_t := \sum_{i=1}^{t} \gamma_i\widehat{\mathbf{\Phi}}_t^{i-1}, \tag{148a}$$

$$\widehat{\mathbf{G}}_t(\lambda) := \sum_{i=1}^{t} H_i(\lambda)\widehat{\mathbf{\Phi}}_t^{i-1}. \tag{148b}$$

From (145c), the following equation holds

$$(\mathbf{I}_t - \lambda\widehat{\mathbf{\Phi}}_t)\widehat{\mathbf{G}}_t(\lambda) - \lambda\mathbf{I}_t + \mathbf{C}_t = \mathbf{0}_{t\times t}. \tag{149}$$

Using the polynomial representations of $\mathbf{C}_t$ and $\widehat{\mathbf{G}}_t(\lambda)$ in (148), we write the LHS of the above equation as

$$(\mathbf{I}_t - \lambda\widehat{\mathbf{\Phi}}_t)\widehat{\mathbf{G}}_t(\lambda) - \lambda\mathbf{I}_t + \mathbf{C}_t \tag{150a}$$

$$= \left(\mathbf{I}_t - \lambda\widehat{\mathbf{\Phi}}_t\right)\sum_{i=1}^{\infty} H_i(\lambda)\widehat{\mathbf{\Phi}}_t^{i-1} - \lambda\mathbf{I}_t + \sum_{i=1}^{\infty}\gamma_i\widehat{\mathbf{\Phi}}_t^{i-1} \tag{150b}$$

$$= \sum_{i=1}^{\infty} H_i(\lambda)\widehat{\mathbf{\Phi}}_t^{i-1} - \sum_{i=1}^{\infty}\lambda H_i(\lambda)\widehat{\mathbf{\Phi}}_t^{i} - \lambda\mathbf{I}_t + \sum_{i=1}^{\infty}\gamma_i\widehat{\mathbf{\Phi}}_t^{i-1} \tag{150c}$$

$$= \sum_{i=1}^{\infty} H_i(\lambda)\widehat{\mathbf{\Phi}}_t^{i-1} - \sum_{i=2}^{\infty}\lambda H_{i-1}(\lambda)\widehat{\mathbf{\Phi}}_t^{i-1} - \lambda\mathbf{I}_t + \sum_{i=1}^{\infty}\gamma_i\widehat{\mathbf{\Phi}}_t^{i-1} \tag{150d}$$

$$= (H_1(\lambda) - \lambda + \gamma_1)\mathbf{I}_t + \sum_{i=2}^{\infty}\left(H_i(\lambda) - \lambda H_{i-1}(\lambda) + \gamma_i\right)\widehat{\mathbf{\Phi}}_t^{i-1}. \tag{150e}$$

Since the above equation is identically zero for any $\widehat{\mathbf{\Phi}}_t$, we must have

$$H_1(\lambda) = \lambda - \gamma_1, \tag{151a}$$

$$H_i(\lambda) = \lambda H_{i-1}(\lambda) - \gamma_i, \quad \forall i \geq 2. \tag{151b}$$

This provides a recursive definition of $(H_i(\lambda))_{i\geq 1}$ in terms of the sequence $(\gamma_i)_{i\geq 1}$. On the other hand, $(\gamma_i)_{i\geq 1}$ are set such that $(H_i(\Lambda))_{i\geq 1}$ have zero means w.r.t. $\Lambda \in \mu$. Hence,

$$\gamma_1 = \mathbb{E}[\Lambda], \tag{152a}$$

$$\gamma_i = \mathbb{E}[\lambda H_{i-1}(\Lambda)], \quad \forall i \geq 2. \tag{152b}$$

To summarize, we have the following recursive representations of $(H_i(\lambda))_{i\geq 1}$:

$$H_i(\lambda) = \lambda H_{i-1}(\lambda) - \mathbb{E}\left[\lambda H_{i-1}(\lambda)\right], \quad \forall i \geq 1, \tag{153}$$

where $H_0(\lambda) := 1$. The sequences $(\gamma_i)_{i\geq 1}$ are given by

$$\gamma_i = \mathbb{E}[\lambda H_{i-1}(\Lambda)], \quad \forall i \geq 1. \tag{154}$$

**Proof of item (2):** The state evolution of RI-AMP-DF readily follows from the master state evolution result of OAMP. Its proof is similar to that of Theorem 2 and omitted.

## D.2 Reduction of GFOM to RI-AMP-DF

We first recall the definition of generalized first order method (GFOM) introduced in [10, 39].

**Definition 9** (GFOM [10, 39])**.** A generalized first order method (GFOM) generates the iterates $(\boldsymbol{x}_t)_{t\geq 1}$ via

$$\boldsymbol{x}_t = \boldsymbol{W}\phi_t(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{t-1};\boldsymbol{a}) + \psi_t(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{t-1};\boldsymbol{a}), \quad \forall t\geq 1. \tag{155}$$

At each iteration, the output $\widehat{\boldsymbol{x}}_t$ is generated by further applying a post-processing:

$$\widehat{\boldsymbol{x}}_t = h_t\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t;\boldsymbol{a}\right). \tag{156}$$

In the above equations, the functions $\phi_t : \mathbb{R}^{t-1}\times\mathbb{R}^k \mapsto \mathbb{R}$, $\psi_t : \mathbb{R}^{t-1}\times\mathbb{R}^k \mapsto \mathbb{R}$ and $h_t : \mathbb{R}^{t}\times\mathbb{R}^k \mapsto \mathbb{R}$, are all continuously-differentiable and Lipschitz. Further, they all act on the $N$ components of their input vectors separately. Moreover, these functions do not depend on the dimension $N$.

For the purpose of establishing a precise reduction result, we make a minor change to the RI-AMP-DF algorithm: we replace the divergence terms and the de-biasing terms by their limiting deterministic equivalents. We call it deterministic RI-AMP-DF.

**Definition 10** (Deterministic RI-AMP-DF)**.** Let $\boldsymbol{u}_1 = \bar{\boldsymbol{u}}_1 \in \mathbb{R}^N$ and generate $(\boldsymbol{r}_t)_{t\geq 1}$ through

$$\boldsymbol{r}_t = \boldsymbol{W}\boldsymbol{u}_t - \left(\mathsf{c}_{t,1}\bar{\boldsymbol{u}}_1 + \mathsf{c}_{t,2}\bar{\boldsymbol{u}}_2 + \cdots + \mathsf{c}_{t,t}\bar{\boldsymbol{u}}_t\right), \quad \forall t\geq 1, \tag{157a}$$

$$\boldsymbol{u}_{t+1} = \eta_{t+1}(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_t), \tag{157b}$$

$$\bar{\boldsymbol{u}}_{t+1} = \eta_{t+1}(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_t) - \left(\mathsf{d}_{t,1}\boldsymbol{r}_1 + \cdots + \mathsf{d}_{t,t}\boldsymbol{r}_t\right), \tag{157c}$$

where, for all $t\geq 1$, $(\mathsf{c}_{t,i})_{1\leq t,1\leq i\leq t}$ and $(\mathsf{d}_{t,1})_{1\leq t,1\leq i\leq t}$ are deterministic constants that only depends on the function $(\eta_i)_{i\in[t]}$ and the limiting spectrum $\mu$, but not on the dimension $N$.

Following the idea in [10, 39], we show in the following proposition that (deterministic) RI-AMP-DF can implement any GFOM through a proper change of variables. Since RI-AMP-DF is itself a GFOM, this implies that RI-AMP-DF and GFOM belong to the same class of algorithms. It is straightforward to show that the same claim applies to the original RI-AMP algorithm.

**Proposition 4** (Deterministic RI-AMP-DF can implement any GFOM)**.** *Let $(\boldsymbol{x}_t)_{t\geq 1}$ be generated by any GFOM. There exists a deterministic RI-AMP-DF algorithm, whose iterates are denoted as $(\boldsymbol{r}_t)_{t\geq 1}$, and a post-processing function $\varphi_t : \mathbb{R}^t\times\mathbb{R}^k \mapsto \mathbb{R}$, such that the following holds:*

$$\boldsymbol{x}_t = \varphi_t(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_t;\boldsymbol{a}), \quad \forall t\geq 1.$$

*Proof.* The proof is essentially identical to [39, Lemma 4.1], and we include it here for completeness. We prove by induction. The claim clearly holds for $t=1$. Suppose it holds up to iteration $t\geq 1$, we prove that it also holds for iteration $t+1$. From (155), the new iterate $\boldsymbol{x}_{t+1}$ of GFOM reads:

$$\boldsymbol{x}_{t+1} = \boldsymbol{W}\phi_{t+1}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t;\boldsymbol{a}) + \psi_{t+1}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t;\boldsymbol{a}) \tag{158a}$$

$$\overset{\text{(a)}}{=} \boldsymbol{W}\phi_{t+1}\left(\varphi_1(\boldsymbol{r}_1;\boldsymbol{a}),\ldots,\varphi_t(\boldsymbol{r}_{\leq t};\boldsymbol{a});\boldsymbol{a}\right) + \psi_{t+1}\left(\varphi_1(\boldsymbol{r}_1;\boldsymbol{a}),\ldots,\varphi_t(\boldsymbol{r}_{\leq t};\boldsymbol{a});\boldsymbol{a}\right) \tag{158b}$$

$$\overset{\text{(b)}}{=} \boldsymbol{W}f_{t+1}(\boldsymbol{r}_{\leq t};\boldsymbol{a}) + \psi_{t+1}\left(\varphi_1(\boldsymbol{r}_1;\boldsymbol{a}),\ldots,\varphi_t(\boldsymbol{r}_{\leq t};\boldsymbol{a});\boldsymbol{a}\right) \tag{158c}$$

$$\overset{\text{(c)}}{=} \underbrace{\boldsymbol{W}\boldsymbol{u}_{t+1} - \sum_{i=1}^{t}\mathsf{c}_{t,i}\boldsymbol{r}_i}_{\boldsymbol{r}_{t+1}} + \sum_{i=1}^{t}\mathsf{c}_{t,i}\boldsymbol{r}_i + \psi_{t+1}\left(\varphi_1(\boldsymbol{r}_1;\boldsymbol{a}),\ldots,\varphi_t(\boldsymbol{r}_{\leq t};\boldsymbol{a});\boldsymbol{a}\right) \tag{158d}$$

$$\overset{\text{(d)}}{=} \varphi_{t+1}(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_{t+1};\boldsymbol{a}), \tag{158e}$$

where

- Step (a) is from the induction hypothesis;
- Step (b) is from the definition of the new denoising function $f_{t+1}$ for deterministic RI-AMP-DF;
- Step (c) is from the definition of the new iterate $\boldsymbol{r}_{t+1}$ in RI-AMP-DF. Note that $(\mathsf{c}_{t,i})_{1\leq t,1\leq i\leq t}$ are deterministic constants that depend on the denoising functions in previous iterations and the spectrum $\mu$;
- Step (d) is a definition of the new post-processing function $\varphi_{t+1}$.

The proof is now complete. □