# Bayes-Optimal Joint Channel-and-Data Estimation for Massive MIMO with Low-Precision ADCs

Chao-Kai Wen, Chang-Jen Wang, Shi Jin, Kai-Kit Wong, and Pangan Ting

***Abstract*—This paper considers a multiple-input multiple-output (MIMO) receiver with very low precision analog-to-digital convertors (ADCs), motivated by the interest of massive MIMO antenna systems operating with low cost and power requirements. In this case, prior work demonstrated that the training duration is required to be *very large* just to obtain an acceptable channel state information (CSI). To tackle this, we adopt joint channel-and-data (JCD) estimation based on the Bayes-optimal inference which results in the minimal mean-square-error (MSE) with respect to (w.r.t.) the channels as well as the payload data. We realize the Bayes-optimal JCD estimator using a recent technique based on approximate message passing and present an analytical framework to study its theoretical performances in the large-system limit. Simulation results confirm our analytical results, which allow efficient evaluation of the performance for the quantized massive MIMO systems and provide insights to system design.**

***Keywords*—Bayes-optimal inference, joint channel-and-data estimation, low precision ADC, massive MIMO, replica method.**

## I. Introduction

The fifth-generation mobile communications, widely known as the 5G, is anticipated to obtain $1,000$-fold gains in capacity, 10-fold increase in spectral and energy efficiencies, and also 25-fold gains in average cell throughput [1]. The large-scale multiple-input multiple-output (MIMO) antenna systems, a.k.a. "massive MIMO" are being considered as a key enabler for delivering these promises, e.g., [1–4]. Such systems employ numerous number of antennas at the base station (BS) (e.g., hundreds or thousands) to serve multiple user terminals (tens or hundreds) in the same time-frequency resource. As such, the array gain is expected to grow unboundedly with the number of antennas at the BSs so that the radiated energy-efficiency shall increase dramatically and multiuser interference shall be eliminated completely.

The high dimensionality however considerably increases the hardware cost and power consumption. In particular, the hardware complexity and power consumption of an analog-to-digital converter (ADC) increase exponentially in the number of bits per sample [5], and will be a major obstacle. This has motivated the use of low-cost low-precision ADCs (e.g., 1-3 bits) at the antennas, resulting in the *quantized* MIMO systems.[1] With such coarse quantization, all communication theories as well as signal processing techniques dedicating to high-resolution quantization fail [6–9]. Some aspects of the quantized MIMO systems have been studied in the literature covering capacity analysis [10–12], energy efficiency analysis [13, 14], feedback codebook design [15], data detection [16–24], and channel estimation [18, 20, 23–26].

This paper's focus is on data detection and channel estimation for the quantized MIMO systems. Previous work in this regard mostly assumed perfect channel state information (CSI) at the receiver (CSIR) or considered problems merely to channel estimation. The use of coarse quantization greatly reduces the number of *effective* measurements, and therefore the acquisition of CSIR becomes more challenging in the quantized MIMO systems than in its unquantized counterpart. In [18], it was shown that a one-bit quantized MIMO system requires an extremely long training sequence (e.g., approximately $50$ times the number of users) to achieve the same performance as in the full CSI case. The requirement of long training sequence motivates us to consider joint channel-and-data (JCD) estimation in which the estimated payload data are utilized to aid channel estimation. A major advantage of JCD estimation is that relatively few pilot symbols are required to achieve the equivalent channel and data estimation performances [27, 28].

Though performance enhancement by using the JCD technique is expected, its performance in *quantized* MIMO systems is not understood.[2] The most related work appears to be [20] where the achievable throughput was investigated in the one-bit quantized single-input single-output (SISO) channel using JCD estimation (i.e., least-squares channel estimation jointly on pilot and data symbols). For the one-bit quantized MIMO system, [20] just considered a pilot-only scheme with least-squares channel estimation followed by data detection utilizing maximal-ratio combining. Although it was found that high-order constellation such as 16-QAM can also be supported by the one-bit quantized MIMO system, which outperforms the ones reported in [18] for QPSK, the problem of requiring long training sequence remains. Hence, there is strong desire to study the fundamental performance limits on quantized MIMO systems imposed by the JCD estimation.

In this paper, we present a framework to analyze the achievable performance of the quantized MIMO system with

C.-K. Wen and C.-J. Wang are with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan (e-mail: chaokai.wen@mail.nsysu.edu.tw). S. Jin is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, P. R. China. K. Wong is with the Department of Electronic and Electrical Engineering, University College London, United Kingdom. P. Ting is with the Industrial Technology Research Institute, Hsinchu, Taiwan.

[1] In practice, an ADC, typically with a precision of 8-12 bits, is used in modern communication systems so that the received signals can be processed in the digital domain. In this paper, the "quantized" MIMO system specifically represents a MIMO system equipped with very low-precision ADCs (e.g., 1-3 bits).

[2] In the context of *unquantized* MIMO system, several aspects of the JCD estimation have already been widely studied, see e.g., [27, 28].

JCD estimation. Unlike other JCD estimation schemes based on suboptimal criteria [20, 27, 28], we use the Bayes-optimal inference for JCD estimation because this approach gives the minimum mean-square-error (MMSE) with respect to (w.r.t.) the channels and data symbols. In the conference version of this work [29], our simulation results have demonstrated that the Bayes-optimal JCD estimator provides a significant gain over the pilot-only schemes in the quantized MIMO system. Besides the derivations that were omitted in [29], the main contributions of this paper are summarized as follows:

- To implement the Bayes-optimal JCD estimator, a variant of belief propagation (BP) to approximate the marginal distributions of each data and channel components is used. We modify the bilinear generalized approximate message passing (BiG-AMP) algorithm in [30] and adapt it to the quantized MIMO system by providing the corresponding closed-form expressions for the nonlinear steps. We refer to this scheme as the GAMP-based JCD algorithm.[3]
- By large-system analysis based on the replica method from statistical physics, we show the *decoupling principle* for the Bayes-optimal JCD estimator. That is, in the large-system regime, the input-output relationship of the quantized MIMO system using the Bayes-optimal JCD estimator is decoupled into a bank of scalar additive white Gaussian noise (AWGN) channels w.r.t. the data symbols and the channel response, respectively. This allows the characterization of several system performances of interest in an intuitive way. In particular, the average symbol error rate (SER) w.r.t. the data symbols as well as the average MSE w.r.t. the channel estimate for the Bayes-optimal JCD estimator are determined.
- Finally, computer simulations are provided to verify the efficiency of the proposed GAMP-based JCD algorithm and the accuracy of our analysis. The high accuracy of our results allows a quick and efficient way to evaluate the performances of the quantized MIMO system. Several useful observations on aiding the system design are made from the analysis.

*Notations*—Throughout, for any matrix $\mathbf{A}$, $A_{ij}$ refers to the $(i,j)$th entry of $\mathbf{A}$, $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$, $\mathbf{A}^H$ is the conjugate transpose of $\mathbf{A}$, and $\mathsf{tr}(\mathbf{A})$ denotes its trace. Also, $\mathbf{I}$ denotes the identity matrix, $\mathbf{0}$ is the zero matrix, $\|\cdot\|_{\mathsf{F}}$ denotes the Frobenius norm, $\mathsf{E}[\cdot]$ represents the expectation operator, $\log(\cdot)$ is the natural logarithm, and $\mathrm{sign}(\cdot)$ is the signum function. In addition, a random vector $\mathbf{z}$ drawn from the proper complex Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Omega}$ is described by the probability density function:

$$\mathcal{N}_{\mathbb{C}}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Omega}) = \frac{1}{\det(\pi\boldsymbol{\Omega})} e^{-(\mathbf{z}-\boldsymbol{\mu})^H\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\mu})},$$

where $\det(\cdot)$ returns the determinant. We write $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Omega})$. With $\mathrm{D}\mathbf{z}$ denoting the real (or complex) Gaussian integration measure, for an $n\times 1$ real valued vector $\mathbf{z}$, we have

$$\mathrm{D}\mathbf{z} = \prod_{i=1}^{n} \phi(z_i)\,\mathrm{d}z_i \quad \text{with} \quad \phi(z_i) = \frac{e^{-\frac{z_i^2}{2}}}{\sqrt{2\pi}};$$

or $\mathrm{D}\mathbf{z} = \prod_{i=1}^{n} \frac{e^{-(\mathrm{Re}(z_i))^2-(\mathrm{Im}(z_i))^2}}{\pi}\mathrm{dRe}(z_i)\mathrm{dIm}(z_i)$ for the complex valued vector, where $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ extracts the real and imaginary components, respectively. Finally,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,\mathrm{d}t.$$

denotes the cumulative Gaussian distribution function [31].

[3]In this paper, the Bayes-optimal JCD estimator is regarded as the *theoretical* optimal estimator, while the GAMP-based JCD algorithm can be thought of as a *practical method* to approximate the theoretical optimal estimator.

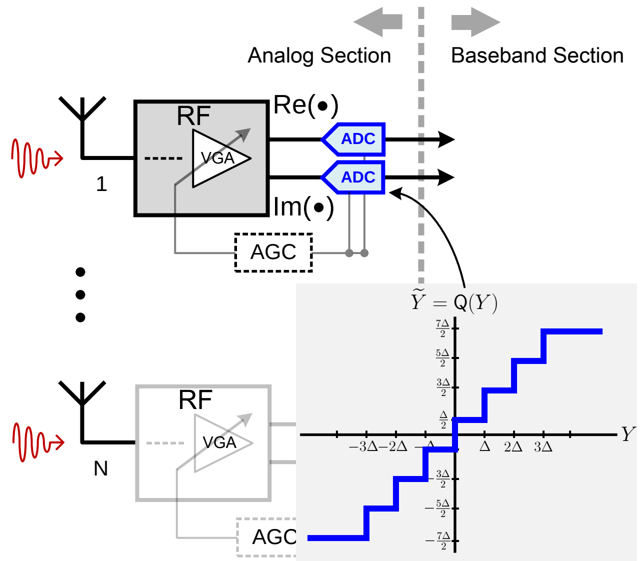


Fig. 1. The quantized MIMO antenna system.

## II. System Model

We consider a MIMO uplink system where one BS equipped with $N$ receive antennas serves $K$ single-antenna users. The channel is assumed to be flat block fading, wherein the channel remains constant over $T$ consecutive symbol intervals (i.e., a block). The received signal $\mathbf{Y} = [Y_{nt}] \in \mathbb{C}^{N\times T}$ over the block interval can be written in matrix form as

$$\mathbf{Y} = \frac{1}{\sqrt{K}}\mathbf{H}\mathbf{X} + \mathbf{W} = \mathbf{Z} + \mathbf{W}, \tag{1}$$

where $\mathbf{X} = [X_{kn}] \in \mathbb{C}^{K\times T}$ denotes the transmit symbols in the block, $\mathbf{H} = [H_{nk}] \in \mathbb{C}^{N\times K}$ denotes the channel matrix containing the fading coefficients between the transmit antennas and the receive antennas, $\mathbf{W} = [W_{nt}] \in \mathbb{C}^{N\times T}$ represents the additive temporally and spatially white Gaussian noise with zero mean and element-wise variance $\sigma_w^2$, and we define $\mathbf{Z} = [Z_{nt}] = \frac{1}{\sqrt{K}}\mathbf{H}\mathbf{X} \in \mathbb{C}^{N\times T}$.

On the receiver side, as illustrated in Figure 1, each received signal is down-converted into analog baseband $Y_{nt}$ and then discretized using a *complex-valued* quantizer $\mathsf{Q}_c$. Each complex-valued quantizer $\mathsf{Q}_c(\cdot)$ is defined as $\widetilde{Y}_{nt} = \mathsf{Q}_c(Y_{nt}) \triangleq \mathsf{Q}(\mathrm{Re}\{Y_{nt}\}) + \mathrm{j}\mathsf{Q}(\mathrm{Im}\{Y_{nt}\})$, i.e., the real and imaginary parts are quantized separately. In practice, a variable gain amplifier

(VGA) with an automatic gain control (AGC) is used prior to the quantization to ensure the analog baseband within a proper range, e.g., $(-1,+1)$. It is assumed that $Y_{nt}$ has included the AGC gain and thus is in a proper range. The resulting quantized signal $\widetilde{\mathbf{Y}} = [\widetilde{Y}_{nt}] \in \mathbb{C}^{N\times T}$ is therefore given by

$$\widetilde{\mathbf{Y}} = \mathsf{Q}_c(\mathbf{Y}) = \mathsf{Q}_c(\mathbf{Z}+\mathbf{W}), \tag{2}$$

where the quantization is applied element-wise.

Specifically, each complex-valued quantizer $\mathsf{Q}_c$ consists of two real-valued B-bit quantizers $\mathsf{Q}$. Each real-valued quantizer maps a real-valued input to one of the $2^{\mathsf{B}}$ bins, which are characterized by the set of $2^{\mathsf{B}}-1$ thresholds $[r_1, r_2, \ldots, r_{2^{\mathsf{B}}-1}]$, such that $-\infty < r_1 < r_2 < \cdots < r_{2^{\mathsf{B}}-1} < \infty$. For notational convenience, we define $r_0 = -\infty$ and $r_{2^{\mathsf{B}}} = \infty$. The output is assigned a value in $(r_{b-1}, r_b]$ when the quantizer input falls in the interval $(r_{b-1}, r_b]$ (namely, the $b$-th bin). For example, the threshold of a typical uniform quantizer with the quantization step-size $\Delta$ is given by

$$r_b = \left(-2^{\mathsf{B}-1}+b\right)\Delta, \quad \text{for} \quad b=1,\ldots,2^{\mathsf{B}}-1, \tag{3}$$

and the quantization output is assigned the value $r_b - \frac{\Delta}{2}$ when the input falls in the $b$-th bin.[4] Figure 1 shows an example of the 3-bit uniform quantizer. Notice that in practice, we can adjust the VGA gain to attain the desired step-size $\Delta$.

Since the channel matrix $\mathbf{H}$ needs to be estimated at the receiver, we make the first $T_{\mathsf{t}}$ symbols of the block of $T$ symbols serve as pilot sequences. The remaining $T_{\mathsf{d}} = T - T_{\mathsf{t}}$ symbols are used for data transmissions. The training and data phases are referred to as t-phase and d-phase, respectively. This setting is equivalent to partitioning $\mathbf{X}$ and $\widetilde{\mathbf{Y}}$ as

$$\mathbf{X} = \left[\mathbf{X}_{\mathsf{t}}\ \mathbf{X}_{\mathsf{d}}\right], \text{ with } \mathbf{X}_{\mathsf{t}} \in \mathbb{C}^{K\times T_{\mathsf{t}}},\ \mathbf{X}_{\mathsf{d}} \in \mathbb{C}^{K\times T_{\mathsf{d}}}, \tag{4a}$$

$$\widetilde{\mathbf{Y}} = \left[\widetilde{\mathbf{Y}}_{\mathsf{t}}\ \widetilde{\mathbf{Y}}_{\mathsf{d}}\right], \text{ with } \widetilde{\mathbf{Y}}_{\mathsf{t}} \in \mathbb{C}^{N\times T_{\mathsf{t}}},\ \widetilde{\mathbf{Y}}_{\mathsf{d}} \in \mathbb{C}^{N\times T_{\mathsf{d}}}. \tag{4b}$$

We assume that $\mathbf{X}_{\mathsf{t}}$ (or $\mathbf{X}_{\mathsf{d}}$) is composed of independent and identically distributed (i.i.d.) random variables $\mathsf{X}_{\mathsf{t}}$ ($\mathsf{X}_{\mathsf{d}}$) drawn from a known probability distribution $\mathsf{P}_{\mathsf{X}_{\mathsf{t}}}$ (or $\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}$), i.e.,

$$\mathsf{P}_{\mathsf{X}}(\mathbf{X}) = \underbrace{\left(\prod_{k=1}^{K}\prod_{t=1}^{T_{\mathsf{t}}} \mathsf{P}_{\mathsf{X}_{\mathsf{t}}}(X_{\mathsf{t},kt})\right)}_{=\mathsf{P}_{\mathsf{X}_{\mathsf{t}}}(\mathbf{X}_{\mathsf{t}})} \underbrace{\left(\prod_{k=1}^{K}\prod_{t=1}^{T_{\mathsf{d}}} \mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(X_{\mathsf{d},kt})\right)}_{=\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(\mathbf{X}_{\mathsf{d}})}. \tag{5}$$

Since the pilot and data symbols should appear on constellation points uniformly, the ensemble averages of $\{X_{\mathsf{t},kt}\}$ and $\{X_{\mathsf{d},kt}\}$ are assumed to be zero. In addition, we let $\sigma_{x_{\mathsf{t}}}^2$ and $\sigma_{x_{\mathsf{d}}}^2$ be the transmit powers during the t-phase and d-phase, respectively, i.e., $\mathsf{E}\{|\mathsf{X}_{\mathsf{t},kt}|^2\} = \sigma_{x_{\mathsf{t}}}^2$ and $\mathsf{E}\{|\mathsf{X}_{\mathsf{d},kt}|^2\} = \sigma_{x_{\mathsf{d}}}^2$. For ease of notation, we refer an entry of $\mathbf{X}$ to $X_{kt}$ instead of $X_{\mathsf{t},kt}$ or $X_{\mathsf{d},kt}$. Therefore, we use $\mathcal{T}_{\mathsf{t}} = \{1,\ldots,T_{\mathsf{t}}\}$ and $\mathcal{T}_{\mathsf{d}} = \{T_{\mathsf{t}}+1,\ldots,T\}$ to denote the sets of symbol indices in t-phase and d-phase, respectively.

Similarly, we assume that each entry $H_{nk}$ is drawn from a complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}(0,\sigma_h^2)$, where $\sigma_h^2$ is the large-scale fading coefficient. Let $\mathsf{P}_{\mathsf{H}}(H_{nk}) \equiv \mathcal{N}_{\mathbb{C}}(0,\sigma_h^2)$. Then

$$\mathsf{P}_{\mathsf{H}}(\mathbf{H}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathsf{P}_{\mathsf{H}}(H_{nk}). \tag{6}$$

To prevent the key features of our results from being obfuscated by complex notations, we consider the case where all the users have the same large-scale fading factor in the main text. A generalized version of our main result is presented in Appendix E, where the users have different large-scale fading factors. This generalization can be easily achieved by plugging user index $k$ into $\sigma_h^2$.

## III. Bayes-Optimal JCD Estimation

We consider the case where the receiver knows the distributions of $\mathbf{H}$ and $\mathbf{X}$ but not their realizations. In contrast to the pilot-only scheme, we consider JCD estimation, where the BS estimates both $\mathbf{H}$ and $\mathbf{X}_{\mathsf{d}}$ from $\widetilde{\mathbf{Y}}$ given $\mathbf{X}_{\mathsf{t}}$. In Section III-A, we will treat the problem under the framework of Bayesian inference, which provides a foundation for achieving the best MSE estimates [32]. Then, we explain explain this inference in a simple SISO system in Section III-B. In Section III-C, we present an algorithm to realize the Bayesian inference by an approximation of the sum-product algorithm.

### A. Theoretical Foundation

We define the likelihood, i.e., the distribution of the received signals under (2) conditional on the unknown parameters, as

$$\mathsf{P}_{\mathsf{out}}(\widetilde{\mathbf{Y}}|\mathbf{H},\mathbf{X}) \triangleq \prod_{n=1}^{N}\prod_{t=1}^{T} \mathsf{P}_{\mathsf{out}}\left(\widetilde{Y}_{nt}\middle|Z_{nt}\right), \tag{7}$$

where

$$\mathsf{P}_{\mathsf{out}}\left(\widetilde{Y}\middle|Z\right) = \left(\frac{1}{\sqrt{\pi\sigma_w^2}}\int_{r_{b-1}}^{r_b} e^{-\frac{(y-\mathrm{Re}(Z))^2}{\sigma_w^2}}\,\mathrm{d}y\right) \times \left(\frac{1}{\sqrt{\pi\sigma_w^2}}\int_{r_{b'-1}}^{r_{b'}} e^{-\frac{(y-\mathrm{Im}(Z))^2}{\sigma_w^2}}\,\mathrm{d}y\right) \tag{8}$$

when $\mathrm{Re}(\widetilde{Y}) \in (r_{b-1}, r_b]$ and $\mathrm{Im}(\widetilde{Y}) \in (r_{b'-1}, r_{b'}]$. Based on the cumulative Gaussian distribution function (see the definition in Notations), (8) becomes

$$\mathsf{P}_{\mathsf{out}}\left(\widetilde{Y}\middle|Z\right) = \Psi_b(\mathrm{Re}(Z))\Psi_{b'}(\mathrm{Im}(Z)), \tag{9}$$

where

$$\Psi_b(x) \triangleq \Phi\left(\frac{\sqrt{2}(r_b - x)}{\sigma_w}\right) - \Phi\left(\frac{\sqrt{2}(r_{b-1} - x)}{\sigma_w}\right). \tag{10}$$

The prior distributions of $\mathbf{X}$ and $\mathbf{H}$ are given by (5) and (6), respectively. Then the posterior probability can be computed according to Bayes' rule as

$$\mathsf{P}(\mathbf{H},\mathbf{X}|\widetilde{\mathbf{Y}}) = \frac{\mathsf{P}_{\mathsf{out}}(\widetilde{\mathbf{Y}}|\mathbf{H},\mathbf{X})\mathsf{P}_{\mathsf{H}}(\mathbf{H})\mathsf{P}_{\mathsf{X}}(\mathbf{X})}{\mathsf{P}(\widetilde{\mathbf{Y}})}, \tag{11}$$

where

$$\mathsf{P}(\widetilde{\mathbf{Y}}) = \int_{\mathbf{H}}\int_{\mathbf{X}} \mathsf{P}(\widetilde{\mathbf{Y}}|\mathbf{H},\mathbf{X})\mathsf{P}_{\mathsf{H}}(\mathbf{H})\mathsf{P}_{\mathsf{X}}(\mathbf{X})\,\mathrm{d}\mathbf{H}\mathrm{d}\mathbf{X} \tag{12}$$

[4] This output assignment is only true for $b=1,\ldots,2^{\mathsf{B}}-1$. If $b=2^{\mathsf{B}}$, the quantization output is assigned the value $\left(2^{\mathsf{B}-1}-2^{-1}\right)\Delta$.

is the marginal likelihood.

Given the posterior probability, an estimator for $H_{nk}$ can be obtained by the posterior mean

$$\widehat{H}_{nk} = \int H_{nk} \mathscr{P}(H_{nk})\, \mathrm{d}H_{nk}, \tag{13}$$

where

$$\mathscr{P}(H_{nk}) = \int_{\mathbf{H}\setminus H_{nk}} \mathrm{d}\mathbf{H} \int_{\mathbf{X}} \mathrm{d}\mathbf{X}\, \mathsf{P}(\mathbf{H}, \mathbf{X}|\widetilde{\mathbf{Y}})$$

denotes the marginal posterior probability of $H_{nk}$. Here, the notation $\int_{\mathbf{H}\setminus H_{nk}} \mathrm{d}\mathbf{H}$ denotes the integration over all the variables in $\mathbf{H}$ except for $H_{nk}$. Similarly, the estimator for $X_{o,kt}$ for $o \in \{\mathsf{t}, \mathsf{d}\}$ can be obtained by the posterior mean

$$\widehat{X}_{o,kt} = \int \mathscr{P}(X_{o,kt}) X_{\mathsf{d},kt}\, \mathrm{d}X_{o,kt}, \tag{14}$$

where

$$\mathscr{P}(X_{o,kt}) = \int_{\mathbf{H}} \mathrm{d}\mathbf{H} \int_{\mathbf{X}\setminus X_{o,kt}} \mathrm{d}\mathbf{X}\, \mathsf{P}(\mathbf{H}, \mathbf{X}|\widetilde{\mathbf{Y}})$$

is the marginal posterior probability of $X_{o,kt}$. Also, the notation $\int_{\mathbf{X}\setminus X_{o,kt}} \mathrm{d}\mathbf{X}$ denotes the integration over all the variables in $\mathbf{X}$ except for $X_{o,kt}$. The posterior mean estimators (13) and (14) minimize the (Bayesian) MSE [32] defined as

$$\mathsf{mse}(\mathbf{H}) = \frac{1}{NK} \mathsf{E}\left\{ \|\widehat{\mathbf{H}} - \mathbf{H}\|_{\mathsf{F}}^2 \right\}, \tag{15a}$$

$$\mathsf{mse}(\mathbf{X}_o) = \frac{1}{KT_o} \mathsf{E}\left\{ \|\widehat{\mathbf{X}}_o - \mathbf{X}_o\|_{\mathsf{F}}^2 \right\}, \text{ for } o \in \{\mathsf{t}, \mathsf{d}\}, \tag{15b}$$

where the expectation operator is w.r.t. $\mathsf{P}(\widetilde{\mathbf{Y}}, \mathbf{H}, \mathbf{X}_o)$, and we have defined $\widehat{\mathbf{H}} = [\widehat{H}_{nk}]$ and $\widehat{\mathbf{X}}_o = [\widehat{X}_{o,kt}]$. We refer to (13) and (14) as the Bayes-optimal estimator.

*Remark 1:* Given a *known* pilot matrix $\underline{\mathbf{X}}_{\mathsf{t}}$, which by definition is given by $\mathsf{P}_{\mathsf{X}_{\mathsf{t}}}(\mathbf{X}_{\mathsf{t}}) = \delta(\mathbf{X}_{\mathsf{t}} - \underline{\mathbf{X}}_{\mathsf{t}})$, we obtain $\widehat{X}_{\mathsf{t},kt} = \underline{X}_{\mathsf{t},kt}$ from (14), and therefore $\mathsf{mse}(\mathbf{X}_{\mathsf{t}}) = 0$. For the case of our interest, we always have $\mathsf{mse}(\mathbf{X}_{\mathsf{t}}) = 0$. The algorithm as well as the analytical results still work even if the pilots are unknown, we can express the MSE as in (15b).

### *B. Bayes-Optimal Estimator in SISO Channel*

To better understand the Bayes-optimal estimator, we first explain it in a simple SISO system. We consider a SISO version of the system (1) given by

$$Y = Z + W. \tag{16}$$

Recall that $W$ is the additive white Gaussian noise with zero mean and variance $\sigma_w^2$. After a complex-valued quantizer, $\widetilde{Y} = \mathsf{Q}_c(Y)$ is obtained. Based on the system model (1), $Z = HX$ should be kept. However, to facilitate interpretation, we first let $Z$ be a random variable with distribution $\mathsf{P}_{\mathsf{Z}}$. According to Bayes' rule (11), the posterior probability can be computed as

$$\mathsf{P}(Z|\widetilde{Y}) = \frac{\mathsf{P}_{\mathsf{out}}(\widetilde{Y}|Z)\mathsf{P}_{\mathsf{Z}}(Z)}{\mathsf{P}(\widetilde{Y})}, \tag{17}$$

where $\mathsf{P}(\widetilde{Y}) = \int \mathsf{P}_{\mathsf{out}}(\widetilde{Y}|z)\mathsf{P}_{\mathsf{Z}}(z)\,\mathrm{d}z$ is the marginal likelihood. Then, from (13) or (14), the posterior mean estimator for $Z$ is given by

$$\widehat{Z} = \int z \mathsf{P}(z|\widetilde{Y})\, \mathrm{d}z. \tag{18}$$

To specify the estimator, we further assume that $Z$ is a proper complex Gaussian with mean $\hat{p}$ and variance $v^p$, i.e., $\mathsf{P}_{\mathsf{Z}}(Z) = \mathcal{N}_{\mathbb{C}}(Z; \hat{p}, v^p)$. Then, we derive the estimator (18) under the two channels, unquantized and quantized, in the following examples.

**Example 1** (Unquantized Channel). In this case, we have $\widetilde{Y} = Y$ and $\mathsf{P}_{\mathsf{out}}(\widetilde{Y}|Z) = \frac{1}{\pi\sigma_w^2} e^{-|\widetilde{Y}-Z|^2/\sigma_w^2}$. Using these distributions, we obtain

$$\begin{aligned} \mathsf{P}_{\mathsf{out}}(\widetilde{Y}|Z)\mathsf{P}_{\mathsf{Z}}(Z) &= \mathcal{N}_{\mathbb{C}}(Z; \widetilde{Y}, \sigma_w^2)\mathcal{N}_{\mathbb{C}}(Z; \hat{p}, v^p) \\ &= D \cdot \mathcal{N}_{\mathbb{C}}\left( Z; \frac{v^p \widetilde{Y} + \sigma_w^2 \hat{p}}{\sigma_w^2 + v^p}, \frac{\sigma_w^2 v^p}{\sigma_w^2 + v^p} \right), \end{aligned} \tag{19}$$

where $D = \mathcal{N}_{\mathbb{C}}(0; \widetilde{Y} - \hat{p}, \sigma_w^2 + v^p)$, and the second equality follows the *Gaussian reproduction property* [33, (A.7)].[5] Substituting (19) into (17), we obtain

$$\mathsf{P}(Z|\widetilde{Y}) = \mathcal{N}_{\mathbb{C}}\left( Z; \frac{v^p \widetilde{Y} + \sigma_w^2 \hat{p}}{\sigma_w^2 + v^p}, \frac{\sigma_w^2 v^p}{\sigma_w^2 + v^p} \right). \tag{20}$$

The estimator (18), which is the mean of $\mathsf{P}(Z|\widetilde{Y})$ after *rearranging* is determined as

$$\widehat{Z} = \hat{p} + \frac{v^p}{\sigma_w^2 + v^p}(\widetilde{Y} - \hat{p}). \tag{21}$$

The MSE of the estimator, which is the variance of $\mathsf{P}(Z|\widetilde{Y})$, is

$$v^z = v^p - \frac{(v^p)^2}{\sigma_w^2 + v^p}. \tag{22}$$

**Example 2** (Quantized Channel). If $\mathrm{Re}(\widetilde{Y}) \in (r_{b-1}, r_b]$ and $\mathrm{Im}(\widetilde{Y}) \in (r_{b'-1}, r_{b'}]$, then the likelihood of the quantized measurement $\widetilde{Y}$ is given by (9). The calculation of the posterior mean and variance in the quantized channel is technical, but it basically follows a procedure similar to that in the unquantized channel. A derivation is given in Appendix A, which turns out to yield

$$\widehat{Z} = \hat{p} + \frac{\mathsf{sign}(\widetilde{Y}) v^p}{\sqrt{(\sigma_w^2 + v^p)/2}} \left( \frac{\phi(\eta_1) - \phi(\eta_2)}{\Phi(\eta_1) - \Phi(\eta_2)} \right), \tag{23}$$

$$\begin{aligned} v^z = \frac{v^p}{2} + \frac{(v^p)^2}{2(\sigma_w^2 + v^p)} \times \\ \left( \frac{\eta_1 \phi(\eta_1) - \eta_2 \phi(\eta_2)}{\Phi(\eta_1) - \Phi(\eta_2)} + \left( \frac{\phi(\eta_1) - \phi(\eta_2)}{\Phi(\eta_1) - \Phi(\eta_2)} \right)^2 \right), \end{aligned} \tag{24}$$

[5] The product of two Gaussians gives another Gaussian [33, (A.7)]:

$$\mathcal{N}_{\mathbb{C}}(x; a, A)\mathcal{N}_{\mathbb{C}}(x; b, B) = D \cdot \mathcal{N}_{\mathbb{C}}(x; c, C),$$

where $c = C(A^{-1}a + B^{-1}b)$, $C = (A^{-1} + B^{-1})^{-1}$, and $D = \mathcal{N}_{\mathbb{C}}(0; a - b, A + B)$.

where

$$\eta_1 = \frac{\mathsf{sign}(\widetilde{Y})\hat{p} - \min\{|r_{b-1}|, |r_b|\}}{\sqrt{(\sigma_w^2 + v^{\mathrm{p}})/2}}, \tag{25a}$$

$$\eta_2 = \frac{\mathsf{sign}(\widetilde{Y})\hat{p} - \max\{|r_{b-1}|, |r_b|\}}{\sqrt{(\sigma_w^2 + v^{\mathrm{p}})/2}}. \tag{25b}$$

The real and imaginary parts are quantized separately, and each complex-valued channel can be decoupled into two real-valued channels. The expressions (23)–(25) are the estimators only for the real part of $Z$. To facilitate notation, we have abused $\widetilde{Y}$ and $\widehat{Z}$ in (23)–(25) to denote $\mathrm{Re}(\widetilde{Y})$ and $\mathrm{Re}(\widehat{Z})$, respectively. The estimator for the imaginary part $\mathrm{Im}(\widehat{Z})$ can be obtained analogously as (23) and (24), while $\widetilde{Y}$ and $b$ should be replaced by $\mathrm{Im}(\widetilde{Y})$ and $b'$, respectively.

*Remark 2:* Recall $r_0 = -\infty$ and $r_{2^{\mathsf{B}}} = \infty$. Therefore, if $b = 1$ or $b = 2^{\mathsf{B}}$, we obtain $\phi(\eta_2) = 0$, $\eta_2\phi(\eta_2) = 0$, and $\Psi(\eta_2) = 0$. Additionally, for a special case of $\mathsf{B} = 1$ (i.e., one-bit quantizer), the expressions of (23) and (24) agree with those reported in [34].

*Remark 3:* For another extremely case of $\mathsf{B} \to \infty$ and $\Delta \to 0$, we return to the unquantized channel as that in Example 1. Instead of using the procedure in Example 1, we show how can the expressions (21)–(22) be obtained from (23)–(24). Recall that $r_{b-1}$ and $r_b$ are the upper and the lower bin boundary positions w.r.t. the $b$-th bin. Let $r_{b-1} = r$ and $r_b = r_{b-1} + \mathrm{d}r$. As $\mathsf{B} \to \infty$ and $\Delta \to 0$, we have $\mathrm{d}r \to 0$ which results in $r_b \to r$ and $\eta_1 \to \eta_2 \triangleq \eta$. Furthermore, we obtain $\Phi(\eta_1) - \Phi(\eta_2) \to \frac{\mathrm{d}}{\mathrm{d}r}\Phi(\eta)$, $\phi(\eta_1) - \phi(\eta_2) \to \frac{\mathrm{d}}{\mathrm{d}r}\phi(\eta)$, and $\eta_1\phi(\eta_1) - \eta_2\phi(\eta_2) \to \frac{\mathrm{d}}{\mathrm{d}r}\eta\phi(\eta)$. By substituting these in (23)–(24) and applying the facts that $\frac{\mathrm{d}}{\mathrm{d}r}\Phi(\eta) = \phi(\eta)\frac{\mathrm{d}}{\mathrm{d}r}\eta$, $\frac{\mathrm{d}}{\mathrm{d}r}\phi(\eta) = -\eta\phi(\eta)\frac{\mathrm{d}}{\mathrm{d}r}\eta$, and $\frac{\mathrm{d}}{\mathrm{d}r}\eta\phi(\eta) = (1-\eta^2)\phi(\eta)\frac{\mathrm{d}}{\mathrm{d}r}\eta$, we recover the same expressions as given in (21) and (22) for the real part of $\widehat{Z}$. The imaginary part for $\widehat{Z}$ can be obtained analogously.

The aforementioned example is the estimator for $Z$. The same concept can be easily applied to the estimate of $H$ or $X$, if $Z$ is replaced by $H$ or $X$ in (16). However, if $Z = HX$ and both $H$ and $X$ are unknown, the Bayes-optimal estimator increases in complexity. In this case, the posterior probability in (17) becomes $\mathsf{P}(H, X|\widetilde{Y}) = \frac{\mathsf{P}_{\mathsf{out}}(\widetilde{Y}|H,X)\mathsf{P}_{\mathsf{H}}(H)\mathsf{P}_{\mathsf{X}}(X)}{\mathsf{P}(\widetilde{Y})}$, which involves two prior distributions for $H$ and $X$ as that in (11). To implement the posterior mean estimator for $H$ and $X$, we need the marginal posterior probabilities $\mathscr{P}(H) = \int \mathsf{P}(H, X|\widetilde{Y})\,\mathrm{d}X$ and $\mathscr{P}(X) = \int \mathsf{P}(H, X|\widetilde{Y})\,\mathrm{d}H$, respectively. The posterior probability $\mathsf{P}(H, X|\widetilde{Y})$ could not be found in closed form. Although we can resort to numerical integration to implement the estimator, the computational complexity is high. Therefore, one might consider an alternative technique; that is, the estimate of $H$ is performed with fixed $X$ and vice versa.

In the next subsection, we develop a practical algorithm for the Bayes-optimal estimator. Before proceeding, we intend to provide an intuition on the algorithm. A representation of the algorithm is shown in Fig. 2, which seems to operate in the alternative manner. Conceptually, when the posterior mean and variance of $Z$ are obtained from the quantized observation $\widetilde{Y}$, we can *reconstruct* $Y$ and then approximate $\mathsf{P}_{\mathsf{out}}(Y|Z)$ as a Gaussian distribution. Then, the posterior mean estimator for $H$ (or $X$) can be conducted through $Y$, which is an *AWGN* channel rather than a quantized channel, in an alternative manner. This representation is merely for an intuition. The accurate algorithm development takes a different route.

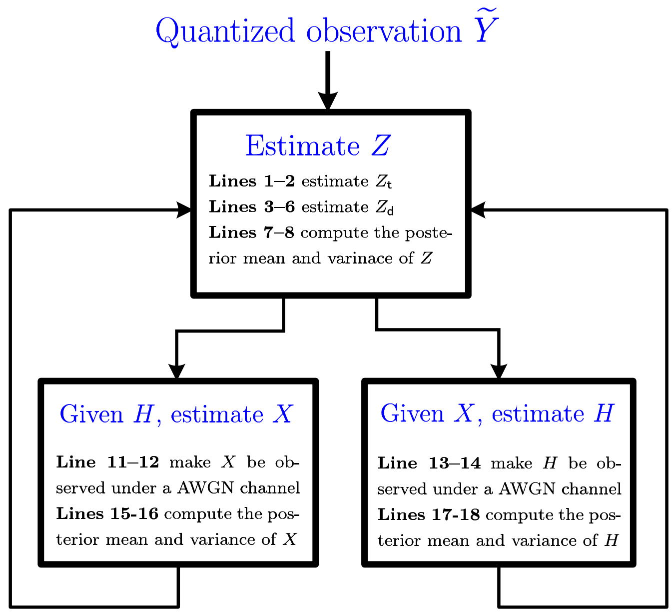


Fig. 2. A representation of the GAMP-based JCD algorithm.

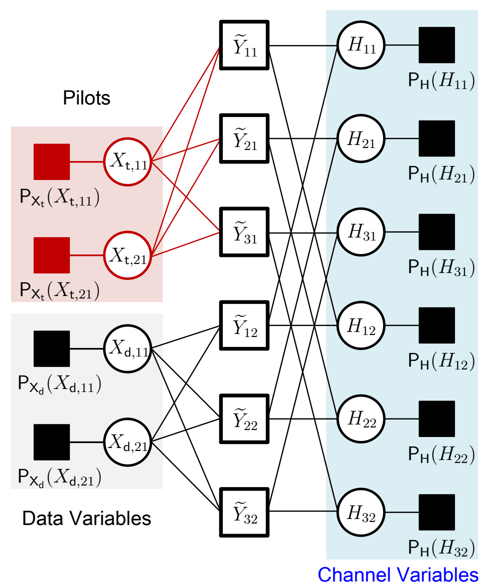


Fig. 3. Factor graph representation of the integrant of (26), where $N = 3$, $K = 2$, and $T = 2$.

### B. GAMP-Based JCD Algorithm

From the discussions above, direct computations of (13) and (14) are intractable due to high-dimensional integrals in the marginal posteriors $\mathscr{P}(X_{kt})$ and $\mathscr{P}(H_{nk})$. To be tractable, we first note that by combining (5)–(7), the posterior probability

(11) can be factored into

$$\frac{1}{\mathsf{P}(\widetilde{\mathbf{Y}})}\prod_{n=1}^{N}\prod_{t=1}^{T}\mathsf{P}_{\mathsf{out}}\Big(\widetilde{Y}_{nt}\Big|Z_{nt}\Big)\times\prod_{n=1}^{N}\prod_{k=1}^{K}\mathsf{P}_{\mathsf{H}}(H_{nk})$$
$$\times\prod_{k=1}^{K}\prod_{t=1}^{T_{\mathsf{t}}}\mathsf{P}_{\mathsf{X}_{\mathsf{t}}}(X_{\mathsf{t},kt})\times\prod_{k=1}^{K}\prod_{t=1}^{T_{\mathsf{d}}}\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(X_{\mathsf{d},kt}). \quad (26)$$

An example factor graph for (26) is shown in Figure 3, where a square represents a factor node associated with the sub-constraint function $\mathsf{P}_{\mathsf{out}}(\widetilde{Y}_{nt}|Z_{nt})$ in (26), while a circle shows a variable node associated with $H_{nk}$, $X_{\mathsf{d},kt}$, or $X_{\mathsf{t},kt}$. The factor graph suggests the use of the canonical sum-product algorithm to compute the marginal posterior probabilities. The algorithm uses a set of message passing equations which go from factor nodes to variable nodes and vice versa.

However, the computational complexity of the sum-product algorithm is still infeasible in the case of our interest because it still involves high-dimensional integration and summation. Hence, we resort to a recently developed approximation: the so-called AMP (approximate message passing) algorithm [35] and generalized AMP (GAMP) [36]. In particular, AMP is a variant of the sum-product algorithm, which was initially proposed by Donoho *et al.* [35] in order to solve a linear inverse problem in the context of compressive sensing. Using GAMP to our MIMO system means that given $\mathbf{H}$ known, GAMP can provide a tractable way to approximate the marginal posteriors $\mathscr{P}(X_{kt})$'s. This part corresponds to addressing the message passing equations among $\widetilde{Y}_{nt}$ and $X_{kt}$, i.e., the left-hand side of Figure 3. For the study see, e.g., [19, 37]. More recently, Parker *et al.* in [30] applied the same strategy of GAMP to the problem of reconstructing matrices from bilinear noisy observations (i.e., reconstructing $\mathbf{H}$ and $\mathbf{X}$ from $\mathbf{Y}$), which is referred to as BiG-AMP.

BiG-AMP for JCD estimation is presented in Algorithm 1 for a given instantiation of the quantized observations $\widetilde{\mathbf{Y}}$, the pilot matrix $\mathbf{X}_{\mathsf{t}}$, as well as the likelihood $\mathsf{P}_{\mathsf{out}}(\widetilde{\mathbf{Y}}|\mathbf{Z})$, and the variable distributions $\mathsf{P}_{\mathsf{H}}(\mathbf{H})$ and $\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(\mathbf{X}_{\mathsf{d}})$. We refer to this scheme as the GAMP-based JCD algorithm, which follows the same structure as BiG-AMP [30] except for the part dealing with the known pilots, i.e., $t\in\mathcal{T}_{\mathsf{t}}$ in Algorithm 1. We refer the interested reader to [30] for derivation details of BiG-AMP.

To better understand the algorithm, we provide some intuition on each step of Algorithm 1. Also see the representation in Fig. 2. Lines 3–6 compute an estimate $\hat{\mathbf{P}}_{\mathsf{d}}=[\hat{p}_{nt}]$ of the matrix product $\mathbf{Z}_{\mathsf{d}}=\mathbf{H}\mathbf{X}_{\mathsf{d}}$ and the corresponding variances $\{v^{p}_{nt}:t\in\mathcal{T}_{\mathsf{d}}\}$. Here, $\bar{\mathbf{P}}_{\mathsf{d}}=[\bar{p}_{nt}]$ and $\{\bar{v}^{p}_{nt}:t\in\mathcal{T}_{\mathsf{d}}\}$ in lines 3–4 can be regarded as auxiliary variables.[6] Similarly, lines 1–2 do the same matter but for the matrix product $\mathbf{Z}_{\mathsf{t}}=\mathbf{H}\mathbf{X}_{\mathsf{t}}$. Because pilot matrix $\mathbf{X}_{\mathsf{t}}$ is known, the corresponding variances for $\mathbf{X}_{\mathsf{t}}$ are zero, i.e., $v^{x}_{kt}=0$ for $t\in\mathcal{T}_{\mathsf{t}}$. With $v^{x}_{kt}=0$, we thus have plugged $\bar{p}_{nt}$ and $\bar{v}^{p}_{nt}$ into $\hat{p}_{nt}$ and $v^{p}_{nt}$ for $t\in\mathcal{T}_{\mathsf{t}}$ to get lines 1–2. Using $\{\hat{p}_{nt},v^{p}_{nt}\}$, lines 7–8 then yield the posterior means $\widehat{\mathbf{Z}}=[\widehat{Z}_{nt}]$ and variances $\{v^{z}_{nt}\}$ of $\mathbf{Z}$. Then lines 9–10 use these posterior results to compute the residual $\hat{\mathbf{S}}=[\hat{s}_{nt}]$ and the inverse residual variances $\{v^{s}_{nt}\}$. Lines 11–12 then use these residual terms to compute $\hat{\mathbf{R}}=[\hat{r}_{kt}]$ and $\{v^{r}_{kt}\}$, where $\hat{r}_{kt}$ can be interpreted as an observation of $X_{\mathsf{d},kt}$ under an AWGN channel with zero-mean and variance of $v^{r}_{kt}$. Similarly, lines 13–14 evaluate $\hat{\mathbf{Q}}=[\hat{q}_{nk}]$ and $v^{q}_{nk}$, where $\hat{q}_{nt}$ can be interpreted as an observation of $H_{nk}$ under an AWGN channel with noise variance of $v^{q}_{nk}$. Finally, lines 15–16 estimate the posterior mean $\widehat{\mathbf{X}}=[\widehat{X}_{kt}]$ and variances $\{v^{x}_{kt}\}$ by taking into account the prior $\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}$; lines 17–18 perform the same for $H_{nk}$.

[6] $\bar{\mathbf{P}}_{\mathsf{d}}$ is a plug-in estimate of $\mathbf{Z}_{\mathsf{d}}$ while $\hat{\mathbf{P}}_{\mathsf{d}}=[\hat{p}_{nt}]$ provides a refinement by introducing "Onsager" correction in the context of AMP. See [30] for details.

**Algorithm 1:** GAMP-based JCD Algorithm

**input** : Quantized observations $\widetilde{\mathbf{Y}}$, pilot matrix $\mathbf{X}_{\mathsf{t}}$, likelihood $\mathsf{P}_{\mathsf{out}}(Y|Z)$, and variable distributions $\mathsf{P}_{\mathsf{H}}(H)$ and $\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(X_{\mathsf{d}})$

**output** : $\widehat{\mathbf{H}}$, $\widehat{\mathbf{X}}_{\mathsf{d}}$

**definition**: $\sum_k\triangleq\sum_{k=1}^{K}$, $\sum_n\triangleq\sum_{n=1}^{N}$

**initialize** : $\forall n,t$: $\hat{s}_{nt}(0)=0$, $v^{z}_{nt}(0)=1$, $\widehat{Z}_{nt}(0)=0$; $\forall n,k,t$: $v^{x}_{kt}(1)=1$, $\widehat{X}_{\mathsf{d},kt}(1)=0$, $v^{h}_{nk}(1)=1$, $\widehat{H}_{nk}(1)=0$.

$\xi\leftarrow 1$;

**while** $\frac{\sum_{n,t}|\widehat{Z}_{nt}(\xi)-\widehat{Z}_{nt}(\xi-1)|^2}{\sum_{n,t}|\widehat{Z}_{nt}(\xi-1)|^2}>\epsilon$ **and** $\xi<\xi_{\max}$ **do**

**if** $t\in\mathcal{T}_{\mathsf{t}}$ **then**

1 $\forall n$: $v^{p}_{nt}(\xi)=\sum_k v^{h}_{nk}(\xi)|X_{kt}|^2$;

2 $\forall n$: $\hat{p}_{nt}(\xi)=\sum_k\widehat{H}_{nk}(\xi)X_{kt}-\hat{s}_{nt}(\xi-1)v^{p}_{nt}(\xi)$;

**if** $t\in\mathcal{T}_{\mathsf{d}}$ **then**

3 $\forall n$: $\bar{v}^{p}_{nt}(\xi)=\sum_k|\widehat{H}_{nk}(\xi)|^2v^{x}_{kt}(\xi)+v^{h}_{nk}(\xi)|\widehat{X}_{kt}(\xi)|^2$;

4 $\forall n$: $\bar{p}_{nt}(\xi)=\sum_k\widehat{H}_{nk}(\xi)\widehat{X}_{kt}(\xi)$;

5 $\forall n$: $v^{p}_{nt}(\xi)=\bar{v}^{p}_{nt}(\xi)+\sum_k v^{h}_{nk}(\xi)v^{x}_{kt}(\xi)$;

6 $\forall n$: $\hat{p}_{nt}(\xi)=\bar{p}_{nt}(\xi)-\hat{s}_{nt}(\xi-1)\bar{v}^{p}_{nt}(\xi)$;

7 $\forall n,t$: $v^{z}_{nt}(\xi)=\mathsf{Var}\{Z_{nt}|\hat{p}_{nt}(\xi),v^{p}_{nt}(\xi)\}$;

8 $\forall n,t$: $\widehat{Z}_{nt}(\xi)=\mathsf{E}\{Z_{nt}|\hat{p}_{nt}(\xi),v^{p}_{nt}(\xi)\}$;

9 $\forall n,t$: $v^{s}_{nt}(\xi)=(1-v^{z}_{nt}(\xi)/v^{p}_{nt}(\xi))/v^{p}_{nt}(\xi)$;

10 $\forall n,t$: $\hat{s}_{nt}(\xi)=(\widehat{Z}_{nt}(\xi)-\hat{p}_{nt}(\xi))/v^{p}_{nt}(\xi)$;

11 $\forall k,t$: $v^{r}_{kt}(\xi)=\Big[\sum_n|\widehat{H}_{nk}(\xi)|^2v^{s}_{kt}(\xi)\Big]^{-1}$;

12 $\forall k,t$: $\hat{r}_{kt}(\xi)=\widehat{X}_{kt}(\xi)\left(1-v^{r}_{kt}(\xi)\sum_n v^{h}_{nk}(\xi)v^{s}_{nt}(\xi)\right)+v^{r}_{kt}(\xi)\sum_n\widehat{H}^{*}_{nk}(\xi)\hat{s}_{nt}(\xi)$ ;

13 $\forall n,k$: $v^{q}_{nk}(\xi)=\Big[\sum_{t\in\mathcal{T}_{\mathsf{t}}}|X_{kt}|^2v^{s}_{nt}(\xi)\Big]^{-1}$;

14 $\forall n,k$: $\hat{q}_{nk}(\xi)=\widehat{H}_{nk}(\xi)\Big(1-v^{q}_{nk}(\xi)\sum_{t\in\mathcal{T}_{\mathsf{d}}}v^{x}_{kt}(\xi)v^{s}_{nt}(\xi)\Big)+v^{q}_{nk}(\xi)\Big(\sum_{t\in\mathcal{T}_{\mathsf{t}}}X^{*}_{kt}\hat{s}_{nt}(\xi)+\sum_{t\in\mathcal{T}_{\mathsf{d}}}\widehat{X}^{*}_{kt}(\xi)\hat{s}_{nt}(\xi)\Big)$ ;

15 $\forall k,t\in\mathcal{T}_{\mathsf{d}}$: $v^{x}_{kt}(\xi+1)=\mathsf{Var}\{X_{kt}|\hat{r}_{kt}(\xi),v^{r}_{kt}(\xi)\}$;

16 $\forall k,t\in\mathcal{T}_{\mathsf{d}}$: $\widehat{X}_{kt}(\xi+1)=\mathsf{E}\{X_{kt}|\hat{r}_{kt}(\xi),v^{r}_{kt}(\xi)\}$;

17 $\forall n,k$: $v^{h}_{nk}(\xi+1)=\mathsf{Var}\{H_{nk}|\hat{q}_{nk}(\xi),v^{q}_{nk}(\xi)\}$;

18 $\forall n,k$: $\widehat{H}_{nk}(\xi+1)=\mathsf{E}\{H_{nk}|\hat{q}_{nk}(\xi),v^{q}_{nk}(\xi)\}$;

$\xi\leftarrow\xi+1$ ;

### C. Nonlinear Steps

Algorithm 1 gives a high-level description of BiG-AMP to perform JCD estimation. Lines 7–8, 15–16, and 17–18 of Algorithm 1 perform the posterior mean and variance estimators for $Z_{nt}$, $X_{kt}$, and $H_{nk}$, respectively. A remarkable feature of the algorithm is that at each iteration, the estimate of $Z_{nt}$, $X_{kt}$, and $H_{nk}$ can perform separably as the estimators over a bank of scalar channels. Next, we detail these nonlinear steps. For brevity, we omit the subscript indexes $n, k, t$ hereafter.

First, we notice that lines 7–8 compute the posterior mean and variance of $Z$, where the expectation operator is w.r.t.

$$\mathscr{P}(Z) = \frac{\mathsf{P}_{\text{out}}(\widetilde{Y}|Z)\mathsf{P}_{\mathsf{Z}}(Z)}{\int \mathsf{P}_{\text{out}}(\widetilde{Y}|z')\mathsf{P}_{\mathsf{Z}}(z')\,\mathrm{d}z'},$$

where $\mathsf{P}_{\text{out}}(\widetilde{Y}|Z)$ is given by (9), and $\mathsf{P}_{\mathsf{Z}}(Z) = \mathcal{N}_{\mathbb{C}}(Z; \hat{p}, v^p)$. This process is exactly identical to what we have done in Example 2 of Section III.B. As a result, lines 7–8 of Algorithm 1 for each real-valued channel can be computed using the expressions in (23) and (24).

Next, we discuss the nonlinear steps used to compute $(\widehat{X}, v^x)$ and $(\widehat{H}, v^h)$ in lines 15–16 and 17–18 of Algorithm 1. Specifically, the expectations and variances in lines 15-16 and 17-18 are taken w.r.t. the marginal posterior

$$\mathscr{P}(X_{\mathsf{d}}) = \frac{\mathcal{N}(X_{\mathsf{d}}; \hat{r}, v^r)\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(X_{\mathsf{d}})}{\int \mathcal{N}(x'_{\mathsf{d}}; \hat{r}, v^r)\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(x'_{\mathsf{d}})\,\mathrm{d}x'_{\mathsf{d}}}, \tag{27}$$

$$\mathscr{P}(H) = \frac{\mathcal{N}(H; \hat{q}, v^q)\mathsf{P}_{\mathsf{H}}(H)}{\int \mathcal{N}(h'; \hat{q}, v^q)\mathsf{P}_{\mathsf{H}}(h')\,\mathrm{d}h'}. \tag{28}$$

These posterior probabilities are similar to that of $(\widehat{Z}, v^z)$ except that $\mathsf{P}_{\text{out}}$ is replaced by a Gaussian distribution, and the corresponding priors $\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}$ and $\mathsf{P}_{\mathsf{H}}$ are used in place of $\mathsf{P}_{\mathsf{Z}}$. In fact, the former change results in a fundamentally difference estimator than in the case of $Z$. Recall from Example 1 of Section III.B that if $\mathsf{P}_{\text{out}}$ is a Gaussian distribution, the estimator is operated in an *unquantized* channel. That is, the estimates of $H$ and $X$ in Algorithm 1 are based on AWGN channels.

To specify $(\widehat{X}, v^x)$, we consider the square QAM constellation with $2\nu \times 2\nu$ points

$$\mathcal{X} = \Big\{ X_{\mathrm{R}} + \mathrm{j} X_{\mathrm{I}} : X_{\mathrm{R}}, X_{\mathrm{I}} \in \{-(2\nu-1)\zeta, \ldots, -3\zeta, -\zeta, \zeta, 3\zeta, \ldots, (2\nu-1)\zeta\} \Big\}, \tag{29}$$

where $\zeta = 1/\sqrt{2((2\nu)^2-1)/3}$ is the power normalization factor. If $X$ is drawn from the constellation points uniformly, i.e., $\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(X_{\mathsf{d}}) = 1/(2\nu)^2$ for $X_{\mathsf{d}} \in \mathcal{X}$, lines 15–16 of Algorithm 1 can be computed using

$$\widehat{X}_{\mathsf{d}} = \frac{\sum_{i=1}^{\nu}(2i-1)\zeta F_i^{\mathsf{s}}(\mathrm{Re}(\hat{r}))}{\sum_{i=1}^{\nu} F_i^{\mathsf{c}}(\mathrm{Re}(\hat{r}))} + \mathrm{j}\frac{\sum_{i=1}^{\nu}(2i-1)\zeta F_i^{\mathsf{s}}(\mathrm{Im}(\hat{r}))}{\sum_{i=1}^{\nu} F_i^{\mathsf{c}}(\mathrm{Im}(\vartheta_i))}, \tag{30}$$

$$v^x = \frac{\sum_{i=1}^{\nu}(2i-1)^2\zeta^2 F_i^{\mathsf{s}}(\mathrm{Re}(\hat{r}))}{\sum_{i=1}^{\nu} F_i^{\mathsf{c}}(\mathrm{Re}(\hat{r}))} + \frac{\sum_{i=1}^{\nu}(2i-1)^2\zeta^2 F_i^{\mathsf{s}}(\mathrm{Im}(\hat{r}))}{\sum_{i=1}^{\nu} F_i^{\mathsf{c}}(\mathrm{Im}(\hat{r}))} - |\widehat{X}_{\mathsf{d}}|^2, \tag{31}$$

where

$$F_i^{\mathsf{s}}(x) = e^{-\frac{(2i-1)^2\zeta}{v^r}} \sinh\left(\frac{2(2i-1)\zeta}{v^r}x\right),$$
$$F_i^{\mathsf{c}}(x) = e^{-\frac{(2i-1)^2\zeta}{v^r}} \cosh\left(\frac{2(2i-1)\zeta}{v^r}x\right).$$

Finally, recall that $\mathsf{P}_{\mathsf{H}}(H_{nk}) = \mathcal{N}_{\mathbb{C}}(0, \sigma_h^2)$. Then lines 17–18 of Algorithm 1 can be computed using

$$\widehat{H} = \frac{\sigma_h^2}{\sigma_h^2 + v^q}\,\hat{q} \quad \text{and} \quad v^h = v^q - \frac{(v^q)^2}{\sigma_h^2 + v^q}. \tag{32}$$

The derivation of (32) is identical to that in Example 2.

Using the above nonlinear steps (23)–(24) and (30)–(32), the GAMP-based JCD algorithm has been implemented based on the open-source "GAMPmatlab" software suite. The code for the GAMP-based JCD algorithm is available upon request.

## IV. Performance Analysis

Here, we present a framework to analyze the Bayes-optimal JCD estimator. First, recall the definitions of the MSEs of $\mathbf{H}$ and $\mathbf{X}_{\mathsf{d}}$ and from (15). The key strategy for analyzing $\mathsf{mse}(\mathbf{H})$ and $\mathsf{mse}(\mathbf{X}_{\mathsf{d}})$ is through the average free entropy

$$\mathcal{F} \triangleq \frac{1}{K^2}\mathsf{E}_{\widetilde{\mathbf{Y}}}\Big\{\log \mathsf{P}(\widetilde{\mathbf{Y}})\Big\}, \tag{33}$$

where $\mathsf{P}(\widetilde{\mathbf{Y}})$ denotes the marginal likelihood in (12), namely the partition function. Following the argument of [38, 39], it can be shown that $\mathsf{mse}(\mathbf{X}_{\mathsf{d}})$ and $\mathsf{mse}(\mathbf{H})$ are saddle points of the average free entropy. Thus, our goal reduces to finding (33).

Our analysis is based on the large-system limit. That is, when $N, K, T \to \infty$ but the ratios

$$N/K = \alpha, \quad T/K = \beta, \quad T_{\mathsf{t}}/K = \beta_{\mathsf{t}}, \quad T_{\mathsf{d}}/K = \beta_{\mathsf{d}}, \tag{34}$$

are fixed and finite. For convenience, we simply use $K \to \infty$ to denote this large-system limit. Even in the large-system limit, the computation of (33) is hard. The major difficulty in computing (33) is the expectation of the logarithm of $\mathsf{P}(\widetilde{\mathbf{Y}})$, which, nevertheless, can be facilitated by rewriting $\mathcal{F}$ as[7]

$$\mathcal{F} = \frac{1}{K^2}\lim_{\tau\to 0}\frac{\partial}{\partial\tau}\log \mathsf{E}_{\widetilde{\mathbf{Y}}}\Big\{\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\Big\}. \tag{35}$$

Note that the expectation operator is now moved inside the log-function. We first evaluate $\mathsf{E}_{\widetilde{\mathbf{Y}}}\{\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\}$ for an integer-valued $\tau$, and then generalize the result to any positive real number $\tau$. This technique, called *the replica method*, is from the field of statistical physics [40], which is not mathematically rigorous. Nevertheless, the replica method has proved successful in a number of highly difficult problems in statistical physics [40] and information theory, e.g., [27, 41–47]. Some of the results originally obtained by the replica method have been subsequently validated by other approaches, e.g., [48, 49]. Under the assumption of $K \to \infty$ and replica symmetry (RS), an asymptotic free entropy can be obtained later in Proposition

[7] We use the formula

$$\lim_{\tau\to 0}\frac{\partial}{\partial\tau}\log \mathsf{E}\{\mathsf{A}^{\tau}\} = \lim_{\tau\to 0}\frac{\mathsf{E}\{\mathsf{A}^{\tau}\log \mathsf{A}\}}{\mathsf{E}\{\mathsf{A}^{\tau}\}} = \mathsf{E}\{\log \mathsf{A}\},$$

where $\mathsf{A}$ is any positive random variable.

1. We check the accuracy of the replica-based analysis via simulations. Proposition 1 involves several new parameters, we find it helpful to introduce them first.

### A. Parameters of Proposition 1

Most parameters (except for some auxiliary parameters) of Proposition 1 can be illustrated systematically by the scalar AWGN channels:

$$Y_{X_{\mathsf{d}}} = \sqrt{\tilde{q}_{X_{\mathsf{d}}}} X_{\mathsf{d}} + W_{X_{\mathsf{d}}}, \tag{36a}$$
$$Y_H = \sqrt{\tilde{q}_H} H + W_H, \tag{36b}$$

where $W_H, W_{X_{\mathsf{d}}} \sim \mathcal{N}_{\mathbb{C}}(0,1)$, $H \sim \mathsf{P}_H$, and $X_{\mathsf{d}} \sim \mathsf{P}_{X_{\mathsf{d}}}$. We shall specify how the parameters $\tilde{q}_H$ and $\tilde{q}_{X_{\mathsf{d}}}$ are related to the asymptotic free entropy later in Proposition 1. Here, we know that the parameters $\tilde{q}_{X_{\mathsf{d}}}$ and $\tilde{q}_H$ serve as the signal-to-noise ratios (SNRs) of the above AWGN channels. The likelihoods under (36a) and (36b) are, respectively, given by

$$\mathsf{P}(Y_{X_{\mathsf{d}}}|X_{\mathsf{d}}) = \frac{1}{\pi} e^{-|Y_{X_{\mathsf{d}}} - \sqrt{\tilde{q}_{X_{\mathsf{d}}}} X_{\mathsf{d}}|^2}, \tag{37a}$$
$$\mathsf{P}(Y_H|H) = \frac{1}{\pi} e^{-|Y_H - \sqrt{\tilde{q}_H} H|^2}, \tag{37b}$$

and then we get the posteriors

$$\mathsf{P}(X_{\mathsf{d}}|Y_{X_{\mathsf{d}}}) = \frac{\mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(X_{\mathsf{d}})\mathsf{P}(Y_{X_{\mathsf{d}}}|X_{\mathsf{d}})}{\int \mathrm{d}x'_{\mathsf{d}}\, \mathsf{P}_{\mathsf{X}_{\mathsf{d}}}(x'_{\mathsf{d}})\mathsf{P}(Y_{X_{\mathsf{d}}}|x'_{\mathsf{d}})}, \tag{38a}$$
$$\mathsf{P}(H|Y_H) = \frac{\mathsf{P}_{\mathsf{H}}(H)\mathsf{P}(Y_H|H)}{\int \mathrm{d}h'\, \mathsf{P}_{\mathsf{H}}(h')\mathsf{P}(Y_H|h')}. \tag{38b}$$

With the posteriors, some important quantities are obtained. For example, the posterior mean estimators for $X_{\mathsf{d}}$ and $H$ read

$$\widehat{X}_{\mathsf{d}} = \int \mathrm{d}X_{\mathsf{d}}\, X_{\mathsf{d}} \mathsf{P}(X_{\mathsf{d}}|Y_{X_{\mathsf{d}}}), \tag{39a}$$
$$\widehat{H} = \int \mathrm{d}H\, H \mathsf{P}(H|Y_H). \tag{39b}$$

The MSEs of the estimators are thus given by

$$\mathsf{mse}_{X_{\mathsf{d}}} = \mathsf{E}\left\{|X_{\mathsf{d}} - \widehat{X}_{\mathsf{d}}|^2\right\}, \tag{40a}$$
$$\mathsf{mse}_H = \mathsf{E}\left\{|H - \widehat{H}|^2\right\}, \tag{40b}$$

in which the expectations are taken over $\mathsf{P}(Y_{X_{\mathsf{d}}}, X_{\mathsf{d}})$ and $\mathsf{P}(Y_H, H)$, respectively. Also see (30)–(32) for an example of explicit expressions of the above quantities. In addition, the mutual information between $Y_{X_{\mathsf{d}}}$ and $X_{\mathsf{d}}$ reads [50]

$$I(X_{\mathsf{d}}; Y_{X_{\mathsf{d}}}|\tilde{q}_{X_{\mathsf{d}}}) = -\mathsf{E}_{Y_{X_{\mathsf{d}}}}\left\{\log \mathsf{E}_{X_{\mathsf{d}}}\left\{e^{-|Y_{X_{\mathsf{d}}} - \sqrt{\tilde{q}_{X_{\mathsf{d}}}} X_{\mathsf{d}}|^2}\right\}\right\} - 1 \tag{41}$$

and the mutual information between $Y_H$ and $H$ is

$$I(H; Y_H|\tilde{q}_H) = -\mathsf{E}_{Y_H}\left\{\log \mathsf{E}_H\left\{e^{-|Y_H - \sqrt{\tilde{q}_H} H|^2}\right\}\right\} - 1. \tag{42}$$

From (36), one would infer that there exists another scalar AWGN channel w.r.t. the t-phase; i.e.,

$$Y_{X_{\mathsf{t}}} = \sqrt{\tilde{q}_{X_{\mathsf{t}}}} X_{\mathsf{t}} + W_{X_{\mathsf{t}}}, \tag{43}$$

where $W_{X_{\mathsf{t}}} \sim \mathcal{N}_{\mathbb{C}}(0,1)$ and $X_{\mathsf{t}} \sim \mathsf{P}_{X_{\mathsf{t}}}$. As the pilot is known, we can easily obtain $\mathsf{mse}_{X_{\mathsf{t}}} = 0$ following the argument in Remark 1; and the mutual information between $Y_{X_{\mathsf{t}}}$ and $X_{\mathsf{t}}$ is 0. As all the performances relating to (43) are trivial, we will not use (43) in the following discussions.

### B. Analytical Results

*Proposition 1:* As $K \to \infty$, the asymptotic free entropy is

$$\begin{aligned}\mathcal{F} = \alpha &\sum_{o\in\{\mathsf{t},\mathsf{d}\}} \beta_o \left( \sum_{b=1}^{2^{\mathsf{B}}} \int \mathrm{D}v\, \Psi_b(V_o) \log \Psi_b(V_o) \right) \\ &- \alpha I(H; Y_H|\tilde{q}_H) - \beta_{\mathsf{d}} I(X_{\mathsf{d}}; Y_{X_{\mathsf{d}}}|\tilde{q}_{X_{\mathsf{d}}}) \\ &+ \alpha(c_H - q_H)\tilde{q}_H + \sum_{o\in\{\mathsf{t},\mathsf{d}\}} \beta_o (c_{X_o} - q_{X_o})\tilde{q}_{X_o},\end{aligned} \tag{44}$$

where

$$\begin{aligned}\Psi_b(V_o) \triangleq \Phi&\left( \frac{\sqrt{2} r_b - V_o}{\sqrt{\sigma_w^2 + c_H c_{X_o} - q_H q_{X_o}}} \right) \\ &- \Phi\left( \frac{\sqrt{2} r_{b-1} - V_o}{\sqrt{\sigma_w^2 + c_H c_{X_o} - q_H q_{X_o}}} \right);\end{aligned} \tag{45}$$

$V_o \triangleq \sqrt{q_H q_{X_o}} v$ for $o \in \{\mathsf{t}, \mathsf{d}\}$; $I(\cdot)$'s are given by (41) and (42); and $c_{X_o} \triangleq \mathsf{E}\{|X_o|^2\} = \sigma_{x_o}^2$, $c_H \triangleq \mathsf{E}\{|H|^2\} = \sigma_h^2$. In (44), the other parameters $\{q_{X_o}, q_H, \tilde{q}_{X_o}, \tilde{q}_H\}$ are obtained from the solutions to the following fixed-point equations

$$\tilde{q}_H = \beta_{\mathsf{t}} q_{X_{\mathsf{t}}} \chi_{\mathsf{t}} + \beta_{\mathsf{d}} q_{X_{\mathsf{d}}} \chi_{\mathsf{d}}, \qquad q_H = c_H - \mathsf{mse}_H, \tag{46a}$$
$$\tilde{q}_{X_{\mathsf{t}}} = \alpha q_H \chi_{\mathsf{t}}, \qquad q_{X_{\mathsf{t}}} = c_{X_{\mathsf{t}}} - \mathsf{mse}_{X_{\mathsf{t}}}, \tag{46b}$$
$$\tilde{q}_{X_{\mathsf{d}}} = \alpha q_H \chi_{\mathsf{d}}, \qquad q_{X_{\mathsf{d}}} = c_{X_{\mathsf{d}}} - \mathsf{mse}_{X_{\mathsf{d}}}, \tag{46c}$$

in which $\mathsf{mse}_{X_{\mathsf{t}}} = 0$, and $\mathsf{mse}_H$ and $\mathsf{mse}_{X_{\mathsf{d}}}$ are given by (40). Also, in (46), we have defined

$$\chi_o \triangleq \sum_{b=1}^{2^{\mathsf{B}}} \int \mathrm{D}v \frac{\left( \Psi'_b\left(\sqrt{q_H q_{X_o}} v\right) \right)^2}{\Psi_b\left(\sqrt{q_H q_{X_o}} v\right)}, \text{ for } o \in \{\mathsf{t}, \mathsf{d}\} \tag{47}$$

with $\Psi_b(\cdot)$ given by (45) and

$$\begin{aligned}\Psi'_b(V_o) &\triangleq \frac{\partial \Psi_b(V_o)}{\partial V_o} \\ &= \frac{e^{-\frac{(\sqrt{2} r_b - V_o)^2}{2(\sigma_w^2 + c_H c_{X_o} - q_H q_{X_o})}} - e^{-\frac{(\sqrt{2} r_{b-1} - V_o)^2}{2(\sigma_w^2 + c_H c_{X_o} - q_H q_{X_o})}}}{\sqrt{2\pi(\sigma_w^2 + c_H c_{X_o} - q_H q_{X_o})}}.\end{aligned} \tag{48}$$

*Proof:* See Appendix B. □

As mentioned earlier, the asymptotic MSEs of $\mathbf{X}_{\mathsf{d}}$ and $\mathbf{H}$ are saddle points of the free entropy. Clearly, from Proposition 1, they are $\mathsf{mse}_{X_{\mathsf{d}}}$ and $\mathsf{mse}_H$, respectively. Note that the MSEs are associated with the scalar AWGN channels (36a) and (36b). Therefore, one would infer that performances of the quantized MIMO system can be fully characterized by the scalar AWGN channels (36). The following proposition formulates such intuition.

*Proposition 2:* Let $X_{\mathsf{d},kt}$, $H_{nk}$, $\widehat{X}_{\mathsf{d},kt}$, and $\widehat{H}_{nk}$ denote the $(k,t)$-th and the $(n,k)$-th entries of $\mathbf{X}_{\mathsf{d}}$, $\mathbf{H}$, $\widehat{\mathbf{X}}_{\mathsf{d}}$, and $\widehat{\mathbf{H}}$. As $K \to \infty$, the joint distribution of $(X_{\mathsf{d},kt}, H_{nk}, \widehat{X}_{\mathsf{d},kt}, \widehat{H}_{nk})$ of channels (2), (13), and (14) converges to the joint distribution $(X_{\mathsf{d}}, H, \widehat{X}_{\mathsf{d}}, \widehat{H})$, for the scalar channels (36a) and (36a).

*Proof:* See Appendix C. □

Proposition 2 shows that in the large-system limit, the input-output of the quantized MIMO system employing the Bayes-optimal JCD estimator is decoupled into equivalently a bank of the scalar AWGN channels (36a) and (36b). This characteristic is known as the decoupling principle, which was introduced by [43] for treading an *unquantized* MIMO system with *perfect* CSIR. If perfect CSIR is available, then we will not need (36b) for treating the channel estimation quality. Clearly, Proposition 2 extends the decoupling principle to a very general setting. In particular, we allow the *JCD* estimator so that the decoupled AWGN channels involve not only the data symbol [i.e, (36a)] but also the channel response [i.e., (36b)] as well.

*Remark 4:* The equivalent channels (36) are the scalar AWGN channels with $\tilde{q}_{\mathsf{d}}$ and $\tilde{q}_H$ being the equivalent SNRs. From (46) and (47), we see that the quantization effect is included in $\tilde{q}_{\mathsf{d}}$ and $\tilde{q}_H$ through $\chi_o$ for $o \in \{\mathsf{t}, \mathsf{d}\}$. Consider the extremely case of $\mathsf{B} \to \infty$ and $\Delta \to 0$, i.e., the unquantized channel. In this case, the Riemann sums $\sum_{b=1}^{2^{\mathsf{B}}}$ in (47) becomes the Riemann integral over the interval $(-\infty, \infty)$. Applying the technique in Remark 3 to (47) and evaluating the integrals, $\chi_o$ can be simplified to

$$\chi_o = \frac{1}{\sigma_w^2 + c_H c_{X_o} - q_H q_{X_o}}. \tag{49}$$

Substituting (46) for $q_H$ and $q_{X_o}$ in the demonstrator of (49), we obtain $\sigma_w^2 + c_H \mathsf{mse}_{X_o} + (c_{X_o} - \mathsf{mse}_{X_o})\mathsf{mse}_H$. The quantity in this form can be understood as the noise pulse the residual interference due to estimation errors of the data symbol and channel response.

To gain more insight into Proposition 2, we particularize our interest to some special cases in the following examples.

**Example 3** (Constellation-like Inputs). By Proposition 2, the asymptotic MSEs w.r.t. $\mathbf{X}_{\mathsf{d}}$ and $\mathbf{H}$ can be determined by the MSEs of the scalar AWGN channels (36a) and (36b), respectively. Thus, if the data symbol is drawn from a quadrature phase shift keying (QPSK) constellation, we will have

$$\mathsf{mse}_{X_{\mathsf{d}}} = 1 - \int \mathrm{D}z \tanh\Big(\tilde{q}_{X_{\mathsf{d}}} + \sqrt{\tilde{q}_{X_{\mathsf{d}}}}z\Big), \tag{50}$$

$$\mathsf{mse}_H = \frac{\sigma_h^2}{1 + \sigma_h^2 \tilde{q}_H}. \tag{51}$$

Besides, the SER w.r.t. $\mathbf{X}_{\mathsf{d}}$ can also be evaluated through the scalar AWGN channel (36a), which is given by [51, p.269]

$$\mathsf{SER} = 2\mathcal{Q}\Big(\sqrt{\tilde{q}_X}\Big) - \Big[\mathcal{Q}\Big(\sqrt{\tilde{q}_X}\Big)\Big]^2, \tag{52}$$

where $\mathcal{Q}(x) = \int_x^\infty \mathrm{D}z$ is the Q-function.

In fact, all these performances w.r.t. $\mathbf{X}_{\mathsf{d}}$ can be determined based on the knowledge merely of the scalar AWGN channel with SNR $\tilde{q}_X$. Thus, if the data symbol is drawn from other square QAM constellations, the corresponding SER can be easily obtained by the closed-form SER expression in [51, p.279].

**Example 4** (Perfect CSIR). If the channel matrix $\mathbf{H}$ is known perfectly, then the t-phase will not be required so that

$$\beta_{\mathsf{t}} = 0 \quad \text{and} \quad \beta_{\mathsf{d}} = \beta. \tag{53}$$

Because $\mathbf{H}$ is perfectly known, $\mathsf{mse}_H = 0$. Plugging this into (46a), we immediately obtain $q_H = c_H = \sigma_h^2$, which gives

$$q_H q_{X_{\mathsf{d}}} = c_H q_{X_{\mathsf{d}}}, \tag{54}$$

$$c_H c_{X_{\mathsf{d}}} - q_H q_{X_{\mathsf{d}}} = c_H \mathsf{mse}_{X_{\mathsf{d}}}, \tag{55}$$

in which (55) follows from the result that $c_H c_{X_{\mathsf{d}}} - q_H q_{X_{\mathsf{d}}} = c_H(c_{X_{\mathsf{d}}} - q_{X_{\mathsf{d}}})$ and (46c). Substituting (53)–(55) into (45), (47) and (48), we get more concise expressions for $\chi_{\mathsf{d}}$, $\Psi_{\mathsf{b}}(\cdot)$, and $\Psi'_{\mathsf{b}}(\cdot)$. Interestingly, when particularizing our results to the case with the QPSK inputs, we recover the same asymptotic MSE expression as given in [16, (7) & (8)]. More precisely, in [16], the real-valued system with BPSK signal was considered. In such case, $\sqrt{2}r_b$ in our paper should be replaced by $r_b$.

**Example 5** (Pilot-Only Scheme). In the conventional pilot-only scheme, the receiver solely uses $\tilde{\mathbf{Y}}_{\mathsf{t}}$ and $\mathbf{X}_{\mathsf{t}}$ to generate an estimate of $\mathbf{H}$ and subsequently uses the estimated channel for estimating the data $\mathbf{X}_{\mathsf{d}}$ from $\tilde{\mathbf{Y}}_{\mathsf{d}}$ [18]. The analysis of the asymptotic MSE w.r.t. $\mathbf{H}$ is the same as that in Example 4 but the roles of $\mathbf{H}$ and $\mathbf{X}_{\mathsf{t}}$ are exchanged. Specifically, during the t-phase, we have $\beta_{\mathsf{d}} = 0$ and $\mathsf{mse}_{X_{\mathsf{t}}} = 0$ because no data symbol is involved and the pilot matrix $\mathbf{X}_{\mathsf{t}}$ is known. After substituting these parameters into (46) and simplification, we obtain the following *self-contained* fixed-point equations

$$\mathsf{mse}_H = \frac{\sigma_h^2}{1 + \sigma_h^2 \tilde{q}_H}, \tag{56}$$

$$\tilde{q}_H = \beta_{\mathsf{t}} \sigma_{x_{\mathsf{t}}}^2 \chi_{\mathsf{t}} \tag{57}$$

with

$$\chi_{\mathsf{t}} = \sum_{b=1}^{2^{\mathsf{B}}} \int \mathrm{D}v \frac{\left(\Psi'_b\left(\sqrt{\sigma_{x_{\mathsf{t}}}^2(\sigma_h^2 - \mathsf{mse}_H)}v\right)\right)^2}{\Psi_b\left(\sqrt{\sigma_{x_{\mathsf{t}}}^2(\sigma_h^2 - \mathsf{mse}_H)}v\right)}. \tag{58}$$

Here, $\mathsf{mse}_H$ in (56) represents the asymptotic MSE w.r.t. $\mathbf{H}$ for the pilot-only scheme, which is also the MSE w.r.t. $H$ for the scalar AWGN channel (36b). Recall that $\tilde{q}_H$ serves as the SNR of the AWGN channel (36b). Comparing $\tilde{q}_H$ in (57) with that in (46a), we realize that the second term of (46a) is the gain due to data-aided channel estimation.

Before proceeding with the analysis of data estimation in the pilot-only scheme, we provide the following proposition to get a better understanding on $\mathsf{mse}_H$ in (56).

*Proposition 3:* Let the channel gain and the transmit pilot power be normalized, i.e., $\sigma_h^2 = 1$ and $\sigma_{x_{\mathsf{t}}}^2 = 1$. In the high-SNR regime and $\beta_{\mathsf{t}} = T_{\mathsf{t}}/K \gg 1$, $\mathsf{mse}_H$ of the pilot-only scheme can be approximately expressed as

$$\mathsf{mse}_H \approx -20\log_{10}(\beta_{\mathsf{t}}) + C_{\mathsf{B}} \ (\mathrm{dB}), \tag{59}$$

where $C_{\mathsf{B}}$ is a quantizer-dependent (e.g., $\Delta$ and B) constant.

*Proof:* See Appendix D. □

As an example, Table I provides the corresponding value of $C_{\mathsf{B}}$ for a uniform B-bit quantizer with $\Delta = \sqrt{\mathsf{B}}2^{-\mathsf{B}}$. In this case, we plot the MSEs that use the approximate expression (59) as well as its analytical form (56) in Figure 4. We see that for $\beta_{\mathsf{t}} > 2$, the approximation (59) matches the theoretical result (56) perfectly. Interestingly from Table I, the constant $C_{\mathsf{B}}$ satisfies $C_{\mathsf{B}} \approx -6.02\mathsf{B} + 4.4895$ in high resolution cases, indicating that $\mathsf{mse}_H$ decreases 6 dB for each 1-bit increase of rate. Interestingly, this property coincides with the well-

TABLE I
$C_{\mathsf{B}}$ FOR UNIFORM B-BIT QUANTIZER WITH $\Delta = \sqrt{\mathsf{B}}2^{-\mathsf{B}}$.

| B | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $C_{\mathsf{B}}$ (in dB) | 2.8731 | $-5.9852$ | $-13.0201$ | $-19.4804$ | $-25.7065$ | $-31.8265$ | $-37.6547$ |

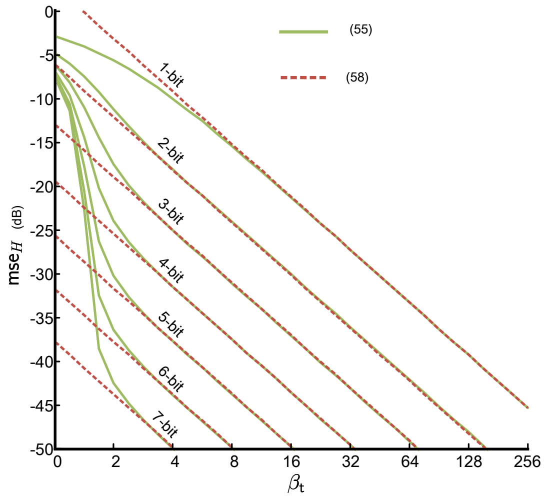


Fig. 4. The asymptotic MSE w.r.t. $\mathbf{H}$ of the pilot-only scheme versus the pilot ratio $\beta_{\mathsf{t}} = T_{\mathsf{t}}/K$ for different B-bit quantizer.

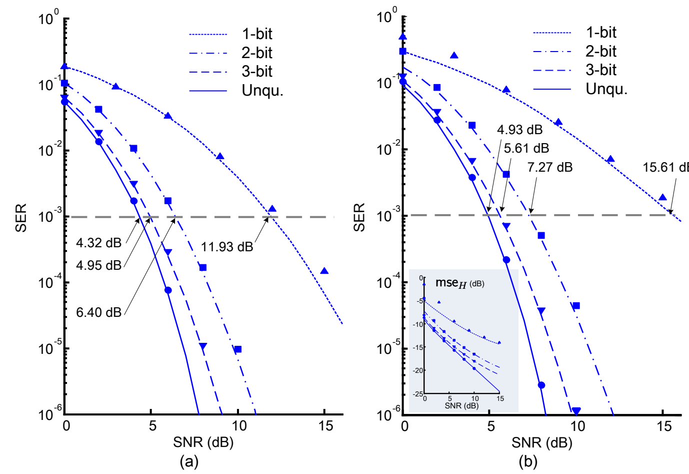


Fig. 5. SER versus SNR for QPSK constellations. In the results, the JCD estimation scheme is used under the settings with a) perfect CSIR and b) no CSIR. Curves denote analytical results and markers denote Monte-Carlo simulation results achieved by the GAMP-based JCD algorithm. The MSEs w.r.t. $\mathbf{H}$ of the JCD estimator are plotted as a subfigure.

known figure of merit in quantization.[8] In addition, from (59), given a fixed quantizer (i.e., fixed $C_{\mathsf{B}}$), $\mathsf{mse}_H$ improves 6 dB for each doubling of training length $\beta_{\mathsf{t}}$. Consequently, doubling the length for training plays the same effect as increasing an extra bit on every ADC at the massive MIMO receiver.

The proceeding observation provides a useful guideline on the trade-off between the training length and the ADC word length. For instance, if we target $\beta_{\mathsf{t}}$ to that attained by the pilot-only scheme at $\mathsf{mse}_H = -30$dB, the 4-bit receiver requires $\beta_{\mathsf{t}} = 4$ from Figure 4. If we intend to reduce the ADC word length of each ADC to 1-bit, the training length increases $2^{4-1} = 8$ times than that in the 4-bit case. This argument shows the special importance of the JCD technique in the *quantized* MIMO system. With the JCD technique, the estimated payload data are utilized to aid channel estimation so that the *effective* training length virtually increases.

Next, we return to the analysis of data estimation. If the channel estimate is subsequently used for data estimation via the Bayes-optimal approach, we can get the corresponding *self-contained* fixed-point equations for the d-phase similar to (56)-(57). Specifically, we have (46c) given a *fixed* $q_H = \sigma_h^2 - \mathsf{mse}_H$ with $\mathsf{mse}_H$ given by (56). Notice that since there is no iteration processing between the pilots and data symbols, (46a) and (46b) are not involved in the d-phase. If the JCD technique is employed, $\mathsf{mse}_H$ can be further reduced. Any reduction in channel estimation error $\mathsf{mse}_H$ results in a increasing of $q_H$ and thus increases $\tilde{q}_{X_{\mathsf{d}}} = \alpha q_H \chi_{\mathsf{d}}$.

## V. Discussions and Numerical Results

### A. Accuracy of Analytical Results

Computer simulations are carried out to verify the accuracy of our analytical results. In particular, we compare the SER expression (52) and the analytical MSE w.r.t. $\mathbf{H}$ (51) with those obtained by simulations. The simulation results are obtained by averaging over 10,000 channel realizations wherein the GAMP-based JCD algorithm (Algorithm 1) are performed with tolerance $\epsilon = 10^{-8}$ and the maximum number of iterations $\xi_{\max} = 100$.[9] The system parameters are set as follows: $K = 50$, $N = 200$, $T_{\mathsf{t}} = 50$, and $T_{\mathsf{d}} = 450$. The SNR of the system is defined by $\mathrm{SNR} = 1/\sigma_w^2$. The pilot matrix $\mathbf{X}_{\mathsf{t}} \in \mathbb{C}^{K\times T_{\mathsf{t}}}$ consists of statistically independent QPSK constellations. In the simulations, we use the typical uniform quantizer with a fixed quantization step-size $\Delta = 1/2$. Note that this quantization step-size is not optimal. The optimal step-size will be discussed in the next subsection. As QPSK constellations are used for data transmission, Figure 5 shows the corresponding SER results for the cases of 1) perfect CSIR and b) no CSIR. The corresponding MSEs w.r.t. $\mathbf{H}$ of the JCD estimator are plotted as a subfigure in Figure 5(b). We observe that the GAMP-based JCD algorithm can generally achieve the performances of the *theoretical* Bayes-optimal estimator

[8]The property of 6 dB improvement in signal-to-quantization-noise ratio for each extra bit is a well-known figure of merit in the ADC literature [52, p.248].

[9]Due to space limitation, we do not show the convergence of the GAMP-based JCD algorithm. In most cases, the GAMP-based JCD algorithm converges after 20–30 iterations although it shows very slow convergence at low SNRs.

whose performances can be described by our analytical expressions. Note that the GAMP-based JCD algorithm is only an approximation to the Bayes-optimal JCD estimator whose implementation is prohibitive. For the case with no CSIR, the GAMP-based JCD algorithm cannot work as well as that predicted by the analytical result at low SNRs. This is because at low SNRs, the GAMP-based JCD algorithm shows very slow convergence so that the adopted maximum number of iterations is not sufficient.[10] This gap has motivated the search for other improved estimators in the future.

From Figure 5(b), we see that the performance degradation due to low-precision quantization is small. For instance, if we target the SNR to that attained by the unquantized system at BER$= 10^{-3}$, the 3-bit Bayes-optimal JCD estimator only incurs a loss of $4.98 - 4.40 = 0.58\,$dB. Even with 2-bit quantization, the loss of $6.59 - 4.40 = 2.19\,$dB remains acceptable.

### B. Optimal Step-Size

In the one-bit ADC (i.e., $\mathsf{B} = 1$), the quantization output is assigned the value $\frac{\Delta}{2}$ if the input is a positive number and $-\frac{\Delta}{2}$ otherwise. For the Bayes-optimal estimator, the performances are *irrelevant* to any particular value of $\Delta$.[11] This property can be easily realized by reviewing the likelihood in (8), wherein $r_b = \{-\infty, 0, \infty\}$ for $b = 0, 1, 2$. Clearly, $\Delta$ is not involved at the very beginning of the estimate development. Therefore, we shall focus on the cases with $\mathsf{B} > 1$.

Recall that $Y_{nt}$ is the input signal to the quantizer. Direct application of the central limit theorem results in that $Y_{nt}$ can be approximated as Gaussian distribution with variance $\mathsf{E}\{|Y_{nt}|^2\} = 1 + \sigma_w^2$. For a Gaussian signal with unit variance, the optimal step size for minimizing the quantization distortion is computed in [53] and is $1.008/\sqrt{2} \approx 0.7128$ if $\mathsf{B} = 2$.[12] Under the same setting as previously, i.e., $\alpha = N/K = 4$, $\beta = T/K = 10$, $\beta_{\mathsf{t}} = T_{\mathsf{t}}/K = 1$, Figure 6 gives the BERs of the Bayes-optimal estimator as a function of the *normalized* step size $\Delta/\sqrt{\mathsf{E}\{|Y_{nt}|^2\}}$ for $\mathsf{B} = 2$. It turns out that the step size optimized in terms of the BER for the Bayes-optimal estimator is quite different from that for minimizing its distortion.

Figure 7 shows the optimal step sizes for different input signals $\mathbf{X}_{\mathsf{d}}$ including QPSK, 16QAM, 64AM, and Gaussian inputs. The optimal step size seems to vary slightly for different input signals, while all become smaller with increasing SNR. We observe from other simulations that the optimal step size varies only very slightly for different setting of $\alpha$ and $\beta$. We thus conclude that the optimal step size for the Bayes-optimal estimator is mainly dominated by the SNR.

To get a general expression, we fit the optimal step sizes for different input signals by a first degree polynomial equation

$$\Delta_{\mathsf{opt}}(\mathsf{snr}_{\mathsf{dB}}) = a_0 + a_1 \mathrm{snr}_{\mathsf{dB}}, \tag{60}$$

where $\mathsf{snr}_{\mathsf{dB}}$ represents SNR in dB scale, and the (least-squares fit) coefficients $a_0, a_1$ are listed in Table II. The optimal step

[10] At low SNRs, we observe a good result by increasing the maximum number of iterations.

[11] This property is not true for other estimators such as linear estimators [54].

[12] The optimal step size [53] is divided by $\sqrt{2}$ here because the signal power of the real or imaginary part is $1/\sqrt{2}$.

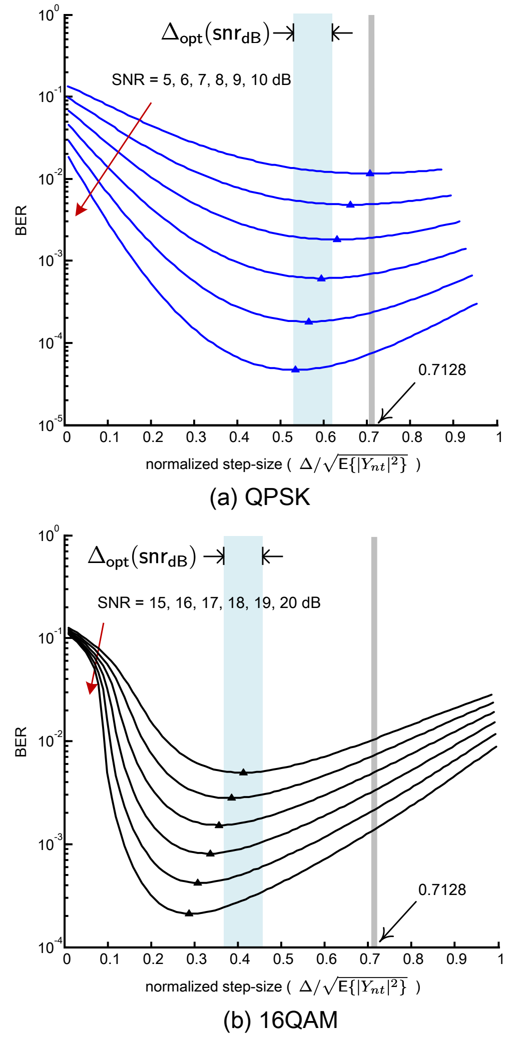


Fig. 6. BERs versus the normalized step size under the quantized MIMO system with a) QPSK and b) 16QAM constellations for $\alpha = 4$, $\beta = 10$, $\beta_{\mathsf{t}} = 1$. Markers correspond to the lowest BER w.r.t. the normalized step size. The optimal step size determined by minimizing the distortion of a Gaussian signal [53], i.e., $\Delta = 0.7128$, is plotted as the vertical axis.

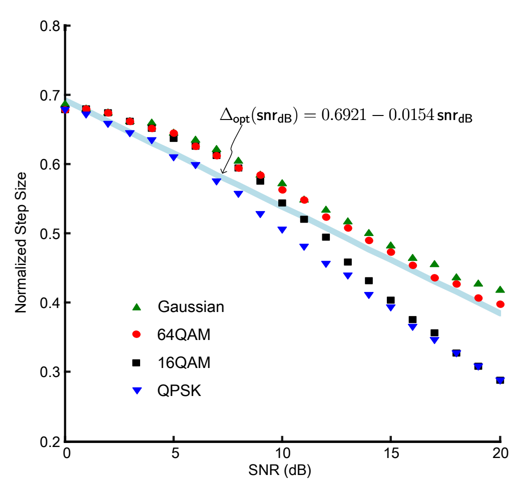


Fig. 7. The optimal step size (normalized by $\mathsf{E}\{|Y_{nt}|^2\}$).

TABLE II
THE COEFFICIENTS $a_0$ AND $a_1$ OF $\Delta_{\text{opt}}(\text{snr}_{\text{dB}})$ FOR $\mathsf{B} = 2, 3, 4$.

| B | $a_0$ | $a_1$ |
|---|---|---|
| 2 | 0.6921 | −0.0154 |
| 3 | 0.4364 | −0.0118 |
| 4 | 0.2559 | −0.0071 |

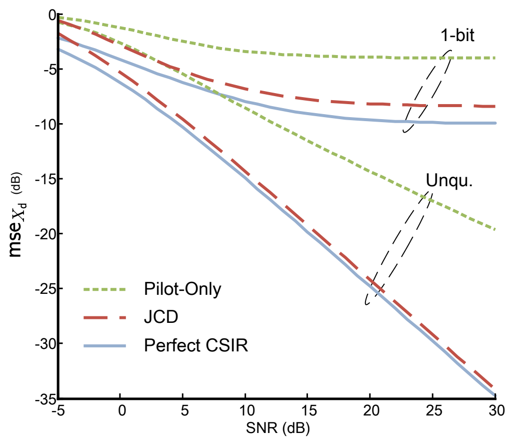


Fig. 8. $\text{mse}_{X_\text{d}}$ versus SNR for the pilot-only scheme and the Bayes-optimal JCD estimator with and without perfect CSIR under the 1-bit quantization and unquantized receivers. $\alpha = N/K = 4$, $\beta = T/K = 10$, $\beta_\text{t} = T_\text{t}/K = 1$, and $X_{\text{d},kt} \sim \mathcal{N}_\mathbb{C}(0, 1)$.

sizes determined by $\Delta_{\text{opt}}$ are also indicated by a shadow drawn in Figure 6. We observe that although $\Delta_{\text{opt}}$ is not optimal for each specific input, their corresponding performances remain affordable. Following the same argument above, we find the corresponding polynomial equation $\Delta_{\text{opt}}(\text{snr}_{\text{dB}})$ for different quantization bits, with their coefficients listed in Table II.

### *C. Effects Due to Absence of CSIR*

Comparing Figures 5(a) and 5(b), we see that the loss due to no CSIR is small for the Bayes-optimal JCD estimator. To have a better understanding on the effects of CSIR over the quantized MIMO system, we then discuss the performances of the Bayes-optimal JCD estimator *with* and *without* the perfect CSIR under various system settings. Unlike the QPSK signals used in pervious simulations, we focus on the Gaussian inputs, i.e., $X_\text{d} \sim \mathcal{N}_\mathbb{C}(0, 1)$, in the following experiments. The other system parameters are the same as before, i.e., $\alpha = N/K = 4$, $\beta = T/K = 10$, $\beta_\text{t} = T_\text{t}/K = 1$. Showing in Figure 8 is the asymptotic MSE $\text{mse}_{X_\text{d}}$ for the Bayes-optimal JCD estimator with and without perfect CSIR. Also shown is the MSE for the pilot-only scheme. It can be seen that the Bayes-optimal JCD estimator shows a large improvement over the pilot-only scheme in both the 1-bit and unquantized cases. The gap between the Bayes-optimal JCD estimator with and without perfect CSIR is very small in the unquantized case while the gap is enlarged in the case of a 1-bit quantizer. By observing the 1-bit and unquantized cases, we can expect that the gap can be smaller with increasing the ADC resolution.

A straightforward way to reduce the gap of the 1-bit case

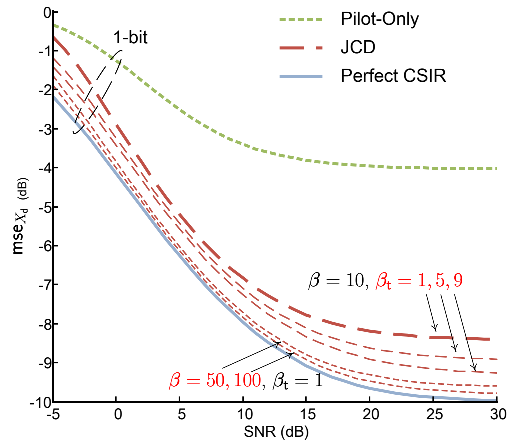


Fig. 9. $\text{mse}_{X_\text{d}}$ versus SNR for the Bayes-optimal JCD estimator with 1-bit receivers under various setting of $\beta$ and $\beta_\text{t}$. $\alpha = 4$ and $X_{\text{d},kt} \sim \mathcal{N}_\mathbb{C}(0, 1)$.

is increasing the training length. To verify this intuition, we provide the MSE results in Figure 9 for $\beta_\text{t} = 5$ and $\beta_\text{t} = 9$. However, the improvement by increasing the training length is limited even if $\beta_\text{t} = 9$, leaving only $\beta_\text{d} = 1$ for data. Alternatively, we may consider a greater $\beta$ although it may not be a controllable factor. Note that $\beta$ is determined by the coherent time. If $\beta = 100$ and $\beta_\text{t} = 1$, the Bayes-optimal JCD estimator without perfect CSIR can perform very close to the (fundamental limit) case with perfect CSIR. Nevertheless, such a long coherent time is usually unavailable in practice.

## VI. CONCLUSION

We developed a framework for studying the *best* possible estimation performance of the quantized MIMO system, namely, the massive MIMO system with very low-precision ADCs. In particular, we used the Bayes-optimal inference for the JCD estimation and achieve this estimation by applying the BiG-AMP technique. The asymptotic performances (e.g., MSEs and SERs) w.r.t. the channels and the payload data were derived and shown as simply characterized by scalar AWGN channels. Monte-Carlo simulations were conducted to demonstrate the accuracy of our analytical results.

The high accuracy of the analytical expressions enable us to quickly and efficiently assess the performance of the quantized MIMO system. Thus, we obtained the following useful guidelines for the system design:

- We showed that the asymptotic MSE of the channel estimate in the conventional pilot-only scheme decreases by approximately 6 dB for each bit added to the ADCs or each doubling of training length. This finding supports the importance of the JCD technique, especially in the *quantized* MIMO system.
- The optimal step size for minimizing BERs of the Bayes-optimal estimator were shown to be highly different from that for minimizing the distortion of a Gaussian signal and, fortunately, can be quickly determined by our analytical expressions.
- The Bayes-optimal estimator already exhibits the best possible estimation performance. Even so, we showed

that the performance gap between the Bayes-optimal JCD estimator with and without perfect CSIR still cannot be negligible in the *quantized* MIMO system. We also discussed the ways to reduce the gap and then concluded that achieving the same performance as the full CSIR case in the quantized MIMO system is very difficult.

Many potential directions for future work are available. The GAMP-based JCD algorithm presented in this paper is a first step toward achieving the *optimal* JCD estimate under the quantized MIMO system. The computational complexity of the GAMP-based JCD algorithm may still be too high to be affordable in a commercial system. One possible solution is to adopt other suboptimal schemes such as linear estimators. Another feasible solution is using mixed-ADC receiver architecture [12] wherein a small number of high-resolution ADCs is available. Thus, CSIR gains high accuracy and facilitates the JCD procedure. For a development in this direction, see [54].

## Appendix A: Derivations of (23) and (24)

In this appendix, we derive the expressions (23) and (24), by applying the techniques in [33, Chapter 3.9]. The derivations below are only dedicated for the real part of the estimator because the imaginary part of the estimator can be obtained analogously. Note that the signal power and noise power are $v^p/2$ and $\sigma_w^2/2$, respectively, per real and imaginary part. For ease of notation, we have abused $\widetilde{Y}$, $Z$, and $\hat{p}$ to denote $\mathrm{Re}(\widetilde{Y})$, $\mathrm{Re}(Z)$, and $\mathrm{Re}(\hat{p})$, respectively.

To get (23), we begin by deriving the denominator of (17). First, recall from (8) that if $\widetilde{Y} \in (r_{b-1}, r_b]$ and $\widetilde{Y} \le 0$, the likelihood is given by

$$\mathsf{P}_{\mathsf{out}}(\widetilde{Y}|Z) = \Phi\left(\frac{r_b - Z}{\sqrt{\sigma_w^2/2}}\right) - \Phi\left(\frac{r_{b-1} - Z}{\sqrt{\sigma_w^2/2}}\right). \tag{61}$$

Note that for the special case $b = 1$, we have $r_0 = -\infty$, and the second term of (61) will disappear. Substituting (61) into the denominator of (17), it can be shown that[13]

$$\int \mathsf{P}_{\mathsf{out}}(\widetilde{Y}|z)\mathcal{N}(z;\hat{p}, v^p/2)\,\mathrm{d}z = \Phi\left(\frac{\mathsf{sign}(\widetilde{Y})\hat{p} - |r_b|}{\sqrt{(\sigma_w^2 + v^p)/2}}\right) - \Phi\left(\frac{\mathsf{sign}(\widetilde{Y})\hat{p} - |r_{b-1}|}{\sqrt{(\sigma_w^2 + v^p)/2}}\right) \triangleq C. \tag{62}$$

Differentiating w.r.t. $\hat{p}$ on both sides of (62) yields

$$\int \left(\frac{z - \hat{p}}{v^p/2}\right)\mathsf{P}_{\mathsf{out}}(\widetilde{Y}|z)\mathcal{N}(z;\hat{p}, v^p/2)\,\mathrm{d}z = \frac{\mathsf{sign}(\widetilde{Y})}{\sqrt{(\sigma_w^2 + v^p)/2}}\left(\phi\left(\frac{\mathsf{sign}(\widetilde{Y})\hat{p} - |r_b|}{\sqrt{(\sigma_w^2 + v^p)/2}}\right) - \phi\left(\frac{\mathsf{sign}(\widetilde{Y})\hat{p} - |r_{b-1}|}{\sqrt{(\sigma_w^2 + v^p)/2}}\right)\right), \tag{63}$$

where we have used $\partial\Phi(x)/\partial\hat{p} = \phi(x)\partial x/\partial\hat{p}$. Using (62), (63) can be rearranged as

$$\int z\,\mathsf{P}_{\mathsf{out}}(\widetilde{Y}|z)\mathcal{N}(z;\hat{p}, v^p/2)\,\mathrm{d}z = \hat{p}C + \frac{\mathsf{sign}(\widetilde{Y})v^p}{\sqrt{2(\sigma_w^2 + v^p)}}\left(\phi(\eta_1) - \phi(\eta_2)\right), \tag{64}$$

where $\eta_2$ and $\eta_1$ are given by (25). Multiplying both sizes by $1/C$, we obtain the marginal posterior mean given in (23).

Similarly, (24) can be calculated by differentiating (62) twice as

$$\int \left(\frac{z^2}{(v^p/2)^2} - \frac{2\hat{p}z}{(v^p/2)^2} + \frac{\hat{p}^2}{(v^p/2)^2} - \frac{1}{v^p/2}\right) \times \mathsf{P}_{\mathsf{out}}(\widetilde{Y}|z)\mathcal{N}(z;\hat{p}, v^p/2)\,\mathrm{d}z = \frac{-1}{\sqrt{(\sigma_w^2 + v^p)/2}}\Big(\eta_1\phi(\eta_1) - \eta_2\phi(\eta_2)\Big), \tag{65}$$

which then can be rearranged as

$$\mathsf{E}\big\{Z^2\big|\,\hat{p}, v^p/2\big\} = 2\hat{p}\,\mathsf{E}\big\{Z\big|\,\hat{p}, v^p/2\big\} + \big(v^p/2 - \hat{p}^2\big) - \frac{1}{C}\frac{(v^p)^2}{\sqrt{2(\sigma_w^2 + v^p)}}\Big(\eta_1\phi(\eta_1) - \eta_2\phi(\eta_2)\Big). \tag{66}$$

We also note that

$$\mathsf{Var}\big\{Z\big|\,\hat{p}, v^p/2\big\} = \mathsf{E}\big\{Z^2\big|\,\hat{p}, v^p/2\big\} - \big(\mathsf{E}\big\{Z\big|\,\hat{p}, v^p/2\big\}\big)^2. \tag{67}$$

After plugging (66) and (23) into (67), we obtain (24). In the above derivations, we have assumed $\widetilde{Y} \le 0$. For $\widetilde{Y} > 0$, the above derivations can be used in the same way.

## Appendix B: Proof of Proposition 1

Using the replica method, we first compute the replicate partition function $\mathsf{E}_{\widetilde{\mathbf{Y}}}\{\mathsf{P}^\tau(\mathbf{Y})\}$ in (35), which with the definition of (11) can be expressed as

$$\mathsf{E}_{\widetilde{\mathbf{Y}}}\Big\{\mathsf{P}^\tau(\widetilde{\mathbf{Y}})\Big\} = \mathsf{E}_{\boldsymbol{\mathcal{H}},\boldsymbol{\mathcal{X}}}\left\{\int \mathrm{d}\widetilde{\mathbf{Y}}\prod_{a=0}^{\tau}\mathsf{P}_{\mathsf{out}}\left(\widetilde{\mathbf{Y}}\,\middle|\,\mathbf{Z}^{(a)}\right)\right\}, \tag{68}$$

where we define $\mathbf{Z}^{(a)} \triangleq \mathbf{H}^{(a)}\mathbf{X}^{(a)}/\sqrt{K}$ with $\mathbf{H}^{(a)}$ and $\mathbf{X}^{(a)}$ being the $a$-th replica of $\mathbf{H}$ and $\mathbf{X}$, respectively, and $\boldsymbol{\mathcal{X}} \triangleq \{\mathbf{X}^{(a)}, \forall a\}$ and $\boldsymbol{\mathcal{H}} \triangleq \{\mathbf{H}^{(a)}, \forall a\}$. Here, $(\mathbf{H}^{(a)}, \mathbf{X}^{(a)})$ are random matrices taken from the distribution $(\mathsf{P}_{\mathsf{H}}, \mathsf{P}_{\mathsf{X}})$ for $a = 0, 1, \ldots, \tau$. In addition, $\int \mathrm{d}\widetilde{\mathbf{Y}}$ denotes the integral w.r.t. a discrete measure because the quantized output $\widetilde{\mathbf{Y}}$ is a finite set. We will calculate the right-hand side of (68), by applying the techniques in [38, 39] after additional manipulations.[14]

To average over $(\boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{X}})$, we introduce two $(\tau+1)\times(\tau+1)$ matrices $\mathbf{Q}_H = [Q_H^{ab}]$ and $\mathbf{Q}_{X_o} = [Q_{X_o}^{ab}]$ for $o \in \{\mathsf{t}, \mathsf{d}\}$ whose elements are defined by $Q_H^{ab} = \mathbf{h}_n^{(b)}(\mathbf{h}_n^{(a)})^\dagger/K$ and $Q_{X_o}^{ab} = (\mathbf{x}_t^{(a)})^\dagger\mathbf{x}_t^{(b)}/K$ for $t \in \mathcal{T}_o$, where, $\mathbf{h}_n^{(a)}$ is the $n$th row vector of $\mathbf{H}^{(a)}$, and $\mathbf{x}_t^{(a)}$ is the $t$th column vector of $\mathbf{X}^{(a)}$ for $t \in \mathcal{T}_{\mathsf{t}}$ or

[13] The calculation can be done by using the Gaussian reproduction property given by footnote 5.

[14] Details are omitted here. The interested reader is referred to the longer version of this paper from ArXiv.

$\mathcal{T}_{\mathsf{d}}$. The definitions of $\mathbf{Q}_H$ and $\mathbf{Q}_{X_{\mathsf{d}}}$ are equivalent to

$$1=\int\prod_{n=1}^{N}\prod_{0\le a\le b}^{\tau}\delta\Big(\mathbf{h}_n^{(b)}(\mathbf{h}_n^{(a)})^{\dagger}-KQ_H^{ab}\Big)\mathrm{d}Q_H^{ab},$$

$$1=\int\prod_{o\in\{\mathsf{t},\mathsf{d}\}}\prod_{t\in\mathcal{T}_o}\prod_{0\le a\le b}^{\tau}\delta\Big((\mathbf{x}_t^{(a)})^{\dagger}\mathbf{x}_t^{(b)}-KQ_{X_o}^{ab}\Big)\mathrm{d}Q_{X_o}^{ab},$$

where $\delta(\cdot)$ denotes Dirac's delta. Let $\boldsymbol{\mathcal{Q}}_X\triangleq\{\mathbf{Q}_{X_o},\forall o\}$ and $\boldsymbol{\mathcal{Z}}\triangleq\{\mathbf{Z}^{(a)},\forall a\}$. Inserting the above into (68) yields

$$\mathsf{E}_{\widetilde{\mathbf{Y}}}\{\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\}=\int e^{K^2\mathcal{G}^{(\tau)}}\mathrm{d}\mu_H^{(\tau)}(\mathbf{Q}_H)\mathrm{d}\mu_X^{(\tau)}(\boldsymbol{\mathcal{Q}}_X),\qquad(69)$$

where

$$\mathcal{G}^{(\tau)}(\boldsymbol{\mathcal{Q}}_Z)\triangleq\frac{1}{K^2}\log\mathsf{E}_{\boldsymbol{\mathcal{Z}}}\bigg\{\int\mathrm{d}\widetilde{\mathbf{Y}}\prod_{a=0}^{\tau}\mathsf{P}_{\mathsf{out}}\Big(\widetilde{\mathbf{Y}}\Big|\mathbf{Z}^{(a)}\Big)\bigg\},$$

$$\mu_H^{(\tau)}(\mathbf{Q}_H)\triangleq\mathsf{E}_{\boldsymbol{\mathcal{H}}}\bigg\{\prod_{n=1}^{N}\prod_{0\le a\le b}^{\tau}\delta\Big(\mathbf{h}_n^{(b)}(\mathbf{h}_n^{(a)})^{\dagger}-KQ_H^{ab}\Big)\bigg\},$$

$$\mu_X^{(\tau)}(\boldsymbol{\mathcal{Q}}_X)\triangleq\mathsf{E}_{\boldsymbol{\mathcal{X}}}\bigg\{\prod_{o,t\in\mathcal{T}_o}\prod_{0\le a\le b}^{\tau}\delta\Big((\mathbf{x}_t^{(a)})^{\dagger}\mathbf{x}_t^{(b)}-KQ_{X_o}^{ab}\Big)\bigg\}.$$

Using the Fourier representation of the $\delta$ function via auxiliary matrices $\tilde{\mathbf{Q}}_H=[\tilde{\mathbf{Q}}_H^{ab}]\in\mathbb{C}^{(\tau+1)\times(\tau+1)}$, $\tilde{\boldsymbol{\mathcal{Q}}}_X\triangleq\{\tilde{\mathbf{Q}}_{X_o}=[\tilde{\mathbf{Q}}_{X_o}^{ab}]\in\mathbb{C}^{(\tau+1)\times(\tau+1)},\forall o\}$ and performing the saddle point method for the integration over $(\mathbf{Q}_H,\boldsymbol{\mathcal{Q}}_X)$, we attain

$$\frac{1}{K^2}\mathsf{E}_{\widetilde{\mathbf{Y}}}\{\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\}=\underset{\mathbf{Q}_H,\boldsymbol{\mathcal{Q}}_X,\tilde{\mathbf{Q}}_H,\tilde{\boldsymbol{\mathcal{Q}}}_X}{\mathsf{Extr}}\Big\{\mathcal{F}^{(\tau)}\Big\}\qquad(70)$$

with

$$\begin{aligned}\mathcal{F}^{(\tau)}\triangleq{}&\frac{1}{K^2}\log\mathsf{E}_{\boldsymbol{\mathcal{Z}}}\bigg\{\prod_{n,o,t\in\mathcal{T}_o}\int\mathrm{d}\widetilde{Y}_{nt}\prod_{a}\mathsf{P}_{\mathsf{out}}\Big(\widetilde{Y}_{nt}\Big|Z_{nt}^{(a)}\Big)\bigg\}\\&+\frac{1}{K^2}\log\mathcal{M}_H^{(\tau)}(\mathbf{Q}_H)-\alpha\mathsf{tr}\Big(\tilde{\mathbf{Q}}_H\mathbf{Q}_H\Big)\\&+\frac{1}{K^2}\log\mathcal{M}_X^{(\tau)}(\tilde{\boldsymbol{\mathcal{Q}}}_X)-\sum_o\beta_o\mathsf{tr}\Big(\tilde{\mathbf{Q}}_{X_o}\mathbf{Q}_{X_o}\Big),\end{aligned}\qquad(71)$$

where $\mathsf{Extr}_x\{f(x)\}$ denotes the extreme value of $f(x)$ w.r.t. $x$;

$$\mathcal{M}_H^{(\tau)}(\tilde{\mathbf{Q}}_H)\triangleq\mathsf{E}_{\boldsymbol{\mathcal{H}}}\bigg\{\prod_{n=1}^{N}e^{\mathsf{tr}\left(\tilde{\mathbf{Q}}_H\mathbf{H}_n^H\mathbf{H}_n\right)}\bigg\},$$

$$\mathcal{M}_X^{(\tau)}(\tilde{\boldsymbol{\mathcal{Q}}}_X)\triangleq\mathsf{E}_{\boldsymbol{\mathcal{X}}}\bigg\{\prod_{o\in\{\mathsf{t},\mathsf{d}\}}e^{\mathsf{tr}\left(\tilde{\mathbf{Q}}_{X_o}\mathbf{X}_o^H\mathbf{X}_o\right)}\bigg\},$$

$\mathbf{H}_n^H\triangleq[\mathbf{h}_n^{(0)T}\,\mathbf{h}_n^{(1)T}\cdots\mathbf{h}_n^{(\tau)T}]^T$, $\mathbf{X}_o\triangleq[\mathbf{x}_o^{(0)}\,\mathbf{x}_o^{(1)}\cdots\mathbf{x}_o^{(\tau)}]$. According to (35), the average free entropy turns out to be $\mathcal{F}=\lim_{\tau\to0}\frac{\partial}{\partial\tau}\mathsf{Extr}_{\mathbf{Q}_H,\boldsymbol{\mathcal{Q}}_X,\tilde{\mathbf{Q}}_H,\tilde{\boldsymbol{\mathcal{Q}}}_X}\left\{\mathcal{F}^{(\tau)}\right\}$.

The saddle points of $\mathcal{F}^{(\tau)}$ can be found by the point of zero gradient w.r.t. $\{\mathbf{Q}_H,\mathbf{Q}_{X_o},\tilde{\mathbf{Q}}_H,\tilde{\mathbf{Q}}_{X_o}\}$ but it is still prohibitive to get explicit expressions about the saddle points. Thus, we assume that the saddle points follow the RS form [39] as $\mathbf{Q}_H=(c_H-q_H)\mathbf{I}+q_H\mathbf{1}\mathbf{1}^T$, $\tilde{\mathbf{Q}}_H=(\tilde{c}_H-\tilde{q}_H)\mathbf{I}+\tilde{q}_H\mathbf{1}\mathbf{1}^T$, $\mathbf{Q}_{X_o}=(c_{X_o}-q_{X_o})\mathbf{I}+q_{X_o}\mathbf{1}\mathbf{1}^T$, and $\tilde{\mathbf{Q}}_{X_o}=(\tilde{c}_{X_o}-\tilde{q}_{X_o})\mathbf{I}+\tilde{q}_{X_o}\mathbf{1}\mathbf{1}^T$. In addition, the application of the central limit theorem suggests that $\mathbf{z}_{nt}\triangleq[Z_{nt}^{(0)}\,Z_{nt}^{(1)}\cdots Z_{nt}^{(\tau)}]^T$ are Gaussian random vectors with $(\tau+1)\times(\tau+1)$ covariance matrix $\mathbf{Q}_{Z_t}$. If $t\in\mathcal{T}_o$, then the $(a,b)$th entry of $\mathbf{Q}_{Z_o}$ is given by

$$(Z_{nt}^{(a)})^*Z_{nt}^{(b)}=Q_H^{ab}Q_{X_o}^{ab}\triangleq Q_{Z_o}^{ab}.\qquad(72)$$

As such, we set $\mathbf{Q}_{Z_o}=(c_Hc_{X_o}-q_Hq_{X_o})\mathbf{I}+q_Hq_{X_o}\mathbf{1}$, which is equivalent to introducing to the Gaussian random variable $\mathbf{z}_{nt}$ for $t\in\mathcal{T}_o$ as $Z_{nt}^{(a)}=\sqrt{c_Hc_{X_o}-q_Hq_{X_o}}\,u_c^{(a)}+\sqrt{q_Hq_{X_o}}v_c$, for $a=0,\dots\tau$, where $u_c^{(a)}$ and $v_c$ are independent standard complex Gaussian random variables.

With RS, the problem of seeking the extremum w.r.t.

$$\{\mathbf{Q}_H,\mathbf{Q}_{X_o},\tilde{\mathbf{Q}}_H,\tilde{\mathbf{Q}}_{X_o}\}$$

is reduced to seeking the extremum over

$$\{c_H,q_H,c_{X_o},q_{X_o},\tilde{c}_H,\tilde{q}_H,\tilde{c}_{X_o},\tilde{q}_{X_o}\},$$

which can be obtained by equating the corresponding partial derivatives of the RS expression $\mathcal{F}^{(\tau)}$ to zero. In doing so, as $\tau\to0$, it is easy to get that $\tilde{c}_H=0$, $\tilde{c}_{X_o}=0$, $c_H=\mathsf{E}\{|H|^2\}$, and $c_{X_o}=\mathsf{E}\{|X_o|^2\}$. Also, we obtain the fixed-point equations given in (46). Finally, taking the partial derivatives of $\mathcal{F}^{(\tau)}$ at $\tau=0$, and applying the parameters introduced in Section IV-A, we obtain (44).

## Appendix C: Proof of Proposition 2

Consider the $(n,k)$-th and $(k,t)$-th entries of $\mathbf{H}$ and $\mathbf{X}_{\mathsf{d}}$, respectively. We will show that the joint moments of the joint distribution of $(H_{nk},X_{\mathsf{d},kt},\widehat{H}_{nk},\widehat{X}_{\mathsf{d},kt})$ for some indices $(n,k)$ and $(k,t)$ converges to the joint distribution of

$$\mathsf{P}(H)\mathsf{P}(Y_H|H)\mathsf{P}(H|Y_H)\mathsf{P}(X_{\mathsf{d}})\mathsf{P}(Y_{X_{\mathsf{d}}}|X_{\mathsf{d}})\mathsf{P}(X_{\mathsf{d}}|Y_{X_{\mathsf{d}}}),\quad(73)$$

independent of $(n,k)$ and $(k,t)$. Following [43], we proceed to calculate the joint moments

$$\begin{aligned}&\mathsf{E}\{\mathrm{Re}(H_{nk})^{i_{\mathrm{R}_h}}\mathrm{Im}(H_{nk})^{i_{\mathrm{I}_h}}\mathrm{Re}(\widehat{H}_{nk})^{j_{\mathrm{R}_h}}\mathrm{Im}(\widehat{H}_{nk})^{j_{\mathrm{I}_h}}\\&\mathrm{Re}(X_{\mathsf{d},kt})^{i_{\mathrm{R}_x}}\mathrm{Im}(X_{\mathsf{d},kt})^{i_{\mathrm{I}_x}}\mathrm{Re}(\widehat{X}_{\mathsf{d},kt})^{j_{\mathrm{R}_x}}\mathrm{Im}(\widehat{X}_{\mathsf{d},kt})^{j_{\mathrm{I}_x}}\}\end{aligned}\quad(74)$$

for arbitrary non-negative integers $i_{\mathrm{R}_h}$, $i_{\mathrm{I}_h}$, $j_{\mathrm{R}_h}$, $j_{\mathrm{I}_h}$, $i_{\mathrm{I}_x}$, $j_{\mathrm{R}_x}$, $j_{\mathrm{I}_x}$, $j_{\mathrm{I}_x}$. To proceed, we define

$$\begin{aligned}f_h={}&\sum_{n,k}\Big(\mathrm{Re}(H_{nk}^{(0)})\Big)^{i_{\mathrm{R}_h}}\Big(\mathrm{Im}(H_{nk}^{(0)})\Big)^{i_{\mathrm{I}_h}}\\&\times\Big(\mathrm{Re}(H_{nk}^{(a_{\mathrm{R}})})\Big)^{j_{\mathrm{R}_h}}\Big(\mathrm{Im}(H_{nk}^{(a_{\mathrm{I}})})\Big)^{j_{\mathrm{I}_h}},\\f_x={}&\sum_{k,t}\Big(\mathrm{Re}(X_{\mathsf{d},kt}^{(0)})\Big)^{i_{\mathrm{R}_x}}\Big(\mathrm{Im}(X_{\mathsf{d},kt}^{(0)})\Big)^{i_{\mathrm{I}_x}}\\&\times\Big(\mathrm{Re}(X_{\mathsf{d},kt}^{(b_{\mathrm{R}})})\Big)^{j_{\mathrm{R}_x}}\Big(\mathrm{Im}(X_{\mathsf{d},kt}^{(b_{\mathrm{I}})})\Big)^{j_{\mathrm{I}_x}},\end{aligned}\quad(75)$$

with $a_{\mathrm{R}},a_{\mathrm{I}}\in\{1,\dots,\tau\}$, $a_{\mathrm{R}}\neq a_{\mathrm{I}}$ and $b_{\mathrm{R}},b_{\mathrm{I}}\in\{1,\dots,\tau\}$, $b_{\mathrm{R}}\neq b_{\mathrm{I}}$. If we define the *generalized* free entropy as

$$\tilde{\mathcal{F}}=\frac{1}{K^2}\lim_{\tau\to0^+}\frac{\partial^2}{\partial\varepsilon_h\partial\varepsilon_x}\ln\mathsf{E}_{\widetilde{\mathbf{Y}}}\Big\{e^{\varepsilon_hf_h\varepsilon_xf_x}\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\Big\}\bigg|_{\varepsilon_h=0,\varepsilon_x=0},\qquad(76)$$

it exactly provides the joint moments of interest.

As $\varepsilon_h=0$ and $\varepsilon_x=0$, $\mathsf{E}_{\widetilde{\mathbf{Y}}}\{e^{\varepsilon_hf_h\varepsilon_xf_x}\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\}$ reduces to $\mathsf{E}_{\widetilde{\mathbf{Y}}}\{\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\}$ given in (68). Therefore, proceeding with the same steps as in Appendix B from (68) to (70), we get

$$\frac{1}{K^2}\mathsf{E}_{\widetilde{\mathbf{Y}}}\left\{e^{\varepsilon_h f_h \varepsilon_x f_x}\mathsf{P}^{\tau}(\widetilde{\mathbf{Y}})\right\} = \underset{\mathbf{Q}_H,\boldsymbol{\mathcal{Q}}_X,\tilde{\mathbf{Q}}_H,\tilde{\boldsymbol{\mathcal{Q}}}_X}{\mathsf{Extr}}\left\{\tilde{\mathcal{F}}^{(\tau)}\right\}, \quad (77)$$

where $\tilde{\mathcal{F}}^{(\tau)}$ is exactly identical to (71) while $\mathcal{M}_H^{(\tau)}(\tilde{\mathbf{Q}}_H)$ and $\mathcal{M}_X^{(\tau)}(\tilde{\boldsymbol{\mathcal{Q}}}_X)$ should be replaced by

$$\tilde{\mathcal{M}}_H^{(\tau)}(\tilde{\mathbf{Q}}_H) = \mathsf{E}_{\boldsymbol{\mathcal{H}}}\left\{e^{\varepsilon_h f_h}\prod_{n=1}^{N} e^{\mathrm{tr}\left(\tilde{\mathbf{Q}}_H \mathbf{H}_n^H \mathbf{H}_n\right)}\right\},$$

$$\tilde{\mathcal{M}}_X^{(\tau)}(\tilde{\boldsymbol{\mathcal{Q}}}_X) = \mathsf{E}_{\boldsymbol{\mathcal{X}}}\left\{e^{\varepsilon_x f_x}\prod_{o\in\{\mathsf{t},\mathsf{d}\}} e^{\mathrm{tr}\left(\tilde{\mathbf{Q}}_{X_o}\mathbf{X}_o^H\mathbf{X}_o\right)}\right\}.$$

Thus, except for $\tilde{\mathcal{M}}_H^{(\tau)}(\tilde{\mathbf{Q}}_H)$ and $\tilde{\mathcal{M}}_X^{(\tau)}(\tilde{\boldsymbol{\mathcal{Q}}}_X)$, the RS expressions for the other parts of $\tilde{\mathcal{F}}^{(\tau)}$ can be obtained as in Appendix B. We now only need to obtain the RS expressions for $\log\tilde{\mathcal{M}}_H^{(\tau)}(\tilde{\mathbf{Q}}_H)$ and $\log\tilde{\mathcal{M}}_X^{(\tau)}(\tilde{\boldsymbol{\mathcal{Q}}}_X)$. The generalized free energy in (76) becomes

$$\begin{aligned}\tilde{\mathcal{F}} = \int \mathrm{d}Y_H \mathrm{d}Y_{X_\mathsf{d}}\ & \mathsf{E}_H\{(\mathrm{Re}(H))^{i_{\mathrm{R}h}}(\mathrm{Im}(H))^{i_{\mathrm{I}h}}\mathsf{P}(Y_H|H)\}\\ &\times \underbrace{\mathsf{E}_H\{(\mathrm{Re}(H))^{j_{\mathrm{R}h}}(\mathrm{Im}(H))^{j_{\mathrm{I}h}}\mathsf{P}(H|Y_H)\}}_{\mathrm{Re}(\widehat{H})^{j_{\mathrm{R}h}}\,\mathrm{Im}(\widehat{H})^{j_{\mathrm{I}h}}}\\ &\times \mathsf{E}_{X_\mathsf{d}}\{(\mathrm{Re}(X_\mathsf{d}))^{i_{\mathrm{R}x}}(\mathrm{Im}(X_\mathsf{d}))^{i_{\mathrm{I}x}}\mathsf{P}(Y_{X_\mathsf{d}}|X_\mathsf{d})\}\\ &\times \underbrace{\mathsf{E}_{X_\mathsf{d}}\{(\mathrm{Re}(X_\mathsf{d}))^{i_{\mathrm{R}x}}(\mathrm{Im}(X_\mathsf{d}))^{i_{\mathrm{I}x}}\mathsf{P}(X_\mathsf{d}|Y_{X_\mathsf{d}})\}}_{\mathrm{Re}(\widehat{X}_\mathsf{d})^{i_{\mathrm{R}x}}\,\mathrm{Im}(\widehat{X}_\mathsf{d})^{i_{\mathrm{I}x}}}\end{aligned} \quad (78)$$

which is the joint moments of $(H, X_\mathsf{d}, \widehat{H}, \widehat{X}_\mathsf{d})$. Consequently, the joint moment of interest is thus uniquely determined by (78) due to the Carleman theorem.

## Appendix D: Proof of Proposition 3

In this derivation, we consider the case at infinity SNR, i.e. $\sigma_w^2 = 0$, and we let $\sigma_h^2 = 1$ and $\sigma_{x_\mathsf{t}}^2 = 1$ without loss of generality. From (56), as $\beta_\mathsf{t} \to \infty$, we have $\tilde{q}_H \to \infty$. An application of the Taylor expansion yields $1 - \mathsf{mse}_H = (1 + 1/\tilde{q}_H)^{-1} \approx 1 - 1/\tilde{q}_H$, and thus we have

$$\mathsf{mse}_H \approx 1/\tilde{q}_H. \quad (79)$$

Let $u = \frac{\sqrt{2}r_b - \sqrt{1-\mathsf{mse}_H}v}{\sqrt{\mathsf{mse}_H}}$. We then evaluate $\chi_\mathsf{t}$ in (58) by changing the integration variable from $v$ to $u$, which yields

$$\chi_\mathsf{t} = \frac{c_\mathsf{B}}{\sqrt{\mathsf{mse}_H(1-\mathsf{mse}_H)}}, \quad (80)$$

where

$$c_\mathsf{B} = \sum_{b=1}^{2^\mathsf{B}} \int \frac{e^{-\frac{(\sqrt{\mathsf{mse}_H}z - \sqrt{2}r_b)^2}{2(1-\mathsf{mse}_H)}}}{\sqrt{2\pi}} \times \frac{\left(\phi(z) - \phi\left(z - \frac{\sqrt{2}(r_b - r_{b-1})}{\sqrt{\mathsf{mse}_H}}\right)\right)^2}{\Phi(z) - \Phi\left(z - \frac{\sqrt{2}(r_b - r_{b-1})}{\sqrt{\mathsf{mse}_H}}\right)}\mathrm{d}z. \quad (81)$$

As $\mathsf{mse}_H \to 0$, $c_\mathsf{B}$ can be approximated by

$$c_\mathsf{B} \approx \frac{1}{(2\pi)^{3/2}}\sum_{b=1}^{2^\mathsf{B}} e^{-r_b^2}\int \frac{e^{-z^2}}{\Phi(z)}\mathrm{d}z, \quad (82)$$

which is a quantizer-dependent constant. Using $\tilde{q}_H = \beta_\mathsf{t}\chi_\mathsf{t}$ given in (58) and combining (79) and (80), we obtain $\mathsf{mse}_H \approx (\beta_\mathsf{t} c_\mathsf{B})^{-2}$ or (59) in dB scale, wherein $C_\mathsf{B} = -20\log_{10}(c_\mathsf{B})$. The values of $C_\mathsf{B}$ in Table I are obtained from (81) numerically.

## Appendix E: A Generalization of Proposition 1

In this Appendix, we extend Proposition 1 into the case where users have different large-scale fading factors $\sigma_{h_k}^2$. This task can be performed by proceeding with the same steps as in Appendix A, and the proof is omitted.

Similar to (36), we define the scalar AWGN channels for this case:

$$Y_{X_{\mathsf{d},k}} = \sqrt{\tilde{q}_{X_{\mathsf{d},k}}}X_{\mathsf{d},k} + W_{X_{\mathsf{d},k}}, \quad (83a)$$

$$Y_{H_k} = \sqrt{\tilde{q}_{H_k}}H_k + W_{H_k}, \quad (83b)$$

where $W_{X_{\mathsf{d},k}}, W_{H_k} \sim \mathcal{N}_\mathbb{C}(0,1)$, $X_{\mathsf{d},k} \sim \mathsf{P}_{X_\mathsf{d}}$, and $H_k \sim \mathsf{P}_{\mathsf{H}_k} \equiv \mathcal{N}_\mathbb{C}(0, \sigma_{h_k}^2)$. For ease of notation, we use $\langle a_k\rangle = \frac{1}{K}\sum_{k=1}^K a_k$ to represent the average over a set $\{a_k : k = 1,\ldots,K\}$.

*Proposition 4:* As $K \to \infty$, the asymptotic free entropy is

$$\begin{aligned}\mathcal{F} = \alpha \sum_{o\in\{\mathsf{t},\mathsf{d}\}}\beta_o\left(\sum_{b=1}^{2^\mathsf{B}}\int \mathrm{D}v\,\Psi_b(V_o)\log\Psi_b(V_o)\right)\\ - \alpha\langle I(H_k; Y_{H_k}|\tilde{q}_{H_k})\rangle - \beta_\mathsf{d}\langle I(X_{\mathsf{d},k}; Y_{X_{\mathsf{d},k}}|\tilde{q}_{X_{\mathsf{d},k}})\rangle\\ + \alpha\langle(c_{H_k} - q_{H_k})\tilde{q}_{H_k}\rangle + \sum_{o\in\{\mathsf{t},\mathsf{d}\}}\beta_o\langle(c_{X_{o,k}} - q_{X_{o,k}})\tilde{q}_{X_{o,k}}\rangle,\end{aligned} \quad (84)$$

where

$$\Psi_b(V_o) \triangleq \Phi\left(\frac{\sqrt{2}r_b - V_o}{\sqrt{\sigma_w^2 + \langle c_{H_k}c_{X_{o,k}} - q_{H_k}q_{X_{\mathsf{d},k}}\rangle}}\right) - \Phi\left(\frac{\sqrt{2}r_{b-1} - V_o}{\sqrt{\sigma_w^2 + \langle c_{H_k}c_{X_{o,k}} - q_{H_k}q_{X_{o,k}}\rangle}}\right); \quad (85)$$

$V_o \triangleq \sqrt{\langle q_{H_k}q_{X_{o,k}}\rangle}\,v$ for $o \in \{\mathsf{t},\mathsf{d}\}$; $I(H_k; Y_{H_k}|\tilde{q}_{H_k})$ is the mutual information between $Y_{H_k}$ and $H_k$; $I(X_{\mathsf{d},k}; Y_{X_{\mathsf{d},k}}|\tilde{q}_{X_{\mathsf{d},k}})$ is the mutual information between $Y_{X_{\mathsf{d},k}}$ and $X_{\mathsf{d},k}$; and $c_{X_o} = \sigma_{x_o}^2$, $c_{H_k} = \sigma_{h_k}^2$. In (84), the other parameters $\{q_{X_{o,k}}, q_{H_k}, \tilde{q}_{X_{o,k}}, \tilde{q}_{H_k}\}$ are obtained from the solutions to the following fixed-point equations

$$\tilde{q}_{H_k} = \beta_{\mathsf{t},k}q_{X_{\mathsf{t},k}}\chi_\mathsf{t} + \beta_\mathsf{d}q_{X_{\mathsf{d},k}}\chi_\mathsf{d}, \quad (86a)$$

$$\tilde{q}_{X_{\mathsf{t},k}} = \alpha q_{H_k}\chi_\mathsf{t}, \quad (86b)$$

$$\tilde{q}_{X_{\mathsf{d},k}} = \alpha q_{H_k}\chi_\mathsf{d}, \quad (86c)$$

$$q_{H_k} = c_{H_k} - \mathsf{mse}_{H_k}, \quad (86d)$$

$$q_{X_{\mathsf{t},k}} = c_{X_\mathsf{t}} - \mathsf{mse}_{X_{\mathsf{t},k}}, \quad (86e)$$

$$q_{X_{\mathsf{d},k}} = c_{X_\mathsf{d}} - \mathsf{mse}_{X_{\mathsf{d},k}}, \quad (86f)$$

in which $\mathsf{mse}_{X_{\mathsf{t},k}} = 0$, and $\mathsf{mse}_{H_k}$ and $\mathsf{mse}_{X_{\mathsf{d},k}}$ are the MSEs of the Bayes-optimal estimators over (83b) and (83a), respectively. Also, in (46), we have defined

$$\chi_o \triangleq \sum_{b=1}^{2^{\mathsf{B}}} \int \mathrm{D}v \frac{\left( \Psi_b' \left( \sqrt{\langle q_{H_k} q_{X_{o,k}} \rangle} v \right) \right)^2}{\Psi_b \left( \sqrt{\langle q_{H_k} q_{X_{o,k}} \rangle} v \right)}, \text{ for } o \in \{\mathsf{t}, \mathsf{d}\} \tag{87}$$

with $\Psi_b(\cdot)$ given by (85) and $\Psi_b'(V_o) = \frac{\partial \Psi_b(V_o)}{\partial V_o}$.

## REFERENCES


[1] 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.

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

[3] E. G. Larsson *et al.*, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

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

[5] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539–550, Apr. 1999.

[6] 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.

[7] T. Koch and A. Lapidoth, "Increased capacity per unit-cost by oversampling," in *Proc. IEEE 26th Conv. Elect. Electron. Eng. Israel*, Eilat, Israel, Nov. 17–20 2010, pp. 684–688.

[8] 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.

[9] 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.

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

[11] J. Mo and R. W. Heath Jr, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, 2015. [Online]. Available: http://arxiv.org/abs/1410.7353

[12] N. Liang and W. Zhang, "Mixed-ADC massive MIMO," preprint, 2015. [Online]. Available: http://arxiv.org/abs/1504.03516

[13] Q. Bai and J. Nossek, "Energy efficiency maximization for 5G multi-antenna receivers," *Trans. Emerging Telecommun. Technol.*, vol. 26, no. 1, pp. 3–14, Jan. 2015.

[14] O. Orhan, E. Erkip, and S. Rangan, "Low power analog-to-digital conversion in millimeter wave systems: Impact of resolution and bandwidth on performance," preprint, 2015. [Online]. Available: http://arxiv.org/abs/1502.01980

[15] J. Mo and R. W. Heath Jr, "Limited feedback in multiple-antenna systems with one-bit quantization," preprint, 2015. [Online]. Available: http://arxiv.org/abs/1505.00484

[16] K. Nakamura and T. Tanaka, "Performance analysis of signal detection using quantized received signals of linear vector channel," in *Proc. Inter. Symp. Inform. Theory and its Applications (ISITA)*, Auckland, New Zealand, Dec. 2008.

[17] A. Mezghani and J. Nossek, "Belief propagation based MIMO detection operating on quantized channel output," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Austin, TX, 13-18 June 2010, pp. 2113–2117.

[18] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," preprint, 2014. [Online]. Available: http://arxiv.org/abs/1404.7736.

[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. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-bit massive MIMO: Channel estimation and high-order modulations," in *2015 IEEE Int. Conf. Communication Workshop (ICCW)*, London, UK, 8-12 June 2015, pp. 1304–1309.

[21] J. Choi, D. J. Love, D. R. Brown III, M. Boutin, "Distributed reception with spatial multiplexing: MIMO systems for the internet of things," preprint, 2014. [Online]. Available: http://arxiv.org/abs/1409.7850

[22] Y. Xu, Y. Kabashima, and L. Zdeborová, "Bayesian signal reconstruction for 1-bit compressed sensing," *J. Stat. Mech.*, no. 11, p. P11015, 2014.

[23] J. Choi, J. Mo, and R. W. Heath Jr, "Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs," preprint, 2015. [Online]. Available: http://arxiv.org/abs/1507.04452

[24] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," preprint, 2015. [Online]. Available: http://arxiv.org/abs/1509.07928

[25] A. Mezghani, F. Antreich, and J. A. Nossek, "Multiple parameter estimation with quantized channel output," in *Int. ITG Workshop Smart Antennas (WSA)*, Bremen, Feb. 2010, pp. 143–150.

[26] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath Jr, "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, CA, USA, Nov. 2014.

[27] K. Takeuchi, R. R. Müller, M. Vehkaperä, and T. Tanaka, "On an achievable rate of large Rayleigh block-fading MIMO channels with no CSI," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6517–6541, Oct. 2013.

[28] J. Ma and L. Ping, "Data-aided channel estimation in large antenna systems," *IEEE Trans. Signal Processing*, vol. 62, no. 12, pp. 3111–3124, June 2014.

[29] C.-K. Wen, S. Jin, K.-K. Wong, C.-J. Wang, and G. Wu, "Joint channel-and-data estimation for large-MIMO systems with low-precision ADCs," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 1237–1241.

[30] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.

[31] M. Hazewinkel, "Normal distribution," *Encyclopedia of Mathematics, Springer*, 2001. [Online]. Available: https://www.encyclopediaofmath.org/index.php/Normal_distribution

[32] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1994.

[33] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[34] J. Ziniel, P. Schniter, and P. Sederberg, "Binary linear classification and feature selection via generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 63, no. 8, pp. 2020–2032, Apr. 2015.

[35] 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, 2009.

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

[37] 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.

[38] F. Krzakala, M. Mézard, and L. Zdeborová, "Phase diagram and approximate message passing for blind calibration and dictionary learning," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Istanbul, Turkey, July 2013, pp. 659–663.

[39] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, "Phase transitions and sample complexity in Bayes-optimal matrix factorization," preprint 2014. [Online]. Available: http://arxiv.org/abs/1402.1298.

[40] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. ser. Number 111 in Int. Series on Monographs on Physics. Oxford U.K.: Oxford Univ. Press, 2001.

[41] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[42] A. L. Moustakas, S. H. Simon, and A. M. Sengupta, "MIMO capacity through correlated channels in the presence of correlated interferers and

noise: a (not so) large N analysis," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2545–2561, Oct. 2003.

[43] D. Guo and S. Verdú , "Randomly spread CDMA: asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 1982–2010, Jun. 2005.

[44] R. R. Müller, "Channel capacity and minimum probability of error in large dual antenna array systems with binary modulation," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2821–2828, Nov. 2003.

[45] C.-K. Wen and K.-K. Wong, "Asymptotic analysis of spatially correlated MIMO multiple-access channels with arbitrary signaling inputs for joint and separate decoding," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 252–268, Jan. 2007.

[46] A. Hatabu, K. Takeda, and Y. Kabashima, "Statistical mechanical analysis of the Kronecker channel model for multiple-input multipleoutput wireless communication," *Phys. Rev. E*, vol. 80, pp. 061 124(1–12), 2009.

[47] M. A. Girnyk, M. Vehkaperä, and L. K. Rasmussen, "Large-system analysis of correlated MIMO channels with arbitrary signaling in the presence of interference," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1536–1276, Apr. 2014.

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

[49] 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.

[50] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[51] J. G. Proakis, *Digital Communications*. McGraw Hill, 4th ed., 1995.

[52] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Prentice Hall, 3rd ed., 2009.

[53] 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.

[54] T.-C. Zhang, C.-K. Wen, S. Jin, T. Jiang, "Mixed-ADC massive MIMO detectors: performance analysis and design optimization," preprint, 2015. [Online]. Available: http://arxiv.org/abs/1509.07928