Mixed-ADC Massive MIMO Detectors: Performance Analysis and Design Optimization

Ti-Cao Zhang, Chao-Kai Wen, *Member, IEEE*, Shi Jin, *Member, IEEE*, and Tao Jiang, *Senior Member, IEEE*

***Abstract*— The hardware cost and power consumption of a massive multiple-input multiple-output (MIMO) system can be remarkably reduced by using a very low-resolution analog-to-digital converter (ADC) unit in each antenna. However, such a pure low-resolution ADC architecture complicates parameter estimation problems. These issues can be resolved and the potential of a pure low-resolution ADC architecture can be achieved by applying a mixed ADC architecture, whose antennas are equipped with low-precision ADCs, while few antennas are composed of high-precision ADCs. In this paper, a unified framework is presented to develop a family of detectors on a massive MIMO uplink system through probabilistic Bayesian inference. Our basic setup comprises an optimal detector, which is developed to provide a minimum mean-squared-error estimate on data symbols. Considering that highly nonlinear steps are involved in quantization, we also investigate the potential for complexity reduction on an optimal detector by postulating a common pseudo-quantization noise model. We provide asymptotic performance expressions, including mean squared error and bit error rate for optimal and suboptimal MIMO detectors. These expressions can be evaluated rapidly and efficiently. Thus, they can be used for system design optimization.**

***Index Terms*— Massive MIMO, MIMO detector, low-resolution ADC, mixed architecture, Bayesian inference.**

I. Introduction

MASSIVE multiple-input multiple-output (MIMO) systems are considered a disruptive technology for next-generation (e.g., 5G) communication systems [1]–[3]. Such systems can substantially reduce cell interference via simple signal processing by equipping a base station (BS) with hundreds or thousands of antennas in a centralized [4], [5] or distributed [6] manner because a channel vector between users and a BS becomes quasi-orthogonal.

A large number of antennas significantly complicate the hardware design for massive MIMO implementation in production. Such systems require an ADC unit for each receiver antenna. Therefore, an equivalent number of ADCs are necessary when many antennas are used. Consequently, costs and power consumption are exponentially increased because of high-speed and high-resolution ADCs,[1] and this drawback impedes the application of massive MIMO systems [8]. Nevertheless, this problem can be resolved by using a very low-resolution ADC (e.g., 1-3 bits) unit in each radio frequency (RF) chain. In massive MIMO systems, robustness against coarse quantization is enhanced as the number of receiving antennas is increased [9], [10].

Previous Work The potential of a 1-bit ADC system has been validated in accordance with information theory. For a single-input single-output channel, the capacity achieving transmit signals is discrete [11], [12]. A 1-bit receiver for QPSK only suffers from 1-3 dB SNR loss compared with the unquantized case [13], and this loss can be recovered by oversampling [14]. Mutual information is $2/\pi$ times smaller in MIMO channels than in an unquantized MIMO case in a low SNR regime [15]. In a high SNR regime, a data rate loss is observed and a tighter capacity bound is provided [16].

Specific algorithms for user detection have been investigated for 1-bit ADC systems. A linear detector with a 1-bit ADC works efficiently for a quantized MIMO system with QPSK and 16-QAM [17], [18]. An uplink multi-user detector was developed to achieve an optimal performance on the basis of a message passing algorithm [19] and convex optimization [20]. Pioneering work focused on frequency-flat channels. In recent studies, frequency-selective channels have been explored [21]–[23].

A *pure* low-resolution ADC system inevitably causes several problems, including phase/frequency synchronization [24], [25] from quantized output, calibrations of in-phase and quadrature-phase (IQ) imbalance [26]–[28], data rate loss in a high SNR regime [16], error floor for linear multi-user detection [19], and complex channel estimations [17].

Manuscript received September 24, 2015; revised March 2, 2016 and July 2, 2016; accepted August 29, 2016. Date of publication September 7, 2016; date of current version November 9, 2016. This work was supported in part by the National Science Foundation (NSFC) for Distinguished Young Scholars of China under Grant 61325004, in part by NSFC under Grant 61301129 and Grant 61428104, in part by the National High Technology Development 863 Program of China under Grant 2015AA01A710, and in part by the Key Project of Hubei Province in China under Grant 2015BAA074. The work of C.-K. Wen was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 103-2221-E-110-029-MY3 and in part by ITRI, Hsinchu, Taiwan. The work of S. Jin was supported by the International Science and Technology Cooperation Program of China under Grant 2014DFT10300. The associate editor coordinating the review of this paper and approving it for publication was J. M. Romero-Jerez. *(Corresponding author: Tao Jiang.)*

T.-C. Zhang and T. Jiang are with the Wuhan National Laboratory for Optoelectronics, School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: ticao.hust@gmail.com; tao.jiang@ieee.org).

C.-K. Wen is with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (e-mail: chaokai.wen@mail.nsysu.edu.tw).

S. Jin is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jinshi@seu.edu.cn).

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

Digital Object Identifier 10.1109/TWC.2016.2606592

[1]High-speed ADCs with a resolution of κ bits typically employ a flash architecture where the input voltage is simultaneously compared with each of the 2^{κ} tap voltages [7].

In particular, the precise estimate of the MIMO channel with a 1-bit ADC requires a long training sequence, such as 50 times the number of users [17], because of the strong nonlinear characteristic of 1-bit quantization. Such a pilot overhead cannot be sustained by using a practical system; therefore, a joint channel-and-data estimation method employing predicted payload data has been proposed to aid channel estimation and to reduce pilot cost [29]. However, this technique exacerbates the computational complexity of receivers.

A mixed ADC architecture was initially proposed on the basis of previous findings [30]. Available high-resolution RF chains can assist in channel estimation and facilitate the establishment of several front-end designs. Furthermore, mixed ADC architectures are economically beneficial because they can be implemented by adding antennas with low-resolution ADCs to existing BS.

Contributions: In this work, a mixed architecture is extended to a general architecture whose antennas are mostly installed with low-resolution ADCs (1-3 bit), while a few antennas are equipped with full-resolution ADCs. This study aims to examine the mixed ADC massive MIMO systems in terms of detection, which is in contrast to another study [30] that emphasized the analysis of mutual information. We demonstrate that the available high-resolution ADCs are practically essential because they can effectively eliminate the error floor of a relaxed Bayes-optimal detector. To the best of our knowledge, studies have yet to elucidate MIMO detection problems in a mixed ADC architecture, and its design and performance are a major concern related to this architecture. This paper presents the following specific contributions:

- Applying probabilistic Bayesian inference, we provide a *unified* framework to develop a family of MIMO detectors. This framework is labeled as a generic Bayes detector. To compute the Bayesian estimate, we must establish a prior distribution for the transmitted data and a likelihood function for the statistical model of receiver signals. With the true prior and likelihood functions, the Bayes detector can achieve the optimal estimate in the mean squared error (MSE) sense. Many low-complexity and popular detectors, such as linear minimum mean-squared-error (LMMSE) and maximal-ratio-combining (MRC) receivers, can be obtained by properly postulating mismatched prior and likelihood functions.
- The exact expression for the likelihood of a receiver signal with quantization is complex because of the highly nonlinear property of a quantizer. The "mixture" architecture further complicates the design of the Bayes detector. As such, the following question should be answered: *how close to the optimum performance can a conventional MIMO detector operate without considering the exact yet annoying nonlinear effect of quantizers?* To answer this question, we apply a traditional heuristic approach that treats quantization noise as additive and independent. This heuristic approach is known as a pseudo-quantization noise (PQN) model [31]. Postulating a mismatched likelihood with this model in the Bayes detector, we can significantly reduce computational costs with a slight performance degradation.
- To develop the Bayes detector, we employ a recently established technique called the generalized approximate message passing (GAMP) algorithm [32]. We apply this approximation for the mixed ADC architecture by specifying the corresponding adjustment in nonlinear steps. Implementing the central-limit theorem (CLT) to a large system limit, we derive an approximate analytical expression for the state evolution (SE) of the Bayes detector. A series of metrics, including MSE and bit error rate (BER), can be predicted, and computer simulations are conducted to verify the accuracy of our analysis. The performance of the mixed ADC massive MIMO receivers can be rapidly and efficiently evaluated. Several useful observations are also performed on the basis of our analysis to optimize the receiver design.

To achieve conciseness and focus on detection techniques, we present the following assumptions. First, we assume that BS provides optimum channel state information (CSI) through which the detector can be implemented. In the mixed receiver architecture, high-quality CSI can be acquired via a round-robin process [30]. In this process, high-resolution ADCs are switched to different antennas at various symbol times to estimate the corresponding channel coefficients.[2] Second, we assume that the channel coefficients among antennas are uncorrelated. For the multi-user MIMO operation considered in this paper, the single-antenna terminals are distributed randomly and are typically separated by hundreds or thousands of wavelengths. Thus, the assumption that the transmit antennas between different terminals are uncorrelated is valid. The receive antennas can also be considered uncorrelated in the massive MIMO system when no strict space constraint exists at the BS side.

In practice, high-resolution ADCs are successively connected to different antennas in the channel estimation stage. Consequently, this process incurs additional costs on hardware, such as switches. Fortunately, switches nowadays can operate at a symbol rate with a tolerable energy consumption [33]. For example, reverse-saturated siliconCgermanium quarter-wave shunt switches [34] can work within less than 200 ps with a power consumption lower than 10 mW. The switch-based transceiver architectures are widely adopted now, such as spatial modulation [35], [36] and antenna selection [37]. Hardware implementation is not a drawback. In the detection stage, our analysis shows that the specific ADC switch mechanism is unnecessary because it unlikely affects the total BER performance under the assumption of Rayleigh fading.

Notation: In this paper, vectors and matrices, such as $\mathbf{x}$ and $\mathbf{X}$, respectively, are presented in bold typeface, while scalars are presented in normal typeface, such as x. We respectively use $\mathbf{X}^T$ and $\mathbf{X}^*$ to represent the transpose and conjugate transpose of a matrix $\mathbf{X}$. $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ respectively denote the real and imaginary parts of a complex matrix (vector). A Gaussian distribution with mean μ and variance σ^2

[2] In round-robin estimation, 1-bit ADC RF chains remain idle. The inefficient use of the existing large amount of RF chains with low-resolution ADCs causes a significant performance degradation. Thus, more accurate CSI can be obtained with less training overhead by designing novel algorithms that jointly exploit high-resolution ADCs and 1-bit ADCs in future research.

is denoted by $\mathcal{N}(\mu, \sigma^2)$ while $\mathcal{CN}(\mu, \sigma^2)$ indicates a complex Gaussian distribution. In particular, we denote $\mathcal{N}(x; \mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ and $\mathcal{CN}(x; \mu, \sigma^2) \triangleq \frac{1}{\pi\sigma^2} e^{-\frac{|x-\mu|^2}{\sigma^2}}$. Finally, let $\Phi(x) = \int_{-\infty}^{x} \mathrm{D}u$ and $\mathrm{D}u \triangleq \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \mathrm{d}u$.

II. System Description

We consider an uplink multiuser massive MIMO system that has K single-antenna users and one BS equipped with an array of N antennas ($N > K$). Under the simplest rayleigh fading channel, the discrete time complex baseband received signal $\mathbf{y} \in \mathbb{C}^N$ is given by

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

where $\mathbf{x} = [x_j] \in \mathbb{C}^K$ contains the transmitted symbols from all the users. $\mathbf{n} = [n_i] \in \mathbb{C}^N$ denotes the additive white Gaussian noise (AWGN). $\mathbf{H} = [h_{ij}] \in \mathbb{C}^{N\times K}$ represents the channel matrix between the BS and the K users. The entries of $\mathbf{H}$ are independent and identically distributed (i.i.d.) random variables with $h_{ij} \sim \mathcal{CN}(0, \frac{1}{K})$.[3] The entries of $\mathbf{x}$ are i.i.d. and are distributed as $P_{\mathsf{in}}(x)$, and the entries of $\mathbf{n}$ are i.i.d. $\mathcal{CN}(0, \sigma_n^2)$. If we suppose that $\mathbb{E}\{|x_j|^2\} = 1, \forall j$, then the signal-to-noise ratio (SNR) of the system is defined by $\mathrm{SNR} = 1/\sigma_n^2$.

The real and imaginary components of the received signal at the ith antenna are quantized separately by the ADC unit. We denote the complex quantizer which operates on the ith antenna as

$$r_i = Q_i^c(y_i), \tag{2}$$

where $Q_i^c(\cdot)$ is applied to the real and imaginary components of the received signal separately such that $r_i = Q_i^c(y_i) \triangleq Q_i(\mathrm{Re}(y_i)) + \mathrm{j}Q_i(\mathrm{Im}(y_i))$. If a statement is for a generic antenna index, we often omit the subscript index i for brevity.

In this paper, we mainly focus on uniform midrise quantizers with a fixed quantization step size Δ. Such a quantizer maps a real-valued[4] input that falls in $(r_i - \frac{\Delta}{2}, r_i + \frac{\Delta}{2}]$ to value r from the discrete set

$$\mathcal{R}_\kappa \triangleq \left\{ \left(-\frac{1}{2} + b\right)\Delta;\ b = -\frac{2^\kappa}{2} + 1, \cdots, \frac{2^\kappa}{2} \right\}, \tag{3}$$

where κ is the number of quantization bits. Figure 1(b) shows a graphical depiction of a 3-bit uniform midrise quantizer. Notice that an input of magnitude larger than $(\frac{2^\kappa}{2} - 1)\Delta$ saturates. For ease of explication, we simply express the lower and upper thresholds associated with r_i as r_i^{low} and r_i^{up}, respectively; specifically, they are

$$r_i^{\mathrm{low}} = \begin{cases} r_i - \frac{\Delta}{2}, & \text{for } r_i \geq -\left(\frac{2^\kappa}{2} - 1\right)\Delta, \\ -\infty, & \text{otherwise,} \end{cases} \tag{4a}$$

[3]We take the variance of the channel coefficient to be $1/K$ to normalize the equivalent channel row and ensure that the analog baseband is within a proper range. This assumption is without loss of generality because in practice, a variable gain amplifier with automatic gain control (AGC) is used before quantization. Moreover, the row normalized channel $\mathbf{H}$ simplifies the discussion in this study.

[4]For ease of notation, we abuse r to denote each real channel although it should be specified as $\mathrm{Re}(r_i)$ or $\mathrm{Im}(r_i)$.

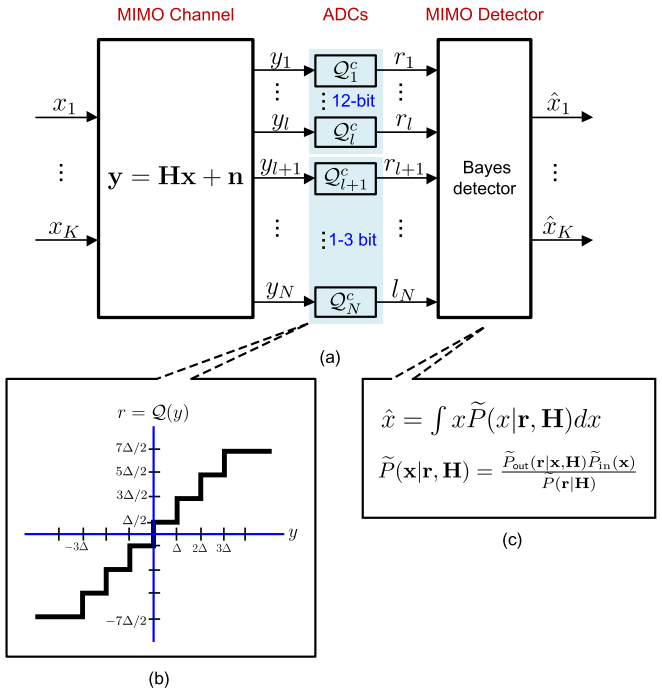

Fig. 1. (a) A mixed-ADC massive MIMO architecture. (b) A 3-bit uniform midrise quantizer. (c) The (generic) Bayes detector.

and

$$r_i^{\mathrm{up}} = \begin{cases} r_i + \frac{\Delta}{2}, & \text{for } r_i \leq \left(\frac{2^\kappa}{2} - 1\right)\Delta, \\ +\infty, & \text{otherwise.} \end{cases} \tag{4b}$$

We assume that most of the antennas adopt low-resolution ADCs (e.g., 1-3 bit) while only a few antennas adopt high-resolution ADCs (e.g., 12 bit). To distinguish the ADC units of different resolution, we use $\mathbf{\Omega}_\kappa$ to indicate the collection of the antennas equipped with ADCs of κ bits. Moreover, the cardinality of $\mathbf{\Omega}_\kappa$ is N_κ, and thus we have $\sum_\kappa N_\kappa = N$.

III. MIMO Detectors

To recover the multiuser signals $\mathbf{x}$ from the quantized measurement $\mathbf{r}$, we adopt the Bayesian inference. The Bayesian inference starts with computing the posterior probability $\widetilde{P}(\mathbf{x}|\mathbf{r}, \mathbf{H})$ according to the Bayes rule

$$\widetilde{P}(\mathbf{x}|\mathbf{r}, \mathbf{H}) = \frac{\widetilde{P}_{\mathsf{out}}(\mathbf{r}|\mathbf{x}, \mathbf{H})\widetilde{P}_{\mathsf{in}}(\mathbf{x})}{\widetilde{P}(\mathbf{r}|\mathbf{H})}, \tag{5}$$

where $\widetilde{P}(\mathbf{r}|\mathbf{H}) = \int \mathrm{d}\mathbf{x}\widetilde{P}_{\mathsf{out}}(\mathbf{r}|\mathbf{x}, \mathbf{H})\widetilde{P}_{\mathsf{in}}(\mathbf{x})$ is the marginal likelihood function, $\widetilde{P}_{\mathsf{out}}(\mathbf{r}|\mathbf{x}, \mathbf{H})$ is a likelihood function, and $\widetilde{P}_{\mathsf{in}}(\mathbf{x})$ is a measure of the input distribution $P_{\mathsf{in}}(\mathbf{x})$. Here, $\widetilde{P}$ (including $\widetilde{P}_{\mathsf{in}}$ and $\widetilde{P}_{\mathsf{out}}$) indicates that the probability is different from the actual one, P.

Using the posterior probability, the Bayes estimate of the unknown vector $\mathbf{x}$ is the mean of the posterior distribution and its element is expressed as

$$\hat{x}_j = \mathbb{E}\{x_j|\mathbf{r}, \mathbf{H}\}, \tag{6}$$

where the expectation over x_j is w.r.t. the marginal posterior probability

$$\widetilde{P}(x_j|\mathbf{r},\mathbf{H}) = \int_{\mathbf{x}\setminus x_j} \mathrm{d}\mathbf{x}\widetilde{P}(\mathbf{x}|\mathbf{r},\mathbf{H}). \tag{7}$$

The notation $\int_{\mathbf{x}\setminus x_j}\mathrm{d}\mathbf{x}$ denotes the integration over all the variables in $\mathbf{x}$ except for x_j.

In the general multiuser detection framework, transmitted signal x_j is typically formed by a constellation modulation scheme such as quadrature amplitude modulation (QAM). In such cases, the final constellation point can be determined from $\hat{x}_j$ via the maximum-likelihood decision rule. However, all our developed MIMO "detectors" are centered around the framework of (6), which should be classified as an estimation problem rather than a detection problem. Therefore, with an abuse of terminology, we call the framework for obtaining (6) the (generic) "Bayes detector".

According to (5), the posterior $\widetilde{P}(\mathbf{x}|\mathbf{r},\mathbf{H})$ is determined by $\widetilde{P}_{\mathsf{out}}(\mathbf{r}|\mathbf{x},\mathbf{H})$ and $\widetilde{P}_{\mathsf{in}}(\mathbf{x})$. Postulating a mismatched measure $\widetilde{P}\neq P$ is due to a particular approximation that can reduce computational complexity [38]. We discuss these properties in the subsequent subsections. With reference to the Bayes detector, we always mean (6) with general $\widetilde{P}_{\mathsf{out}}$ and $\widetilde{P}_{\mathsf{in}}$.

A. De-Quantization (DQ)-Optimal Detector

We begin with the detector developed from the true P_{out} and P_{in}, i.e, no mismatch between P and $\widetilde{P}$. Let $w_i=\mathbf{h}_i^*\mathbf{x}$ with $\mathbf{h}_i^*$ be the i-th row of $\mathbf{H}$. According to (1) and (2), the likelihood function (or the conditional probability distribution) of the quantized output given $\mathbf{x}$ is expressed as

$$P_{\mathsf{out}}(\mathbf{r}|\mathbf{x},\mathbf{H}) = \prod_{i=1}^{N} P_{\mathsf{out},i}(r_i|w_i), \tag{8}$$

where[5]

$$P_{\mathsf{out},i}(r_i|w_i) = \int_{r_i^{\mathrm{low}}}^{r_i^{\mathrm{up}}} \mathrm{d}y\, \mathcal{CN}(y;w_i,\sigma_n^2). \tag{9}$$

Using the true likelihood function (8) in conjunction with the true prior distribution P_{in} to (5), we obtain the exact posterior probability $P(\mathbf{x}|\mathbf{r},\mathbf{H})$. In this case, $\hat{x}_j$ via (6) can achieve the best estimates in terms of MSE. We refer to this detector as the de-quantization (DQ)-optimal detector.

However, the DQ-optimal detector is not computationally tractable because the marginal posterior probability $P(x_j|\mathbf{r},\mathbf{H})$ in (7) involves a high-dimensional integral. The canonical solution to this problem is using *belief propagation* (BP) by expressing the posterior probability $P(\mathbf{x}|\mathbf{r},\mathbf{H})$ as a factor graph [39]. Unfortunately, the complexity of BP is still too high for practical application. To solve this problem, BP was recently simplified as the GAMP algorithm by

[5]Notice that w and y are complex-valued variables. The integral $\int_{r_i^{\mathrm{low}}}^{r_i^{\mathrm{up}}}\mathrm{d}y$ in (9) is given by

$$\int_{\mathrm{Re}(r_i^{\mathrm{low}})}^{\mathrm{Re}(r_i^{\mathrm{up}})}\int_{\mathrm{Im}(r_i^{\mathrm{low}})}^{\mathrm{Im}(r_i^{\mathrm{up}})}\mathrm{dRe}(y)\mathrm{dIm}(y).$$

Algorithm 1: GAMP Algorithm

Input: $\mathbf{r}$, $\mathbf{H}$, $\widetilde{P}_{\mathsf{in},j}(x_j)$, $\widetilde{P}_{\mathsf{out},i}(r_i|w_i)$
Output: $\mathbf{x}$
Definition:

$$\mathscr{P}_{w_i|r_i}(w_i|r_i;\,p_i,v_i^p) = \frac{\widetilde{P}_{\mathsf{out},i}(r_i|w_i)\mathcal{CN}(w_i;p_i,v_i^p)}{\int_{w_i'}\mathrm{d}w_i'\widetilde{P}_{\mathsf{out}}(r_i|w_i')\mathcal{CN}(w_i';p_i,v_i^p)} \quad \text{(D1)}$$

$$g_{\mathsf{out},i}(w_i;\,p_i,v_i^p) = \frac{1}{v_i^p}\left(\mathbb{E}_{w_i|r_i}\left\{w_i|r_i;\,p_i,v_i^p\right\}-p_i\right) \quad \text{(D2)}$$

$$g'_{\mathsf{out},i}(w_i;\,p_i,v_i^p) = \frac{1}{v_i^p}\left(\frac{\mathbb{V}_{w_i|r_i}\left\{w_i|r_i;p_i,v_i^p\right\}}{v_i^p}-1\right) \quad \text{(D3)}$$

$$\mathscr{P}_{x_j}(x_j;\,s_j,v_j^s) = \frac{\widetilde{P}_{\mathsf{in},j}(x_j)\mathcal{CN}(x_j;s_j,v_j^s)}{\int_{x_j'}\mathrm{d}x_j'\widetilde{P}_{\mathsf{in},j}(x_j')\mathcal{CN}(x_j';s_j,v_j^s)} \quad \text{(D4)}$$

$$g_{\mathsf{in},j}(x_j;\,s_j,v_j^s) = \mathbb{E}_{x_j}\{x_j|s_j,v_j^s\} \quad \text{(D5)}$$

$$g'_{\mathsf{in},j}(x_j;\,s_j,v_j^s) = \mathbb{V}_{x_j}\{x_j|s_j,v_j^s\} \quad \text{(D6)}$$

Initialize:

$\forall j: \hat{x}_j(1) = \int \mathrm{d}x\,\widetilde{P}_{\mathsf{in},j}(x)$
$\forall j: v_j^x(1) = \int \mathrm{d}x|x-\hat{x}_j(1)|^2\widetilde{P}_{\mathsf{in},j}(x)$
$\forall i: p_i(0)=0$

for $t=1,\cdots,t_{\max}$ **do**

$\forall i: v_i^p(t) = \sum_{j=1}^K |h_{ij}|^2 v_j^x(t-1)$
$\forall i: p_i(t) = \sum_{j=1}^K h_{ij}x_j(t-1) - v_i^p(t)g_{\mathsf{out},i}(w_i;\,p_i(t-1),v_i^p(t))$
$\forall i: v_i^z(t) = -g'_{\mathsf{out},i}\left(w_i;\,p_i(t),v_i^p(t)\right)$
$\forall i: z_i(t) = g_{\mathsf{out},i}\left(w_i;\,p_i(t),v_i^p(t)\right)$
$\forall j: v_j^s(t) = \left(\sum_{i=1}^N |h_{ij}|^2 v_i^z(t)\right)^{-1}$
$\forall j: s_j(t) = x_j(t-1) + v_j^s(t)\sum_{i=1}^N h_{ij}^* z_i(t)$
$\forall j: v_j^x(t) = g'_{\mathsf{in},j}\left(x;\,s_j(t),v_j^s(t)\right)$
$\forall j: x_j(t) = g_{\mathsf{in},j}\left(x;\,s_j(t),v_j^s(t)\right)$

return $\mathbf{x}$;

using the second-order approximations at measurement nodes. We resort to the GAMP algorithm as an iterative procedure to solve the marginal posterior probability in a recursive way. Algorithm 1 provides a high-level description of GAMP to perform the (generic) Bayes detector under the mixed-ADC architecture. Note that Algorithm 1 can be used not only to achieve the DQ-optimal detector but also to operate other detectors introduced subsequently. The GAMP algorithm has two variants, namely, *max-sum* GAMP, which approximates the MAP estimator, and *sum-product* GAMP, which approximates the MMSE estimator. These two GAMP algorithms are identical in the Gaussian inputs. However, the max-sum GAMP algorithm performs better in practical multi-user detection problems wherein the input may not be Gaussian. We intend to differentiate the performances exhibted by several detectors that align a unified framework. The goal is to determinig the benchmark detector that provides the best estimate in the MSE sense. Thus, we adopt the sum-product GAMP in this paper.

In Algorithm 1, t and $t_{\max}$ represent the current iteration index and the maximum iteration times, respectively. The Expectation $\mathbb{E}$ in (D2) and Variance $\mathbb{V}$ in (D3) are performed with respect to (w.r.t.) the probability $\mathscr{P}_{w|r}$ in (D1).

Since (D1) involves the probability $\widetilde{P}_{\mathsf{out}}(r|w)$, we call lines 3–4 the *output nonlinear* steps. Similarly, the Expectation $\mathbb{E}$ in (D5) and Variance $\mathbb{V}$ in (D6) are taken with repect to the probability $\mathscr{P}_x$ in (D4). Since (D4) involves the probability $\widetilde{P}_{\mathsf{in}}(x)$, we call lines 7–8 the *input nonlinear* steps.

To better understand the algorithm, we provide some intuition on each step of Algorithm 1. Lines 1–2 compute an estimate p_i of w_i and the corresponding variance v_i^p. The first term of p_i is a plug-in estimate of w_i and the second term provides a refinement by introducing Onsager correction in the context of *approximate message passing* (AMP) [40], [41]. Using $\{p_i, v_i^p\}$, lines 3–4 then compute the posterior estimate of the residual $y_i - p_i$ and the inverse residual variances v_i^z, where y_i is the posterior estimate of the un-quantized received signal by considering the likelihood function $\widetilde{P}_{\mathsf{out}}$. Lines 5–6 then use these residual terms to compute s_j and v_j^s, where s_j can be interpreted as an observation of x_j under an AWGN channel with zero mean and variance of v_j^s. Finally, lines 7–8 estimate the posterior mean x_j and variances v_j^x by considering the prior $\widetilde{P}_{\mathsf{in}}$.

Using the likelihood function of (9) to Algorithm 1, we can obtain analytic expressions of $g_{\mathsf{out},i}$ and $g'_{\mathsf{out},i}$ in (D2) and (D3), which are

$$g_{\mathsf{out},i}(w_i;\, p_i, v_i^p) = \frac{\widetilde{r}_i - p_i}{v_i^p + \sigma_n^2}, \tag{10a}$$

$$g'_{\mathsf{out},i}(w_i;\, p_i, v_i^p) = \frac{1}{v_i^p + \sigma_n^2}\left(1 - \frac{\widetilde{\sigma}_i{}^2}{v_i^p + \sigma_n^2}\right), \tag{10b}$$

where

$$\widetilde{r}_i := \frac{\int_{r_i^{\mathrm{low}}}^{r_i^{\mathrm{up}}} \mathrm{d}y\, y\, \mathcal{CN}(y;\, p_i, v_i^p + \sigma_n^2)}{\int_{r_i^{\mathrm{low}}}^{r_i^{\mathrm{up}}} \mathrm{d}y\, \mathcal{CN}(y;\, p_i, v_i^p + \sigma_n^2)}, \tag{11a}$$

$$\widetilde{\sigma}_i{}^2 := \frac{\int_{r_i^{\mathrm{low}}}^{r_i^{\mathrm{up}}} \mathrm{d}y\, |y - \widetilde{r}_i|^2\, \mathcal{CN}(y;\, p_i, v_i^p + \sigma_n^2)}{\int_{r_i^{\mathrm{low}}}^{r_i^{\mathrm{up}}} \mathrm{d}y\, \mathcal{CN}(y;\, p_i, v_i^p + \sigma_n^2)}. \tag{11b}$$

A detailed calculation can be seen in Appendix A.

For the input nonlinear steps, if the QAM constellation with M points is used, the prior is given by $P_{\mathsf{in}}(x) = 1/M$ for x being the constellation points, and $\int \mathrm{d}x'$ denotes the integral w.r.t. the discrete measure.

The calculations of $g_{\mathsf{out},i}(\cdot)$, $g'_{\mathsf{out},i}(\cdot)$, $g_{\mathsf{in},j}(\cdot)$, and $g'_{\mathsf{in},j}(\cdot)$ for the DQ-optimal detector are highly nonlinear. Explicit expressions of the nonlinear functions are provided in [29]. Although the DQ-optimal detector can achieve the best estimate in terms of MSE, the highly nonlinear integration steps involved make the calculation difficult and in turn complicate the circuit. Moreover, in the mixed-ADC massive MIMO system, the integration interval $(r_i^{\mathrm{low}}, r_i^{\mathrm{up}}]$ in (11) varies for different i, which makes the quantization process even more complex. This inconvenience motivates us to import mismatched measures in P_{out} and P_{in} to simplify the calculation.

B. Pseudo-De-Quantization (PDQ)-Optimal Detector

The nonlinear calculations of $g_{\mathsf{out},i}(\cdot)$, $g'_{\mathsf{out},i}(\cdot)$ are due to the integration in (9). To simplify the calculations, one heuristic is used to treat the quantizer as a linear gain $1-\rho$, along with additive noise q_i uncorrelated with the quantizer input y_i, i.e.,

$$r_i = Q_i(y_i) = (1-\rho)y_i + q_i \tag{12}$$

with ρ being the distortion factor. This heuristic is known as the PQN model. The variance of q_i is $\sigma_q^2 = \rho(1-\rho)$. In general the distortion factor ρ will depend on the quantizer input distribution, specific scalar quantizer design, and the number of bits per sample and can therefore be treated as a design constant. For simplicity, we follow the analysis in [31] by choosing $\rho \approx (\pi e/6)2^{-2\kappa}$, where κ is the quantization bits. Consequently, the measure $P_{\mathsf{out},i}$ in (9) can be written as

$$P_{\mathsf{out},i}(r_i|w_i) = \mathcal{CN}(r_i;\, (1-\rho)w_i,\, (1-\rho)^2\sigma_n^2 + \sigma_q^2). \tag{13}$$

Substituting (13) into Algorithm 1, (D1) now becomes

$$\mathscr{P}_{w_i|r_i}(w_i|r_i;\, p_i, v_i^p) = \mathcal{CN}\left(w_i;\, \frac{\frac{r_i}{(1-\rho)\gamma} + \frac{p_i}{v_i^p}}{\frac{1}{\gamma} + \frac{1}{v_i^p}},\, \frac{1}{\frac{1}{\gamma} + \frac{1}{v_i^p}}\right), \tag{14}$$

where $\gamma = \sigma_n^2 + \frac{\sigma_q^2}{(1-\rho)^2}$ and we obtain $g_{\mathsf{out},i}$ and $g'_{\mathsf{out},i}$ in (D2) and (D3) as

$$g_{\mathsf{out},i}(w_i;\, p_i, v_i^p) = \frac{\frac{r_i}{1-\rho} - p_i}{v_i^p + \gamma}, \tag{15a}$$

$$g'_{\mathsf{out},i}(w_i;\, p_i, v_i^p) = \frac{1}{v_i^p + \gamma}. \tag{15b}$$

The highly nonlinear steps in (10) now becomes simple linear steps in (15). Thus, the computational complexity is reduced significantly. We call Algorithm 1 with $g_{\mathsf{out},i}$ and $g'_{\mathsf{out},i}$ in (15) the pseudo-de-quantization(PDQ)-optimal detector.

From the perspective of the algorithm itself, the PDQ-optimal detector is exactly the AMP algorithm [41] although adjusting γ for each antenna is required to reflect different quantization bits for different antennas. Originally, the AMP algorithm was designed to recover a signal from a linearly transformed and additive Gaussian noise corrupted measurement. The resulting PDQ-optimal detector is induced from the GAMP algorithm by postulating the *mismatched* measure in P_{out}, which is mainly due to the low-complexity purpose.

C. Linear Detector

In the PDQ-optimal detector, g and g' are linearized by postulating a mismatched measure in P_{out} while the calculation of $g_{\mathsf{in},j}$ and $g'_{\mathsf{in},j}$ in (D5)–(D6) of Algorithm 1 are still nonlinear. Next, we purpose a further complexity reduction on the nonlinear steps by imposing an additional mismatch in P_{in}. In fact, if we postulate that the input is Gaussian, i.e., $\widetilde{P}_{\mathsf{in},j}(x) = \mathcal{CN}(x_j; 0, 1)$, then the Bayes detector over the PQN model (12) becomes a linear detector. That is, x_j in (6) becomes

$$\hat{\mathbf{x}} = \frac{1}{1-\rho}\left(\mathbf{H}^*\mathbf{H} + \gamma\,\mathbf{I}\right)^{-1}\mathbf{H}^*\mathbf{r}, \tag{16}$$

where the parameter γ is commonly called the regularization factor. In particular, when γ is set to 0 and ∞, respectively,

TABLE I
SPECIAL CASES FOR THE GENERALIZED BAYES ESTIMATOR

Method	$\widetilde{P}_{\mathsf{in}}$		$\widetilde{P}_{\mathsf{out}}$	
	distribution	mismatch	distribution	mismatch
DQ-Opt.	$P_{\mathsf{in}}(x)$	No	eq.(9)	No
PDQ-Opt.	$P_{\mathsf{in}}(x)$	No	eq.(13)	Yes
Linear	Gaussian	Yes	eq.(13)	Yes

we immediately obtain the zero-forcing detector and the MRC detector[6] [38].

The linear estimator (16) can also be implemented via Algorithm 1. To this end, we substitute $\widetilde{P}_{\mathsf{in},j}(x) \sim \mathcal{CN}(x_j; 0, 1)$ into (D4). Then, we get

$$\mathscr{P}_{x_j,j}(x_j; s_j, v_j^s) = \mathcal{CN}\left(x_j; \frac{s_j}{1+v_j^s}, \frac{v_j^s}{1+v_j^s}\right), \quad (17)$$

and (D5)–(D6) in Algorithm 1 become

$$g_{\mathsf{in},j}(x_j; s_j, v_j^s) = \frac{s_j}{1+v_j^s}$$
$$g'_{\mathsf{in},j}(x_j; s_j, v_j^s) = \frac{v_j^s}{1+v_j^s} \quad (18)$$

which are simple linear steps.

The computational complexity order for the large dimensional matrix inversion in (16) is $O(NK^2)$. When the linear estimator is implemented using Algorithm 1, the computational complexity order is $O(NKt_{\max})$. In fact, Algorithm 1 converges very quickly after a constant 8–10 iterations. Thus, the computational complexity by using Algorithm 1 is one order less than the complexity by using the matrix inversion in conventional linear detectors (16). In addition, the iteration steps in Algorithm 1 only involves matrix-vector multiplications that are highly suitable for hardware implementation.

IV. PERFORMANCE ANALYSIS

We have constructed numerous detectors that range from the optimal (but complex) detector to the suboptimal (but simple) linear detector under a unified framework. All of these detectors are induced from the Bayesian inference by a particular choice of $\widetilde{P}_{\mathsf{in}}$ and $\widetilde{P}_{\mathsf{out}}$, the features of which are summarized in Table I. Remarkably, these detectors can be realized by the GAMP algorithm (i.e., Algorithm 1); thus, their performance are amenable to asymptotic analysis. Specifically, the GAMP dynamics can be tracked by a set of state evolution (SE) equations [32].

In Section IV-A, we derive the SE equations for the *generic* Bayes detector in the mixed-ADC architecture. In Section IV-B, we particularize the SE equations for the detectors constructed in Section III. Some performance comparisons among the three detectors are also provided in this subsection.

A. State Evolution Analysis

In this section, we derive the SE equations of Algorithm 1 under the mixed-ADC massive MIMO system. Our derivation is performed in the large-system regime where K and N reach infinity while the ratios

$$N/K \to \lambda, \quad N_\kappa/K \to \lambda_\kappa, \quad \forall \kappa \quad (19)$$

remain fixed.

We perform the analysis starting from the algorithm itself. Recall that output x_j is deduced from line 8 of Algorithm 1. In lines 7–8 of Algorithm 1, the posterior averages are evaluated w.r.t. $\mathscr{P}(x_j; s_j, v_j^s)$, which is determined by (s_j, v_j^s). Thus, we need to determine the asymptotic behaviors for s_j and its associated variance v_j^s at each iteration. Injecting the quantities appearing in s_j term in line 6 of Algorithm 1, we notice that s_j can be represented as a sum of x_j and many other terms. Using the CLT, we can characterize s_j by its mean and variance. A proof is provided in Appendix VI. The result is summarized by the following proposition:

Proposition 1: Given $x_j = x \sim P_{\mathsf{in}}(x)$, the asymptotic behavior of s_j in Algorithm 1 can be described by the *scalar equivalent model*

$$s_j \sim \frac{C}{E}x + \frac{\sqrt{A}}{E}z, \quad (20)$$

where $z \sim \mathcal{CN}(0, 1)$. The parameters A, C, E are independent of index j and are the solutions to the following equations

$$A = \sum_\kappa \lambda_\kappa \left[\sum_{r\in\mathcal{R}_\kappa} \int \mathrm{D}u \Psi\left(\sqrt{\frac{|v_{x\hat{x}}|^2}{v_{\hat{x}}}}u \middle| r\right)\Theta^2(\sqrt{v_{\hat{x}}}u|r)\right], \quad (21a)$$

$$C = \sum_\kappa \lambda_\kappa \left[\sum_{r\in\mathcal{R}_\kappa} \int \mathrm{D}u \Psi'\left(\sqrt{\frac{|v_{x\hat{x}}|^2}{v_{\hat{x}}}}u \middle| r\right)\Theta(\sqrt{v_{\hat{x}}}u|r)\right], \quad (21b)$$

$$E = \sum_\kappa \lambda_\kappa \left[\sum_{r\in\mathcal{R}_\kappa} \int \mathrm{D}u \Psi\left(\sqrt{\frac{|v_{x\hat{x}}|^2}{v_{\hat{x}}}}u \middle| r\right)\Theta'(\sqrt{v_{\hat{x}}}u|r)\right], \quad (21c)$$

$$v_x = \mathbb{E}_x\left\{|x|^2\right\}, \quad (21d)$$

$$c_{\hat{x}} = \mathbb{E}_{x,z}\left\{f_2\left(\frac{C\cdot x + \sqrt{A}z}{E}, \frac{1}{E}\right)\right\}, \quad (21e)$$

$$v_{x\hat{x}} = \mathbb{E}_{x,z}\left\{x^* f_1\left(\frac{C\cdot x + \sqrt{A}z}{E}, \frac{1}{E}\right)\right\}, \quad (21f)$$

$$v_{\hat{x}} = \mathbb{E}_{x,z}\left\{\left|f_1\left(\frac{C\cdot x + \sqrt{A}z}{E}, \frac{1}{E}\right)\right|^2\right\}, \quad (21g)$$

where $\lambda_\kappa = N_\kappa/K$, $x \sim P_{\mathsf{in}}(x)$, $z \sim \mathcal{CN}(0, 1)$, and

$$\Psi(\vartheta|r) = \Phi\left(\frac{\sqrt{2}r^{\mathrm{up}} - \vartheta}{\sqrt{\sigma_n^2 + v_x - \frac{|v_{x\hat{x}}|^2}{v_{\hat{x}}}}}\right) - \Phi\left(\frac{\sqrt{2}r^{\mathrm{low}} - \vartheta}{\sqrt{\sigma_n^2 + v_x - \frac{|v_{x\hat{x}}|^2}{v_{\hat{x}}}}}\right), \quad (22a)$$

$$\Theta(\vartheta|r) = \frac{\partial}{\partial\theta}\log\left(\int \mathrm{d}w \widetilde{P}_{\mathsf{out}}(r|w)\mathcal{CN}(w; \vartheta, c_{\hat{x}} - v_{\hat{x}})\right), \quad (22b)$$

[6]For the MRC detector, a multiplier γ should be added so that $\hat{\mathbf{x}} = \frac{1}{1-\rho}\mathbf{H}^*\mathbf{r}$.

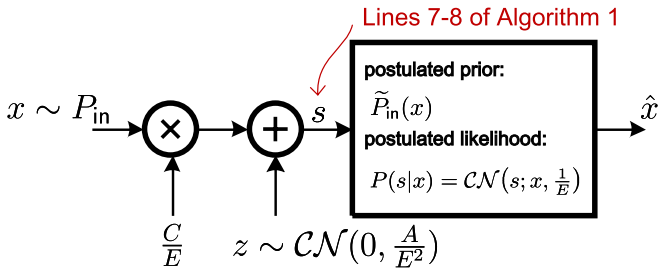

Fig. 2. The equivalent channel sees at lines 7–8 of Algorithm 1.

with $\Psi'(\vartheta|r) = \frac{\partial \Psi(\vartheta|r)}{\partial \vartheta}$, $\Theta'(\vartheta|r) = -\frac{\partial \Theta(\vartheta|r)}{\partial \vartheta}$ and

$$f_1(s,\upsilon) = \frac{\int \mathrm{d}\tilde{x}\, \tilde{x}\, \widetilde{P}_{\mathsf{in}}(\tilde{x}) \mathcal{CN}(\tilde{x}; s, \upsilon)}{\int \mathrm{d}\tilde{x}'\, \widetilde{P}_{\mathsf{in}}(\tilde{x}') \mathcal{CN}(\tilde{x}'; s, \upsilon)}, \tag{23a}$$

$$f_2(s,\upsilon) = \frac{\int \mathrm{d}\tilde{x}\, |\tilde{x}|^2\, \widetilde{P}_{\mathsf{in}}(\tilde{x}) \mathcal{CN}(\tilde{x}; s, \upsilon)}{\int \mathrm{d}\tilde{x}'\, \widetilde{P}_{\mathsf{in}}(\tilde{x}') \mathcal{CN}(\tilde{x}'; s, \upsilon)}. \tag{23b}$$

□

Before verifying the accuracy of the analysis, we closely examine Proposition 1, which has several valuable features.

1) Analytical Tractability: Proposition 1 reveals that the asymptotic behavior of s_j^t in Algorithm 1 can be characterized by the scalar equivalent model (20). The scalar model (20) is independent of index j. This characteristic indicates that lines 7–8 of Algorithm 1 are decoupled into K uncoupled scalar operators like the posterior average is over the scalar equivalent channel (20) as shown in Figure 2. In the scalar equivalent channel, the Bayes detector postulates $\widetilde{P}_{\mathsf{in}}(x)$ and $P(s|x) = \mathcal{CN}(s; x, \frac{1}{E})$ as the prior distribution and the likelihood function, respectively. Note that the noise variance of the scalar channel is $\frac{A}{E^2}$ while the Bayes detector adopts $\frac{1}{E}$ as the noise variance because of the mismatched measure imported in $\widetilde{P}_{\mathsf{out}}$. Using the scalar equivalent channel, we can easily compute a large class performance metrics such as MSE, BER, and mutual information for the Bayes detector. With regard to this performance metric, we are interested in the algorithm at time index $t_{\max} \to \infty$, and thus we omit the superscript index $t_{\max}$ from the related parameters for brevity. We provide two examples as follows:

- If $P_{\mathsf{in}}(x)$ is the QPSK signal, i.e.,

$$P_{\mathsf{in}}(x) = \frac{1}{4}\delta\left(x \pm \frac{1}{\sqrt{2}} \pm \mathrm{j}\frac{1}{\sqrt{2}}\right), \tag{24}$$

  then we can obtain the theoretical BER of the Bayes detector from (20). That is,[7]

$$\mathsf{BER} = Q\left(\frac{C}{\sqrt{A}}\right), \tag{25}$$

  where $Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{+\infty} \mathrm{d}t e^{-\frac{t^2}{2}}$ is the error function.
- For an arbitrary modulation signal, $f_1(\cdot)$ and $f_2(\cdot)$ in (21d)–(21g) are the conditional mean and second moment estimates of x over the scalar channel (20) while postulating mismatched prior and likelihood functions. As a result, the MSE of the Bayes detector can be shown as

$$\mathrm{MSE} = \mathbb{E}\left\{|x - \hat{x}|^2\right\} = \upsilon_x - 2\mathrm{Re}\{\upsilon_{x\hat{x}}\} + \upsilon_{\hat{x}}. \tag{26}$$

2) Generality: The analytical framework can apply to arbitrary postulated measures by substituting different prior distributions $\widetilde{P}_{\mathsf{in}}$ into (23) and likelihood functions $\widetilde{P}_{\mathsf{out}}$ into (22b). Moreover, $\Psi(x|r)$ given in (22a) characterizes the *true* quantization effect on an interval $(r^{\mathrm{low}}, r^{\mathrm{up}}]$ and the summation $\sum_{r \in \mathcal{R}_\kappa}$ in (21a)–(21c) adds all the corresponding interval together. The quantization bins and the quantization level are configurable. Thus, the SE equations can incorporate arbitrary quantization processes, such as nonuniform quantization. Moreover, $\sum_\kappa \lambda_\kappa$ in (21a)-(21c) indicates the effect of the *mixed*-ADC architecture, and the result holds for a fixed allocation of ADCs to antennas. Interestingly, when particularizing our results to the case with a non-mixed ADC architecture, we recover the same asymptotic BER expression (25) as found in [43] with the replica method.

3) Computational Simplicity: The analytical result is computationally simple because the corresponding parameters (i.e., $A, C, E, c_{\hat{x}}, \upsilon_{x\hat{x}}, \upsilon_{\hat{x}}, \upsilon_x$) can be obtained in an iterative way, with each iteration only involving scalar summations (21a)–(21c) and scalar estimation computations (21d)–(21g). In fact, using (20), we can predict the SE in time of Algorithm 1 in the mixed-ADC massive MIMO system. Therefore, instead of performing time-consuming Monte Carlo simulations to obtain the corresponding performance metrics, we can predict the theoretical behavior by SE equations in a very short time.

B. Explicit Expressions

We now show special cases of the SE equations in Section IV-A. To present a relatively intuitive analytical result, all the expressions and discussions in this subsection are centered on a QPSK input distribution although our analysis can be incorporated with arbitrary input distributions.

For the DQ-optimal detector, the postulated measures are the same as the true measures for both P_{in} and P_{out}. Also, recall that the PDQ-optimal detector only suffers a mismatched likelihood measure $\widetilde{P}_{\mathsf{out}}$ in (13) and the linear detector further postulates the input as Gaussian. Substituting the corresponding measures into Proposition 1, we obtain the explicit SE expressions for various detectors. The details of the expressions are listed in Table II.[8] In particular, for the PDQ-optimal detector and the linear detector, the term in the logarithm of (22b) can be simplified as

$$\int \mathrm{d}w\, \widetilde{P}_{\mathsf{out}}(r|w) \mathcal{CN}(w; \vartheta, c_{\hat{x}} - \upsilon_{\hat{x}}) = e^{-\frac{\left(\frac{\sqrt{2}}{1-\rho} r - \vartheta\right)^2}{\gamma + c_{\hat{x}} - \upsilon_{\hat{x}}}}. \tag{28}$$

As a result, the terms $\Theta(y|\vartheta)$ and $\Theta'(y|\vartheta)$ in (21a)–(21c) can be further simplified as (27) shown in Table II.

[7]For other high-order modulations, the closed-form of the BER expression can be derived based on the analysis in [42].

[8]From the previous analysis, the parameters $\upsilon_{\hat{x}}$, υ_x, $c_{\hat{x}}$, and $\upsilon_{x\hat{x}}$ should be the same for the DQ-Opt. and PDQ-Opt. detectors. In this table, we make some approximations so that $\Theta \approx \frac{\Psi'}{\Psi}$. This approximation significantly simplifies the analytical expressions and provides good predictions.

TABLE II
EXPLICIT EXPRESSIONS FOR GENERALIZED BAYES ESTIMATORS FOR QPSK CONSTELLATION SIGNALS

Method	**SE parameters**	
DQ-Opt.	$v_x = c_{\hat{x}} = 1$ $v_{\hat{x}} = v_{x\hat{x}} = \int \mathrm{D}u \tanh(\sqrt{A}u + C)$	$A = \sum_{\kappa} \lambda_\kappa \sum_{r \in \mathcal{R}_\kappa} \int \mathrm{D}u \frac{\left(\Psi'(\sqrt{v_{\hat{x}}}u\|r)\right)^2}{\Psi(\sqrt{v_{\hat{x}}}u\|r)}$ $C = E = A$
PDQ-Opt.	$v_x = c_{\hat{x}} = 1$ $v_{x\hat{x}} = \int \mathrm{D}u \tanh(\sqrt{A}u + C)$ $v_{\hat{x}} = \int \mathrm{D}u \tanh^2(\sqrt{A}u + C)$	A, C, and E are given in (21a)–(21c) without the superscript iteration index t, and the terms $\Theta(\vartheta\|r)$ and $\Theta'(\vartheta\|r)$ can be simplified as $\Theta(\vartheta\|r) = \frac{\frac{\sqrt{2}}{1-\rho}r - \vartheta}{\gamma + c_{\hat{x}} - v_{\hat{x}}}$, (27a) $\Theta'(\vartheta\|r) = \frac{1}{\gamma + c_{\hat{x}} - v_{\hat{x}}}$. (27b)
Linear	$v_x = 1$ $c_{\hat{x}} = \frac{A + C^2}{(1+E)^2} + \frac{1}{1+E}$ $v_{x\hat{x}} = \frac{C}{1+E}$ $v_{\hat{x}} = \frac{A + C^2}{(1+E)^2}$	

We now verify the accuracy of the analytical result presented in Table II. First, we evaluate the simplest non-mixed ADC case, i.e., the quantization function $Q_i^c(\cdot)$ is the same for all antennas. In particular, we compare the BER expression (25) with that obtained by computer simulations. The simulations are conducted over 10,000 channel realizations where Algorithm 1 is performed with the maximum iteration times $t_{\max} = 20$. We set the parameters as $N = 200$ and $K = 50$. In addition, we adopt a fixed step size $\Delta = 0.5$.

As shown in Figure 3, the BER performance for various detectors are compared. The markers denote the numerical results while the curves characterize the analytical behaviors. It can be seen that the markers matches well with the curves, which shows precise predictions by proposition 1. Actually, numerous examination simulations have been conducted to verify the accuracy of the SE equations, including the mixed-ADC architecture. We omit all these plots because of space limitations.

Apart from the accuracy verification, we can also get some insights from Figure 3. With the increase of quantization precision, all detectors achieve a performance improvement. For example, in the 1-bit quantization case, the PDQ-optimal detector and the linear detector suffer from an error floor and almost do not work. However, in the 3-bit quantization case, the PDQ-optimal detector achieves an amazing performance gain; the gap between the DQ-optimal and PDQ-optimal detectors is extremely small. Even the linear detector only incurs a loss of $5.63 - 4.41 = 1.21$ dB to that attained by the DQ-optimal detector when BER $= 10^{-3}$. Recall that the PDQ-optimal detector drops out the integration parts in (9) to reduce the computation cost. Figure 3 shows that the effect of this "dropping" weakens with the increase of the resolution level. When the resolution goes to infinity, the integration effect is eventually eliminated, giving the PDQ-optimal and DQ-optimal detectors a common ultimate performance limit as predicted by the solid curve in Figure 3. Therefore, instead of turning to the more sophisticated DQ-optimal detector, the PDQ-optimal detector with low precision (e.g., 3-bit) ADCs seems sufficient to achieve near optimal performance ($4.41 - 3.83 = 0.58$ dB loss at BER $= 10^{-3}$ is acceptable) as the optimal *unquantized* MIMO detector.

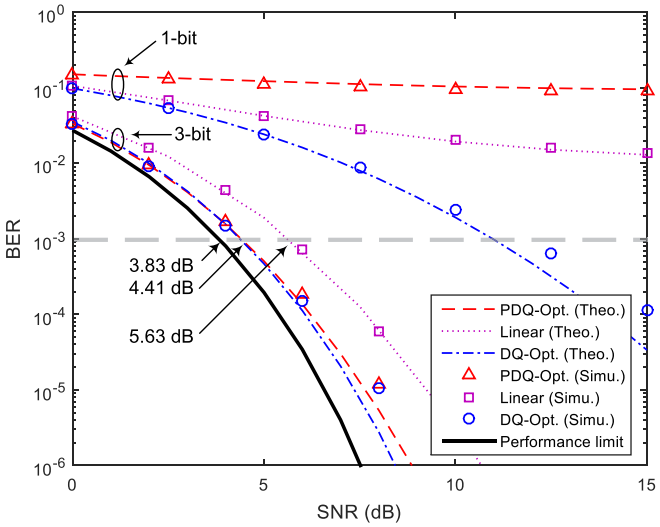

Fig. 3. Comparison of analytical predictions against Monte Carlo simulations for various detectors. $N = 200$, $K = 50$, and $\Delta = 0.5$ with a QPSK input. Dashed curve with color: theoretical prediction. Markers: simulation result. Solid curve: ultimate performance limit achieved by the DQ-optimal detector when the receiver has infinite precision to the received signal.

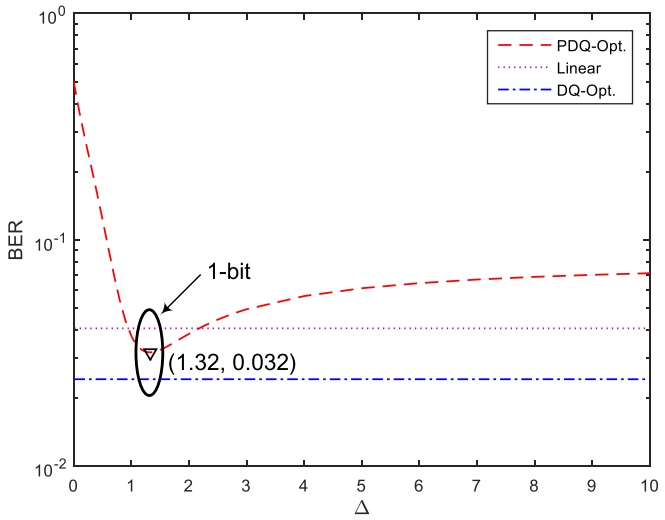

Fig. 4. BER versus step size under 1-bit quantization with $\lambda = 4$, SNR = 5 dB, and a QPSK input.

Finally, we point out an interesting result from Figure 3. Theoretically, the linear detector should have poorer performance than the PDQ-optimal detector because the linear detector further imposes a mismatch in P_{in}. Contrary to our belief, this inference seems to be un-true in the 1-bit quantization case. The reason, as shown in the following section, is that the PDQ-optimal detector is sensitive to the step size Δ, which implies the importance of parameter setting in the MIMO detectors.

V. Discussions

In this section, we discuss how the MIMO detectors are optimized using the analytical expressions presented in Table II. The issues are related to quantization-step size and the mixed receiver architecture.

A. Quantization Step Size

We consider the 1-bit quantization case. Following a parameter setting similar to that shown in Figure 3, Figure 4 presents the average BER versus the step sizes under 1-bit quantization with SNR=5 dB. The performance of the 1-bit DQ-optimal detector is *irrelevant* to the value of the step sizes. The reason for this finding is that the integrals in lines 3–4 in Algorithm 1 [or specifically, (10) and (11)] are considered w.r.t. either an interval $(-\infty, 0]$ or $[0, +\infty)$, and only the sign of the output is of interest. However, this property is *not* true for the other detectors. The performances of the PDQ-optimal detector and the linear detector are affected by the value of Δ, with the former appearing to be more sensitive. For example, when the step size is at an optimal value of 1.32, the PDQ-optimal detector achieves its best performance. This result coincides with our inference that the performance of the PDQ-optimal detector is between that of the DQ-optimal detector and the linear detector.

For multi-bit uniform quantization, the optimum step sizes that minimize the corresponding BER for a QPSK input signal are compared with the input bit resolutions listed in Table III. The corresponding step size values of the PDQ-optimal detector and the linear detector are remarkably close for all bits ($\kappa > 2$). This observation suggests that a prior mismatch may not cause a significant difference in terms of an optimized step size.

TABLE III

Optimal Step Sizes for DQ-Optimal/PDQ-Optimal/ Linear Detectors ($\lambda = 4$)

SNR (dB)	1 bit	2 bit	3 bit
-5	— / 2.724 / 0.490	1.990 / 1.448 / 1.436	1.147 / 0.847 / 0.845
-2.5	— / 2.081 / 0.796	1.611 / 1.190 / 1.174	0.932 / 0.697 / 0.691
0	— / 1.691 / 0.300	1.338 / 1.016 / 0.996	0.781 / 0.591 / 0.586
2.5	— / 1.458 / 0.560	1.132 / 0.901 / 0.880	0.674 / 0.523 / 0.518
5	— / 1.322 / 0.700	0.970 / 0.829 / 0.808	0.594 / 0.481 / 0.475
7.5	— / 1.244 / 1.200	0.835 / 0.786 / 0.764	0.531 / 0.456 / 0.450
10	— / 1.200 / 1.070	0.717 / 0.760 / 0.738	0.474 / 0.441 / 0.435
12.5	— / 1.175 / 1.066	0.609 / 0.745 / 0.724	0.419 / 0.432 / 0.426
15	— / 1.160 / 1.000	0.511 / 0.737 / 0.715	0.366 / 0.427 / 0.421
17.5	— / 1.152 / 1.050	0.423 / 0.732 / 0.710	0.315 / 0.424 / 0.418
20	— / 1.148 / 0.900	0.347 / 0.729 / 0.708	0.267 / 0.423 / 0.416
SNR (dB)	4 bit	5 bit	6 bit
-5	0.651 / 0.484 / 0.484	0.365 / 0.271 / 0.271	0.202 / 0.150 / 0.150
-2.5	0.530 / 0.396 / 0.395	0.298 / 0.222 / 0.222	0.165 / 0.123 / 0.123
0	0.447 / 0.336 / 0.335	0.251 / 0.188 / 0.188	0.145 / 0.104 / 0.104
2.5	0.389 / 0.297 / 0.296	0.220 / 0.167 / 0.166	0.123 / 0.092 / 0.092
5	0.349 / 0.273 / 0.272	0.199 / 0.153 / 0.153	0.111 / 0.084 / 0.084
7.5	0.318 / 0.259 / 0.257	0.184 / 0.145 / 0.145	0.104 / 0.080 / 0.080
10	0.292 / 0.250 / 0.249	0.171 / 0.145 / 0.145	0.098 / 0.077 / 0.077
12.5	0.267 / 0.245 / 0.245	0.160 / 0.145 / 0.145	0.092 / 0.076 / 0.076
15	0.241 / 0.245 / 0.240	0.149 / 0.145 / 0.145	0.087 / 0.075 / 0.075
17.5	0.216 / 0.245 / 0.239	0.137 / 0.145 / 0.145	0.082 / 0.074 / 0.074
20	0.190 / 0.245 / 0.238	0.125 / 0.145 / 0.145	0.077 / 0.074 / 0.074

TABLE IV

Comparison the Optimal Step Sizes for Different Purposes

κ (bits)	Δ_{norm} (PDQ-Opt.)	Δ_{norm} (Linear)	$\Delta_{\text{opt.}}$ from [44]
2	1.0203	0.9957	1.0080
3	0.5926	0.5860	0.5895
4	0.3383	0.3359	0.3360
5	0.1941	0.1940	0.1883
6	0.1041	0.1041	0.1041
7	0.0569	0.0569	0.0569

Recall that the element of an ADC input y_i is the sum across many independent terms, and thus, it can be approximated as a Gaussian distributed variable with variance $1+\sigma_n^2$. Therefore, we normalize the optimal step sizes according to

$$\Delta_{\text{norm}} = \frac{\sqrt{2}\Delta}{\sqrt{1+\sigma_n^2}} \tag{29}$$

and average them over different SNRs. In this case, the step size is multiplied with $\sqrt{2}$ because the signal power is twice the power of the real or imaginary part. The corresponding normalization results are presented in Table IV. The optimal step sizes that attempt to minimize quantization distortion for a Gaussian input signal [44] are provided in Table IV for comparison. The value of Δ_{norm} is extremely close to those

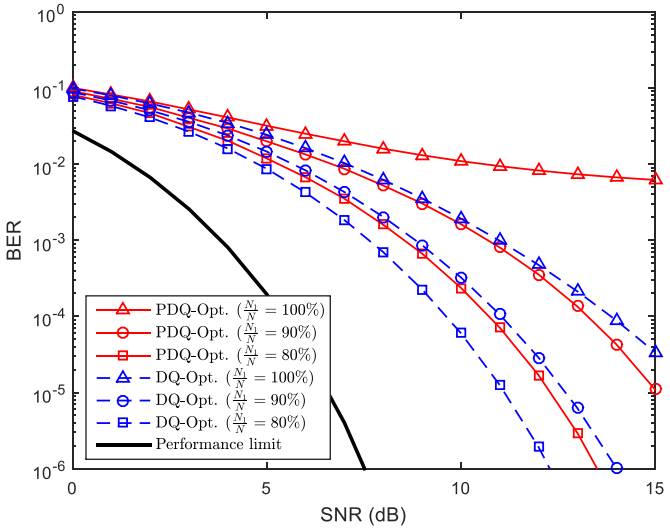

Fig. 5. BER versus SNR under the *mixed 1-bit* architecture for different detectors with QPSK inputs.

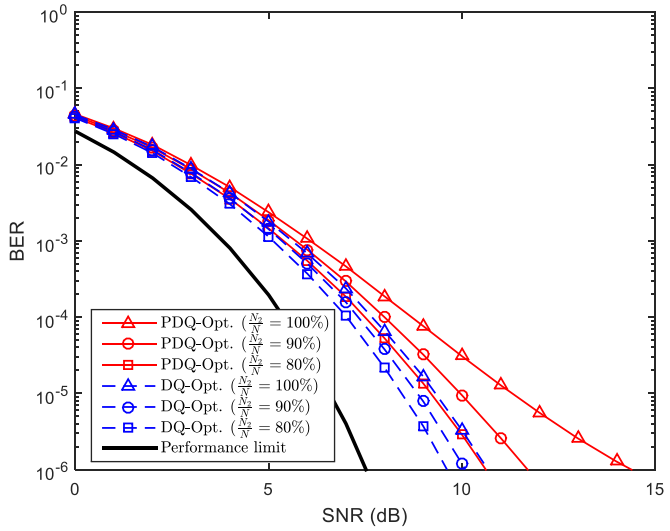

Fig. 6. BER versus SNR under the *mixed 2-bit* architecture for different detectors with QPSK inputs.

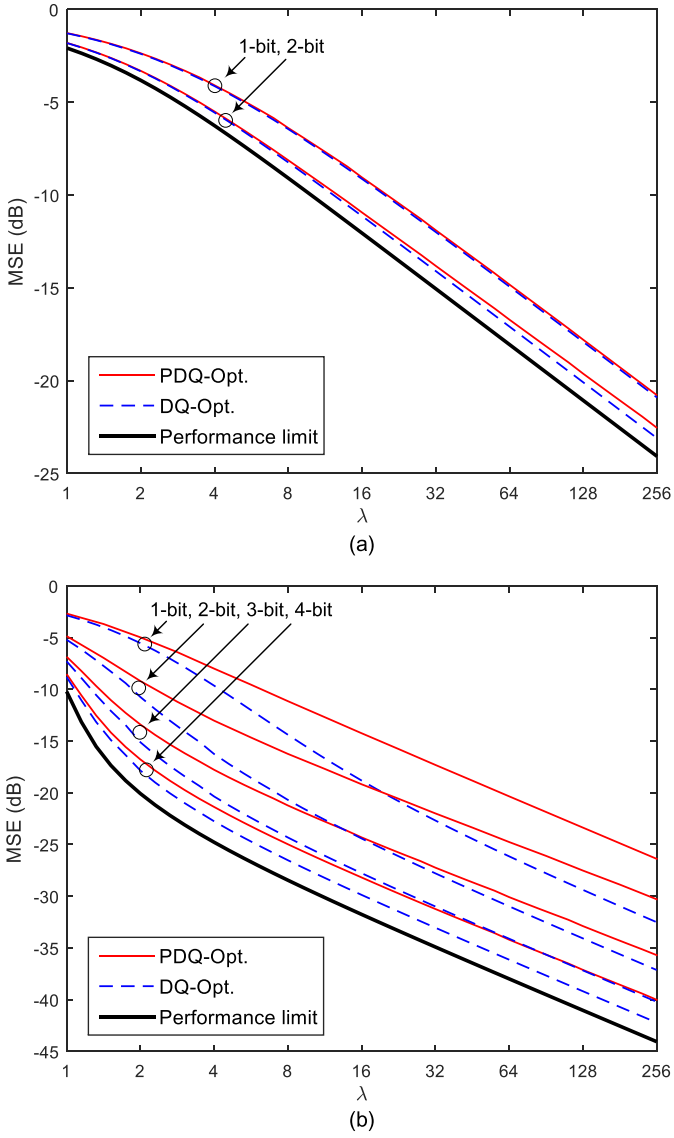

Fig. 7. MSE for a Gaussian input signal versus MIMO configuration $\lambda = N/K$ with a) SNR $= 0$dB and b) SNR $= 20$dB.

presented in [44] for all input bit rates κ that vary from 2 to 7. Through extensive simulations under different antenna ratios, input constellations, and performance metrics (MSE or BER), we determine that this property is general. Consequently, we can safely use the values presented in [44] as a consistent rule to assist in the step size design. The rule of thumb does not apply to the 1-bit quantization case or the DQ-optimal detector. For these special cases, however, we can quickly obtain the optimal step sizes based on the analytical expressions.

B. Mixed Architecture

We have discussed the optimal parameter settings for the quantization step size. Using these optimal parameters, we investigate the performance of the mixed-ADC architecture. Our discussion focuses on the following question: *how close to the best performance can a MIMO detector operate without considering the exact nonlinear effect of the quantizers?* In the following subsections, we refer to a system in which most of the antennas adopt 1-bit resolution ADCs, whereas the remaining antennas have full precision ADCs as the *mixed 1-bit* architecture. We also obtain the *mixed 2-bit* architecture and others.

1) QPSK Input: Figure 5 illustrates the performance of the *mixed 1-bit* architecture. We observe that simply installing 10% high-resolution ADCs (i.e., $\frac{N_1}{N} = 90\%$) eliminates the error floor caused by the distortion of the mismatched measures at the output. In this case, the mixed PDQ-optimal detector can achieve a performance similar to that of a pure 1-bit DQ-optimal detector. If we further increase the fraction of the full-precision ADCs to 20% (i.e.,$\frac{N_1}{N} = 80\%$), then an 80% 1-bit mixed PDQ-optimal detector will achieve a performance similar to that of a 90% DQ-optimal detector.

In contrast to the significant performance gap reduction caused by the *mixed 1-bit* architecture, the *mixed 2-bit* architecture (Figure 6) appears to play a minor role.

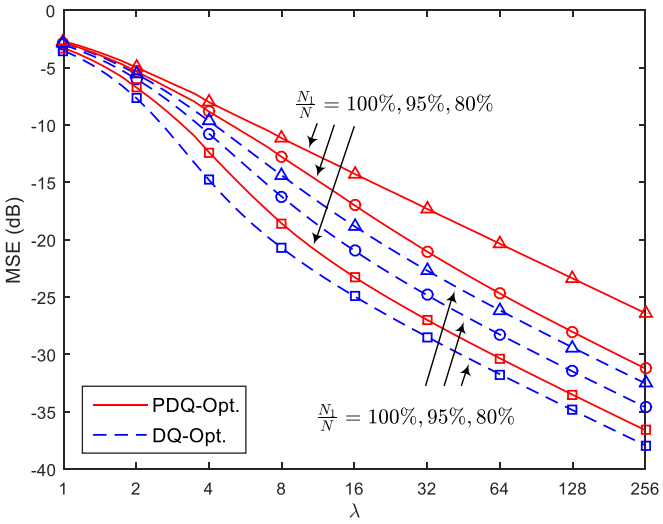

Fig. 8. MSE for a Gaussian input signal versus MIMO configuration $\lambda = N/K$ in the mixed architecture with SNR = 20 dB.

We determine that an 80% 2-bit mixed PDQ-optimal detector achieves a performance similar to that of a pure 2-bit DQ-optimal detector. Furthermore, we can infer that the performance gap will decrease as resolution level increases and will eventually be eliminated within the limit of full precision. Therefore, the mixed architecture can help maintain the promised performance while significantly reducing computational complexity by disregarding the exact but complex quantization process, particularly in extremely low resolution cases (e.g., 1-bit, 2-bit).

2) Gaussian Input: We begin by providing insights into the effect of quantization based on the non-mixed architecture. Figure 7 illustrates the MSEs of the detectors versus the antenna configuration ratios $\lambda = M/N$ for SNR = 0 dB and SNR = 20dB. As shown in Figure 7(a), when SNR = 0 dB, the performance degradation compared with that in the unquantized case caused by 1-bit quantization is approximately 3 dB, which indicates that disregarding the exact quantization effect is feasible in the low-SNR regime. In the high-SNR regime, however, a 1-bit resolution generally incurs significant performance losses, as shown in Figure 7(b). In a typical *massive* MIMO system (e.g., $\lambda = 16$), the PDQ-optimal detector generally causes a 1-bit loss compared with the DQ-optimal detector. For example, the 2-bit PDQ-optimal detector exhibits the same performance as the 1-bit DQ-optimal detector. Alternatively, the loss caused by simplifying the quantization process can be compensated by doubling the number of receiver antennas. Figures 7(a) and 7(b) show that the MSEs of the detectors generally improve by 3–6 dB for each 1-bit rate increase. When the SNR is high, the MSE improvement is close to 6 dB.

Figure 8 illustrates the results under the mixed architecture. The mixed architecture can help narrow down the performance gap that results from the PQN model. For example, in a massive MIMO system with $\lambda > 10$, the pure 1-bit PDQ-optimal detector incurs a loss of approximately 6 dB compared with the pure 1-bit DQ-optimal detector. Their corresponding gaps are respectively reduced to 3 dB and 1 dB when 5% and 20% full precision ADCs are installed.

VI. Conclusion

Using a unified framework, we specified three types of detectors for a mixed ADC massive MIMO receiver by postulating mismatched measures in the Bayes detector. The asymptotic performances of these detectors were analyzed on the basis of SE equations, and their accuracies were validated through Monte Carlo simulations. The SE equations can be rapidly and efficiently evaluated. Thus, they can be used to optimize system designs. We also provided useful observations as a basis for design optimization. Our results indicated that computational burdens can be reduced by regarding complex nonlinear quantization as a PQN model. Disregarding such a complex quantization process may cause 1-bit performance degradation in the high-SNR regime but may not yield a different result in the low-SNR regime. The mixed receiver architecture can also help decrease the performance gap attributed to the PQN model. Although this work has expanded our knowledge on the performance of a mixed-ADC MIMO detector, a general analytic framework that accounts for realistic fading models, including antenna correlations at BS and among users, and imperfect CSI should be considered in future studies.

Appendix A
Derivation of GAMP Functions $g_{\text{out},i}$ and $g'_{\text{out},i}$

In this Appendix, we derive the GAMP quantities $g_{\text{out},i}(w_i; p_i, v_i^p)$ and $g'_{\text{out},i}(w_i; p_i, v_i^p)$ given in (D2)–(D3). For brevity, we omit the index i.

From (D1), we determine that

$$\mathbb{E}_{w|r}\{w|r;\, p, v^p\} = \frac{1}{P_r(r)} \int \mathrm{d}w\, w \widetilde{P}_{\text{out}}(r|w) \mathcal{CN}(w;\, p, v^p), \tag{A.1}$$

$$\mathbb{V}_{w|r}\{w|r;\, p, v^p\} = \frac{1}{P_r(r)} \int \mathrm{d}w \left| w - \mathbb{E}_{w|r}\{w|r;\, p, v^p\} \right|^2 \times \widetilde{P}_{\text{out}}(r|w) \mathcal{CN}(w;\, p, v^p), \tag{A.2}$$

where $P_r(r) \triangleq \int_{w'} \mathrm{d}w' \widetilde{P}_{\text{out}}(r|w') \mathcal{CN}(w';\, p, v^p)$. From the likelihood in (9), we obtain

$$\int \mathrm{d}w\, w \widetilde{P}_{\text{out}}(r|w) \mathcal{CN}(w;\, p, v^p) = \int \mathrm{d}w \int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, w \cdot \mathcal{CN}(w;\, y, \sigma_n^2) \mathcal{CN}(w;\, p, v^p), \tag{A.3}$$

and

$$P_r(r) = \int_w \mathrm{d}w \int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, \mathcal{CN}(y;\, w, \sigma_n^2) \mathcal{CN}(w;\, p, v^p). \tag{A.4}$$

Applying the property that

$$\mathcal{CN}(x;\, a, A) \mathcal{CN}(x;\, b, B) = \mathcal{CN}\left(x;\, \frac{a/A + b/B}{1/A + 1/B}, \frac{1}{1/A + 1/B}\right) \mathcal{CN}(0;\, a - b, A + B),$$

we can write

$$\begin{aligned}&\int \mathrm{d}w\, w \widetilde{P}_{\text{out}}(r|w)\mathcal{CN}(w;\, p, v^p)\\ &= \int \mathrm{d}w \int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, w \cdot \mathcal{CN}\left(w;\, \frac{y/\sigma_n^2 + p/v^p}{1/\sigma_n^2 + 1/v^p}, \frac{1}{1/\sigma_n^2 + 1/v^p}\right)\\ &\quad \times \mathcal{CN}(0;\, y - p, \sigma_n^2 + v^p)\\ &= \int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, \frac{y/\sigma_n^2 + p/v^p}{1/\sigma_n^2 + 1/v^p} \cdot \mathcal{CN}(y;\, p, \sigma_n^2 + v^p). \end{aligned} \tag{A.5}$$

Using a similar procedure, we obtain

$$P_r(r) = \int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, \mathcal{CN}(y;\, p, \sigma_n^2 + v^p). \tag{A.6}$$

Finally, with (A.5) and (A.6), (D2) can be rewritten as

$$g_{\text{out}}(w;\, p, v^p) = \frac{\widetilde{r} - p}{v^p + \sigma_n^2}, \tag{A.7}$$

where

$$\widetilde{r} = \frac{\int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, y \cdot \mathcal{CN}(y;\, p, v^p + \sigma_n^2)}{\int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, \mathcal{CN}(y;\, p, v^p + \sigma_n^2)}. \tag{A.8}$$

Similar to (A.5), we can write

$$\begin{aligned}&\int \mathrm{d}w \left|w - \mathbb{E}_{w|r}\{w|r;\, p, v^p\}\right|^2 \widetilde{P}_{\text{out}}(r|w)\mathcal{CN}(w;\, p, v^p)\\ &= \int \mathrm{d}w \int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y \left|w - \mathbb{E}_{w|r}\{w|r;\, p, v^p\}\right|^2\\ &\quad \times \mathcal{CN}\left(w;\, \frac{y/\sigma_n^2 + p/v^p}{1/\sigma_n^2 + 1/v^p}, \frac{1}{1/\sigma_n^2 + 1/v^p}\right)\\ &\quad \times \mathcal{CN}(0;\, y - p, \sigma_n^2 + v^p). \end{aligned} \tag{A.9}$$

We then define $\bar{w} \triangleq \mathbb{E}_{w|r}\{w|r;\, p, v^p\}$ and $\tilde{w} \triangleq w - \bar{w}$. Then, (A.9) can be written as

$$\begin{aligned}(A.9) = &\int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y \int \mathrm{d}\tilde{w}\, |\tilde{w}|^2 \times \mathcal{CN}(0;\, y - p, \sigma_n^2 + v^p)\\ &\times \mathcal{CN}\left(\tilde{w};\, \frac{y/\sigma_n^2 + p/v^p}{1/\sigma_n^2 + 1/v^p} - \bar{w}, \frac{1}{1/\sigma_n^2 + 1/v^p}\right). \end{aligned} \tag{A.10}$$

Therefore, with (A.6) and (A.10), (D3) can be rewritten as

$$g'_{\text{out}}(w;\, p, v^p) = \frac{1}{v^p + \sigma_n^2}\left(1 - \frac{\widetilde{\sigma}^2}{v^p + \sigma_n^2}\right), \tag{A.11}$$

where

$$\widetilde{\sigma}^2 := \frac{\int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, |y - \widetilde{r}|^2\, \mathcal{CN}(y;\, p, v^p + \sigma_n^2)}{\int_{r^{\text{low}}}^{r^{\text{up}}} \mathrm{d}y\, \mathcal{CN}(y;\, p, v^p + \sigma_n^2)}. \tag{A.12}$$

Appendix B
Proof of Proposition 1

Recall that the GAMP algorithm is an extension of several earlier Gaussian and quadratic approximations of loopy belief propagation (BP), which updates messages between the output nodes and the input nodes iteratively in a *factor* graph as illustrated in Figure 9. Comparing the structure of GAMP with loopy BP, we find that lines 1–4 in the GAMP algorithm resembles the messages transmitted from

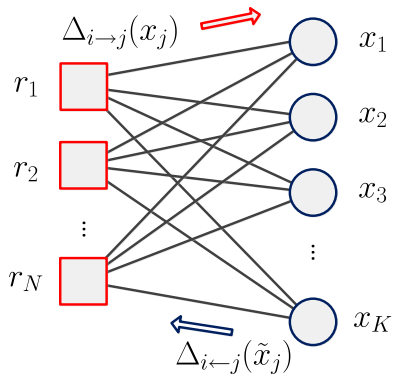

Fig. 9. Factor graph of the linear mixing estimation problem.

the input node x_j $(j = 1, \ldots, K)$ to the output node r_i $(i = 1, \ldots, N)$, whereas lines 5–8 represent the reverse message updates from r_i to x_j, which we denote as $\Delta_{i\to j}^t(x_j)$ and $\Delta_{i\leftarrow j}^{t+1}(\tilde{x}_j)$, respectively.

In this Appendix, we present the asymptotic analysis for Algorithm 1 in the large-system limit (i.e., $N, K \to \infty$) by following [45]. For conceptual clarity and ease of explanation, we restrict the following analysis to a corresponding real-valued system although Proposition 1 is for the complex-valued case. Also, for brevity, we omit the entire iteration index t.

A useful interpretation of Proposition 1 is that it provides a *scalar equivalent model* for the behavior of the GAMP estimates. From lines 7–8 of Algorithm 1 together with the equivalent scalar model illustrated in Figure 2, we observe that the distribution of the estimated signal is identical to s_j being a scaled and noise-corrupted version of x_j. Specifically, the conditional mean and variance of x_j, which we denote as $\hat{x}_j$ and $\hat{v}_j^x$, respectively, are taken w.r.t. the marginal posterior $\mathscr{P}_{x_j}(x_j; s_j, v_j^s)$ given in (D4). Let $\mathbf{h}_i^T$ be the i-th row of $\mathbf{H}$ and $w_{ij} = \sum_{k\neq j} h_{ik}x_k$. Then, we have $\mathbf{h}_i^T\mathbf{x} = w_{ij} + h_{ij}x_j$. For a fixed $\mathbf{h}_i^T$,

$$\mathbb{E}_{x_j|s_j,v_j^s}\{w_{ij}\} = \sum_{k\neq j} h_{ik}\hat{x}_k \triangleq \hat{w}_{ij}, \tag{B.1a}$$

$$\mathbb{V}_{x_j|s_j,v_j^s}\{w_{ij}\} = \sum_{k\neq j} h_{ik}^2\hat{v}_k^x \triangleq \hat{v}_{ij}^w, \tag{B.1b}$$

where the expectation and variance are taken w.r.t. $\mathscr{P}_{x_j}(x_j; s_j, v_j^s)$. According to the CLT in the large system limit, w_{ij} can be regarded as a Gaussian random variable with mean $\hat{w}_{ij}$ and variance $\hat{v}_{ij}^w$. Then, the messages from the output nodes can be expressed as

$$\begin{aligned}\Delta_{i\to j}(x_j) &\approx \int \mathrm{d}w_{ij}\, \widetilde{P}_{\text{out}}(r_i|w_{ij} + h_{ij}x_j)\mathcal{N}(w_{ij};\, \hat{w}_{ij}, \hat{v}_{ij}^w)\\ &\triangleq \Xi(\hat{w}_{ij} + h_{ij}x_j|r_i). \end{aligned} \tag{B.2}$$

Since $h_{ij}x_j$ scales as $\frac{1}{\sqrt{K}}$, the second-order expansion of the messages (B.2) can be expressed as

$$\begin{aligned}\Xi(\hat{w}_{ij} + h_{ij}x_j|r_i) \approx\ &\Xi(\hat{w}_{ij}|r_i) + \Xi'(\hat{w}_{ij}|r_i)h_{ij}x_j\\ &+\Xi''(\hat{w}_{ij}|r_i)h_{ij}^2x_j^2 + O(K^{-3/2}), \end{aligned} \tag{B.3}$$

where we use $\Xi'(w|r)$ and $\Xi''(w|r)$ to denote the first-order and second-order derivatives of $\Xi(w|r)$ w.r.t. w, respectively.

Following a similar approximation in [45], we express the messages as

$$\Delta_{i\to j}(x_j) \propto \exp\left[\alpha_{ij}x_j - \frac{1}{2}\left(\alpha_{ij}^2 - \beta_{ij}\right)x_j^2\right], \tag{B.4a}$$

$$\Delta_{i\leftarrow j}(\tilde{x}_j) \propto \widetilde{P}_{\mathsf{in}}(\tilde{x}_j)\exp\left[\sum_{l\neq i}\left(\alpha_{lj}\tilde{x}_j - \frac{1}{2}\left(\alpha_{lj}^2 - \beta_{lj}\right)\tilde{x}_j^2\right)\right], \tag{B.4b}$$

where

$$\alpha_{ij} = \frac{\Xi'(\hat{w}_{ij}|r_i)}{\Xi(\hat{w}_{ij}|r_i)}h_{ij} \quad \text{and } \beta_{ij} = \frac{\Xi''(\hat{w}_{ij}|r_i)}{\Xi(\hat{w}_{ij}|r_i)}h_{ij}^2. \tag{B.5}$$

Note that we use $\tilde{x}_j$ to denote $\tilde{x}_j \sim \widetilde{P}_{\mathsf{in}}(\tilde{x}_j)$.

We then define

$$A_j = \sum_{l\neq i}\alpha_{lj}^2, \quad B_j = \sum_{l\neq i}\beta_{lj}, \text{ and } a_j = \sum_{l\neq i}\alpha_{lj}, \tag{B.6}$$

the asymptotic (or the large system) behavior of A_j, B_j and a_j follows

$$A_j \to \sum_{l\neq i}\mathbb{E}_{\mathbf{r},\hat{w}_{lj}|\mathbf{H}}\left\{\alpha_{lj}^2\middle|\mathbf{H}\right\} \triangleq \sum_{\kappa}\mathsf{A}_\kappa, \tag{B.7a}$$

$$B_j \to \sum_{l\neq i}\mathbb{E}_{\mathbf{r},\hat{w}_{lj}|\mathbf{H}}\left\{\beta_{lj}\middle|\mathbf{H}\right\} \triangleq \sum_{\kappa}\mathsf{B}_\kappa, \tag{B.7b}$$

$$a_j \to \sum_{l\neq i}\mathbb{E}_{\mathbf{r},\hat{w}_{lj}|\mathbf{H}}\left\{\alpha_{lj}\middle|\mathbf{H}\right\} \triangleq \sum_{\kappa}\mathsf{C}_\kappa, \tag{B.7c}$$

where $\to$ indicates the convergence to the asymptotic limit, and

$$\mathsf{A}_\kappa = \sum_{l\in\Omega_\kappa, l\neq i}\mathbb{E}_{\mathbf{r}_\kappa,\hat{w}_{lj}|\mathbf{H}}\left\{\alpha_{lj}^2\middle|\mathbf{H}\right\}, \tag{B.8a}$$

$$\mathsf{B}_\kappa = \sum_{l\in\Omega_\kappa, l\neq i}\mathbb{E}_{\mathbf{r}_\kappa,\hat{w}_{lj}|\mathbf{H}}\left\{\beta_{lj}\middle|\mathbf{H}\right\}, \tag{B.8b}$$

$$\mathsf{C}_\kappa = \sum_{l\in\Omega_\kappa, l\neq i}\mathbb{E}_{\mathbf{r}_\kappa,\hat{w}_{lj}|\mathbf{H}}\left\{\alpha_{lj}\middle|\mathbf{H}\right\}. \tag{B.8c}$$

Here, $\mathbf{r}_\kappa$ represents the quantized signals that belong to the discrete set $\mathcal{R}_\kappa$. The expectation in (B.8) is taken w.r.t. the joint distribution $P(\mathbf{r}_\kappa, \hat{w}_{ij}|\mathbf{H})$ of the two correlated variables $\mathbf{r}_\kappa$ and $\hat{w}_{ij}$.

Our strategy is to obtain $P(\mathbf{r}_\kappa, \hat{w}_{ij}|\mathbf{H})$ via the marginal of $P(\mathbf{r}_\kappa, \hat{w}_{ij}, w_{ij}|\mathbf{H}) = P(\mathbf{r}_\kappa|w_{ij}, \mathbf{H})P(w_{ij}, \hat{w}_{ij}|\mathbf{H})$. Since w_{ij} and $\hat{w}_{ij}$ are sums over many independent terms of order $O(\frac{1}{\sqrt{K}})$, they are asymptotically joint Gaussian by the CLT. That is,

$$(w_{ij}, \hat{w}_{ij}) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_{w_{ij},\hat{w}_{ij}}). \tag{B.9}$$

Their means are zero because $\{h_{ij}\}$ has zero mean. $\mathbf{V}_{w_{ij},\hat{w}_{ij}}$ is the covariance matrix between w_{ij} and $\hat{w}_{ij}$ with the entries

$$\begin{aligned}\mathbb{E}_{w_{ij}|\mathbf{H}}\left\{w_{ij}^2|\mathbf{H}\right\} &= \mathbb{E}_{x_k|\mathbf{H}}\left\{\left(\sum_{k\neq j}h_{ik}x_k\right)^2\middle|\mathbf{H}\right\}\\ &= \sum_{k\neq j}h_{ik}^2\mathbb{E}_{x_k|\mathbf{H}}\left\{x_k^2|\mathbf{H}\right\}\\ &\to \int x^2\widetilde{P}_{\mathsf{in}}(x)\mathrm{d}x \triangleq \upsilon_x, \end{aligned}\tag{B.10a}$$

$$\begin{aligned}\mathbb{E}_{\hat{w}_{ij}|\mathbf{H}}\left\{\hat{w}_{ij}^2|\mathbf{H}\right\} &= \mathbb{E}_{\hat{x}_k|\mathbf{H}}\left\{\left(\sum_{k\neq j}h_{ik}\hat{x}_k\right)^2\middle|\mathbf{H}\right\}\\ &= \sum_{k\neq j}h_{ik}^2\mathbb{E}_{\hat{x}_k|\mathbf{H}}\left\{\hat{x}_k^2|\mathbf{H}\right\}\\ &\to \left(\int\tilde{x}\mathscr{P}(\tilde{x})\mathrm{d}\tilde{x}\right)^2 \triangleq \upsilon_{\hat{x}}, \end{aligned}\tag{B.10b}$$

$$\begin{aligned}\mathbb{E}_{w_{ij},\hat{w}_{ij}|\mathbf{H}}\left\{w_{ij}\hat{w}_{ij}|\mathbf{H}\right\} &= \sum_{k\neq j}h_{ik}^2\mathbb{E}_{x_k,\hat{x}_k|\mathbf{H}}\left\{x_k\hat{x}_k\middle|\mathbf{H}\right\}\\ &\to \mathbb{E}_{x|\mathbf{H}}\left\{x\left(\int\tilde{x}\mathscr{P}(\tilde{x})\mathrm{d}\tilde{x}\right)\right\} \triangleq \upsilon_{x\hat{x}}. \end{aligned}\tag{B.10c}$$

Here and hereafter, we denote $\mathscr{P}(\tilde{x}) = \mathscr{P}_{\tilde{x}}(\tilde{x}; s, v^s)$ for notation simplicity. Altogether, these notations provide the bivariate Gaussian distribution

$$\begin{aligned}&P_{w_{ij},\hat{w}_{ij}}(w_{ij}, \hat{w}_{ij})\\ &\quad= \mathcal{N}(\hat{w}_{ij}; 0, \upsilon_{\hat{x}})\mathcal{N}\left(w_{ij}; \frac{\upsilon_{x\hat{x}}}{\upsilon_{\hat{x}}}\hat{w}_{ij}, \upsilon_x - \frac{\upsilon_{x\hat{x}}^2}{\upsilon_{\hat{x}}}\right).\end{aligned}\tag{B.11}$$

Now, we are ready to calculate the joint distribution $P(\mathbf{r}_\kappa, \hat{w}_{ij}|\mathbf{H})$. For simplicity, let $P(\mathbf{r}_\kappa, \hat{w}_{ij}) = P(\mathbf{r}_\kappa, \hat{w}_{ij}|\mathbf{H})$. Using (B.11), we can further calculate $P(\mathbf{r}_\kappa, \hat{w}_{ij})$ as

$$\begin{aligned}&P(\mathbf{r}_\kappa, \hat{w}_{ij})\\ &= P(\mathbf{r}_{\kappa\setminus i})\int\mathrm{d}w_{ij}\,P(r_i, \hat{w}_{ij}, w_{ij})\\ &= P(\mathbf{r}_{\kappa\setminus i})\int\mathrm{d}w_{ij}\,P_{\mathsf{out}}(r_i|w_{ij} + h_{ij}x_j)P(w_{ij}, \hat{w}_{ij})\\ &= P(\mathbf{r}_{\kappa\setminus i})\mathcal{N}(\hat{w}_{ij}; 0, \upsilon_{\hat{x}})\int\mathrm{d}w_{ij}\mathcal{N}\left(w_{ij}; \frac{\upsilon_{x\hat{x}}}{\upsilon_{\hat{x}}}\hat{w}_{ij}, \upsilon_x - \frac{\upsilon_{x\hat{x}}^2}{\upsilon_{\hat{x}}}\right)\\ &\quad\times\int_{r_i^{\mathrm{low}}}^{r_i^{\mathrm{up}}}\mathrm{d}y\mathcal{N}(y; w_{ij} + h_{ij}x_j, \sigma_n^2)\\ &\to P(\mathbf{r}_{\kappa\setminus i})\mathcal{N}(\hat{w}_i; 0, \upsilon_{\hat{x}})\Psi\left(\frac{\upsilon_{x\hat{x}}}{\upsilon_{\hat{x}}}\hat{w}_i + h_{ij}x_j\middle|r_i\right),\end{aligned}\tag{B.12}$$

where $\mathbf{r}_{\kappa\setminus i}$ is the vector containing the elements of $\mathbf{r}_\kappa$ excluding r_i, and we define

$$\Psi(x|r_i) = \Phi\left(\frac{r_i^{\mathrm{up}} - x}{\sqrt{\sigma_n^2 + \upsilon_x - \frac{\upsilon_{x\hat{x}}^2}{\upsilon_{\hat{x}}}}}\right) - \Phi\left(\frac{r_i^{\mathrm{low}} - x}{\sqrt{\sigma_n^2 + \upsilon_x - \frac{\upsilon_{x\hat{x}}^2}{\upsilon_{\hat{x}}}}}\right). \tag{B.13}$$

Using the joint distribution (B.12), we can compute the asymptotic behavior of (B.8) as

$$\begin{aligned}\mathsf{A}_\kappa &= \sum_{l\in\Omega_\kappa, l\neq i} \mathbb{E}_{\mathbf{r}_\kappa,\hat{w}_{lj}}\left\{\alpha_{lj}^2|\mathbf{H}\right\} \\ &= \sum_{l\in\Omega_\kappa, l\neq i} \int\int \mathrm{d}\hat{w}_{lj}\mathrm{d}\mathbf{r}_\kappa P(\mathbf{r}_\kappa,\hat{w}_{lj})\left(\frac{\Xi'(\hat{w}_{lj}|r_l)}{\Xi(\hat{w}_{lj}|r_l)}h_{lj}\right)^2 \\ &= \sum_{l\in\Omega_\kappa, l\neq i}\sum_{r_l\in\mathcal{R}_\kappa}\int \mathrm{d}\hat{w}_{lj}\mathcal{N}(\hat{w}_{lj};0,v_{\hat{x}}) \\ &\quad\times\Psi\left(\frac{v_{x\hat{x}}}{v_{\hat{x}}}\hat{w}_{lj}+h_{lj}x_j\Big|r_l\right)\left(\frac{\Xi'(\hat{w}_{lj}|r_l)}{\Xi(\hat{w}_{lj}|r_l)}h_{lj}\right)^2 \\ &\to \lambda_\kappa\sum_{r\in\mathcal{R}_\kappa}\int \mathrm{d}\hat{w}\mathcal{N}(\hat{w};0,v_{\hat{x}})\Psi\left(\frac{v_{x\hat{x}}}{v_{\hat{x}}}\hat{w}\Big|r\right)\left(\frac{\Xi'(\hat{w}|r)}{\Xi(\hat{w}|r)}\right)^2,\end{aligned} \tag{B.14}$$

and, similarly,

$$\mathsf{B}_\kappa \to \lambda_\kappa\sum_{r\in\mathcal{R}_\kappa}\int \mathrm{d}\hat{w}\mathcal{N}(\hat{w};0,v_{\hat{x}})\Psi\left(\frac{v_{x\hat{x}}}{v_{\hat{x}}}\hat{w}\Big|r\right)\frac{\Xi''(\hat{w}|r)}{\Xi(\hat{w}|r)}. \tag{B.15}$$

Note that A_κ and B_κ are asymptotically independent of the indexes i, j in the large system limit. Therefore $\Xi(\hat{w}_{ij}|r_i)$ defined in (B.2) becomes independent of the indexes i, j, and is read as

$$\Xi(\hat{w}|r) = \int \mathrm{d}w\widetilde{P}_{\mathsf{out}}(r|w)\mathcal{N}(w;\hat{w},\hat{v}^w), \tag{B.16}$$

where $\hat{w}$ and $\hat{v}^w$ are the asymptotic limits of $\hat{w}_i$ and $\hat{v}_i^w$, respectively, and $\hat{v}^w$ can be expressed as

$$\begin{aligned}\hat{v}^w &= \mathbb{E}\left\{\sum_{k\neq j}h_{ik}^2\hat{v}_k^x\Big|\mathbf{H}\right\} \\ &\to \int\tilde{x}^2\mathscr{P}(\tilde{x})\mathrm{d}\tilde{x}-\left(\int\tilde{x}\mathscr{P}(\tilde{x})\mathrm{d}\tilde{x}\right)^2 \\ &\triangleq c_{\hat{x}}-v_{\hat{x}}.\end{aligned} \tag{B.17}$$

Together with these results, (B.14) and (B.15) can be further simplified as

$$\mathsf{A}_\kappa = \lambda_\kappa\sum_{r\in\mathcal{R}_\kappa}\int \mathrm{D}u\Psi\left(\sqrt{\frac{v_{x\hat{x}}^2}{v_{\hat{x}}}}u\Big|r\right)\left(\frac{\Xi'(\sqrt{v_{\hat{x}}}u|r)}{\Xi(\sqrt{v_{\hat{x}}}u|r)}\right)^2, \tag{B.18a}$$

$$\mathsf{B}_\kappa = \lambda_\kappa\sum_{r\in\mathcal{R}_\kappa}\int \mathrm{D}u\Psi\left(\sqrt{\frac{v_{x\hat{x}}^2}{v_{\hat{x}}}}u\Big|r\right)\frac{\Xi''(\sqrt{v_{\hat{x}}}u|r)}{\Xi(\sqrt{v_{\hat{x}}}u|r)}. \tag{B.18b}$$

Note that A_κ and B_κ are asymptotic of index j, thus (B.7) can be rewritten as

$$A_j \to \sum_\kappa \mathsf{A}_\kappa = A, \quad B_j \to \sum_\kappa \mathsf{B}_\kappa = B. \tag{B.19}$$

Following similar arguments as those of A and B, we show that a_j defined in (B.6) follows a Gaussian distribution conditioned on $x_j \sim P_{\mathsf{in}}(x)$ with asymptotic mean $C_j\cdot x_j$ and variance A_j, where

$$C_j \to \sum_\kappa \mathsf{C}_\kappa = C, \tag{B.20}$$

$$\mathsf{C}_\kappa = \lambda_\kappa\sum_{r\in\mathcal{R}_\kappa}\int \mathrm{D}u\Psi'\left(\sqrt{\frac{v_{x\hat{x}}^2}{v_{\hat{x}}}}u\Big|r\right)\frac{\Xi'(\sqrt{v_{\hat{x}}}u|r)}{\Xi(\sqrt{v_{\hat{x}}}u|r)}. \tag{B.21}$$

$\Psi'(u|r)$ represents the derivatives of Ψ w.r.t. u. Again, we drop the index j in the large system limit.

Therefore, the input messages $\Delta_{i\leftarrow j}(\tilde{x}_j)$ in (B.4b) can be represented by

$$\begin{aligned}\Delta_{i\leftarrow j}(\tilde{x}_j) &\propto \widetilde{P}_{\mathsf{in}}(\tilde{x}_j)\exp\left[(Cx_j+\sqrt{A}z_j)\tilde{x}_j-\frac{1}{2}(A-B)\tilde{x}_j^2\right] \\ &\propto \widetilde{P}_{\mathsf{in}}(\tilde{x}_j)\exp\left[-\frac{\left(\tilde{x}_j-\frac{Cx_j+\sqrt{A}z_j}{E}\right)^2}{2\frac{1}{E}}\right],\end{aligned} \tag{B.22}$$

where $E = A - B$, $z_j \sim \mathcal{N}(0,1)$, and $x_j \sim P_{\mathsf{in}}(x)$. Note that we drop the index j from A_j, B_j, C_j in (B.22) because they are asymptotically independent of this index.

Comparing the structure of $\Delta_{i\leftarrow j}(\tilde{x}_j)$ in (B.22) with the marginal posterior (D4), we can rewrite the posterior as

$$\mathscr{P}_{\tilde{x}}(\tilde{x};x,z) = \frac{\widetilde{P}_{\mathsf{in}}(\tilde{x})\mathcal{N}\left(\tilde{x};\frac{C\cdot x+\sqrt{A}z}{E},\frac{1}{E}\right)}{\int\widetilde{P}_{\mathsf{in}}(\tilde{x})\mathcal{N}\left(\tilde{x};\frac{C\cdot x+\sqrt{A}z}{E},\frac{1}{E}\right)\mathrm{d}\tilde{x}}, \tag{B.23}$$

where $z \sim \mathcal{N}(0,1)$ and $x \sim P_{\mathsf{in}}(x)$. Again, we have used the asymptotic properties to drop the indexes i and j. Moreover, the equivalent channel can be denoted as

$$s = \frac{C}{E}x+\frac{\sqrt{A}}{E}z. \tag{B.24}$$

Substituting (B.23) into (B.10) and (B.17), we can obtain the parameters v_x, $c_{\hat{x}}$, $v_{\hat{x}}$, and $v_{x\hat{x}}$, which are

$$v_x = \mathbb{E}_x\left\{x^2\right\}, \tag{B.25a}$$

$$c_{\hat{x}} = \mathbb{E}_{x,z}\left\{\int\tilde{x}^2\mathscr{P}(\tilde{x};x,z)\mathrm{d}\tilde{x}\right\}, \tag{B.25b}$$

$$v_{\hat{x}} = \mathbb{E}_{x,z}\left\{\left(\int\tilde{x}\mathscr{P}(\tilde{x};x,z)\mathrm{d}\tilde{x}\right)^2\right\}, \tag{B.25c}$$

$$v_{x\hat{x}} = \mathbb{E}_{x,z}\left\{x\int\tilde{x}\mathscr{P}(\tilde{x};x,z)\mathrm{d}\tilde{x}\right\}. \tag{B.25d}$$

We conclude that the asymptotic behavior GAMP algorithm can be fully described by parameters A, B, C given in (B.19) and (B.20), which corresponds to the output updates in conjunction with parameters $v_x, c_{\hat{x}}, v_{\hat{x}}, v_{x\hat{x}}$ given in (B.25), which account for the input updates.

REFERENCES

[1] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] C.-X. Wang *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[3] L. Sanguinetti, A. Moustakas, and M. Debbah, "Interference management in 5G reverse TDD HetNets with wireless backhaul: A large system analysis," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1187–1200, Jun. 2015.

[4] 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.

[5] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[6] J. Zhang, C.-K. Wen, S. Jin, X. Gao, and K.-K. Wong, "On capacity of large-scale MIMO multiple access channels with distributed sets of correlated antennas," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 133–148, Feb. 2013.

[7] H.-S. Lee and C. G. Sodini, "Analog-to-digital converters: Digitizing the analog world," *Proc. IEEE*, vol. 96, no. 2, pp. 323–334, Feb. 2008.

[8] J. Liu and H. Minn. (Jun. 2015). *The Death of 5G Part 2: Will Analog be the Death of Massive MIMO?* http://www.comsoc.org/ctn

[9] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.

[10] X. Zhang, M. Matthaiou, M. Coldrey, and E. Björnson, "Impact of residual transmit RF impairments on training-based MIMO systems," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2899–2911, Aug. 2015.

[11] A. Mezghani and J. A. Nossek, "Analysis of Rayleigh-fading channels with 1-bit quantized output," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Toronto, ON, Canada, Jul. 2008, pp. 260–264.

[12] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Trans. Commun.*, vol. 57, no. 12, pp. 3629–3639, Dec. 2009.

[13] Z. Wang, H. Yin, W. Zhang, and G. Wei, "Monobit digital receivers for QPSK: Design, performance and impact of IQ imbalances," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3292–3303, Aug. 2013.

[14] W. Zhang, "A general framework for transmission with transceiver distortion and some applications," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 384–399, Feb. 2012.

[15] A. Mezghani and J. A. Nossek, "On ultra-wideband MIMO systems with 1-bit quantized outputs: Performance analysis and input optimization," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Nice, France, Jun. 2007, pp. 1286–1289.

[16] J. Mo and R. W. Heath, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498–5512, Oct. 2015.

[17] C. Risi, D. Persson, and E. G. Larsson. (Apr. 2014). "Massive MIMO with 1-bit ADC." [Online]. Available: https://arxiv.org/abs/1404.7736

[18] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-bit massive MIMO: Channel estimation and high-order modulations," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 1304–1309.

[19] S. Wang, Y. Li, and J. Wang, "Multiuser Detection in Massive Spatial Modulation MIMO With Low-Resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2156–2168, Apr. 2015.

[20] S. Wang, Y. Li, and J. Wang, "Convex optimization based multi-user detection for uplink large-scale MIMO under low-resolution quantization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 4789–4794.

[21] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[22] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "One-bit ADCs in wideband massive MIMO systems with OFDM transmission," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 3386–3390.

[23] N. Liang and W. Zhang. (Jan. 2016). "Mixed-ADC massive MIMO uplink in frequency-selective channels." [Online]. Available: http://arxiv.org/abs/1601.02082

[24] A. Wadhwa and U. Madhow, "Blind phase/frequency synchronization with low-precision ADC: A Bayesian approach,"in *Proc. 51st Allerton Conf. Commun. Control Comput.(Allerton)*, IL, USA, Oct. 2013, pp. 181–188.

[25] A. Wadhwa and U. Madhow, "Near-coherent QPSK performance with coarse phase quantization: A feedback-based architecture for joint phase/frequency synchronization and demodulation," *IEEE Trans. Signal Process.*, vol. 64, no. 17, pp. 4432–4443, Sep. 2016.

[26] A. Tarighat and A. H. Sayed, "MIMO OFDM receivers for systems with IQ imbalances," *IEEE Trans. Signal Process.*, vol. 53, no. 9, pp. 3583–3596, Sep. 2005.

[27] Y.-J. Chiu and S.-P. Hung, "Estimation scheme of the receiver IQ imbalance under carrier frequency offset in communication," *IET Commun.*, vol. 4, no. 11, pp. 1381–1388, Jul. 2010.

[28] Z. Wang, H. Yin, W. Zhang, and G. Wei, "Monobit digital receivers for QPSK: Design, performance and impact of IQ imbalances," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3292–3303, Aug. 2013.

[29] 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.

[30] N. Liang and W. Zhang, "Mixed-ADC massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 983–997, Apr. 2016.

[31] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Robust predictive quantization: Analysis and design via convex optimization," *IEEE J. Sel. Topics Signal Process. (STSP)*, vol. 1, no. 4, pp. 618–632, Dec. 2007.

[32] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Petersburg, Russia, Jul./Aug. 2011, pp. 2168–2172.

[33] R. Méndez-Rial, C. Rusu, A. Alkhateeb, N. González-Prelcic, and R. W. Heath, " Channel estimation and hybrid combining for mmWave: Phase shifters or switches?" in *Proc. Inf. Theory Appl. Workshop*, Feb. 2015, pp. 90–97.

[34] R. L. Schmid, P. Song, C. T. Coen, A. C. Ulusoy, and J. D. Cressler, "On the analysis and design of low-loss single-pole double-throw W-band switches utilizing saturated SiGe HBTs," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 11, pp. 2755–2767, Nov. 2014.

[35] M. D. Renzo, H. Haas, A. Ghrayeb, S. Sugiura, and L. Hanzo, "Spatial modulation for generalized MIMO: Challenges, opportunities, and implementation," *Proc. IEEE*, vol. 102, no. 1, pp. 56–103, Jan. 2014.

[36] P. Yang *et al.*, "Single-carrier spatial modulation: A promising design for large-scale broadband antenna systems," *IEEE Commun. Surv. Tut.*, vol. 18, no. 3, pp. 1687–1716, 3rd. Quart., 2016.

[37] A. Garcia-Rodriguez, C. Masouros, and P. Rulikowski. (May 2016). "Reduced switching connectivity for power-efficient large scale antenna selection." [Online]. Available: https://arxiv.org/abs/1605.01549

[38] D. Guo and S. Verdu, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 1983–2010, Jun. 2005.

[39] D. Guo and C.-C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 421–431, Apr. 2008.

[40] F. Krzakala and M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *J. Statist. Mech.*, vol. 2012, no. 8, pp. P08009, Aug. 2012.

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

[42] K. Cho and D. Yoon, "On the general BER expression of one- and two-dimensional amplitude modulations," *IEEE Trans. Commun.*, vol. 50, no. 7, pp. 1074–1080, Jul. 2002.

[43] K. Nakamura and T. Tanaka, "Performance analysis of signal detection using quantized received signals of linear vector channel," in *Proc. Int. Symp. Inf. Theory Its Appl. (ISITA)*, Auckland, New Zealand, Dec. 2008, pp. 1–5.

[44] N. Al-Dhahir and J. M. Cioffi, "On the uniform ADC bit precision and clip level computation for a Gaussian signal," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 434–438, Feb. 1996.

[45] A. Mezghani and J. A. Nossek, "Belief propagation based MIMO detection operating on quantized channel output," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2010, pp. 2113–2117.

Ti-Cao Zhang, photograph and biography not available at the time of publication.

Chao-Kai Wen, photograph and biography not available at the time of publication.

Shi Jin, photograph and biography not available at the time of publication.

Tao Jiang, photograph and biography not available at the time of publication.