On Orthogonal AMP in Coded Linear Vector Systems

Junjie Ma, Lei Liu, *Member, IEEE*, Xiaojun Yuan, *Senior Member, IEEE*, and Li Ping, *Fellow, IEEE*

***Abstract*— Linear minimum mean square error (LMMSE) estimation based turbo detection has been extensively studied for coded linear systems since the seminal work of Wang and Poor (WP). The WP algorithm operates iteratively between a linear detector (LD) and a nonlinear detector (NLD): the LD suppresses the interference based on LMMSE filtering, and the NLD decodes the data by treating the output of the LD as an observation from an additive white Gaussian noise (AWGN) channel. In WP, the messages exchanged between LD and NLD are required to be *extrinsic*. For the NLD, the extrinsic message comes from the constraint imposed on feedforward error correction (FEC) codes. Therefore, WP does not work in an un-coded linear system. Recently, we proposed an orthogonal approximate message passing (OAMP) algorithm, which only requires the input/output error terms of LD and NLD to be *orthogonal*. We conjectured that for un-coded linear systems that involve certain large random matrices, the dynamics of OAMP can be accurately characterized by state evolution (SE). In this paper, we consider a coded linear system and develop an extrinsic message aided OAMP (EMA-OAMP) algorithm. Similar to the un-coded case, EMA-OAMP relaxes the requirements on output messages to be orthogonal instead of extrinsic. We derive an SE procedure to characterize the performance of OAMP in coded systems. We conjecture that this SE procedure is accurate, which is verified by simulation results. Under this conjecture, we show that EMA-OAMP can outperform WP under certain standard assumptions for iterative decoding. Extensive simulations results are provided to verify the advantages of OAMP in coded MIMO systems.**

***Index Terms*— Approximate message passing (AMP), state evolution, massive MIMO, expectation propagation (EP), orthogonal AMP (OAMP).**

Manuscript received March 25, 2019; revised July 7, 2019; accepted August 19, 2019. Date of publication September 5, 2019; date of current version December 10, 2019. The work of L. Liu and L. Ping was supported in part by the University Grants Committee of the Hong Kong Special Administrative Region, China, under Grant CityU 11280216, Grant CityU 11216817, and Project CityU 11216518. The work of X. Yuan was supported by the National Key Research and Development Program of China under Grant 2018YFB1801105. This article was presented in part at the 2018 IEEE ISTC, Hong Kong, [1]. The associate editor coordinating the review of this article and approving it for publication was C.-K. Wen. *(Corresponding author: Lei Liu.)*

J. Ma was with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. He is now with the Department of Electrical Engineering, Harvard University, Cambridge, MA 02138 USA (e-mail: junjiema2-c@my.cityu.edu.hk).

L. Liu and L. Ping are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: leiliu@cityu.edu.hk; eeliping@cityu.edu.hk).

X. Yuan is with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: xjyuan@uestc.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2938170

I. Introduction

LINEAR vector channels cover a wide range of communication channel models, including frequency-selective fading channels and, in particular, multiple access channels and multiple-input multiple-output (MIMO) channels [2]. Such models have become increasingly important due to their potential applications in 5G cellular systems and beyond. Following the invention of turbo codes [3], turbo detection [4]–[8] has been extensively studied for communication systems over linear vector channels involving forward error control (FEC) codes.

A typical turbo receiver [4] consists of two local processors that exchange extrinsic information iteratively: one for the FEC code and another for the linear vector channel. The former is usually realized via maximum *a posteriori* (MAP) detection. This is, however, not for the latter since the computational complexity of MAP scales exponentially with the dimension for a linear vector channel. To reduce complexity, the *Wang-Poor* (WP) algorithm employs linear minimum mean square error (LMMSE) detection for the local processor handling the channel effect [9]–[11]. The inputs to a communication channel are typically discrete. WP reduces detection complexity using Gaussian approximation on channel inputs. This is accomplished by matching the means and variances of the estimates for different distributions [10]. Due to its excellent performance, WP has been widely studied for various applications; see [12]–[16] and the references therein.

The performance of a turbo detection algorithm can be analyzed by density evolution [17]. Extrinsic transfer information (EXIT) chart [18] analysis provides a relatively simple, though approximate, tool for this purpose. The EXIT technique is useful for visualizing the convergence behavior of a receiver. For WP, a semi-analytic SINR-variance evolution technique, which is a variant of the EXIT method, has been developed in [14], [15].

Recently, approximate message passing (AMP) [19]–[21] has attracted intensive interests in the context of compressed sensing. AMP is derived from belief-propagation (BP) [22] (and therefore closely related to turbo detection) based on Gaussian approximation and first order Taylor approximation. Remarkably, the asymptotic performance of AMP can be described by a scalar recursion called state evolution (SE) [19], [20]. SE is very similar to density evolution, except that SE is developed for dense graphs while density evolution is only accurate for sparse graphs. Although derived in the

context of compressed sensing, AMP can be readily applied to communications systems [23]–[27].

AMP was originally proposed for channel matrices (or sensing matrices in the compressed sensing context [19]) with independent identically distributed (IID) entries, and validated under this condition [20]. For matrices with correlated entries, AMP may perform poorly or even diverge [28]–[30]. To handle this difficulty, an orthogonal AMP (OAMP) algorithm was proposed in [31]; see also closely-related earlier works in [32]–[37].

In OAMP, an orthogonality constraint is imposed such that the so-called Onsager term [19], [20] in AMP vanishes. It was conjectured in [31] that OAMP can be characterized by an SE recursion for general unitarily-invariant matrices (which includes IID Gaussian matrices as special cases). OAMP involves two local processors (a so-called linear detector (LD) and a non-linear detector (NLD)). These two local processors can be constructed in various ways, provided that they meet certain orthogonality constraints. (See Section III.B for details.) In particular, an MMSE-derived OAMP (MMSE-OAMP) can be constructed based on the locally optimal MMSE principle. Such MMSE-OAMP is related to a variant of expectation propagation (EP) algorithm [38] (called diagonally-restricted expectation consistent inference in [32] or *scalar-EP* in [39]), as observed in [40]–[43]. A closely related algorithm, an MMSE derived vector AMP (VAMP) [41], is equivalent to EP in its diagonally-restricted form [32]. The accuracy of SE for such EP type algorithms (including VAMP and MMSE-OAMP) was proved in [40], [41].

OAMP was developed for un-coded linear systems. In this paper, we investigate an extrinsic-message-aided OAMP (EMA-OAMP) for FEC coded system. Our emphasis is on the comparison between EMA-OAMP and WP. The followings are the main contributions of this paper.

- In a message passing process, the so-called *a posteriori* probability (APP) messages are locally optimal, but they may lead to a correlation problem [44]. Extrinsic messages used by WP are less accurate than APP ones but can avoid the correlation problem. EMA-OAMP employs the so-called orthogonal messages. The latter are better estimates than extrinsic ones while, at the same time, avoid the correlation problem of APP ones. Thus EMA-OAMP provides a new approach to receiver design for coded linear vector systems.
- We outline an SE procedure to characterize the behavior of EMA-OAMP in coded systems, based on which we show that EMA-OAMP can potentially outperform WP. We are still unable to prove the SE procedure rigorously, but we observed from simulations that the proposed SE is accurate for large unitarily-invariant channel matrices. The works in [40], [41] provide a promising way towards this direction.
- We will provide extensive simulation results to show that EMA-OAMP can outperform WP. As the two approaches have roughly the same complexity, EMA-OAMP offers an attractive new solution to a wide range of applications that was formally treated by WP. We will also provide comparisons of EMA-OAMP and AMP in coded systems [25]. The latter may not work well in correlated channels whose channel matrices are not IID. We will show by simulations that EMA-OAMP is robust in various channel environments.
- We will discuss the similarity and difference between OAMP and EP.[1] As proved in [40], MMSE-OAMP can be expressed in an equivalent form of EP. However, an MMSE processor can be very costly. For example, linear MMSE detection involves the inversion of a matrix with complexity $O(N^3)$, where N is the signal length. (Such complexity order also applies to VAMP due to the singular value decomposition involved.) When N is large, low-cost alternatives become necessary even though they are not MMSE. As an example, OAMP can be constructed using a low-cost matched filter structure with performance not far away from the MMSE version. Such OAMP, that is different from EP in its straightforward form, provides an attractive option for complexity-performance trade-off.

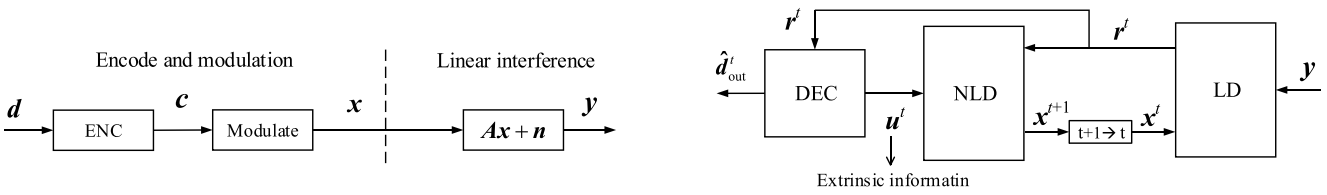
Overall, the concept of extrinsic messages lies at the core of the celebrated turbo principle. WP employing extrinsic messages has been regarded as the de facto solution to coded linear channels. It has been widely reported that WP is nearly optimal in various applications [6]–[8], [13]. It is natural to ask whether there is still room for improvement from WP. The findings in this paper give a positive answer: the improvement is still considerable. The concept of orthogonal messages in EMA-OAMP provides a new perspective to the problem.

Part of the results in this paper have been published in [1]. In this paper, we provide more detailed analysis and numerical results.

This paper is organized as follows. In Section II, the coded linear system and WP algorithm are introduced. The EMA-OAMP for the coded linear systems is introduced in Section III. In section IV, a conjectured SE is proposed for EMA-OAMP involving linear MMSE estimation, based on which we prove that EMA-OAMP outperforms WP. Numerical results are provided in Section V to demonstrates the advantage of EMA-OAMP over WP.

Notations: Boldface lowercase letters represent vectors and boldface uppercase symbols denote matrices. $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ denotes the inner product of the vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}$. $\boldsymbol{I}$ for the identity matrix with a proper size, $\boldsymbol{a}^{\mathrm{T}}$ for the conjugate of $\boldsymbol{a}$, $\|\boldsymbol{a}\|$ for the ℓ_2-norm of the vector $\boldsymbol{a}$, $\mathrm{tr}(\boldsymbol{A})$ for the trace of $\boldsymbol{A}$, $[\eta(\boldsymbol{a}, \boldsymbol{b})]_j \equiv \eta(a_j, b_j)$, a_i for the ith entry of $\boldsymbol{a}$, $[\boldsymbol{A}]_{ij} \equiv a_{ij}$ for the ith-row and jth-column element of $\boldsymbol{A}$. $\mathrm{diag}\{\boldsymbol{A}\}$ for the diagonal part of $\boldsymbol{A}$, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ for Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{C}$, $\mathrm{E}[\cdot]$ for the expectation operation over all random variables involved in the brackets, except when otherwise specified. $\mathrm{E}[a|b]$ for the expectation of a conditional on b, $\mathrm{var}[a]$ for $\mathrm{E}\left[(a - \mathrm{E}[a])^2\right]$, $\mathrm{var}[a|b]$ for $\mathrm{E}\left[(a - \mathrm{E}[a|b])^2 |b\right]$.

[1]Throughout this paper, EP refers to the diagonally-restricted form of EP [32]. See a comparison of diagonally-restricted EP (or scalar-EP) and standard EP in [39].



(a) Coded linear system

(b) Standard Turbo detection: WP

Fig. 1. FEC coded linear system: transmitter and iterative receiver, where red"DEC", "NLD" and "LD" in (b) correspond to "ENC", "modulate" and "$\boldsymbol{Ax}+\boldsymbol{n}$" in (a) respectively.

II. Coded Linear System and Wang-Poor Algorithm

A. System Model

Consider an $M \times N$ linear system modeled by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^{M\times 1}$ is a received signal vector, $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ the channel matrix, $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ an independent Gaussian noise, and $\boldsymbol{x}$ the signal vector to be estimated. The block diagram in Fig. 1 (a) shows the overall system model. In addition, $\boldsymbol{d}$ denotes the information vector, $\boldsymbol{c}$ the corresponding coded data, and $\boldsymbol{x}$ the modulated signals. The entries of $\boldsymbol{x}$ are constrained over a discrete constellation $\mathcal{X} = \{s_1, s_2, \cdots, s_K\}$ which we assume to have zero mean and unit variance. Let the singular value decomposition (SVD) of $\boldsymbol{A}$ be $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$, where $\boldsymbol{U} \in \mathbb{R}^{M\times M}$ and $\boldsymbol{V} \in \mathbb{R}^{N\times N}$ are orthogonal matrices, and $\boldsymbol{\Sigma}$ is a diagonal matrix [31], [41]. We assume that $\boldsymbol{A}$ is unitarily-invariant, i.e., $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\Sigma}$ are mutually independent, and $\boldsymbol{U}$, $\boldsymbol{V}$ are Haar-distributed. We assume that $\boldsymbol{A}$ is known at receiver. For simplicity, we only discuss a real-valued system, based on which the complex version can be easily extended [40].

B. Basic Principles of the Wang-Poor (WP) Algorithm

It is well known that the MAP or ML detector for the above problem is computationally prohibitive when M is large. The linear MMSE based turbo detection algorithm [9], referred to as Wang-Poor (WP) algorithm in this paper, provides a good tradeoff between complexity and performance. WP operates iteratively between two local detectors, namely, *linear detector* (LD) and *non-linear detector* (NLD) as follows:

- The LD handles the linear constraint, and NLD handles the coding-modulation constraint;
- The messages exchanged between LD and NLD are means and variances associated with the variables to be estimated.

Starting with $\boldsymbol{s}^0 = \boldsymbol{0}$, WP proceeds as

$$\text{LD}: \quad \boldsymbol{r}^t = f^{\text{WP}}(\boldsymbol{s}^t), \tag{2a}$$

$$\text{NLD}: \quad \boldsymbol{s}^{t+1} = \eta^{\text{WP}}(\boldsymbol{r}^t), \tag{2b}$$

where $t = 0, 1, \ldots$ denotes the iteration index. At the last iteration, the NLD generates an APP estimate of $\boldsymbol{d}$.

The use of extrinsic messages in WP avoids the correlation problem during the iterative process. These messages, i.e., $\boldsymbol{r}^t$ and $\boldsymbol{s}^t$, are generated by f^{WP} and η^{WP} in (2). Here f^{WP} is "extrinsic" in that its ith output is not a function of its ith input. This follows the turbo-principle developed in [3]. Specifically, the extrinsic LD is defined as

$$r_i^t = [f^{\text{WP}}(\boldsymbol{s}^t)]_i \equiv f(\boldsymbol{s}^t_{\sim i}), \quad \forall i, \tag{3}$$

where $[f^{\text{WP}}(\boldsymbol{s}^t)]_i$ denotes the ith entry of $f^{\text{WP}}(\boldsymbol{s}^t)$, and $\boldsymbol{s}^t_{\sim i}$ denotes a vector formed by $\boldsymbol{s}^t$ with the ith entry excluded. The extrinsic NLD $\eta^{\text{WP}}(\boldsymbol{r}^t)$ is defined similarly. In the following subsections, we discuss the detailed operations of LD and NLD for WP.

C. Linear Detector for WP

The linear MMSE detection involving *a priori* information has many equivalent forms. The exposition below is based on [6, Section II-D]. Let $\boldsymbol{s}^t$ and v^t be the *a priori* mean and variance of $\boldsymbol{x}$, respectively. The linear MMSE estimate of $\boldsymbol{x}$ is

$$\hat{\boldsymbol{r}}^t \equiv f^{\text{MMSE}}(\boldsymbol{s}^t) = \boldsymbol{s}^t + \boldsymbol{W}^{\text{MMSE}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t), \tag{4a}$$

where $\hat{\boldsymbol{r}}^t = [\hat{r}_1^t, \cdots, \hat{r}_N^t]$ and

$$\boldsymbol{W}^{\text{MMSE}} \equiv v^t\boldsymbol{A}^{\text{T}}(v^t\boldsymbol{A}\boldsymbol{A}^{\text{T}} + \sigma^2\boldsymbol{I})^{-1}. \tag{4b}$$

The corresponding covariance matrix is

$$\boldsymbol{V}^t = v^t\boldsymbol{I} - (v^t)^2\boldsymbol{A}^{\text{T}}(v^t\boldsymbol{A}\boldsymbol{A}^{\text{T}} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{A}. \tag{4c}$$

The direct use of (4) in an iterative receiver may lead to the correlation problem. The following are some treatments inspired by the turbo processing principles [4], [11].

In linear MMSE, we treat $\hat{r}_i^t$ and s_i^t as two Gaussian observations of x_i, and their variances are given by V_{ii}^t and v^t. In turbo detection, we compute the extrinsic estimate to ensure that the output r_i^t is independent of the input s_i^t. For the ith entry, the extrinsic estimate r_i^t is computed as [6], [10]:

$$r_i^t = \tau_i^t \cdot \left(\frac{\hat{r}_i^t}{V_{i,i}^t} - \frac{s_i^t}{v^t}\right), \tag{5a}$$

with

$$\tau_i^t = \left(\frac{1}{V_{i,i}^t} - \frac{1}{v^t}\right)^{-1}, \tag{5b}$$

where $V_{i,i}^t$ is (i,i)th entry of $\boldsymbol{V}^t$ in (4c). Let $\boldsymbol{V}^t_{\text{diag}}$ be a diagonal matrix that has the same diagonal elements as $\boldsymbol{V}^t$. Furthermore, we define

$$\boldsymbol{\Lambda}^{\text{WP}} = \left(\boldsymbol{I} - (v^t)^{-1}\boldsymbol{V}^t_{\text{diag}}\right)^{-1}. \tag{6}$$

The extrinsic linear MMSE estimate in (5) can be written more concisely as

$$f^{\mathrm{WP}}(\boldsymbol{s}^t) = \boldsymbol{\Lambda}^{\mathrm{WP}} f^{\mathrm{MMSE}}(\boldsymbol{s}^t) + \left(\boldsymbol{I} - \boldsymbol{\Lambda}^{\mathrm{WP}}\right)\boldsymbol{s}^t. \tag{7}$$

Denote $\boldsymbol{W}^{\mathrm{WP}} = \boldsymbol{\Lambda}^{\mathrm{WP}}\boldsymbol{W}^{\mathrm{MMSE}}$. From (7) and (4a), we have

$$f^{\mathrm{WP}}(\boldsymbol{s}^t) = \boldsymbol{s}^t + \boldsymbol{W}^{\mathrm{WP}}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t\right). \tag{8}$$

D. Nonlinear Detector for WP

The extrinsic decoder generates the following probabilities:

$$u_i^t(k) \equiv P_{\mathrm{SM}}\left(X_i = s_k | \boldsymbol{r}_{\sim i}^t\right), \tag{9}$$

where $i \in \mathcal{N} = \{1, 2, \cdots, N\}$, $k \in \mathcal{K} = \{1, 2, \cdots, K\}$, s_k is the kth constellation point, as defined below (1). The subscript "SM" denotes an approximate posterior probability using sum-product decoding. The decoder calculates the conditional probabilities in (9) based on the coding constraint by using message passing algorithms [45], [46]. Let $\boldsymbol{u}^t = [u_1^t, \cdots, u_N^t]$, where $u_i^t = [u_i^t(1), \cdots, u_i^t(K)]^T$. In this paper, we will assume that u_i^t and r_i^t are two independent information sources of x_i. Such approximation was widely used in turbo-type processors [47].

We first consider the MMSE NLD, which will be used in the APP iterative receiver (see discussions in Section V-A). From the extrinsic probabilities in (9), the MMSE estimate and variance are calculated as

$$\eta^{\mathrm{MMSE}}(r_i^t, u_i^t) = \mathrm{E}\left[x_i | r_i^t, u_i^t\right] = \sum_{k\in\mathcal{K}} s_k \beta_i^t(k), \tag{10a}$$

$$\mathrm{var}\left[x_i | r_i^t, u_i^t\right] = \sum_{k\in\mathcal{K}} \left(s_k - \eta^{\mathrm{MMSE}}(r_i^t, u_i^t)\right)^2 \beta_i^t(k), \tag{10b}$$

where u_i^t denotes the extrinsic probability defined in (9), and $\beta_i^t(k)$ the APP of the decoder

$$\beta_i^t(k) \equiv P\left(x_i = s_k | r_i^t, u_i^t\right) = \frac{u_i^t(k)\alpha_i^t(k)}{\sum\limits_{k\in\mathcal{K}} u_i^t(k)\alpha_i^t(k)}, \tag{10c}$$

and $\alpha_i^t(k)$ denotes the *a priori* probability:

$$\alpha_i^t(k) \equiv P\left(r_i^t | x_i = s_k\right). \tag{10d}$$

In WP, the MMSE estimate is not directly feedback to the LD. Instead, the NLD generates the following extrinsic estimate and variance (for x_i):

$$s_i^{t+1} = \eta^{\mathrm{WP}}(u_i^t) = \mathrm{E}\left[x_i | u_i^t\right], \tag{11a}$$

$$v^{t+1} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{var}\left[x_i | u_i^t\right], \tag{11b}$$

where u_i^t is defined in (9), and the expectation and variance are calculated as

$$\mathrm{E}\left[x_i | u_i^t\right] = \sum_{k\in\mathcal{K}} s_k u_i^t(k), \tag{11c}$$

$$\mathrm{var}\left[x_i | u_i\right] = \sum_{k\in\mathcal{K}} \left(s_k - \mathrm{E}\left[x_i | u_i^t\right]\right)^2 u_i^{\mathrm{t}}(k). \tag{11d}$$

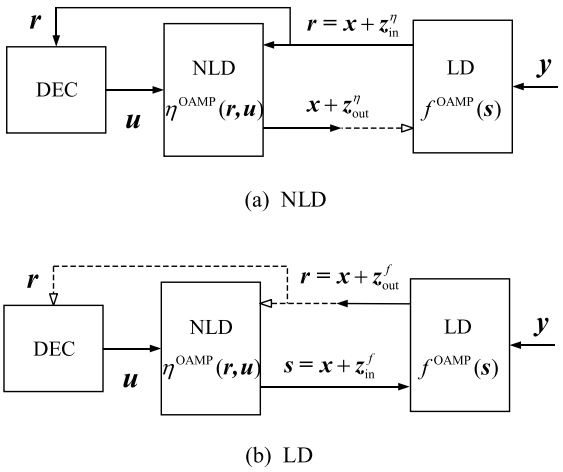


Fig. 2. LD and extrinsic-message-aided NLD process.

E. Discussions

From (4)-(7), we could verify that $\boldsymbol{W}^{\mathrm{WP}}$ satisfies

$$[\boldsymbol{I} - \boldsymbol{W}^{\mathrm{WP}}\boldsymbol{A}]_{ii} = 0, \quad \forall i. \tag{12}$$

For the NLD, since r_i^t is excluded from computing $[\eta^{\mathrm{WP}}(\boldsymbol{r}^t)]_i$, we have

$$\frac{\partial[\eta^{\mathrm{WP}}(\boldsymbol{r}^t)]_i}{\partial r_i^t} = 0, \quad \forall r_i^t. \tag{13}$$

In the next section, we will compare the above properties to those imposed on $\boldsymbol{W}$ and η for an OAMP algorithm.

Finally, we noted that the extrinsic information u_i does not exist for an uncoded system. In this case, we would have

$$\eta^{\mathrm{WP}}(u_i^t) = 0, \quad \forall i, t. \tag{14}$$

This implies that the iterative cannot provide any improvement for uncoded systems. Recently, it is shown that meaningful iterative detection can be devised for uncoded systems using AMP and the related OAMP algorithms [19], [31]. In what follows, we will generalize the results in [31] to coded systems and show that the new approach can noticeably outperform WP in some situations.

III. Extrinsic-Message-Aided OAMP

As shown in Section II, an iterative receiver for the coded linear system consists of two components: LD with a linear constraint $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$ and NLD with a non-linear code constraint $\boldsymbol{x} \in \mathcal{C}$. In Fig. 2, LD is represented by a function $f^{\mathrm{OAMP}}(\boldsymbol{s})$ and NLD is by $\eta^{\mathrm{OAMP}}(\boldsymbol{r}, \boldsymbol{u})$, where $\boldsymbol{z}_{\mathrm{in}}^f$ and $\boldsymbol{z}_{\mathrm{out}}^f$ represent respectively, the input and output errors of LD, and $\boldsymbol{z}_{\mathrm{in}}^{\eta}$ and $\boldsymbol{z}_{\mathrm{out}}^{\eta}$ are defined similarly for NLD.

We now extend the original OAMP principle [31] to coded systems. The main difference here is that the NLD exploits the extrinsic messages generated by the decoder, hence the name extrinsic message aided OAMP (EMA-OAMP).

A. Orthogonality Property

Definition 1 (Orthogonality): We say that $f^{\mathrm{OAMP}}(\cdot)$ and $\eta^{\mathrm{OAMP}}(\cdot)$ are orthogonal detectors if

$$\mathrm{E}\left[\frac{1}{N}\langle \boldsymbol{z}_{\mathrm{in}}^{f}, \boldsymbol{z}_{\mathrm{out}}^{f}\rangle\right] = \mathrm{E}\left[\frac{1}{N}\langle \boldsymbol{z}_{\mathrm{in}}^{\eta}, \boldsymbol{z}_{\mathrm{out}}^{\eta}\rangle\right] = 0, \tag{15}$$

i.e., the input and output errors are statistically uncorrelated, or simply orthogonal.

B. EMA-OAMP Algorithm

Starting with $\boldsymbol{s}^0 = \boldsymbol{0}$, a general EMA-OAMP is given by

$$\text{LD}: \ \boldsymbol{r}^t = f^{\mathrm{OAMP}}(\boldsymbol{s}^t), \tag{16a}$$
$$\text{NLD}: \ \boldsymbol{s}^{t+1} = \eta^{\mathrm{OAMP}}\left(\boldsymbol{r}^t, \boldsymbol{u}^t\right), \tag{16b}$$

where $\boldsymbol{u}^t$ is the output extrinsic information at the decoder. Here, "extrinsic" means that u_i^t is a function of $\{r_j^t, j \neq i\}$. f^{OAMP} can be written as

$$f^{\mathrm{OAMP}}(\boldsymbol{s}^t) = \boldsymbol{s}^t + \boldsymbol{W}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t\right) \tag{16c}$$

with the de-correlated constraint

$$\frac{1}{N}\mathrm{Tr}(\boldsymbol{I}\text{-}\boldsymbol{W}\boldsymbol{A}) = 0, \tag{16d}$$

and η^{OAMP} is subject to the divergence-free constraint:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \eta^{\mathrm{OAMP}}(r_i^t, u_i^t)}{\partial r_i^t} = 0. \tag{16e}$$

The LD in (16c) is the same as that in the uncoded OAMP [31], i.e. $\boldsymbol{W} = \frac{N}{\mathrm{tr}(\hat{\boldsymbol{W}}\boldsymbol{A})}\hat{\boldsymbol{W}}$ with $\hat{\boldsymbol{W}} = \boldsymbol{V}^T\boldsymbol{G}\boldsymbol{U}^T$, where $\boldsymbol{G}$ can be any matrix with proper size. Therefore, we obtain

$$f^{\mathrm{OAMP}}(\boldsymbol{x}^t) = \boldsymbol{s}^t + \frac{N}{\mathrm{tr}(\hat{\boldsymbol{W}}\boldsymbol{A})}\hat{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t). \tag{17a}$$

The NLD in (16b) with constraint (16e) can be constructed as[2]

$$\eta^{\mathrm{OAMP}}\left(\boldsymbol{r}^t, \boldsymbol{u}^t\right) = C_t\left(\eta(\boldsymbol{r}^t, \boldsymbol{u}^t) - \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \eta(r_i^t, u_i^t)}{\partial r_i^t}\right)\cdot \boldsymbol{r}^t\right), \tag{17b}$$

where η is an arbitrary function in a general OAMP framework. A general EMA-OAMP algorithm is presented in Algorithm 1.

Note that the matrix $\hat{\boldsymbol{W}}$ can depend on the parameter v^t (e.g., the linear MMSE detector), and the function η will depend on the parameter τ^t in general. For notational brevity, we do not explicitly write $\hat{\boldsymbol{W}}$ and η as functions of v^t and τ^t. This paper focuses on communication systems with FEC codes, and we will only consider an η function that is an extrinsic-message aided MMSE decoder (including approximate versions of it). More specifically, we will focus on the following choices of $\hat{\boldsymbol{W}}$, $\eta(\boldsymbol{r}^t, \boldsymbol{u}^t)$, and C_t in (16c) and (16b):

$$\hat{\boldsymbol{W}} = \boldsymbol{W}^{\mathrm{MMSE}} \ \text{ and } \eta(r, u) = \eta^{\mathrm{MMSE}}(r, u), \tag{18a}$$
$$C_t = \frac{\tau^t}{\tau^t - \frac{1}{N}\sum\limits_{i=1}^{N}\mathrm{var}\left[x_i | r_i^t, u_i^t\right]}, \tag{18b}$$

[2]This is true under a heuristic assumption that $\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \eta(r_i^t, u_i^t)}{\partial r_i^t}$ converges to a constant invariant to each individual r_i^t.

Algorithm 1 General EMA-OAMP

Require: $\boldsymbol{A}$, $\boldsymbol{y}$, σ^2, $\hat{\boldsymbol{W}}^t$, T, $\mathrm{Dec}^{\mathrm{ext}}(\cdot)$, $\eta(\cdot)$, c^t
Initialization: $\boldsymbol{s}^0 = \boldsymbol{0}$
for $t = 0, 1, \ldots, T$ **do**
  $v^t = \frac{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t\|^2 - M\cdot\sigma^2}{\mathrm{Tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$
  $\tau^t = \frac{\mathrm{Tr}(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}})}{N}\cdot v^t + \frac{\mathrm{Tr}(\hat{\boldsymbol{W}}\hat{\boldsymbol{W}}^{\mathrm{T}})}{N}\sigma^2$, $\boldsymbol{B} := \boldsymbol{I} - \hat{\boldsymbol{W}}\boldsymbol{A}$
  $\boldsymbol{r}^t = \boldsymbol{s}^t + \frac{N}{\mathrm{tr}(\hat{\boldsymbol{W}}^t\boldsymbol{A})}\hat{\boldsymbol{W}}^t(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t)$ //LD
  $\boldsymbol{u}^t = \mathrm{Dec}^{\mathrm{ext}}(\boldsymbol{r}^t)$ //EXT DEC
  $\boldsymbol{s}^{t+1} = c^t\left[\eta(\boldsymbol{r}^t, \boldsymbol{u}^t) - \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \eta(r_i^t, u_i^t)}{\partial r_i^t}\right)\cdot \boldsymbol{r}^t\right]$ //NLD
end for
Return: $\hat{\boldsymbol{x}} = \eta(\boldsymbol{r}^t, \boldsymbol{u}^t)$.

where $\boldsymbol{W}^{\mathrm{MMSE}}$ and $\eta^{\mathrm{MMSE}}(r, u)$ are defined in (4) and (10) respectively.

In this paper, the EMA-OAMP algorithm with the above choices of $\hat{\boldsymbol{W}}$, $\eta(\boldsymbol{r}^t, \boldsymbol{u}^t)$, and C_t in (16c) is referred to as the *MMSE-derived OAMP* (MMSE-OAMP).

C. Assumptions and Properties

The following propositions establish the orthogonal properties of EMA-OAMP.

1) Assumptions and Properties of LD:

Assumption 1: For the LD, $\boldsymbol{z}_{\mathrm{in}}^{f}$ has IID entries with $\mathrm{E}[(z_{\mathrm{in},i}^{f})^2] = v, \forall i$.

Proposition 1: If Assumption 1 holds, LD in (16c) has the following properties:

a) $\mathrm{E}\left[\frac{1}{N}\langle \boldsymbol{z}_{\mathrm{in}}^{f}, \boldsymbol{z}_{\mathrm{out}}^{f}\rangle\right] = 0$,
b) the entries of $\boldsymbol{z}_{\mathrm{out}}^{f}$ are mutually uncorrelated with zero-mean and identical variance, and
c) the entries of $\boldsymbol{z}_{\mathrm{out}}^{f}$ are uncorrelated with those of $\boldsymbol{x}$.

For the proof of Proposition 1, see [31].

2) Assumptions and Properties of NLD:

Assumption 2: The input of the NLD is an observation of $\boldsymbol{x}$ from an additive Gaussian channel, i.e. $\boldsymbol{r} = \boldsymbol{x} + \boldsymbol{z}_{\mathrm{in}}^{\eta}$, where $\boldsymbol{z}_{\mathrm{in}}^{\eta} \sim \mathcal{N}(\boldsymbol{0}, \tau\boldsymbol{I})$ is independent of $\boldsymbol{x}$.

Proposition 2: If Assumption 2 holds, the EMA orthogonal NLD given in (17b) satisfies $\mathrm{E}\left[\frac{1}{N}\langle \boldsymbol{z}_{\mathrm{in}}^{\eta}, \boldsymbol{z}_{\mathrm{out}}^{\eta}\rangle\right] = 0$, as $N \to \infty$.

Proof: From Assumption 2 and Stein's lemma [48], we have

$$\frac{1}{N}\mathrm{E}\left[\langle \boldsymbol{z}_{\mathrm{in}}^{\eta}, \eta^{\mathrm{OAMP}}(\boldsymbol{r}, \boldsymbol{u})\rangle\right] = \frac{\tau}{N}\sum_{i=1}^{N}\mathrm{E}\left[\frac{\partial \eta^{\mathrm{OAMP}}(r_i, u_i)}{\partial r_i}\right]. \tag{19}$$

The law of large numbers implies that η asymptotically satisfies

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \eta(r_i, u_i)}{\partial r_i} \to \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[\frac{\partial \eta(r_i, u_i)}{\partial r_i}\right] \tag{20}$$

as N goes to infinity, where the expectation is taken over r_i and u_i. Substituting (20) and (17b) into (19), we obtain

$$\frac{1}{N}\mathrm{E}\left[\left\langle \boldsymbol{\eta}_{\mathrm{in}}^{\eta}, \eta^{\mathrm{OAMP}}(\boldsymbol{r}, \boldsymbol{u})\right\rangle\right] = 0, \tag{21}$$

as $N \to \infty$. Since $\boldsymbol{z}_{\rm in}^{\eta}$ is independent of $\boldsymbol{x}$, we have

$$\mathrm{E}\left[\frac{1}{N}\langle \boldsymbol{z}_{\rm in}^{\eta}, \boldsymbol{z}_{\rm out}^{\eta}\rangle\right] = \mathrm{E}\left[\frac{1}{N}\left\langle \boldsymbol{z}_{\rm in}^{\eta}, \eta^{\rm OAMP}(\boldsymbol{r}, \boldsymbol{u}-\boldsymbol{x})\right\rangle\right] \quad (22)$$

$$= \mathrm{E}\left[\frac{1}{N}\left\langle \boldsymbol{z}_{\rm in}^{\eta}, \eta^{\rm OAMP}(\boldsymbol{r}, \boldsymbol{u})\right\rangle\right] = 0. \quad (23)$$

This proves the proposition. ∎

The orthogonal properties of EMA-OAMP are guaranteed by Propositions 1 and 2, which respectively rely on Assumptions 1 and 2. We are currently working on rigorous justifications of Assumptions 1 and 2.

D. Connections of WP and EMA-OAMP

Comparing (2) and (16), we can see that WP and EMA-OAMP have the same LD and NLD expressions but are subject to different constraints. WP is a special case of OAMP since:

- for LD, the constraint of WP in (12) (i.e. $[\boldsymbol{I} - \boldsymbol{W}^{\rm WP}\boldsymbol{A}]_{ii} = 0, \forall i$) is a special case of that of EMA-OAMP (16d) (i.e. $\frac{1}{N}\mathrm{tr}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{A}) = 0$); and
- for NLD, the constraint of WP in (13) (i.e. $\partial[\eta^{\rm WP}(\boldsymbol{r}^t)]_i/\partial r_i^t = 0, \forall i$ and $\forall r_i^t$) is a special case of that of EMA-OAMP (16e) (i.e. $\frac{1}{N}\sum_{i=1}^{N}\partial\eta^{\rm OAMP}(r_i^t, u_i^t)/\partial r_i^t = 0$).

From the above discussions, OAMP is less restrictive than WP, so OAMP can potentially achieve lower MSE (and so better performance) than WP. However, we should carefully treat the correlation issue, which will be discussed in Section IV.

E. Connections of OAMP and EP

OAMP is closely related to the expectation propagation (EP) algorithm [38], [40]. In this paper, we will compare OAMP with the diagonally-restricted version of EP introduced in [32] (see the first bullet point on page 2184), referred to as *scalar-EP* (S-EP) [39]. S-EP was independently re-discovered in [33]–[36] under certain heuristic Gaussian message passing approximations. S-EP was originally presented as a heuristic algorithm. It was shown in [33]–[36] that this S-EP algorithm (for signal detection under a linear channel model) could be described by a state evolution (SE) recursion for certain channel matrices. Further, the dynamics of the SE was analyzed in [31], [35] which was recently rigorously proved in [40], [41]. *In the rest of this paper, S-EP will be simply called "EP" for notational simplicity.*

The MMSE derived EP and OAMP (i.e., MMSE-EP and MMSE-OAMP) are identical when the local processors are derived using MMSE principles, as detailed in Appendix B-A. OAMP is more general than S-EP in that the LD and NLD in OAMP do not have to be Bayesian estimators; they only need to satisify certain orthogonality conditions. Such general OAMP algorithms have been considered in [44] for non-scalar denoising and in [49] for low-rank matrix recovery.

IV. Analysis and Comparison for WP and EMA-OAMP

We now consider the analysis and comparison of WP and OAMP based on the state evolution (SE) technique [20], [40], [41]. SE is a recursive procedure that tracks the MSE of local processors during iterative processing. The SE formulas developed below are based on the assumptions that the input and output errors of a local processor are independent of each other. Such independence might be justified for WP in which the output message of x_i is calculated by excluding the input message of x_i. The problem is more complicated for OAMP. It boils down to the justification for Assumptions 1 and 2 in Section III-B. In what follows, the accuracy of the SE for OAMP will be stated as a conjecture.

A. State Evolution of WP and EMA-OAMP

SE was developed in [19], [20] for the analysis of AMP. SE is in principle similar to density evolution [17], EXIT-chart [18] or SINR-variance evolution [14], [15] for the semi-analytical performance evolution for iterative detections. Denote the MSE of $\boldsymbol{r}^t$ and $\boldsymbol{s}^t$ as τ^t and v^t, respectively:

$$\tau^t \equiv \tfrac{1}{N}\mathrm{E}\left[\left\|\boldsymbol{r}^t - \boldsymbol{x}\right\|^2\right], \quad v^t \equiv \tfrac{1}{N}\mathrm{E}\left[\left\|\boldsymbol{s}^t - \boldsymbol{x}\right\|^2\right]. \quad (24)$$

SE refers to the following recursions of $\{\tau^t\}$ and $\{v^t\}$:

$$\tau^t = \phi(v^t), \quad v^{t+1} = \psi(\tau^t). \quad (25)$$

The transfer functions $\phi(\cdot)$ and $\psi(\cdot)$ for WP and EMA-OAMP are respectively given by (**WP**)

$$\phi^{\rm WP}(v) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{V_{i,i}} - \frac{1}{v}\right)^{-1}, \quad (26a)$$

$$\psi^{\rm WP}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{mmse}\,(x_i|u_i). \quad (26b)$$

and (**EMA-OAMP**)[3]

$$\phi^{\rm OAMP}(v) = \left(\frac{1}{\frac{1}{N}\mathrm{tr}(\boldsymbol{V})} - \frac{1}{v}\right)^{-1}, \quad (27a)$$

$$\psi^{\rm OAMP}(\tau) = \left(\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\mathrm{mmse}\,(x_i|r_i, u_i)} - \frac{1}{\tau}\right)^{-1}. \quad (27b)$$

In the above, $\mathrm{mmse}(x_i|u_i) \equiv \mathrm{E}[\mathrm{var}[x_i|u_i]]$ and $\mathrm{mmse}(x_i|u_i) \equiv \mathrm{E}[\mathrm{var}[x_i|r_i, u_i]]$, where r_i is the ith entry of $\boldsymbol{r} = \boldsymbol{x} + \mathcal{N}(\boldsymbol{0}, \tau\boldsymbol{I})$ and u_i represents the extrinsic information (generated by the decoder) for x_i. Note that $\mathrm{mmse}(x_i|u_i)$ and $\mathrm{mmse}(x_i|r_i, u_i)$ are implicit functions of τ because u_i is a function of r_i whose noise variance is τ. The transfer function (27a) has been derived in [31]. The derivations of (27b) can be found in Appendix A.

B. Discussions

It is straightforward to verify that the SE procedure in Section IV-A holds for both WP and OAMP, provided that Assumptions 1 and 2 in Section III-B hold. These assumptions, however, require careful scrutiny. The results in [40], [41], [50]

[3]Notice that $\mathrm{mmse}(x_i|r_i, u_i)$ can be zero when τ is smaller than the decoding threshold. In such cases, it is understood that $\psi^{\rm OAMP}(\tau) = 0$.

provide useful tools for this purpose, but a detailed treatment is beyond the scope of this paper. Instead, we will present the rest of this paper based on the following conjecture.

Conjecture 1: The SE based on (26) and (27) are, respectively, accurate for WP and OAMP.

Based on the above conjecture, we will compare the performances of WP and OAMP using numerical results in Section V-B. Our numerical results show that the SE is very accurate for both WP and OAMP for certain large dense random matrices.

C. Performance Comparison of WP and EMA-OAMP

Theorem 1 below states that the transfer functions of EMA-OAMP outperform those of WP. This means that EMA-OAMP outperforms WP if the SE predictions of EMA-OAMP and WP are accurate (Conjecture 1).

Theorem 1: For any $\tau > 0$, we have $\phi^{\mathrm{OAMP}}(\tau) < \phi^{\mathrm{WP}}(\tau)$, and $\psi^{\mathrm{OAMP}}(\tau) < \psi^{\mathrm{WP}}(\tau)$. Further, under Conjecture 1, the final BER of EMA-OAMP is lower than that of WP.

Proof: First, we proved in Appendix C that $\phi^{\mathrm{OAMP}}(\tau) < \phi^{\mathrm{WP}}(\tau)$ and $\psi^{\mathrm{OAMP}}(\tau) < \psi^{\mathrm{WP}}(\tau)$. We next prove that, under Conjecture 1, the final BER performance of EMA-OAMP is better than that of WP. It is easy to show that all the MSE functions ϕ^{OAMP}, ψ^{OAMP}, ϕ^{WP}, ψ^{WP}, and ψ^{out} are monotonically increasing. (The proof is similar to that of [31, Lemma 2].) According to the SE process, the BER MSEs of WP and EMA-OAMP can be expressed as

$$\begin{aligned}&\mathrm{BER}^{\mathrm{WP}}\\&= \Gamma_{dec}^{\mathrm{post}}\left(\phi^{\mathrm{WP}}\left(\psi^{\mathrm{WP}}\left(\phi^{\mathrm{WP}}\left(\cdots\left(\phi^{\mathrm{WP}}\left(v^0\right)\right)\right)\right)\right)\right),\\&\mathrm{BER}^{\mathrm{opt}}\\&= \Gamma_{dec}^{\mathrm{post}}\left(\phi^{\mathrm{OAMP}}\left(\psi^{\mathrm{OAMP}}\left(\phi^{\mathrm{OAMP}}\left(\cdots\left(\phi^{\mathrm{OAMP}}\left(v^0\right)\right)\right)\right)\right)\right).\end{aligned}$$

With the monotonicity of the APP decoding BER functions $\Gamma_{dec}^{\mathrm{post}}(\cdot)$, $\phi^{\mathrm{OAMP}}(\tau) < \phi^{\mathrm{WP}}(\tau)$ and $\psi^{\mathrm{OAMP}}(\tau) < \psi^{\mathrm{WP}}(\tau)$, we readily have $\mathrm{BER}^{\mathrm{OAMP}} < \mathrm{BER}^{\mathrm{WP}}$. This concludes the proof of Theorem 1. ∎

Intuitively speaking, WP deals the correlation problem of an iterative process by requiring the input and output errors to be independent. On the other hand, OAMP only requires the input and output errors to be orthogonal. The independence constraint is more stringent than the orthogonal constraint. Therefore, we expect that OAMP outperforms WP as it has a more relaxed constraint.

V. Simulation Results

In this section, we will provide simulation results to verify the findings obtained so far. In particular, we will show the following:

- OAMP can indeed achieve improved performance compared with WP.
- SE is quite accurate for OAMP in large coded systems. This provides numerical evidence for Conjecture 1.

The similarity and difference between OAMP and EP are discussed in different settings. *Within this section, "EMA-OAMP" refers to the MMSE-derived version of OAMP (MMSE-OAMP) unless when otherwise stated (e.g., in Fig. 8).*

A. Simulation Model

Let the singular value decomposition (SVD) of $\boldsymbol{A}$ be $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$. The system model in (1) can be rewritten as [29], [31], [41], [51]:

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{x} + \boldsymbol{n}. \tag{28}$$

Note that $\boldsymbol{U}^{\mathrm{H}}\boldsymbol{n}$ has the same distribution as $\boldsymbol{n}$. Thus, we can assume $\boldsymbol{U} = \boldsymbol{I}$ without loss of generality. We conjecture that the SE in (27) is accurate for a Haar distributed $\boldsymbol{V}$. However, to reduce the calculation complexity, we approximate a large random unitary matrix by $\boldsymbol{V} = \boldsymbol{\Pi}\boldsymbol{F}$, where $\boldsymbol{\Pi}$ is a random permutation matrix and $\boldsymbol{F}$ is a discrete Fourier transform (DFT) matrix. Note that the all the algorithms involved here admit fast implementation for this channel model. Similar models have been studied in [34], [52] for precoding purposes. The eigenvalues $\{d_i\}$ are generated as [30]: $d_i/d_{i+1} = \kappa^{1/N}$ for $i = 1, \ldots, N-1$ and $\sum_{i=1}^{N} d_i = N$. Here, $\kappa \geq 1$ controls the condition number of $\boldsymbol{A}$.

For reference, we also include the performance of a heuristic *a posteriori* probability (APP) based algorithm (abbreviated as APP hereafter). Specifically, we keep the linear estimator in EMA-OAMP unchanged, and replace the update in (16b) by

$$\boldsymbol{s}^{t+1} = \eta^{\mathrm{MMSE}}\left(\boldsymbol{r}^t, \boldsymbol{u}^t\right), \tag{29a}$$

and the variance update is changed to

$$v^{t+1} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{var}\left[x_i | r_i^t, u_i^t\right], \tag{29b}$$

with which we can establish a similar SE procedure for APP. Note that APP can achieve MMSE if the input message can be indeed modeled as an observation of $\boldsymbol{x}$ corrupted with IID Gaussian noise. However, unlike WP and OAMP, APP does not impose any additional constraint on the independency or orthogonality of input and output errors. The correlation between input and output errors may build up during iterative processing, which implies the following:

- The performance predicted by SE for APP can be better than those for WP and OAMP.
- The SE for APP may not be accurate due to the correlation problem.
- Thus the actual performance of APP can be worse than those of WP and OAMP.

The above will be verified by numerical results below.

B. MSE Transfer Functions for FEC Codes

Fig. 3 compares the transfer functions $\psi^{\mathrm{OAMP}}(\tau)$ and $\psi^{\mathrm{WP}}(\tau)$:

$$\psi^{\mathrm{OAMP}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left[\left|\eta^{\mathrm{OAMP}}(r_i, u_i) - x_i\right|^2\right], \tag{30a}$$

$$\psi^{\mathrm{WP}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left[\left|\eta^{\mathrm{WP}}(u_i) - x_i\right|^2\right], \tag{30b}$$

where $\boldsymbol{r} = \boldsymbol{x} + \mathcal{CN}(\boldsymbol{0}, \tau\boldsymbol{I})$, and the expectations are approximated via Monte Carlo simulations. We provide simulation results for both a convolutional code (left panel) and

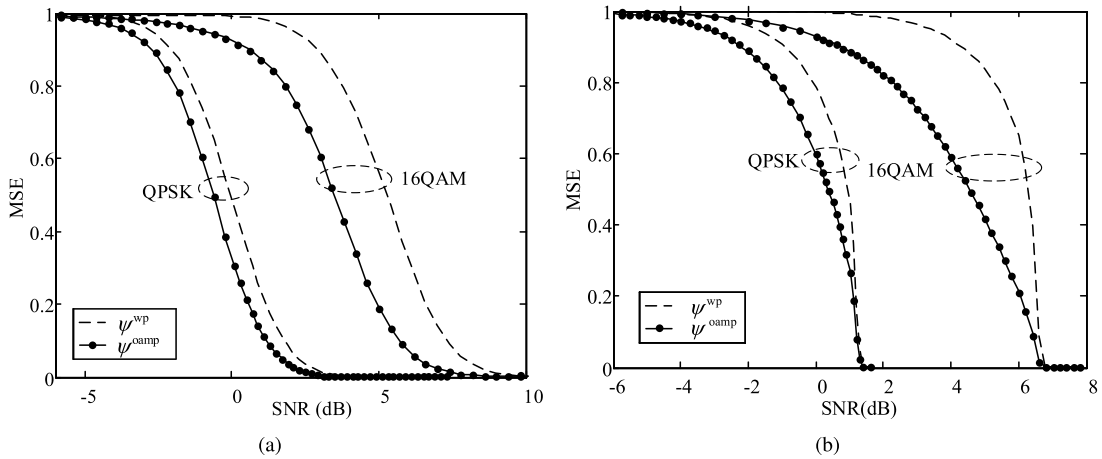


Fig. 3. Simulated MSE transfer functions for (a) a convolutional code; (b) a regular (3, 6) LDPC code with 30 inner BP iterations. The codeword lengths are 16384 bits. The number of inner iterations between the binary decoder and the demapper is 15.

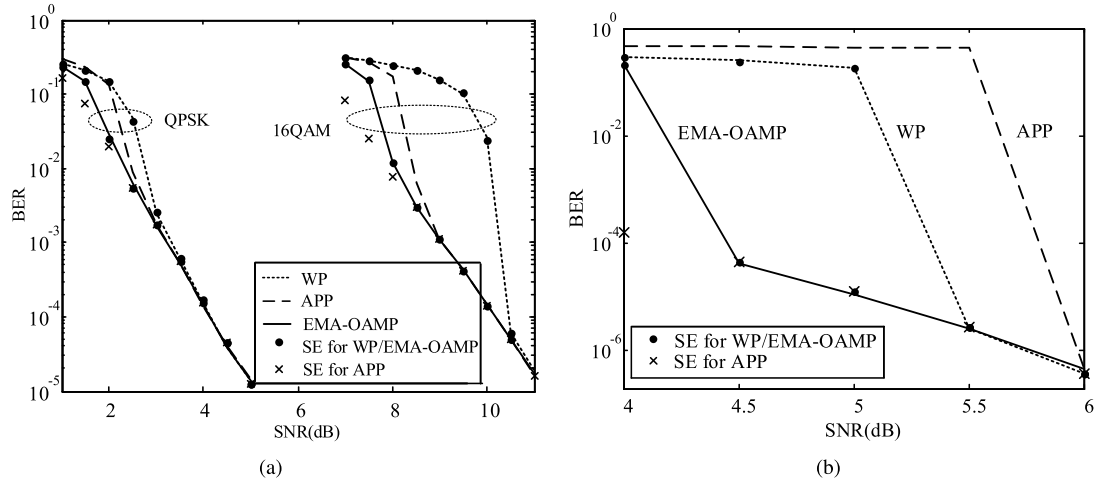


Fig. 4. **(a):** Comparison of simulation and SE predictions for $(23, 35)_8$ convolutional code. The codeword length is 32768. The channel condition number is 10. The number of iterations is 10. QPSK: $M = N = 16384$; 16QAM: $M = N = 8192$. **(b):** BER performances for condition number $\kappa = 50$. The $(23, 35)_8$ convolutional code is employed with QPSK modulation. Codeword length $= 131072$. The number of iterations is 15. $M = N = 65536$.

an LDPC code (right panel). For both figures, we employ bit-interleaved coded modulation (BICM) with Gray-mapped QPSK or 16QAM modulation.[4] For 16QAM constellations, the decoder (DEC) involves an inner iteration between the binary decoder and the demapper.

Fig. 3 shows that $\psi^{\mathrm{OAMP}}(\tau) \leq \psi^{\mathrm{WP}}(\tau)$ for both QPSK/16QAM modulations and convolution/LDPC codes, and this observation is consistent with Theorem 1. We can also see that the gain of EMA-OAMP over WP is more significant for 16QAM at a higher transmission rate. Intuitively, as the coding rate increases, the coding constraint becomes weaker, and so the reliability of the extrinsic information relative to the *a priori* information decreases. In this case, the gain of relaxing the constraints from "extrinsic" to "orthogonal" (the latter partially includes *a priori* information) is more noticeable.

[4]Our previous discussions focus on real-valued systems. For complex-valued systems, EMA-OAMP can be applied with minor modifications: (i) for the LD, the matrix transpose involved in $\boldsymbol{W}$ is replaced by conjugate transpose, and; (ii) for the NLD we assume that $\boldsymbol{r}^t$ is an observation of $\boldsymbol{x}$ corrupted by circularly-symmetric complex Gaussian noise. Based on this assumption, we can view the input of the decoder $[\boldsymbol{r}_R^t; \boldsymbol{r}_I^t]$ as is an AWGN observation of $[\boldsymbol{x}_R; \boldsymbol{x}_I]$ where $\boldsymbol{x}_R$, $\boldsymbol{x}_I$ and $\boldsymbol{r}_R^t$, $\boldsymbol{r}_I^t$ are the real and imaginary parts of $\boldsymbol{x}$ and $\boldsymbol{r}^t$ respectively. Further, in this paper, we consider square QAM signal constellations which have symmetric real and imaginary parts. Our discussions about the NLD for the real-valued case can be extended to the complex-valued case straightforwardly (by viewing $[\boldsymbol{x}_R; \boldsymbol{x}_I]$ as an effective codeword).

C. Accuracy of State Evolution

Fig. 4a compares the simulated and predicted BER performances for convolutional code. The optimal BCJR decoding is used at the receiver, i.e., $u_i^t(k) = P(x_i = s_k | \boldsymbol{r}_{\sim i}^t), \forall i, k$. From Fig. 4a, we have the following observations.

- For both WP and EMA-OAMP, the predicted BERs (based on SE) agree well with the simulated BERs. In contrast, the SE for APP deviates noticeably from simulation.
- EMA-OAMP consistently outperforms WP, and the performance gain is more noticeable for 16QAM modulation.

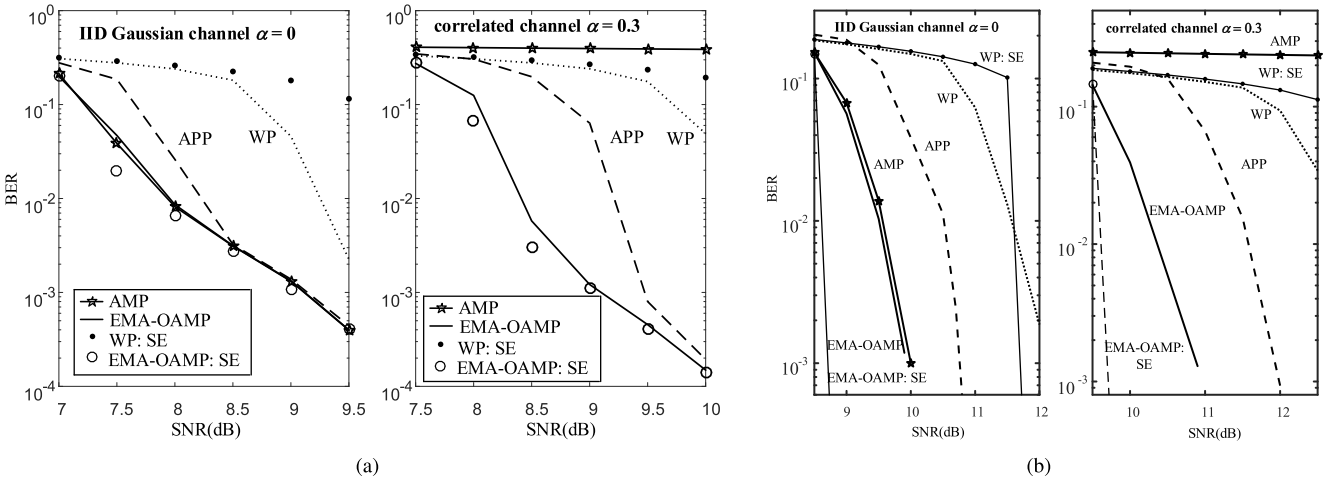


Fig. 5. BER performances for various algorithms for medium sized MIMO channels. 16QAM is used modulation with random interleaving. $m = n = 64$. **(a):** $\alpha = 0$ (IID Gaussian channel). **(b):** $\alpha = 0.3$. Number of iterations: 15 for WP, APP, and EMA-OAMP, and 30 for AMP. (a) $(23, 35)_8$ convolutional code, codeword length $= 8192$. (b) regular $(3, 6)$ LDPC code with 30 inner BP iterations, and codeword length $= 4096$. WP and APP use symbol-wise variance update for both the linear estimator and the decoder.

- APP can outperform WP.

Although APP outperforms WP in Fig. 4a, it can be inferior in other cases. In Fig. 4b, we consider a larger channel condition number. In this case, WP outperforms APP. Comparing the results in Fig. 4a and Fig. 4b, we see that the performance of APP is difficult to predict and not as reliable as WP (which can be tracked via SE).

D. Applications of EMA-OAMP for MIMO Systems

The SE analysis for EMA-OAMP is developed for a relatively large $\boldsymbol{A}$. We now provide simulation results for applications to medium-sized (smaller than one hundred) MIMO systems. For such system sizes, there will be noticeable deviations between the predicted and actual performances of OAMP; see Fig. 6 of [31]. Nevertheless, we will see that EMA-OAMP can still provide significant performance gain over WP, especially for highly correlated MIMO systems.

We assume that a codeword spans L channel uses. The overall channel can be written as

$$\boldsymbol{A} = \text{diag}\,\{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_L\}, \tag{31a}$$

where $\boldsymbol{A}_i \in \mathbb{C}^{m\times n}$ is the channel matrix at the ith channel use with $m = M/L$ and $n = N/L$. We model each $\boldsymbol{A}_i$ using the Kronecker channel model [53]:

$$\boldsymbol{A}_i = \boldsymbol{C}_R^{\frac{1}{2}} \tilde{\boldsymbol{A}}_i \boldsymbol{C}_T^{\frac{1}{2}}, \tag{31b}$$

where $\tilde{\boldsymbol{A}}_i$ consists of independent IID Gaussian elements with zero mean and variance $1/n$, and $\boldsymbol{C}_R$ and $\boldsymbol{C}_T$ are the receive and transmit correlation matrices, respectively. We use the following exponential correlation matrix for $\boldsymbol{C}_R$ (and also for $\boldsymbol{C}_T$): $(C_R)_{m,n} = \alpha^{|m-n|}$, which is a reasonable model for uniform linear arrays. Here, $\alpha \in [0, 1)$ is the correlation coefficient, and a larger α corresponds to stronger antenna correlation. For simplicity, we set the correlation coefficients for $\boldsymbol{C}_R$ and $\boldsymbol{C}_T$ to be the same.

In Fig. 5, we compare the BER performances of WP, EMA-OAMP, APP (see (29)), and AMP. For AMP, we replace $\eta(\cdot)$ in [19] by a decoder. Fig. 5a shows the simulation results for convolutional code and Fig. 5b for LDPC code. In both Fig. 5a and Fig. 5b, the left subfigures are for $\alpha = 0$ (i.e., IID Gaussian channel) and the right subfigures are for $\alpha = 0.3$. Some comments are in order.

- Accuracy of SE: there exists a noticeable mismatch between simulation and SE predictions (except for EMA-OAMP with convolutional codes). This is mainly due to the fact that the blocks of $\boldsymbol{A}$ are relatively small (i.e., 64×64). (In contrast, the results in Fig. 4 are based on a single large dense matrix.) Please refer to [31, Fig. 6] for comparison of simulations and SE predictions of OAMP under different system sizes;
- OAMP versus AMP: in both Fig. 5a and Fig. 5b, AMP and EMA-OAMP have similar BER performances when $\alpha = 0$, corresponding to IID Gaussian channels. Note that AMP slightly outperforms EMA-OAMP in the left subfigure of Fig. 5a since a larger number of iterations are used in AMP. In fact, it can be proved that the fixed points of the SEs of AMP and EMA-OAMP are the same for large IID Gaussian channels. A numerical example for the uncoded system can be found in [31, Fig. 3]. Further, as shown in the right subfigures of Figs. 5a and 5b, AMP is not as robust as EMA-OAMP for correlated MIMO channels. This is consistent with the observations in [31, Fig. 4];
- OAMP versus WP and APP: EMA-OAMP outperforms both WP and APP, similar to the results in Fig. 4. Also, we found that for highly correlated channels (e.g., $\alpha = 0.8$), WP outperforms APP (but is still worse than EMA-OAMP). The situation is similar to that of Fig. 4b.

Notice that WP and APP in Fig. 5a and 5b use symbol-wise variance (termed WP-diagonal) update for both the linear estimator and the decoder. The uniform variance treatment

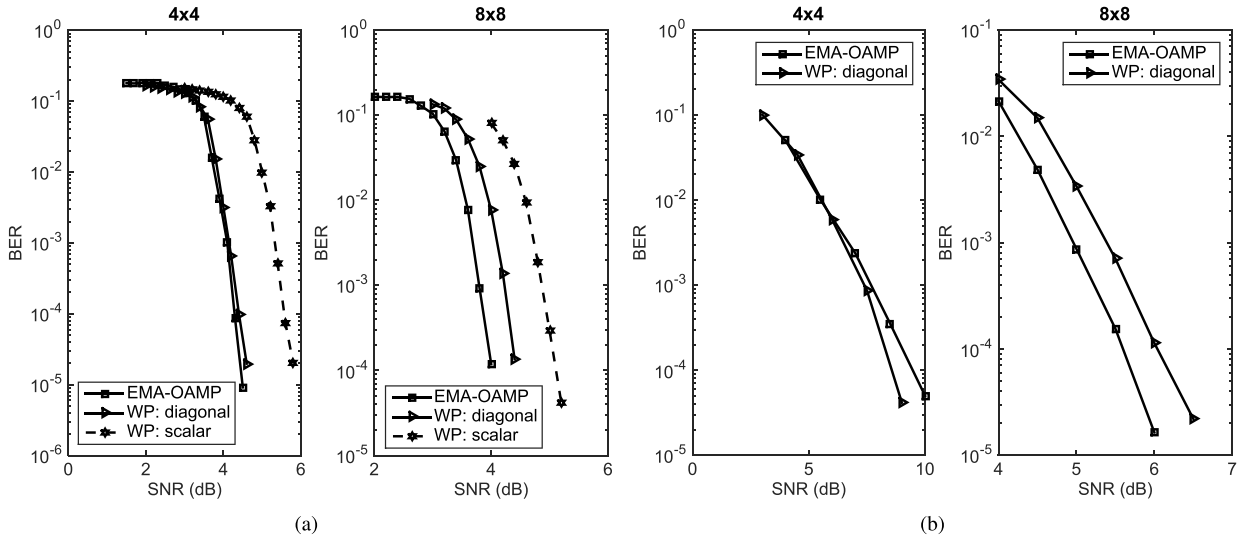


Fig. 6. Performances of EMA-OAMP and WP for small MIMO channels. The (3, 6) regular LDPC code with QPSK modulation is used. Number of EMA-OAMP/WP iterations: 30. Number of inner BP decoding iterations: 30. **(a):** $\{\boldsymbol{A}_i\}_{i=1}^{L}$ are independent. $\alpha = 0.3$. $L = 512$ for the 4×4 MIMO channel, and $L = 256$ for the 8×8 MIMO channel. **(b):** The settings are the same as those of (a), except that $\{\boldsymbol{A}_i\}_{i=1}^{L}$ are identical.

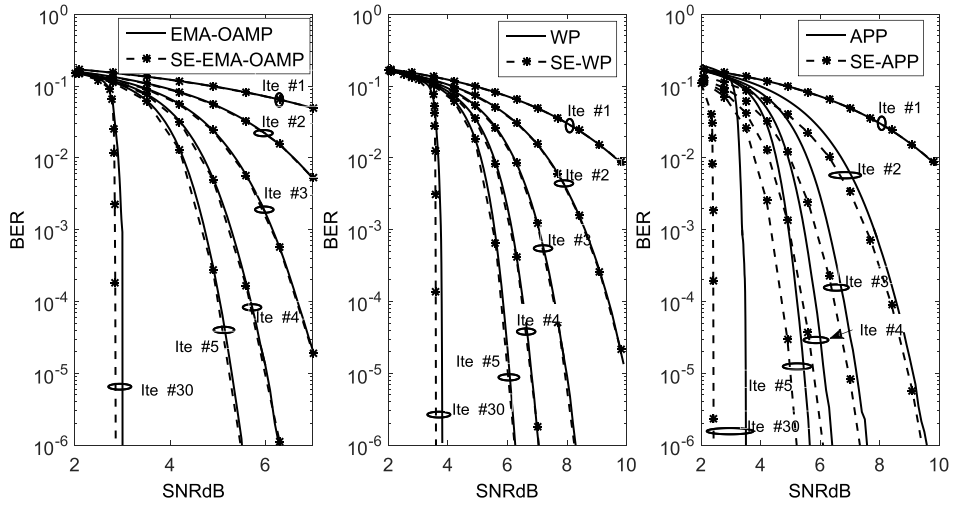


Fig. 7. Comparison of simulation and SE predictions for a regular (3, 6) LDPC code with 1 inner BP iteration. QPSK modulation and the memory sub-optimal decoding with one inner iteration at the decoder per NLD is used. The curves from right to left correspond to iterations $t = [1 \sim 5, 30]$. Other parameters are the same as those in Fig. 4.

(termed WP-scalar) in (29b) is for the purpose of establishing the connection between WP and EMA-OAMP. For properly randomized and sufficiently large $\boldsymbol{A}$ (such as that considered in Figs. 4), WP-diagonal and WP-scalar have indistinguishable differences. For the medium and small-sized system in Figs. 5, WP-diagonal outperforms WP-scalar and thus used in Fig. 5. A comparison of WP-scalar and WP-diagonal will be provided in Fig. 6.

Fig. 5 shows that EMA-OAMP outperforms WP for medium-sized MIMO channels. We emphasize that EMA-OAMP can be worse than WP for small MIMO channels. Fig. 6 shows the performances of EMA-OAMP and WP for 4×4 and 8×8 MIMO channels. We considered two different settings: for the figures on the left panel, a codeword spans L independent channels; for the figures on the right panel, the L channel matrices are identical. Comparing Fig. 4a and Fig. 6 we see that the performance gain of EMA-OAMP over WP reduces for small MIMO channels, and in some cases (Fig. 6b) WP can even outperform EMA-OAMP.

E. Memory Decoding

All our previous simulation results involving an LDPC code are based on an inner belief-propagation decoding that exchange messages between check nodes and variable nodes. In this approach, the LDPC decoder is re-initialized (namely, reset all messages to zero) in each outer iteration. In what follows, we will call this approach non-memory decoding. Clearly, this strategy has relatively high decoding complexity since it involves a large number of inner iterations. An alternative approach is to initialize the LDPC decoder by the

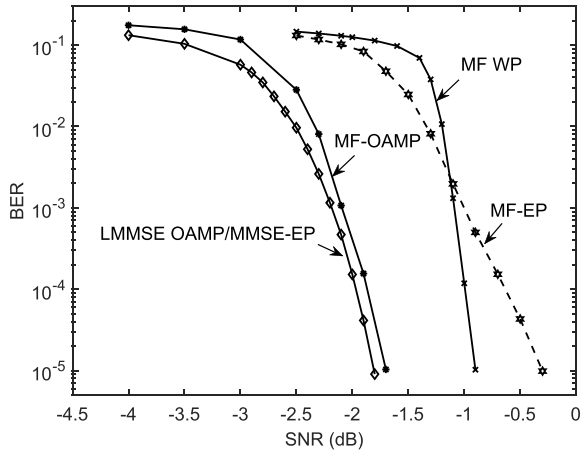


Fig. 8. BER comparisons of MF-OAMP, MF-EP and MMSE-OAMP. The (3,6) LDPC code with 30 inner BP iterations is used together with QPSK modulation with random interleaving. The codeword length is 4096. $m = 100$, $n = 32$, and $\alpha = 0.3$.

previous states and only update the LDPC decoder once at each outer iteration. This approach has a much lower decoding complexity than the non-memory decoder, and will be referred to as memory decoding.

Fig. 7 compares the simulated and predicted BER performances for a regular (3, 6) LDPC code with memory decoding. We see that the SEs of WP and EMA-OAMP agree well with the simulated BERs while the SE for APP is not accurate. Furthermore, EMA-OAMP consistently outperforms both WP and APP. Importantly, the BER performance of memory iterative receivers (with a single inner iteration at decoder) converges to that of non-memory receivers (with large inner iterations at decoder). Note that the complexity per iteration of memory decoding is greatly reduced since the inner iteration at the decoder is set to one.

F. Comparisons of MF-OAMP, MF-EP and MMSE-OAMP

Fig. 8 compares the performances of matched filter derived OAMP (MF-OAMP), matched filter derived EP (MF-EP) and MMSE derived OAMP (MMSE-OAMP), in a MIMO system characterized by (31) with $m = 100$, $n = 32$ and $\alpha = 0.3$. (Detailed discussions on the MF derived approached can be found in Appendix B.) We can observe the following:

- MMSE-OAMP (which is equivalent to MMSE-EP; see Appendix B for details) has the best performance. However, it involves high-complexity matrix inversion.
- The straightforward MF-EP (see Appendix B for details) does not work well.
- The performance gap between MF-OAMP and MMSE-OAMP is only about 0.1 dB. The former does not involve matrix inversion, so it provides an attractive low-cost alternative in massive MIMO systems.

We noticed that the performance gap between MF-OAMP and MMSE-OAMP increases when n increases towards m. We are now working on more computationally efficient solutions for such system settings.

VI. Conclusion and Future Work

This paper proposed an extrinsic message aided OAMP (EMA-OAMP) algorithm for FEC coded linear systems. Compared with the well-known Wang-Poor (WP) algorithm, EMA-OAMP relaxes the extrinsic message requirement to orthogonality one. Based on a conjectured state evolution (SE), we showed that EMA-OAMP outperforms WP. Numerical results are provided to confirm the advantages of EMA-OAMP.

The accuracy of SE recursions of WP and EMA-OAMP are conjectured based on several assumptions. The recent progress related to AMP and its variants for non-separable denoisers (with a FEC decoder being a special case) [50], [54], [55] may shed some light on this issue.

Another interesting future work is to compare the performances of the conventional diagonal-EP [56], [57] with scalar-EP/MMSE-OAMP. In a diagonal-EP algorithm, the variances produced by the decoder/demapper module can be negative. This is not a numerical issue and does not disappear even for very large random systems. A common remedy for the negative variance problem is to introduce damping and other heuristic tricks (e.g., no update whenever a negative variance occurs [56]). On the contrary, the variances in scalar-EP/MMSE-OAMP are guaranteed to be positive for unitarily-invariant $\boldsymbol{A}$ in the asymptotic regime. An extensive comparison of the performances of diagonal-EP and scalar-EP/MMSE-OAMP under various channel conditions is interesting future work.

Finally, a long-memory OAMP (LM-OAMP) algorithm was proposed in [58]. It showed that AMP is a special case of LM-OAMP, unveiling deep connections between AMP and OAMP. Extending the LM-OAMP algorithm to FEC coded systems is another interesting future direction.

Appendix A
Appendix: Derivation of (27b)

Denote $\hat{\boldsymbol{s}}^{\mathrm{MMSE}} = \eta^{\mathrm{MMSE}}(\boldsymbol{r}, \boldsymbol{u})$, where $\boldsymbol{r} = \boldsymbol{x} + \boldsymbol{z}_{\mathrm{in}}$, $\boldsymbol{x}$ is a codeword, $\boldsymbol{z}_{\mathrm{in}} \sim \mathcal{N}(\boldsymbol{0}, \tau \boldsymbol{I})$ an IID Gaussian noise, and $\boldsymbol{u} = [u_1, \ldots, u_n]$ represents the extrinsic information. Based on the law of large numbers, C_2^t converges to a deterministic constant almost surely. For notational brevity, we will omit the iteration index and write C_2^t as C.

$\eta^{\mathrm{OAMP}}(\boldsymbol{r}, \boldsymbol{u})$ in (17b) (with η and C given in (18b)) can be written as

$$\eta^{\mathrm{OAMP}}(\boldsymbol{r}, \boldsymbol{u}) = C \cdot \hat{\boldsymbol{s}}^{\mathrm{MMSE}} + (1 - C) \cdot \boldsymbol{r}, \tag{32}$$

where $\hat{\boldsymbol{s}}^{\mathrm{MMSE}} = \eta^{\mathrm{MMSE}}(\boldsymbol{r}, \boldsymbol{u})$ denotes the MMSE estimate. The MSE of $\eta^{\mathrm{OAMP}}(\boldsymbol{r}, \boldsymbol{u})$ is given by

$$\frac{1}{N}\mathrm{E}\left[\left\|\eta^{\mathrm{OAMP}}(\boldsymbol{r}, \boldsymbol{u}) - \boldsymbol{x}\right\|^2\right] \tag{33a}$$

$$= \frac{1}{N}\mathrm{E}\left[\left\|C\left(\hat{\boldsymbol{s}}^{\mathrm{MMSE}} - \boldsymbol{x}\right) + (1 - C)(\boldsymbol{r} - \boldsymbol{x})\right\|^2\right] \tag{33b}$$

$$= \frac{1}{N}\mathrm{E}\left[\left\|C\left(\hat{\boldsymbol{s}}^{\mathrm{MMSE}} - \boldsymbol{x}\right) + (1 - C)\boldsymbol{z}_{\mathrm{in}}\right\|^2\right] \tag{33c}$$

$$= C^2 \frac{1}{N}\sum_{i=1}^{N} \mathrm{mmse}\left(x_i | r_i, u_i\right) + (1 - C)^2 \tau + 2C(1 - C)\frac{1}{N}\mathrm{E}\left[\boldsymbol{z}_{\mathrm{in}}^{\mathrm{T}}\left(\hat{\boldsymbol{z}}^{\mathrm{MMSE}} - \boldsymbol{x}\right)\right] \tag{33d}$$

where $\boldsymbol{z}_{\rm in} = \boldsymbol{r} - \boldsymbol{x}$, and we have used

$$\frac{1}{N}\mathrm{E}\left[\left\|\hat{\boldsymbol{s}}^{\rm MMSE} - \boldsymbol{x}\right\|^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathrm{mmse}\left(x_i|r_i,u_i\right), \quad (34)$$

and $N^{-1}\mathrm{E}[\|\boldsymbol{z}_{\rm in}\|^2] = 1$. Now consider the last term in (33)

$$\begin{aligned}&\frac{1}{N}\mathrm{E}\left[\boldsymbol{z}_{\rm in}^{\rm T}\left(\hat{\boldsymbol{s}}^{\rm MMSE} - \boldsymbol{x}\right)\right]\\ &\quad\overset{(a)}{=} \frac{1}{N}\mathrm{E}\left[\boldsymbol{z}_{\rm in}^{\rm T}\hat{\mathbf{s}}^{\rm MMSE}\right] \overset{(b)}{=} \tau\cdot\frac{1}{N}\mathrm{E}\left[\sum_{i=1}^{N}\frac{\partial\hat{s}_i^{\rm MMSE}}{\partial r_i}\right] \quad (35a)\\ &\quad\overset{(c)}{=} \frac{1}{N}\mathrm{E}\left[\sum_{i=1}^{N}\mathrm{var}\left[x_i|r_i,u_i\right]\right] \overset{(d)}{=} \frac{1}{N}\sum_{i=1}^{N}\mathrm{mmse}\left(x_i|r_i,u_i\right)\end{aligned} \quad (35b)$$

where step (a) is from the assumption that $\boldsymbol{x}$ and $\boldsymbol{z}_{\rm in}$ are independent, step(b) from the multivariate Stein's lemma [48], step (c) from a property of MMSE estimator [21, Eq. (123)], and step (d) from the identity $\mathrm{E}\left[\mathrm{var}\left[z_i|r_i,u_i\right]\right] = \mathrm{mmse}\left(x_i|r_i,u_i\right)$. Substituting (35) into (33), we finally have (27b).

APPENDIX B
APPENDIX: COMPARISON BETWEEN EP AND OAMP

The scalar-EP algorithm and OAMP are equivalent when the optimal MMSE derived LD and NLD are used. For general non-MMSE LDs and/or NLDs, OAMP and EP can be different. We will discuss the detailed differences between EP and OAMP when a matched filter (MF) LD is used.

A. Scalar EP Algorithm

1) Linear Detector: The *scalar-EP* algorithm [32], [39] (simply called EP below) can be formulated as an iterative process involving an LD and an NLD [41]. The MMSE-LD for EP is given by [41]

$$f^{\rm EP}(\boldsymbol{s}^t) = \tau^t\cdot\left(\frac{f^{\rm MMSE}(\boldsymbol{s}^t)}{\frac{1}{N}\mathrm{Tr}(\boldsymbol{V}^t)} - \frac{\boldsymbol{s}^t}{v^t}\right), \quad (36a)$$

where

$$\tau^t = \left(\frac{1}{\frac{1}{N}\mathrm{Tr}(\boldsymbol{V}^t)} - \frac{1}{v^t}\right)^{-1}. \quad (36b)$$

Using (4a) and after some manipulations, we can express (36a) into the following equivalent form:

$$f^{\rm EP}(\boldsymbol{s}^t) = \boldsymbol{s}^t + \boldsymbol{W}^{\rm EP}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t\right), \quad (37)$$

with $\boldsymbol{W}^{\rm EP} = \frac{\boldsymbol{V}^t\boldsymbol{A}^{\rm T}\sigma^{-2}}{1-\frac{1}{v}\cdot\frac{1}{N}\mathrm{Tr}(\boldsymbol{V}^t)}$. Notice that the MMSE derived LD for WP is equivalent to the LD of a standard EP (instead of scalar-EP) algorithm. This can be seen by comparing [6, Eq. (5)] and [56, Eqs. (31) and (32)].

2) Nonlinear Detector: The non-linear constraint defined below (9) involves symbol-by-symbol modulation over $X_i = \{s_1, s_2, \ldots, s_K\}$. The related EP operation is a scalar version of (36a):

$$\eta^{\rm EP}(r_i^t,u_i^t) = \frac{v^{t+1}}{\frac{1}{N}\sum_{i=1}^{N}\mathrm{var}[x_i|r_i^t,u_i^t]}\eta^{\rm MMSE}(r_i^t,u_i^t) - \frac{v^{t+1}}{\tau^t}r_i^t, \quad (38a)$$

where

$$v^{t+1} = \left(\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\mathrm{var}[x_i|r_i^t,u_i^t]} - \frac{1}{\tau^t}\right)^{-1}. \quad (38b)$$

Using the matrix inversion lemma, it can be shown that $f^{\rm OAMP}(\cdot)$ and $f^{\rm EP}(\cdot)$ are equivalent. It is also straightforward to show that $\eta^{\rm EP}(\cdot)$ and $\eta^{\rm OAMP}(\cdot)$ are equivalent. A similar scheme based on diagonal-EP has been reported in [59].

B. MF-OAMP

For convenience of discussions, in this appendix we assume $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^{\rm T}) = N$. Recall from (4) that the MMSE LD for OAMP involves inverting an $N\times N$ matrix, which has complexity $O(N^3)$. This can be a serious burden when N is large. The low-cost matched filter (MF) LD, given by $\boldsymbol{W} = \frac{N}{\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^{\rm T})}\cdot\boldsymbol{A}^{\rm T} \approx \boldsymbol{A}^{\rm T}$, can be an useful alternative.

Using an MF detector, the LD of OAMP becomes (see line 5 of Algorithm 1) becomes

$$\boldsymbol{r}^t = \boldsymbol{s}^t + \boldsymbol{A}^{\rm T}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t). \quad (39)$$

The output variance τ_t, which is treated as the variance of the "effective noise" in the FEC decoder, is given by [31]:

$$\begin{aligned}\tau^t &= \frac{1}{N}\mathrm{Tr}\left(\mathrm{E}\left[\boldsymbol{B}^2\right]\right)\cdot v^t + \frac{1}{N}\mathrm{Tr}\left(\mathrm{E}\left[\boldsymbol{A}^{\rm H}\boldsymbol{A}\right]\right)\cdot\sigma^2 \quad (40a)\\ &= \frac{1}{N}\mathrm{E}\left[\|\boldsymbol{B}\|_F^2\right]\cdot v^t + \frac{1}{N}\mathrm{E}[\|\boldsymbol{A}\|_F^2]\cdot\sigma^2, \quad (40b)\end{aligned}$$

where the second step is from the definition $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{A}^{\rm H}\boldsymbol{A}$ and v^t is the MSE of $\boldsymbol{s}^t$. For the special case where $A_{ij}\sim\mathcal{N}(0,1/M)$, we have

$$\tau^t = \frac{N}{M}\cdot v^t + \sigma^2. \quad (41)$$

In the general case where we do not know the distributions of $\boldsymbol{A}$, we could approximate τ^t as

$$\hat{\tau}^t = \frac{1}{N}\|\boldsymbol{B}\|_F^2\cdot v^t + \frac{1}{N}\|\boldsymbol{A}\|_F^2\cdot\sigma^2. \quad (42)$$

To evaluate the above expression we need to compute the matrix $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{A}^{\rm H}\boldsymbol{A}$, which has complexity $O(N^3)$. Nevertheless, we only need to compute $\boldsymbol{B}$ once and the overall complexity of MF-OAMP is still much lower than LMMSE-OAMP.

C. MF-EP

To the best of our knowledge, no MF based EP has been derived in the literature. The key is how to approximate $f^{\mathrm{MMSE}}(\boldsymbol{s}^t)$ and $\mathrm{Tr}(\boldsymbol{V}^t)$ in (36a) without resorting to high cost matrix inversion. One possible way is to consider the connection between *mismatched* linear MMSE detector and MF detector [60, Section II-C] as follows. We rewrite (36a) using (4) as

$$f^{\mathrm{EP}}(\boldsymbol{s}^t) = \boldsymbol{s}^t + \frac{\tau^t v^t}{\frac{1}{N}\mathrm{Tr}(\boldsymbol{V}^t)}\boldsymbol{A}^{\mathrm{T}}(v^t\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t), \tag{43}$$

where $\tau^t = \left(\frac{1}{\mathrm{Tr}(\boldsymbol{V}^t)/N} - \frac{1}{v^t}\right)^{-1}$ and $\boldsymbol{V}^t$ is defined in (4). Recall from the beginning of this appendix that we assumed $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}) \approx N$ in the large system limit. Under this assumption, we have $v^t/(\mathrm{Tr}(\boldsymbol{V}^t)/N) \to 1$ when $v^t \to 0$. In this case, $f^{\mathrm{EP}}(\boldsymbol{s}^t)$ in (43) reduces to the MF detector in (39), i.e.,

$$\boldsymbol{r}^t \to \boldsymbol{s}^t + \boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}^t\right). \tag{44}$$

Thus, $f^{\mathrm{EP}}(\boldsymbol{s}^t)$ in (43) can be seen as an approximation of the MMSE counterpart in (36a) when $v^t \to 0$. It is also straightforward to verify that

$$\tau^t = \left(\frac{1}{\mathrm{Tr}(\boldsymbol{V}^t)/N} - \frac{1}{v^t}\right)^{-1} \to \sigma^2, \tag{45}$$

which is different from its counterpart in MF-OAMP; see (40) and (45). Notice that τ^t is treated as the variance of the effective noise by the decoder. The variance τ^t produced by OAMP is close to the true effective noise variance (provided that our conjecture that the SE of OAMP is accurate).

Appendix C
Appendix: Proof of $\phi^{\mathrm{OAMP}}(\tau) < \phi^{\mathrm{WP}}(\tau)$ and $\psi^{\mathrm{OAMP}}(\tau) < \psi^{\mathrm{WP}}(\tau)$

First, $\phi^{\mathrm{OAMP}}(\tau) < \phi^{\mathrm{WP}}(\tau)$ can be proved straightforwardly by applying Jensen's inequality on (26a) and (27a). We omit the details here. The proof of $\psi^{\mathrm{OAMP}}(\tau) < \psi^{\mathrm{WP}}(\tau)$ is more involved. Below, we sketch the basic steps of our proof:

1) We treat r_i as a Gaussian observation of x_i, i.e., $r_i = x_i + z_i$ where $z_i \sim \mathcal{N}(0, \tau)$.
2) $\psi^{\mathrm{OAMP}}(\cdot)$ is an increasing function of τ_i, while $\psi^{\mathrm{WP}}(\cdot)$ is independent of τ_i.
3) When $\tau_i \to \infty$, the contribution of r_i is negligible. In this case, $\psi^{\mathrm{OAMP}} = \psi^{\mathrm{WP}}$.
4) For any $0 < \tau_i < \infty$, we have $\psi^{\mathrm{OAMP}} < \psi^{\mathrm{WP}}$ from points 2 and 3.

Next, we give the details of the above steps.

A. Auxiliary Lemmas

We first introduce two technical lemmas that are crucial for our proof.

Lemma 1: Consider the following Gaussian channel

$$r_i = x_i + z_i, \tag{46}$$

where $z_i \sim \mathcal{N}(0, \rho_i^{-1})$ is independent of $\{x_i\}$, and $\{z_i\}$ are mutually independent for different i.[5] Let u_i be an extrinsic statistic of x_i, namely, it is a deterministic function of $\{r_j, j \neq i\}$. Then, the following holds for all i:

$$\frac{1}{\mathrm{mmse}\left(x_i|r_i, u_i\right)} - \rho_i \geq \frac{1}{\mathrm{mmse}\left(x_i|u_i\right)}. \tag{47}$$

Proof: Notice that $\mathrm{mmse}(x_i|r_i, u_i)$ is a function of all $\{\rho_i\}$, since r_i is a function of ρ_i and u_i is a function of $\{\rho_j, j \neq i\}$. For convenience of discussions, let us denote the left hand side of (47) as

$$g\left(\rho_1, \ldots, \rho_N\right) \equiv \frac{1}{\mathrm{mmse}\left(x_i|r_i, u_i\right)} - \rho_i. \tag{48}$$

In the following, we keep $\{\rho_j, j \neq i\}$ fixed and study the impact of ρ_i. We first note that, when $\rho_i \to 0$, the contribution of r_i in $\mathrm{mmse}(x_i|r_i, u_i)$ becomes negligible and thus

$$g\left(\rho_1, \ldots, \rho_N\right) = \frac{1}{\mathrm{mmse}\left(x_i|r_i, u_i\right)} - \rho_i \to \frac{1}{\mathrm{mmse}\left(x_i|u_i\right)}. \tag{49}$$

From (49), we can rewrite our objective in (47) as

$$g\left(\rho_1, \ldots, \rho_N\right) \geq g\left(\rho_i = 0, \{\rho_j, j \neq i\}\right). \tag{50}$$

To prove (50), it is sufficient to prove that $g(\rho_1, \ldots, \rho_N)$ is a non-decreasing function of ρ_i. To this end, we will use the following identity [61]:

$$\frac{\mathrm{dmmse}\left(x|r, u\right)}{\mathrm{d}\rho} = -\mathrm{E}\left[\left(\mathrm{var}\left[x|r, u\right]\right)^2\right], \tag{51}$$

where $r = x + \mathcal{N}(0, \rho^{-1})$ is an AWGN observation of x, and u denotes some side information of x whose distribution is not a function of ρ. As mentioned earlier, the extrinsic statistic u_i is a function of $\{r_j, j \neq i\}$ and does not depend on r_i. Together with the modeling in (46), we see that the extrinsic information u_i in $\mathrm{mmse}(x_i|r_i, u_i)$ is not a function of ρ_i. Hence, we could apply (51) to get

$$\frac{\partial}{\partial \rho_i}\mathrm{mmse}\left(x_i|r_i, u_i\right) = -\mathrm{E}\left[\left(\mathrm{var}\left[x_i|r_i, u_i\right]\right)^2\right]. \tag{52}$$

Then, the partial derivative of $g(\rho_1, \ldots, \rho_N)$ is given by

$$\frac{\partial g\left(\rho_1, \ldots \rho_N\right)}{\partial \rho_i} \tag{53a}$$

$$= \frac{\mathrm{E}\left[\left(\mathrm{var}\left[x_i|r_i, u_i\right]\right)^2\right] - \left[\mathrm{mmse}\left(x_i|r_i, u_i\right)\right]^2}{\left(\mathrm{mmse}\left(x_i|r_i, u_i\right)\right)^2} \tag{53b}$$

$$= \frac{\mathrm{E}\left[\left(\mathrm{var}\left[x_i|r_i, u_i\right]\right)^2\right] - \left(\mathrm{E}\left[\mathrm{var}\left[x_i|r_i, u_i\right]\right]\right)^2}{\left(\mathrm{E}\left[\mathrm{var}\left[x_i|r_i, u_i\right]\right]\right)^2} \tag{53c}$$

$$\geq 0, \tag{53d}$$

where the second step follows from $\mathrm{mmse}\left(x_i|r_i, u_i\right) = \mathrm{E}\left[\mathrm{var}\left[x_i|r_i, u_i\right]\right]$, and the inequality in (53d) is due to Jensen's inequality. ∎

Lemma 1 considers a set of AWGN channels with different SNRs. Lemma 2 below is a special case of Lemma 1 with $\rho_1 = \cdots = \rho_N = \rho$.

[5] In contrast, $\{x_i\}$ are correlated due to the coding constraint.

Lemma 2: Consider the following AWGN model:

$$r_i = x_i + z_i,$$

where $z_i \sim \mathcal{N}(0, \rho^{-1})$ is independent of $\{x_i\}$, and $\{z_i\}$ are mutually independent for different i. Let u_i be a deterministic function of $\{r_j, j \neq i\}$. Then, the following holds for each i:

$$\frac{1}{\operatorname{mmse}(x_i|r_i, u_i)} - \rho \geq \frac{1}{\operatorname{mmse}(x_i|u_i)}. \tag{54}$$

B. Proof of $\psi^{\mathrm{OAMP}}(\tau) < \psi^{\mathrm{WP}}(\tau)$

We prove $\psi^{\mathrm{OAMP}}(\tau) < \psi^{\mathrm{WP}}(\tau)$ using Lemma 2 below. We have

$$\frac{1}{\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\operatorname{mmse}(x_i|r_i,u_i)} - \rho} \overset{(a)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\frac{1}{\operatorname{mmse}(x_i|r_i,u_i)} - \rho} \tag{55a}$$

$$\overset{(b)}{<} \frac{1}{N}\sum_{i=1}^{N}\operatorname{mmse}(x_i|u_i), \tag{55b}$$

where step (b) is from Lemma 2, and step (a) is due to Jensen's inequality (More specifically, from a basic property of MMSE [61], we have $\operatorname{mmse}(x_i|r_i, u_i) < \frac{1}{\rho}$. Then, step (a) can be proved by applying Jensen's inequality to the convex function $h(x, \rho) \equiv \left(x^{-1} - \rho\right)^{-1}$, where $x \in (0, 1/\rho)$.)

REFERENCES

[1] J. Ma, L. Liu, X. Yuan, and L. Ping, "Iterative detection in coded linear systems based on orthogonal amp," in *Proc. IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, Dec. 2018, pp. 1–5.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Trans. Commun.*, vol. 44, no. 10, pp. 1261–1271, Oct. 1996.

[4] C. Douillard *et al.*, "Iterative correction of intersymbol interference: Turbo-equalization," *Eur. Trans. Telecommun.*, vol. 6, no. 5, pp. 507–511, 1995.

[5] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing for overloaded massive MIMO-NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 210–226, Jan. 2019.

[6] X. Yuan, L. Ping, C. Xu, and A. Kavcic, "Achievable rates of MIMO systems with linear precoding and iterative LMMSE detection," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7073–7089, Nov. 2014.

[7] L. Liu, Y. Chi, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving MIMO-NOMA: Iterative lmmse detection," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1758–1773, Apr. 2019.

[8] Y. Chi, L. Liu, G. Song, C. Yuen, Y. L. Guan, and Y. Li, "Practical MIMO-NOMA: Low complexity and capacity-approaching solution," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6251–6264, Sep. 2018.

[9] X. Wang and H. V. Poor, "Iterative (Turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1046–1061, Jul. 1999.

[10] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proc. IEEE*, vol. 95, no. 6, pp. 1295–1322, Jun. 2007.

[11] M. Tuchler, R. Koetter, and A. C. Singer, "Turbo equalization: Principles and new results," *IEEE Trans. Commun.*, vol. 50, no. 5, pp. 754–767, May 2002.

[12] R. Visoz, A. O. Berthet, and M. Lalam, "Semi-analytical performance prediction methods for iterative MMSE-IC multiuser MIMO joint decoding," *IEEE Trans. Commun.*, vol. 58, no. 9, pp. 2576–2589, Sep. 2010.

[13] A. Sanderovich, M. Peleg, and S. Shamai (Shitz), "LDPC coded MIMO multiple access with iterative joint decoding," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1437–1450, Apr. 2005.

[14] X. Yuan, Q. Guo, X. Wang, and L. Ping, "Evolution analysis of low-cost iterative equalization in coded linear systems with cyclic prefixes," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 301–310, Feb. 2008.

[15] J. Boutros and G. Caire, "Iterative multiuser joint decoding: Unified framework and asymptotic analysis," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1772–1793, Jul. 2002.

[16] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and Y. Su, "Convergence analysis and assurance for Gaussian message passing iterative detector in massive MU-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6487–6501, Sep. 2016.

[17] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[18] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

[19] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[20] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[21] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," 2010, *arXiv:1010.5141*. [Online]. Available: https://arxiv.org/abs/1010.5141

[22] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[23] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A, Math. Gen.*, vol. 36, no. 43, p. 11111, 2003.

[24] J. P. Neirotti and D. Saad, "Improved message passing for inference in densely connected systems," *Europhys. Lett.*, vol. 71, no. 5, p. 866, 2005.

[25] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 902–915, Oct. 2014.

[26] C.-K. Wen, S. Jin, K.-K. Wong, J.-C. Chen, and P. Ting, "Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1356–1368, Mar. 2015.

[27] C.-K. Wen, C. J. Wang, S. Jin, K. K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

[28] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Sparse estimation with the swept approximated message-passing algorithm," 2014, *arXiv:1406.4311*. [Online]. Available: https://arxiv.org/abs/1406.4311

[29] S. Rangan, A. K. Fletcher, P. Schniter, and U. S. Kamilov, "Inference for generalized linear models via alternating directions and Bethe free energy minimization," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 676–697, Jan. 2017.

[30] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2021–2025.

[31] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[32] M. Opper and O. Winther, "Expectation consistent approximate inference," *J. Mach. Learn. Res.*, vol. 6, pp. 2177–2204, Dec. 2005.

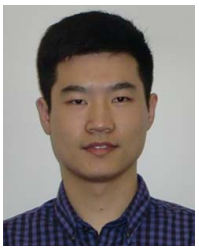
[33] X. Yuan and J. Ma, "Iterative equalization for MIMO systems: Algorithm design and evolution analysis," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 3974–3979.

[34] X. Yuan, J. Ma, and L. Ping, "Energy-spreading-transform based MIMO systems: Iterative equalization, evolution analysis, and precoder optimization," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5237–5250, Sep. 2014.

[35] J. Ma, X. Yuan, and L. Ping, "Turbo compressed sensing with partial DFT sensing matrix," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 158–161, Feb. 2015.

[36] J. Ma, X. Yuan, and L. Ping, "On the performance of turbo signal recovery with partial DFT sensing matrices," *IEEE Trans. Signal Process.*, vol. 22, no. 10, pp. 1580–1584, Oct. 2015.

[37] T. Liu, C.-K. Wen, S. Jin, and X. You, "Generalized turbo signal recovery for nonlinear measurements and orthogonal sensing matrices," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 2883–2887.
[38] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 362–369.
[39] B. Çakmak and M. Opper, "Expectation propagation for approximate inference: Free probability framework," 2018, *arXiv:1801.05411*. [Online]. Available: https://arxiv.org/abs/1801.05411
[40] K. Takeuchi, "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements," 2017, *arXiv:1701.05284*. [Online]. Available: https://arxiv.org/abs/1701.05284
[41] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," 2016, *arXiv:1610.03082*. [Online]. Available: https://arxiv.org/abs/1610.03082
[42] X. Meng, S. Wu, L. Kuang, and J. Lu, "An expectation propagation perspective on approximate message passing," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1194–1197, Aug. 2015.
[43] X. Meng, S. Wu, and J. Zhu, "A unified Bayesian inference framework for generalized linear models," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 398–402, Mar. 2018.
[44] Z. Xue, J. Ma, and X. Yuan, "D-OAMP: A denoising-based signal recovery algorithm for compressed sensing," 2016, *arXiv:1610.05991*. [Online]. Available: https://arxiv.org/abs/1610.05991
[45] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 2, May 1993, pp. 1064–1070.
[46] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *IEEE Electron. Lett.*, vol. 32, no. 18, pp. 1645–1646, Aug. 1996.
[47] S. ten Brink, G. Kramer, and A. Ashikhmin, "Design of low-density parity-check codes for modulation and detection," *IEEE Trans. Commun.*, vol. 52, no. 4, pp. 670–678, Apr. 2004.
[48] C. Stein, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, vol. 2, pp. 583–602, 1972.
[49] Z. Xue, X. Yuan, and J. Ma, "TARM: A turbo-type algorithm for low-rank matrix recovery," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4614–4618.
[50] A. K. Fletcher, S. Rangan, S. Sarkar, and P. Schniter, "Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis," 2018, *arXiv:1806.10466*. [Online]. Available: https://arxiv.org/abs/1806.10466
[51] Q. Guo and J. Xi, "Approximate message passing with unitary transformation," 2015, *arXiv:1504.04799*. [Online]. Available: https://arxiv.org/abs/1504.04799
[52] T. Hwang and Y. Li, "Novel iterative equalization based on energy-spreading transform," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 190–203, Jan. 2006.
[53] D.-S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.
[54] Y. Ma, C. Rush, and D. Baron, "Analysis of approximate message passing with a class of non-separable denoisers," 2017, *arXiv:1705.03126*. [Online]. Available: https://arxiv.org/abs/1705.03126
[55] R. Berthier, A. Montanari, and P.-M. Nguyen, "State evolution for approximate message passing with non-separable functions," 2017, *arXiv:1708.03950*. [Online]. Available: https://arxiv.org/abs/1708.03950
[56] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional MIMO systems," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2840–2849, Aug. 2014.
[57] I. Santos and J. J. Murillo-Fuentes, "Self and turbo iterations for MIMO receivers and large-scale systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1095–1098, Aug. 2019.
[58] K. Takeuchi, "A unified framework of state evolution for message-passing algorithms," 2019, *arXiv:1901.03041*. [Online]. Available: https://arxiv.org/abs/1901.03041
[59] N. Wu, W. Yuan, Q. Guo, and J. Kuang, "A hybrid BP-EP-VMP approach to joint channel estimation and decoding for FTN signaling over frequency selective fading channels," *IEEE Access*, vol. 5, pp. 6849–6858, May 2017.
[60] D. Guo and S. Verdû, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, Jun. 2005.
[61] D. Guo, Y. Wu, S. Shamai (Shitz), and S. Verdu, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, Apr. 2011.



Junjie Ma received the B.E. degree from Xidian University, China, in 2010, and the Ph.D. degree from the City University of Hong Kong, in 2015.

He was a Post-Doctoral Researcher with the City University of Hong Kong, from 2015 to 2016, and with Columbia University, from 2016 to 2019. Since July 2019, he has been with the Department of Electrical Engineering, Harvard University, as a Post-Doctoral Fellow. His current research interest includes message passing approaches for high dimensional signal processing.

Lei Liu (S'15–M'17) received the Ph.D. degree in communication and information system from Xidian University in 2017.

He was an Exchange Ph.D. Student with Nanyang Technological University, Singapore. From 2016 to 2017, he was a Research Assistant with the Singapore University of Technology and Design, Singapore, where he was a Post-Doctoral Research Fellow, in 2017. He is currently a Post-Doctoral Fellow with the City University of Hong Kong, Hong Kong. His current research interests include message passing, massive MIMO, NOMA, information theory, compressed sensing, and channel coding. He received the Ph.D. National Scholarship in China, in 2015, and the State Scholarship Fund from the China Scholarship Council, from 2014 to 2016.

Xiaojun Yuan (S'04–M'09–SM'15) received the Ph.D. degree in electrical engineering from the City University of Hong Kong, in 2008.

From 2009 to 2011, he was a Research Fellow with the Department of Electronic Engineering, City University of Hong Kong. He was also a Visiting Scholar with the Department of Electrical Engineering, University of Hawaii, Manoa, in 2009, and in the same period of 2010. From 2011 to 2014, he was a Research Assistant Professor with the Institute of Network Coding, The Chinese University of Hong Kong. From 2014 to 2017, he was an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University. He is currently a Professor with the Center for Intelligent Networking and Communications, University of Electronic Science and Technology of China, supported by the Thousand Youth Talents Plan, China. He has published over 130 peer-reviewed research articles in the leading international journals and conferences in the related areas. His research interests cover a broad range of signal processing, machine learning, and wireless communications, including but not limited to multi-antenna and cooperative communications, sparse and structured signal recovery, Bayesian approximate inference, and network coding. He has served on a number of technical programs for international conferences. He was a co-recipient of the Best Paper Award of the IEEE International Conference on Communications (ICC) 2014, and also a co-recipient of the Best Journal Paper Award of the IEEE Technical Committee on Green Communications and Computing (TCGCC) 2017. He has been an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, since 2017; and has also been an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, since 2018.

Li Ping (S'87–M'91–SM'06–F'10) received the Ph.D. degree with Glasgow University, in 1990.

He was a Lecturer with the Department of Electronic Engineering, Melbourne University, from 1990 to 1992; and was a Research Staff with the Telecom Australia Research Laboratories from 1993 to 1995. Since January 1996, he has been with the Department of Electronic Engineering, City University of Hong Kong, where he is currently a Chair Professor of information engineering.

Prof. Ping served as a member for the Board of Governors for the IEEE Information Theory Society from 2010 to 2012. He was a recipient of the IEE J J Thomson premium in 1993, the Croucher Foundation Award in 2005, and the British Royal Academy of Engineering Distinguished Visiting Fellowship in 2010.