Low-Complexity Precoding Design for Massive Multiuser MIMO Systems Using Approximate Message Passing

Jung-Chieh Chen, *Member, IEEE*, Chang-Jen Wang, Kai-Kit Wong, *Senior Member, IEEE*, and Chao-Kai Wen, *Member, IEEE*

Abstract

A practical challenge in the precoding design for massive multiuser multiple-input multiple-output (MIMO) systems is to facilitate hardware-friendly implementation. To achieve this, we propose a low peak-to-average power ratio (PAPR) precoding based on approximate message passing (AMP) algorithm to minimize multiuser interference (MUI) in massive multiuser MIMO systems. The proposed approach exhibits fast convergence and low complexity characteristics. Compared with conventional constant envelope precoding and annulus-constrained precoding, simulation results demonstrate that the proposed AMP precoding is superior both in terms of computational complexity and the average running time. In addition, the proposed AMP precoding exhibits a much desirable tradeoff between MUI suppression and PAPR reduction. These findings indicate that the proposed AMP precoding is a suitable candidate for hardware implementation, which is very appealing for massive MIMO systems.

Index Terms

Massive MIMO, message passing, PAPR.

I. Introduction

Massive multiple-input multiple-output (MIMO) antenna systems are regarded as one of the key technologies for next-generation (i.e., 5G) wireless communications systems due to their potential to improve data rate and link reliability, as well as to simplify the required signal processing [1, 2]. The distinguishing feature of such systems is its use of a *very large* antenna array at the base station (BS) using perhaps tens, even hundreds, of antenna elements to serve simultaneously a small number of user terminals (UTs). However, equipping a BS with a large number of antennas will significantly increase hardware cost and power consumption when highly linear but expensive power amplifiers are installed. Thus, to enable practical implementation of massive MIMO systems, nonlinear but power-efficient power amplifiers must be used. In this case, the transmit signals should have *low* peak-to-average power ratio (PAPR) to alleviate the impact of amplifier nonlinearities.

Recently, constant envelope (CE) precoding was proposed by Mohammed and Larsson [3–5] to reduce the PAPR of the transmit signal, thereby enabling the use of cheap and highly power-efficient power amplifiers

J. C. Chen is with the Department of Optoelectronics and Communication Engineering, National Kaohsiung Normal University, Kaohsiung 802, Taiwan (e-mail: jcchen@nknucc.nknu.edu.tw).

C. J. Wang and C. K. Wen are with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan (e-mail: dkman0988@gmail.com, ckwen@ieee.org).

K. Wong is with the Department of Electronic and Electrical Engineering, University College London, United Kingdom (e-mail: kitwong@ieee.org).

for massive multiuser MIMO systems. In CE precoding, the transmit signals are strictly limited by a fixed amplitude, and their phases are optimized to minimize multiuser interference (MUI) at all the UTs by solving a nonlinear least squares (NLS) problem. Nonetheless, the NLS problem is non-convex and difficult to solve explicitly. To tackle the NLS problem of CE precoding, the work in [4] devised a sequential gradient descent (SGD) method. Recently, [6] proposed an efficient precoder algorithm for achieving exact phase recovery while only focusing on *single-user* systems. Following the same multiuser setting as [4], Chen *et al.* [7] handled the NLS precoding problem by applying a cross-entropy optimization technique. This proposed method can provide better MUI suppression than the algorithm in [4], but requires high computational complexity, which may *not* be suitable for practical implementation.

Motivated by the fact that *relaxing* the lower bound of amplitude constraints in CE precoding does *not* distort transmit signals while introducing an additional degree of freedom to improve system performance, Mollén [8] proposed an annulus-constrained (AC) precoding scheme that allows amplitudes to vary inside a predefined interval to further improve MUI performance. Building upon the work in [4], Mollén also developed a similar SGD search method to find the optimal AC precoding weights, but with a larger searching space than that in CE precoding, which makes the developed method more computationally expensive. Also, AC precoding will increase the PAPR of the transmit signal due to the relaxation of amplitude constraints. Moreover, the computational complexity of SGD-based methods adopted in [4, 8] is still very high.

Given that the computational complexity of precoding increases with the number of BS antennas, the adoption of low-complexity precoding is preferred to facilitate the deployments of massive MIMO systems. Mathematically, however, the formulated design problems for the CE and AC precodings are *non*-convex and thus computationally intractable. The purpose of this work is *not* to obtain globally optimal solutions of these non-convex problems in a computationally intractable manner, but to achieve a *practical* solution that can reduce computational complexity without much compromise in performance. Based on this, we propose a low-complexity precoding based on the approximate message passing (AMP)[1] algorithm [11, 12] for massive multiuser MIMO systems. The key results and our contributions are summarized as follows.

- Inspired by the considerable success of the AMP algorithm in performing rapid inference within the context of compressed sensing, we have transformed the massive multiuser MIMO precoding design problem into a probabilistic inference problem through a factor graph in this study. The AMP algorithm exhibits excellent performance in terms of both precision and speed in the compressed sensing problem, while maintaining low complexity. Through this technique, we propose a low-complexity precoding scheme for massive multiuser MIMO systems.
- In the proposed framework, a factor graph decomposes the considered estimation problem into several simple mutually interactive local problems, each of which is handled by a local node. These local constraints are solved in parallel and interactively among local nodes. The message, which describes the probability distribution function of the signal component, is then efficiently updated among the local

[1] The AMP was recently proposed in the context of compressive sensing and has been revealed to be very efficient in terms of both precision and speed for digital communications and signal processing applications [9, 10].

nodes via the AMP algorithm to optimize iteratively the precoding design for massive multiuser MIMO systems.

- The simulation results reveal that the proposed AMP precoding is superior to the CE and AC precodings in terms of the required number of complex multiplications and the time it takes to solve the problem. In addition, the proposed AMP precoding can provide a trade-off between MUI suppression and PAPR reduction with desirable results.

Notations—Throughout the paper, we use $\mathbb{R}$ and $\mathbb{C}$ to represent the set of real numbers and complex numbers, respectively. The superscripts $(\cdot)^{\mathsf{T}}$ and $(\cdot)^{*}$ denote the transpose and conjugate transpose, respectively. $\mathsf{CN}(0,1)$ denotes the complex Gaussian distribution with zero mean and unit variance. $\mathfrak{Re}\{\cdot\}$ returns the real part of its input argument. $\mathrm{Arg}(\cdot)$ returns the principal argument of its input complex number. Finally, $\mathrm{j} \triangleq \sqrt{-1}$.

II. System Model and Problem Formulation

A. System Model

We consider the downlink transmission of a massive multiuser MIMO system, where a BS is equipped with N antennas and communicates with M single-antenna UTs. Let $\mathbf{x} \in \mathbb{C}^N$ denote a transmitted vector whose i-th element, $x_i \triangleq \alpha_i\, e^{\mathrm{j}\theta_i}$, is the complex signal/value transmitted from the i-th BS antenna, where $\alpha_i \in \mathbb{R}^+$ is the amplitude of x_i and $\theta_i \in [-\pi, \pi)$ is the phase of x_i. The collectively received vector, denoted by $\mathbf{y} \in \mathbb{C}^M$, at the M UTs is given as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{M\times N}$ contains the flat fading channel coefficients with the (μ, i)-th entry being the complex-Gaussian channel tap between the i-th BS antenna and the μ-th UT; $\mathbf{z} \in \mathbb{C}^M$ is a noise vector, with z_μ being the additive noise at the μ-th UT; and the entries of $\mathbf{H}$ and $\mathbf{z}$ are independent and identically distributed (i.i.d.) as $\mathsf{CN}(0,1)$. The received signal $y_\mu \in \mathbb{C}$ at the μ-th UT can then be expressed as

$$y_\mu = \sum_{i=1}^{N} H_{\mu i}\, x_i + z_\mu, \;\text{ for } \mu = 1, 2, \ldots, M. \tag{2}$$

B. Problem Formulation

To minimize the MUI at all the UTs while facilitating the use of power-efficient amplifiers at the BS, Mohammed and Larsson [4] proposed a low-PAPR precoding scheme that uses CE signals for massive multiuser MIMO systems. The main idea of [4] is to constrain the transmit signal $x_i = \alpha\, e^{\mathrm{j}\theta_i}$ of each antenna to have a CE $\alpha = \frac{1}{\sqrt{N}}$. The objective of the CE precoder at the BS is to find a transmit phase vector $\boldsymbol{\theta} = [\theta_1 \,\ldots\, \theta_N]^{\mathsf{T}}$, such that the noiseless signal received at each UT forms a desired information signal.

The degrees of freedom of the CE precoding design problem are clearly $N - M$. Instead of directly increasing the number of BS antennas to produce the *extra* degrees of freedom of the null space to achieve the desired MUI level, Mollén [8] proposed an AC precoding scheme, which *relaxes* amplitude constraints in the CE precoding to improve MUI performance. By allowing the transmit amplitude α_i of each antenna to vary in the interval $\mathcal{D}_\varepsilon \triangleq [\alpha-\varepsilon, \alpha]$, the AC precoder needs to *simultaneously* find a transmit phase vector $\boldsymbol{\theta} = [\theta_1 \ldots \theta_N]^\mathsf{T}$ and a transmit amplitude vector $\boldsymbol{\alpha} = [\alpha_1 \ldots \alpha_N]^\mathsf{T}$ to solve the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\alpha},\boldsymbol{\theta}}{\text{minimize}} \quad & \mathcal{G}\left(\boldsymbol{\alpha},\boldsymbol{\theta}\right) \\ \text{subject to} \quad & |\theta_i| \le \pi, \quad \text{for } i = 1,2,\ldots,N, \\ & \alpha_i \in \mathcal{D}_\varepsilon, \quad \text{for } i = 1,2,\ldots,N, \end{aligned} \tag{3}$$

where

$$\mathcal{G}\left(\boldsymbol{\alpha},\boldsymbol{\theta}\right) \triangleq \frac{1}{M}\sum_{\mu=1}^{M}\left|\sum_{i=1}^{N} H_{\mu i}\,\alpha_i\, e^{\mathrm{j}\theta_i} - u_\mu\right|^2 \tag{4}$$

is the average MUI energy over all the UTs, and u_μ is the symbol intended for the μ-th UT. For $\alpha = \frac{1}{\sqrt{N}}$ and $\varepsilon = 0$, the AC precoding reduces to the CE precoding.

Both the CE and AC precoding problems are *non-convex* and are very difficult to solve. To deal with the AC precoding problem, Mollén developed a SGD-based search method, which is in fact built on the same SGD idea used in the search algorithm for the CE precoding [4] while searching the optimal amplitudes and the optimal phases concurrently. Owing to the relaxation of amplitude constraints, AC precoding can achieve better MUI performance than CE precoding, but at the cost of computational complexity and the PAPR performance.

III. Proposed Algorithm

A. AMP-Based Precoding

In this subsection, we propose to adopt the AMP algorithm [11, 12] to search the optimal $\mathbf{x}^\star$ that minimizes the MUI power $\|\mathbf{H}\mathbf{x}-\mathbf{u}\|^2$ subject to certain PAPR constraints on $\mathbf{x}$, where $\mathbf{u} = [u_1 \ldots u_M]^\mathsf{T} \in \mathbb{C}^M$. Because the AMP algorithm is a tool for probabilistic inference, we have to recast the optimization problem into an estimation problem. Toward this end, we introduce the following *virtual* model [13]

$$\mathbf{u} = \mathbf{H}\mathbf{x}. \tag{5}$$

By this virtual model, we can think of the precoding design as an *estimate* of $\mathbf{x}$ based on the received signal $\mathbf{u}$. Next, we shall describe how $\mathbf{x}$ can be estimated based the probabilistic inference approach.

Based on (5), the conditional probability distribution function (pdf) of the unknown signal $\mathbf{x}$ given the measurements $\mathbf{u}$ and the knowledge of matrix $\mathbf{H}$ can be obtained by using Bayes' rule as

$$\mathscr{P}(\mathbf{x}|\mathbf{H},\mathbf{u}) \propto \mathscr{P}(\mathbf{u}|\mathbf{H},\mathbf{x})\mathscr{P}(\mathbf{x}) \propto \underbrace{\prod_{\mu=1}^{M}\mathscr{P}(u_\mu|\mathbf{H},\mathbf{x})}_{\text{the likelihood function}}\underbrace{\prod_{i=1}^{N}\mathscr{P}(x_i)}_{\text{the prior pdf}}, \tag{6}$$

where $\propto$ denotes the identity after normalization to unity, $\mathscr{P}(x_i)$ is the prior distribution of the i-th element of the signal $\mathbf{x}$, and $\mathscr{P}(u_\mu|\mathbf{H},\mathbf{x})$ stands for the μ-th receiver constraint that is given by

$$\mathscr{P}(u_\mu|\mathbf{H},\mathbf{x}) = \delta\left(u_\mu - \sum_{i=1}^{N} H_{\mu i}x_i\right), \tag{7}$$

where $\delta(\cdot)$ denotes the Dirac delta function. Once the posterior pdf is defined, the Bayes optimal way of estimating $\mathbf{x} = \{x_i\}_{i=1}^{N}$ is given by

$$\widehat{x}_i = \int \mathrm{d}x_i \, x_i \mathscr{Q}(x_i), \ \forall i, \tag{8}$$

where $\mathscr{Q}(x_i)$ is the marginal probability distribution of x_i, which is defined as

$$\mathscr{Q}(x_i) \triangleq \int \prod_{k \neq i} \mathrm{d}x_k \, \mathscr{P}(\mathbf{x}|\mathbf{H}, \mathbf{u}). \tag{9}$$

If $\mathscr{P}(\mathbf{x})$ is assumed to be the standard complex Gaussian distribution, i.e., $\mathscr{P}(\mathbf{x}) = \frac{1}{\pi^N} e^{-\|\mathbf{x}\|^2}$, then the solution of (8) is $\widehat{\mathbf{x}} = \mathbf{H}^*(\mathbf{H}\mathbf{H}^*)^{-1}\mathbf{u}$, which is, in fact, the conventional zero-forcing (ZF) precoding which has a high PAPR value. Clearly, by adjusting the prior distribution $\mathscr{P}(\mathbf{x})$, we can restrict the amplitude variation of $\mathbf{x}$ and thus reduce the PAPR of the transmit signal. In this work, we adopt $\mathscr{P}(x_i = \alpha\, e^{\mathrm{j}\theta_i}) = \frac{1}{2\pi}$ with $\theta_i \in (0,\, 2\pi]$, $\forall i$, which indicates that θ_i is regarded as independent and uniformly distributed within the range of $(0,\, 2\pi]$. In this case, x_i will be a random point on a circle of radius α in the complex plane. These conditions are selected to constrain the transmit signal $x_i = \alpha\, e^{\mathrm{j}\theta_i}$ of each antenna to achieve a *constant envelope*, similar to that in the CE precoding. However, although the argument x_i is restricted to the points on the circle of radius α, the output $\widehat{x}_i$ in (8), which is a combination of all points on the circle can no longer be restricted on the circle. Fortunately, based on our observations made through numerical simulations, we find that most of the $\widehat{x}_i$'s remain close to the circle. Thanks to this property[2], the PAPR values of the AMP-based precoding are expected to be lower than the AC precoding but higher than the CE precoding.

Although the precoding design through (8) seems promising, the exact evaluation of (9) involves high-dimensional integrals that make (8) intractable. Thus, we have to resort to computing *approximate* marginal posterior pdfs $m_i(x_i) \approx \mathscr{Q}(x_i)$, called "beliefs," which provide a practical alternative to the direct computation of marginal posteriors. By inspecting (6), we observe that the likelihood function is decomposed into a product of M factors relative to the constraint over each u_μ, and the prior pdf is decomposed into a product of N factors relative to what is expected of each x_i. This decomposition can be represented by the *factor graph* [14], as shown in Fig. 1. The graph is a bipartite graph that contains N variable nodes and $M + N$ factor nodes, connected by an edge when the corresponding variable appears in the corresponding factor. As long as a joint posterior pdf can be represented by a factor graph model, an approximate solution for the posterior marginals of the variables can be obtained through an iterative message-passing procedure among variable nodes and factor nodes based on belief propagation (BP) [12, 15].[3]

BP involves two types of messages: (1) messages from variable nodes to factor nodes, denoted by $m_{i\to\mu}(x_i)$, and (2) messages from factor nodes to variable nodes, denoted by $m_{\mu\to i}(x_i)$. Here, "messages" are probability distribution functions. Two message update rules are available, which are described as follows [15]:

[2] The proposed AMP precoding attempts to find $\widehat{x}_i$'s and allow them to remain near the circle. Notably, if *all* $\widehat{x}_i$'s are *on the circle*, then the proposed AMP precoding achieves the same PAPR performance as that of the CE precoding.

[3] For convenience, the similar notations to those in [12] are used.

- *Variable-to-factor message update rule*. The messages in BP being passed from variable x_i to factor u_μ is given by
$$m_{i\to\mu}(x_i) = \frac{1}{Z_{i\to\mu}} \mathscr{P}(x_i) \prod_{\gamma\neq\mu} m_{\gamma\to i}(x_i), \tag{10}$$
where $Z_{i\to\mu}$ is the normalization factor that guarantees $\int \mathrm{d}x_i\, m_{i\to\mu}(x_i) = 1$. That is, the message from variable node x_i to factor node u_μ is simply the product of all the incoming messages at variable node x_i *excluding* the message from target factor node u_μ.
- *Factor-to-variable message update rule*. The messages in BP being passed from factor u_μ to variable x_i is updated as
$$m_{\mu\to i}(x_i) = \frac{1}{Z_{\mu\to i}} \int \prod_{k\neq i} \mathrm{d}x_k \left[\mathscr{P}(u_\mu|\mathbf{H},\mathbf{x}) \prod_{k\neq i} m_{k\to\mu}(x_k) \right], \tag{11}$$
where $Z_{\mu\to i}$ is the scale factor to ensures $\int \mathrm{d}x_i\, m_{\mu\to i}(x_i) = 1$. That is, the message from factor node u_μ to variable node x_i is given by the integral over the product of the factor function itself and all the incoming messages *excluding* the messages from target variable node x_i.

According to these two rules, messages are passed iteratively between the variable nodes and the factor nodes. After messages converge, the approximate marginal distribution or *belief* at variable node x_i is computed as
$$m_i(x_i) = \frac{1}{Z_i} \mathscr{P}(x_i) \prod_{\mu=1}^{M} m_{\mu\to i}(x_i), \tag{12}$$
where Z_i is a normalization constant for $\int \mathrm{d}x_i\, m_i(x_i) = 1$.

However, obtaining an exact evaluation of (11) is still practically intractable due to the high-dimensional integrals involved. To enhance computational tractability, we have to simplify the expression of (11). First, we rewrite $\mathscr{P}(u_\mu|\mathbf{H},\mathbf{x})$ in (7) using the following identity[4]
$$\delta\left(u_\mu - \sum_{i=1}^{N} H_{\mu i} x_i\right) = \frac{1}{4\pi^2} \int \mathrm{d}\tilde{u}_\mu\, e^{-\mathrm{j}\mathfrak{Re}\left\{\tilde{u}_\mu^*\left(u_\mu - \sum_{i=1}^{N} H_{\mu i} x_i\right)\right\}}. \tag{13}$$
Inserting this identity into (11) yields
$$m_{\mu\to i}(x_i) = \frac{1}{\widetilde{Z}_{\mu\to i}} \int \mathrm{d}\tilde{u}_\mu\, e^{-\mathrm{j}\mathfrak{Re}\left\{\tilde{u}_\mu^*(u_\mu - H_{\mu i} x_i)\right\}} \int \prod_{k\neq i} \mathrm{d}x_k m_{k\to\mu}(x_k) e^{\mathrm{j}\mathfrak{Re}\left\{\tilde{u}_\mu^* H_{\mu k} x_k\right\}} \tag{14}$$
with $\widetilde{Z}_{\mu\to i}$ being a normalization factor ensuring that $\int \mathrm{d}x_i\, m_{\mu\to i}(x_i) = 1$. Before proceeding, we introduce the means and variances of the variable-to-factor messages as
$$a_{i\to\mu} = \int \mathrm{d}x_i\, x_i\, m_{i\to\mu}(x_i), \tag{15}$$
$$v_{i\to\mu} = \int \mathrm{d}x_i\, |x_i|^2\, m_{i\to\mu}(x_i) - |a_{i\to\mu}|^2. \tag{16}$$

[4] This identity follows the inverse Fourier transform of the Dirac delta function.

Using the above definitions, we expand the Taylor series of the last exponential in (14) up to the second order of $\mathrm{j}\mathfrak{Re}\left\{\tilde{u}_\mu^* H_{\mu k} x_k\right\}$, which yields

$$m_{\mu\to i}(x_i) = \frac{1}{Z_{\mu\to i}} \int \mathrm{d}\tilde{u}_\mu \, e^{-\mathrm{j}\mathfrak{Re}\left\{\tilde{u}_\mu^*(u_\mu - H_{\mu i}x_i)\right\}} \times e^{\mathrm{j}\mathfrak{Re}\left\{\tilde{u}_\mu^*\left(\sum_{k\neq i} H_{\mu k} a_{k\to\mu}\right)\right\} - |\tilde{u}_\mu|^2\left(\sum_{k\neq i}|H_{\mu k}|^2 v_{k\to\mu}\right)}. \tag{17}$$

Performing the integral over $\tilde{u}_\mu$, we obtain

$$m_{\mu\to i}(x_i) = \frac{1}{\widetilde{Z}_{\mu\to i}} \, e^{-A_{\mu\to i}\,|x_i|^2 + 2\,\mathfrak{Re}\{B_{\mu\to i}\,x_i\}} \tag{18}$$

with

$$A_{\mu\to i} \triangleq \frac{|H_{\mu i}|^2}{\sum_{k\neq i}|H_{\mu k}|^2 v_{k\to\mu}}, \tag{19}$$

$$B_{\mu\to i} \triangleq \frac{H_{\mu i}^*\left(u_\mu - \sum_{k\neq i} H_{\mu k} a_{k\to\mu}\right)}{\sum_{k\neq i}|H_{\mu k}|^2 v_{k\to\mu}}. \tag{20}$$

Next, substituting (18) into (10), we have

$$m_{i\to\mu}(x_i) \propto \mathscr{P}(x_i)\, e^{-|x_i|^2 \sum\limits_{\gamma\neq\mu} A_{\gamma\to i} + 2\,\mathfrak{Re}\left\{x_i \sum\limits_{\gamma\neq\mu} B_{\gamma\to i}\right\}} \propto \mathscr{M}\left(x_i; \frac{\sum_{\gamma\neq\mu} B_{\gamma\to i}}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}, \frac{1}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}\right), \tag{21}$$

in which we define

$$\mathscr{M}(x; R, \Sigma^2) \triangleq \frac{1}{\widehat{Z}(R,\Sigma^2)} \mathscr{P}(x) \frac{1}{\pi\Sigma} e^{-\frac{|x-R|^2}{\Sigma^2}} \tag{22}$$

with $\widehat{Z}(R,\Sigma^2) = \int \mathrm{d}x \, \mathscr{P}(x) \frac{1}{\pi\Sigma} e^{-\frac{|x-R|^2}{\Sigma^2}}$ being the normalization constant. To simplify the notation, we also define the mean and the variance of (22) by the functions f_a and f_c, respectively, as

$$f_a(R,\Sigma^2) \triangleq \int \mathrm{d}x \, x \, \mathscr{M}(x; R, \Sigma^2) = \alpha\, e^{\mathrm{j}\,\mathrm{Arg}(R)} \left[\frac{\mathcal{I}_1\left(\frac{2\alpha|R|}{\Sigma^2}\right)}{\mathcal{I}_0\left(\frac{2\alpha|R|}{\Sigma^2}\right)}\right], \tag{23}$$

$$f_c(R,\Sigma^2) \triangleq \int \mathrm{d}x \, |x|^2 \, \mathscr{M}(x; R, \Sigma^2) - \left|f_a(R,\Sigma^2)\right|^2 = \alpha^2 - \left|f_a(R,\Sigma^2)\right|^2, \tag{24}$$

where $\mathcal{I}_\nu(\cdot)$ is the ν-th order modified Bessel function of the first kind. In addition, the expressions (23) and (24) that are used to approximately compute the functions f_a and f_c are based on the derivations of [16] with a slightly modification.

With these definitions, the equations for the mean $a_{i\to\mu}$ (15) and variance $v_{i\to\mu}$ (16) can be expressed as

$$a_{i\to\mu} = f_a\left(\frac{\sum_{\gamma\neq\mu} B_{\gamma\to i}}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}, \frac{1}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}\right), \tag{25}$$

$$v_{i\to\mu} = f_c\left(\frac{\sum_{\gamma\neq\mu} B_{\gamma\to i}}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}, \frac{1}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}\right). \tag{26}$$

Then (19) and (20) together with (25) and (26) lead to closed iterative message-passing equations, which are referred to as relaxed BP (r-BP) equations. The message passing procedures are repeated until some termination conditions are satisfied. When a fixed point of this iteration is reached, the mean a_i and variance

v_i of $m_i(x_i)$ should be computed. These can be obtained by simply adding back the μ dependent part to (25) and (26) as

$$a_i = f_a\left(R_i, \Sigma_i^2\right), \tag{27}$$

$$v_i = f_c\left(R_i, \Sigma_i^2\right), \tag{28}$$

where

$$R_i \triangleq \frac{\sum_\mu B_{\mu\to i}}{\sum_\mu A_{\mu\to i}} \text{ and } \Sigma_i^2 \triangleq \frac{1}{\sum_\mu A_{\mu\to i}}. \tag{29}$$

The above description clearly indicates that even when only the two first moments of all messages are kept in the r-BP equations, the number of messages that needs to be tracked for r-BP is $2M \times N$ per iteration. However, this poses a problem if we want to scale an application to a large M and N. To further reduce the number of messages, we *omit* the negligible terms in the large N limit, and then rewrite the r-BP equations on quantities that correspond to the variables and factors nodes. As a result, the number of r-BP messages to be tracked is *reduced* to only $M+N$ for each iteration. This leads to the so-called AMP (approximated message passing) [11, 12]. To this end, we first define two auxiliary variables, ω_μ and V_μ, for every measurement component μ to ease notation:

$$\omega_\mu \triangleq \sum_i H_{\mu i} a_{i\to\mu}, \tag{30}$$

$$V_\mu \triangleq \sum_i |H_{\mu i}|^2 v_{i\to\mu} \approx \sum_i |H_{\mu i}|^2 v_i. \tag{31}$$

These two quantities, in conjunction with (29), yield

$$R_i = \frac{\sum_\mu \frac{H_{\mu i}^*(u_\mu - \omega_\mu + H_{\mu i} a_{i\to\mu})}{V_\mu - |H_{\mu i}|^2 v_{i\to\mu}}}{\sum_\mu \frac{|H_{\mu i}|^2}{V_\mu - |H_{\mu i}|^2 v_{i\to\mu}}} \overset{(a)}{\approx} a_i + \frac{\sum_\mu H_{\mu i}^* \frac{(u_\mu - \omega_\mu)}{V_\mu}}{\sum_\mu \frac{|H_{\mu i}|^2}{V_\mu}}, \tag{32}$$

$$\Sigma_i^2 = \frac{1}{\sum_\mu \frac{|H_{\mu i}|^2}{V_\mu - |H_{\mu i}|^2 v_{i\to\mu}}} \overset{(a)}{\approx} \left[\sum_\mu \frac{|H_{\mu i}|^2}{V_\mu}\right]^{-1}. \tag{33}$$

In (a) above, we assume that R_i and Σ_i^2 are $\mathcal{O}(1)$, i.e., the magnitudes of these quantities stay *finite* as $N \to \infty$ (see [12] for details). We now turn to approximate $a_{i\to\mu}$ by applying Taylor's expansion to (25) around R_i as

$$a_{i\to\mu} = f_a\left(\frac{\sum_\gamma B_{\gamma\to i} - B_{\mu\to i}}{\sum_\gamma A_{\gamma\to i} - A_{\mu\to i}}, \frac{1}{\sum_\gamma A_{\gamma\to i} - A_{\mu\to i}}\right) \simeq a_i - B_{\mu\to i}\Sigma_i^2 \frac{\partial f_a(R_i, \Sigma_i^2)}{\partial R_i} = a_i - B_{\mu\to i} v_i, \tag{34}$$

where the last equality follows from the fact that $\Sigma^2 \frac{\partial f_a(R,\Sigma^2)}{\partial R} = f_c(R, \Sigma^2)$. Finally, multiplying (34) by $H_{\mu i}$ and summing the resultant expressions over i yields

$$\omega_\mu = \sum_i H_{\mu i} a_i - \frac{(u_\mu - \omega_\mu)}{V_\mu} V_\mu, \tag{35}$$

which allows us to close the equations on the set of a_i, v_i, R_i, Σ_i, V_μ, and ω_μ.

Specifically, the AMP algorithm works as follows. We begin by initializing $\{a_i^{(0)}\}_{i=1}^N$, $\{v_i^{(0)}\}_{i=1}^N$, $\{V_\mu^{(0)}\}_{\mu=1}^M$, and $\{\omega_\mu^{(0)}\}_{\mu=1}^M$ to certain values,[5] where the superindex denotes the iteration index. Next, these values are inserted into (31) and (35) to give $\{V_\mu^{(1)}\}_{\mu=1}^M$ and $\{\omega_\mu^{(1)}\}_{\mu=1}^M$, respectively. The corresponding values are then further inserted into (33) and (32) to yield $\{(\Sigma_i^{(1)})^2\}_{i=1}^N$ and $\{R_i^{(1)}\}_{i=1}^N$. Finally, the *first* iteration is completed by substituting the corresponding values into (27) and (28) to offer $\{a_i^{(1)}\}_{i=1}^N$ and $\{v_i^{(1)}\}_{i=1}^N$. We continue to iterate in this way (i.e., updating all V, ω's, then Σ, R's, and then a, v's) until convergence (i.e., the quantities no longer change). The estimate of the signal component x_i is $a_i^{(l)}$. The pseudocode for the evolution of V_μ, ω_μ, Σ_i, R_i, a_i, and v_i is summarized in Algorithm 1.

B. Complexity Analysis

The computational complexities are compared in terms of the number of *complex* multiplications of CE [4], AC [8] and AMP. First, we analyze the computational complexity of Algorithm 1. For factor nodes, Line 9 of Algorithm 1 involves $2N$ complex multiplications. Line 10 of Algorithm 1 needs $N+2$ complex multiplications. Therefore, Lines 8 to 10 require $3MN+2M$ complex multiplications for each iteration. One the other hand, for variable nodes, Lines 12 and 13 both need $2M+1$ complex multiplications. Lines 14 and 15 require 5 and 2 complex multiplications for the expressions (23) and (24) that are used to approximately compute the functions f_a and f_c, respectively. Hence, Lines 12 to 15 require $4MN+9N$ complex multiplications for each iteration. Then the AMP method requires a total of $7MN+9N+2M$ complex multiplications for each iteration. However, CE and AC need N^2M-NM and $2N^2M-NM$ complex multiplications for each iteration, respectively. Fig. 2 illustrates the variation of the number of complex multiplications of the three precodings as a function of the number of BS antennas for $M=35$ and $M=50$. It is observed that the proposed method not only has the *lowest* complexity among the three, but also is *less* sensitive to N and M. In particular, for the case of $(N,M)=(90,35)$, the number of complex multiplications required in AMP is only 8.18% and 4.07%, respectively, to that required in the CE and AC precodings. For the system configuration with $(N,M)=(130,50)$, the proposed method needs approximately 5.88% and 2.78% of the computational complexity of the CE and AC precodings. Note that the system configurations with $(N,M)=(90,35)$ and $(N,M)=(130,50)$ are enough to achieve the desired MUI level, as will be demonstrated in the following section. Finally, we note that in contrast to the CE and AC precodings, which are found by sequential methods, the AMP precoding exhibits a parallel nature and low computational complexity, and requires low arithmetic precision, which makes it well-suited for hardware implementation [17].

IV. Simulation Results and Discussion

In this section, we evaluate the performance of the proposed AMP-based precoding design for different system configurations. The performance measures[6] used are the average MUI energy over all the UTs (e.g.,

[5] The initial conditions of these values are specified in Lines 2 to 5 of Algorithm 1.

[6] Given that the amplitude constraints of the AC and AMP precodings are *relaxed*, we empirically found that the throughput behaviors of these two precodings are very *similar*. A detailed comparison of the throughputs of the AC, CE, and ZF precodings can be found in [8].

performance at the receiver side) and the PAPR statistics (e.g., performance at the transmitter side). The average MUI energy over all the UTs is given by $\mathbb{E}\{\frac{1}{M}\|\mathbf{H}\widehat{\mathbf{x}}-\mathbf{u}\|^2\}$, where the expectations are computed as averages of $10,000$ independent channel realizations in the simulations unless explicitly stated otherwise. For the AMP-based precoding, we adopt $\mathscr{P}(x_i)$ being a uniform ring distribution with radius $\alpha = \frac{1.07}{\sqrt{N}}$ in the complex plane.[7] For comparison, the CE [4] and AC [8] precodings are tested, where the transmit signals of the AC precoding are constrained to a disk in the complex plane with radius $\alpha = \frac{1}{\sqrt{N}}$ (i.e., $\varepsilon = \alpha$). In addition, the channel gains between the BS and each UT follow a circular Gaussian distribution $\mathsf{CN}(0,1)$ in an i.i.d. manner. The message symbol intended for each UT is chosen to be $u_\mu =$ 16-QAM for $\mu = 1, 2, \ldots, M$.

Given that the aforementioned three precodings are iterative-based algorithms, two important aspects of any iterative method are to check if the algorithm truly converges to a fixed point and to determine when the iterations should be stopped. Therefore, we first fix the system configuration at $(N, M) = (100, 35)$ and determine how the average MUI performance of various precodings are affected by the number of iterations. Fig. 3 shows the average MUI energy over all the UTs against the number of iterations for these precodings. Both the AMP and AC precodings clearly provide *much faster* convergence and perform noticeably better compared to the CE precoding. This finding is expected because the amplitude constraint of the AMP and AC precodings are relaxed, thereby providing extra degrees of freedom compared to the CE precoding. Additionally, the AMP precoding slightly outperforms the AC precoding at the initial stage, but eventually the AC precoding catches up and performs slightly better than the AMP precoding. This is because for AMP, we adopt the uniform ring distribution for $\mathscr{P}(x_i)$, which imposes certain constraints on its outputs $\{\widehat{x}_i\}$, so that most of $\{\widehat{x}_i\}$ are still closed to the ring while the AC precoding can be any points inside the ring. Also because of this restriction, we shall see later that the PAPR value of the AMP precoding is better than that of the AC precoding. In addition, the number of complex multiplications required in the AMP precoding is less than that required in the AC precoding. From the figure, we can also observe that both AMP and AC precodings converge within 200 iterations. Therefore, in the following simulations, the maximum number of iterations of these precodings is set to be 200 for the sake of fairness.

To gauge the computational complexity of these three precodings, their running times are provided for the system configuration $(N, M) = (100, 35)$. Fig. 4 compares the average running time required for 200 iterations for the three precodings. The AMP precoding clearly outperforms the rest in running time. Specifically, it runs about $\frac{0.4651}{0.0448} \approx 10.38$ times and $\frac{0.6103}{0.0448} \approx 13.62$ times faster than the CE precoding and the AC precoding respectively. This result verifies the significant computational efficiency of the AMP precoding. Note that although the core idea used in the iterative search algorithm for the AC and CE precodings is the same, the running time for the AC precoding is about $\frac{0.6103}{0.4651} \approx 1.31$ times longer than that for the CE precoding. This finding is intuitive because the iterative search algorithm for the AC precoding needs to simultaneously find a transmit phase vector and a transmit amplitude vector to minimize the MUI at all

[7] Based on our numerical results, we find that slightly enlarging the radius from $\alpha = \frac{1}{\sqrt{N}}$ to $\alpha = \frac{1.07}{\sqrt{N}}$ can greatly reduce the MUI while maintaining a good PAPR property.

UTs. In comparison, the iterative search algorithm for the CE precoding only has to search a transmit phase vector to achieve the same goal.

To investigate the robustness of the three precodings, we now examine the performance of the three precodings in different system configurations. Fig. 5 shows the average MUI energy over all the UTs for three different precodings, plotted as functions of BS antenna number N, for $M = 35$ and $M = 50$. We observe that increasing N significantly improves MUI performance for all the precoding methods. In addition, for a fixed N, say $N = 110$, as M increases, the performance degrades. This result is expected because increasing M will result in the decrease of the degrees of freedom for MUI suppression. The AC and AMP precodings also outperform the CE precoding, and their performance gains gradually enlarge with the increase of N. Meanwhile, the performance of the proposed AMP precoding is degraded slightly relative to that of the AC precoding. This phenomenon, as observed in Fig. 3, is due to the fact that the uniform ring distribution for the AMP precoding imposes certain constraints on its precoding so the its corresponding MUI is degenerated.

Finally, we compare the performance of different precoding schemes at the transmitter side in terms of *continuous-time* PAPR reduction capabilities. To calculate the PAPR value for the continuous-time signals,[8] the discrete-time signals are shaped with a root-raised-cosine filter with a roll-off factor of 0.3, as was done in [8]. The performance metrics are the complementary cumulative distribution function (CCDF) of the PAPR statistics, which is defined as the probability that the PAPR exceeds a threshold level PAPR_0, i.e.,

$$\mathrm{CCDF} \triangleq \Pr(\mathrm{PAPR} > \mathrm{PAPR}_0). \tag{36}$$

Given that a -30 dB MUI performance is reliable enough for practical application, the considered precodings require approximately $N = 90$ and $N = 130$ to achieve this target MUI performance for $M = 35$ and $M = 50$ respectively, as shown in Fig. 5. Based on this, the system configurations with $(N, M) = (90, 35)$ and $(N, M) = (130, 50)$ are adopted to evaluate the PAPR performance of different precoding schemes. Fig. 6 shows the PAPR-performance characteristics for all considered precoding schemes. As expected, owing to the relaxation of amplitude constraints, the PAPR reduction performance of the AC and AMP precodings is worse than that of the CE precoding. From a detailed inspection, when $\mathrm{CCDF} = 10^{-3}$, for $(N, M) = (90, 35)$ and $(N, M) = (130, 50)$, the PAPR of the CE precoding is 5.43 and 5.41 dB respectively.[9] In addition, we see that the performance loss of the proposed AMP precoding relative to that of the CE precoding is 1.05 and 1.06 dB at $\mathrm{CCDF} = 10^{-3}$, when $(N, M) = (90, 35)$ and $(N, M) = (130, 50)$ respectively. However,

[8] We denote the discrete-time estimated signal $\widehat{x}_i$ after pulse-shape filtering by $\widehat{x}_i(t)$. For the continuous-time signal $\widehat{x}_i(t)$, defined for $t \in [0, T]$, PAPR is the ratio between the maximum power and the average power of $\widehat{x}_i(t)$, which can be express as

$$\mathsf{PAPR} \triangleq \frac{\max |\widehat{x}_i(t)|^2}{\frac{1}{T}\int_0^T |\widehat{x}_i(t)|^2 \,\mathrm{d}t}.$$

[9] It should be noted that although the PAPR of the discrete-time constant envelope precoded signal is zero, the constant-envelope property *does not* hold true for the signals after pulse-shape filtering.

compared with the AC precoding at the same level of CCDF, the proposed AMP precoding offers about 0.36 and 0.33 dB PAPR reduction when $(N, M) = (90, 35)$ and $(N, M) = (130, 50)$ respectively.

V. Conclusion

To suppress the MUI at all the UTs while enabling the use of cheap but highly power-efficient amplifiers at the BS, a computationally efficient precoding using AMP algorithm was proposed for massive multiuser MIMO systems. Simulation results indicated that the proposed AMP precoding can offer a tradeoff between MUI suppression and PAPR reduction with desirable characteristics. Another attraction of AMP is that it has parallel nature and much lower computational complexity but no significant performance degradation. These findings may render it suitable for practical implementation.

References

[1] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[3] S. K. Mohammed and E. G. Larsson, "Single-user beamforming in large-scale MISO systems with per-antenna constant-envelope constraints: The doughnut channel," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3992–4005, Nov. 2012.

[4] ——, "Per-antenna constant envelope precoding for large multi-user MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 1059–1071, Mar. 2013.

[5] ——, "Constant-envelope multi-user precoding for frequency-selective massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 2, no. 5, pp. 547–550, Oct. 2013.

[6] J. Pan and W.-K. Ma, "Constant envelope precoding for single-user large-scale MISO channels: efficient precoding and optimal designs," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 982–995, Oct. 2014.

[7] J.-C. Chen, C.-K. Wen, and K.-K. Wong, "Improved constant envelope multiuser precoding for massive MIMO systems," *IEEE Commun. Lett.*, vol. 18, no. 8, pp. 1311–1314, Aug. 2014.

[8] C. Mollén, "Low-PAR precoding for very-large multi-user MIMO systems," Master's thesis, Linköping University, Sweden, Jun. 2013.

[9] 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.

[10] S. Wang, Y. Li, M. Zhao, and J. Wang, "Energy efficient and low-complexity uplink transceiver for massive spatial modulation MIMO," *IEEE Trans. Veh. Technol.*, 2014.

[11] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[12] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices," *J. Stat. Mech. Theor. Exp.*, vol. 2012, no. 08, p. P08009, Aug. 2012.

[13] B. L. Ng, J. S. Evans, S. V. Hanly, and D. Aktas, "Distributed downlink beamforming with cooperative base stations," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5491–499, Dec. 2008.

[14] 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.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[16] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *arXiv:1405.5618v2*, Oct. 2014.

[17] P. Maechler, C. Studer, D. E. Bellasi, A. Maleki, A. Burg, N. Felber, H. Kaeslin, and R. G. Baraniuk, "VLSI design of approximate message passing for signal restoration and compressive sensing," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 3, pp. 579–590, Sep. 2012.

Algorithm 1: AMP-based precoding

```
begin
    $\mathbf{a}^{(0)} = \{a_i^{(0)}\}_{i=1}^{N} \leftarrow \{0\}$
    $\mathbf{v}^{(0)} = \{v_i^{(0)}\}_{i=1}^{N} \leftarrow \{1\}$
    $\mathbf{V}^{(0)} = \{V_\mu^{(0)}\}_{\mu=1}^{M} \leftarrow \{1\}$
    $\boldsymbol{\omega}^{(0)} = \{\omega_\mu^{(0)}\}_{\mu=1}^{M} \leftarrow \{u_\mu\}$
    $l \leftarrow 1$
    while Predefined number of iterations is met do
        for $\mu = 1$ to $M$ do
            $V_\mu^{(l)} \leftarrow \sum_{i=1}^{N} |H_{\mu i}|^2 \, v_i^{(l-1)}$
            $\omega_\mu^{(l)} \leftarrow \sum_{i=1}^{N} H_{\mu i} a_i^{(l-1)} - \frac{u_\mu - \omega_\mu^{(l-1)}}{V_\mu^{(l-1)}} V_\mu^{(l)}$
        for $i = 1$ to $N$ do
            $(\Sigma_i^{(l)})^2 \leftarrow \left[\sum_{\mu=1}^{M} \frac{|H_{\mu i}|^2}{V_\mu^{(l)}}\right]^{-1}$
            $R_i^{(l)} \leftarrow a_i^{(l-1)} + (\Sigma_i^{(l)})^2 \sum_{\mu=1}^{M} H_{\mu i}^{*} \frac{u_\mu - \omega_\mu^{(l)}}{V_\mu^{(l)}}$
            $a_i^{(l)} \leftarrow f_a \left((R_i^{(l)}, (\Sigma_i^{(l)})^2\right)$
            $v_i^{(l)} \leftarrow f_c \left((R_i^{(l)}, (\Sigma_i^{(l)})^2\right)$
        $l \leftarrow l + 1$
    return $\{a_i^{(l)}\}_{i=1}^{N}$
```

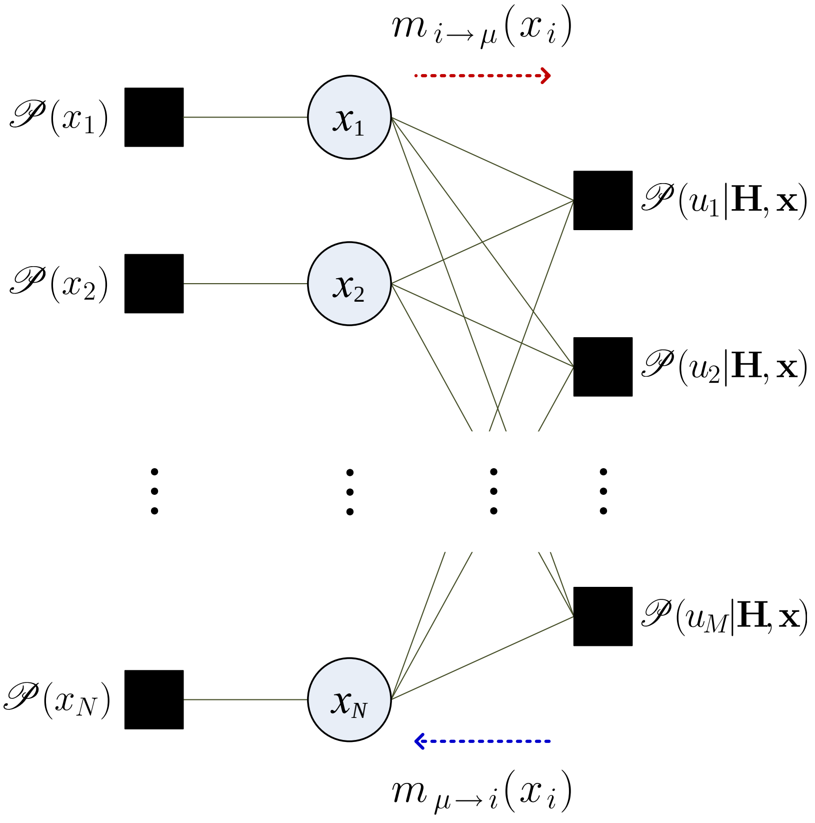
$m_{i \to \mu}(x_i)$
$\mathscr{P}(x_1)$
x_1
$\mathscr{P}(u_1|\mathbf{H}, \mathbf{x})$
$\mathscr{P}(x_2)$
x_2
$\mathscr{P}(u_2|\mathbf{H}, \mathbf{x})$
$\mathscr{P}(u_M|\mathbf{H}, \mathbf{x})$
$\mathscr{P}(x_N)$
x_N
$m_{\mu \to i}(x_i)$

Fig. 1. Factor graph for the considered problem defined in (6). The factor nodes are shown by filled rectangles and variable nodes are depicted as circles.

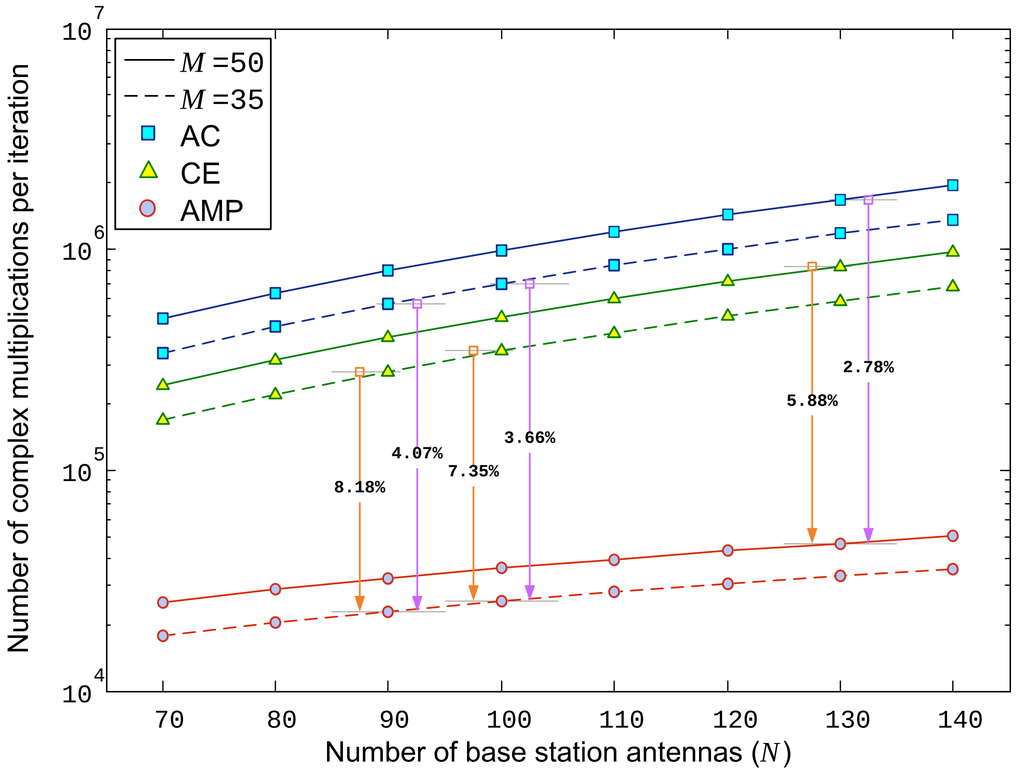

Fig. 2. Number of complex multiplications versus number of BS antennas N for $M = 35, 50$.

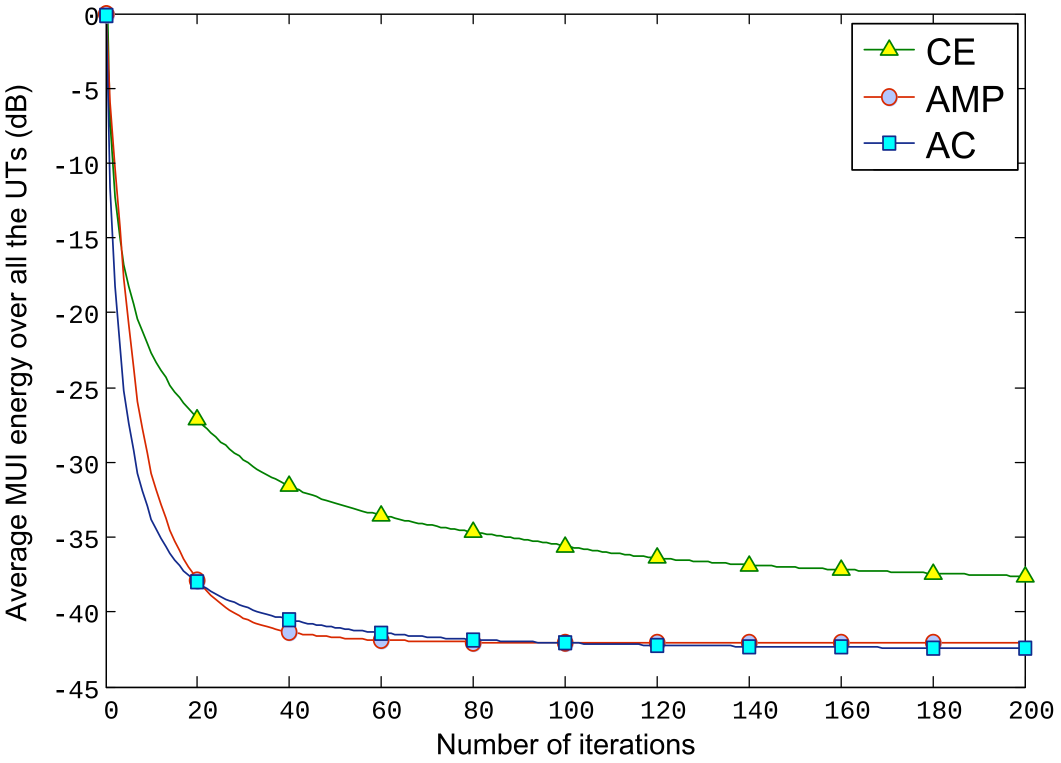

Fig. 3. Average of the MUI energy over all the UTs versus the number of iterations for $(N, M) = (100, 35)$.

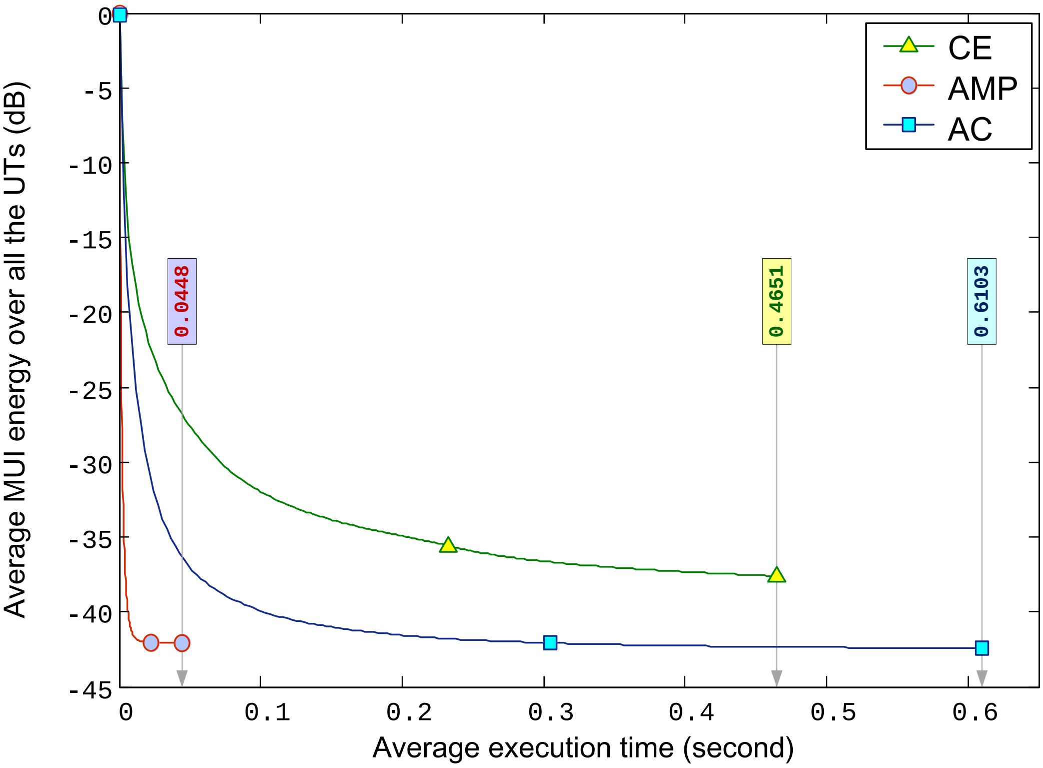

Fig. 4. Average of the MUI energy over all the UTs versus execution time for $(N, M) = (100, 35)$.

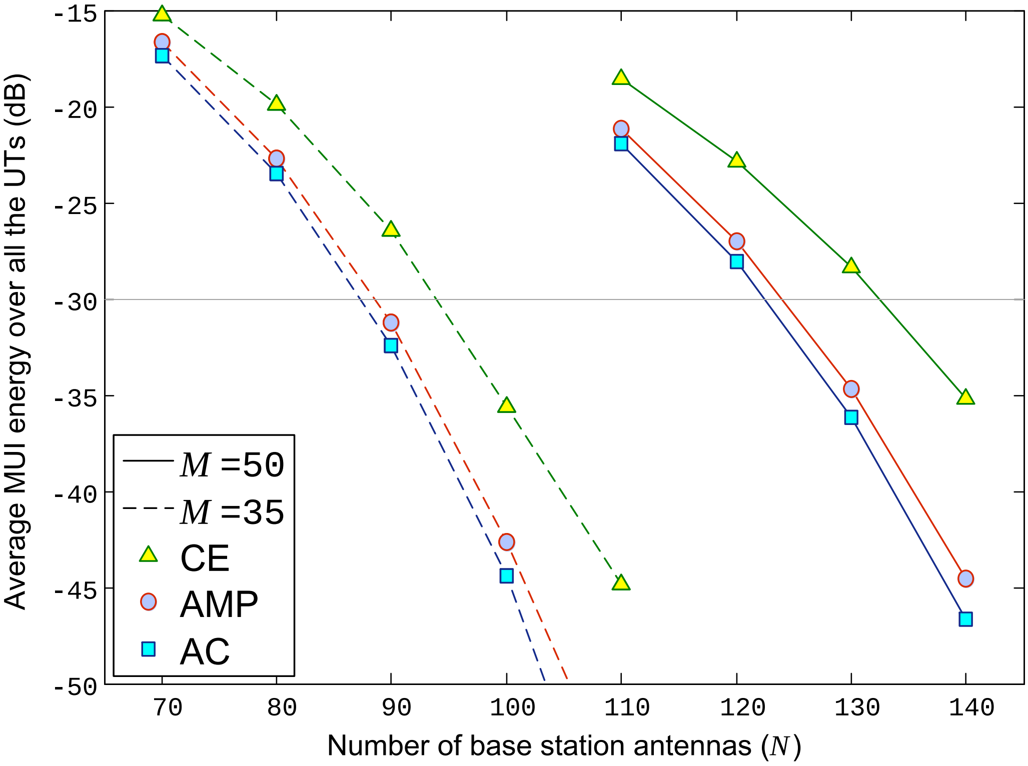

Fig. 5. Average of the MUI energy over all the UTs versus the number of BS antennas for $M = 35, 50$.

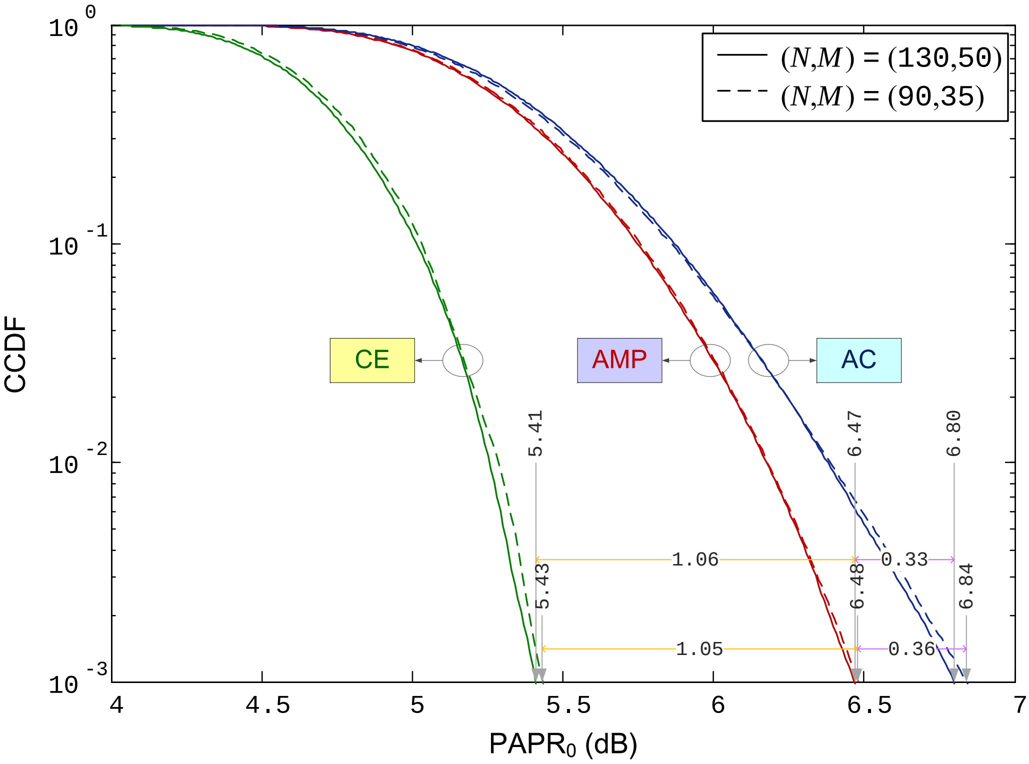

Fig. 6. PAPR performance for various precoding scheme for $(N, M) = (90, 35)$ and $(N, M) = (130, 50)$.