Variational Bayes for Joint Channel Estimation and Data Detection in Few-Bit Massive MIMO Systems

Ly V. Nguyen, A. Lee Swindlehurst, and Duy H. N. Nguyen

Abstract—Massive multiple-input multiple-output (MIMO) communications using low-resolution analog-to-digital converters (ADCs) is a promising technology for providing high spectral and energy efficiency with affordable hardware cost and power consumption. However, the use of low-resolution ADCs requires special signal processing methods for channel estimation and data detection since the resulting system is severely non-linear. This paper proposes joint channel estimation and data detection methods for massive MIMO systems with low-resolution ADCs based on the variational Bayes (VB) inference framework. We first derive matched-filter quantized VB (MF-QVB) and linear minimum mean-squared error quantized VB (LMMSE-QVB) detection methods assuming the channel state information (CSI) is available. Then we extend these methods to the joint channel estimation and data detection (JED) problem and propose two methods we refer to as MF-QVB-JED and LMMSE-QVB-JED. Unlike conventional VB-based detection methods that assume knowledge of the second-order statistics of the additive noise, we propose to float the noise variance/covariance matrix as an unknown random variable that is used to account for both the noise and the residual inter-user interference. We also present practical aspects of the QVB framework to improve its implementation stability. Finally, we show via numerical results that the proposed VB-based methods provide robust performance and also significantly outperform existing methods.

Index Terms—Approximate message passing, Bayesian inference, detection, estimation, massive MIMO, soft interference cancellation, variational Bayesian.

I. INTRODUCTION

Beyond-5G wireless systems will require exploitation of the large bandwidths available at THz frequencies (0.3–3 THz) [1]–[3]. An inherent challenge in operating in these bands is the strong radio frequency (RF) path loss, and while this can be effectively addressed by exploiting the beamforming gain available from large antenna arrays, scaling up existing RF technologies to very large arrays becomes complex, expensive, and demands high power consumption. Therefore, implementing massive antenna arrays for THz communications will require radical simplifications in the RF architecture. Hybrid analog-digital arrays reduce the number of RF chains with respect to (w.r.t.) the number of antenna elements [4], but this approach yields poor spatial multiplexing and does not scale well at higher frequencies and wider

Ly V. Nguyen and A. Lee Swindlehurst are with the Center for Pervasive Communications and Computing, Henry Samueli School of Engineering, University of California, Irvine, CA, USA 92697 (e-mail: vanln1@uci.edu, swindle@uci.edu).

Duy H. N. Nguyen is with the Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA, USA 92182 (e-mail: duy.nguyen@sdsu.edu).

bandwidths due to the need for complex analog circuitry and resource-consuming beam management schemes [5].

An alternative solution is retaining the RF chains for each antenna, but reducing complexity and energy consumption through the use of low-resolution analog-to-digital converters (ADCs). It has been shown that fully digital arrays with lower-resolution data converters (even down to 1 bit) can significantly outperform hybrid analog-digital architectures in terms of beamforming flexibility and spectral/energy efficiency [6]. This is because the use of low-resolution ADCs maintains the high spatial multiplexing gains of massive arrays, and they more easily scale to higher frequencies and bandwidths with significantly reduced hardware cost and power consumption. However, the use of low-resolution quantization requires special signal processing methods for channel estimation and data detection since the resulting system is severely non-linear, and the received signals are significantly distorted.

There has been a plethora of channel estimation and data detection studies for massive MIMO systems with low-resolution ADCs. For example, one-bit ML and near-ML methods were proposed in [7]. The Bussgang decomposition was used to derive different linear channel estimators in [8], [9] and linear data detectors in [9]–[11]. While the ML and near-ML methods are either too complicated for practical implementation or non-robust at high signal-to-noise ratios (SNRs), the linear Bussgang-based receivers have lower complexity and are more robust, but they have limited performance. Several other detection approaches have been proposed in [12]–[15] but they require the use of either a cyclic redundancy check (CRC) or an error correcting code (ECC). The authors in [16] developed a bilinear generalized approximate message passing (BiGAMP) algorithm [17] to solve the joint channel estimation and data detection (JED) problem for few-bit MIMO systems.

Recently, machine learning for low-resolution MIMO channel estimation and data detection has gained interest and there has also been numerous results reported in the literature. In particular, the work in [18] shows how support vector machine (SVM) models can be applied to one-bit massive MIMO channel estimation and data detection. The authors of [11] exploit a deep neural network (DNN) framework to develop a special model-driven detection approach that outperforms the SVM-based methods in [18]. Deep learning-based joint pilot signal and channel estimator designs were proposed in [19] and [20]. While a conventional DNN structure was used in [19], the work in [20] employed a model-driven network similar to [11]. The work in [21] proposed another DNN-based detector but its computational complexity is high since the detection network must be retrained for each new channel

realization. Several learning-based blind detection methods were proposed in [22]–[24] but they are restricted to small-scale systems. In [25], Bayesian inference was used to develop a JED method for quantized single-antenna systems with orthogonal frequency division multiplexing (OFDM) and time-frequency doubly selective (DS) channels where the sparsity of the DS channels was exploited. Another JED method was proposed in [26] based on the variational Bayesian (VB) inference framework, and it was shown to outperforms the BiGAMP-based method in [16] for soft symbol decoding. In a recent work [27], VB inference was also shown to be very efficient in MIMO data detection with infinite-resolution (perfect) ADCs.

In this paper, we develop a VB framework for channel estimation and data detection for massive MIMO systems with low-resolution ADCs. While conventional machine learning models such as SVM and DNN only provide a point estimate of the signal of interest, e.g., the channel or the data symbols, the VB approach can provide the posterior distribution of the estimate, which is important in subsequent signal processing steps such as channel decoding. Another advantage of VB is that it does not require a training process like DNNs which often suffer from performance degradation due to mismatch between the actual model and that used during training. Unlike our previous work in [28] which only considers the data detection problem and assumes perfect channel state information (CSI), we study both channel estimation and data detection in this paper and make the following contributions:

- We devise a matched-filter quantized VB (MF-QVB) detection method for few-bit MIMO systems with known CSI. Unlike the VB-based detection method in [26] that assumes a known noise variance, the proposed MF-QVB method floats the noise variance as a latent variable and uses it to also account for residual inter-user interference. This latent variable is jointly estimated with the transmitted data symbol vector.
- We develop a linear minimum mean-squared error quantized VB (LMMSE-QVB) detector that treats the noise covariance matrix as a latent variable, rather than simply assuming the noise covariance is a scaled identity matrix. The LMMSE-QVB detector offers performance similar to MF-QVB for independent and identically distributed (i.i.d.) channels, but significantly outperforms MF-QVB for spatially correlated channels.
- We study the JED problem for few-bit MIMO systems and develop two methods, referred to as MF-QVB-JED and LMMSE-QVB-JED. The latter algorithm jointly estimates the channel matrix, the symbol data vectors, and the noise variances/covariance matrices. Again, this goes well beyond the prior work in [26] that assumes a known scaled identity noise covariance.
- We also present practical aspects of the VB framework to improve the implementation stability of the algorithms. We show via numerical results that the proposed VB detection algorithms provide much lower symbol error rates (SERs) compared to the conventional VB-based methods in [26]. The proposed QVB-JED algorithms also

outperform FBM-DetNet in [20], particularly for spatially correlated channels.

The rest of this paper is organized as follows. We present the system model and the problem of interest in Section II. Next, a brief introduction to the VB inference framework is given in Section III. Then, in Section IV, we derive the VB-based data detection method when the CSI is known. Section V proposes the VB-based JED methods. We present practical implementation aspects of the VB framework as well as numerical results in Section VI. Finally, Section VII concludes the paper.

Notation: Scalars x_{ij} and $[\mathbf{X}]_{ij}$ both denote the element at the ith row and jth column of a matrix $\mathbf{X}$; vector $\mathbf{x}_i$ denotes the ith column of a matrix $\mathbf{X}$; the operators $\mathrm{Tr}\{\mathbf{X}\}$ and $|\mathbf{X}|$ represent the trace and determinant of a square matrix $\mathbf{X}$, respectively; the Frobenius norm of a matrix $\mathbf{X}$ is represented by $\|\mathbf{X}\|_F$; the distribution of a K-element complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{CN}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\pi^K|\boldsymbol{\Sigma}|}\exp\big(-(\mathbf{x}-\boldsymbol{\mu})^H\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big)$, and is also written as $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu},\boldsymbol{\Sigma})$; the functions $\phi(x)$ and $\Phi(x)$ denote the PDF and cumulative distribution function (CDF) of a standard Gaussian random variable $\mathcal{N}(0,1)$; the operators $\mathbb{E}_{p(x)}[x]$ and $\mathrm{Var}_{p(x)}[x]$ denote the mean and variance of x w.r.t. its distribution $p(x)$; in addition, we use $\langle x \rangle$, τ_x, and $\langle |x|^2 \rangle = |\langle x \rangle|^2 + \tau_x$ to denote the mean, variance, and second moment of x w.r.t. a variational distribution $q(x)$. The symbols $\sim$ and $\propto$ indicate "distributed according to" and "proportional to", respectively. Finally, $\mathbb{1}(\cdot)$ denotes the indicator function which equals one if the argument holds true, or zero otherwise.

II. SYSTEM MODEL AND PROBLEM STATEMENT

We consider an uplink massive MIMO system with K single-antenna users and an M-antenna base station (BS). The symbols transmitted by the users are collected in the vector $\mathbf{x} = [x_1, \ldots, x_K]^T$, where x_i corresponds to user-i and is drawn from a discrete constellation $\mathcal{S}$, e.g., quadrature amplitude modulation (QAM) or phase-shift keying (PSK). The prior distribution of x_i is

$$p(x_i) = \sum_{a \in \mathcal{S}} p_a \delta(x_i - a), \tag{1}$$

where p_a corresponds to a known prior probability of the constellation point $a \in \mathcal{S}$. It is assumed that the symbols in $\mathbf{x}$ are independent of each other, i.e., $p(\mathbf{x}) = \prod_{i=1}^{K} p(x_i)$.

Given $\mathbf{H} \in \mathbb{C}^{M \times K}$ as the uplink channel, the linear uplink MIMO system can be modeled as

$$\mathbf{r} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{2}$$

where $\mathbf{r}$ is the unquantized received signal vector and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I}_M)$ models the independent and identically distributed (i.i.d.) additive white Gaussian noise at the receiver. The channel vector $\mathbf{h}_i$ from user-i to the BS is assumed to be distributed as $p(\mathbf{h}_i) = \mathcal{CN}(\mathbf{h}_i; \mathbf{0}, \mathbf{C}_i)$ where $\mathbf{C}_i \triangleq \mathbb{E}[\mathbf{h}_i\mathbf{h}_i^H]$ is the covariance matrix that describes the spatial correlation between the receive antennas. Finally, we assume that $\mathbb{E}[\mathbf{h}_i\mathbf{h}_j^H] = \mathbf{0}$, if $i \neq j$.

We consider a block fading channel where a pilot matrix $\mathbf{X}_\mathrm{p}$ followed by a data matrix $\mathbf{X}_\mathrm{d}$ are transmitted in each block-fading interval. The unquantized received signals are given by

$$\mathbf{R}_\mathrm{p} = \mathbf{H}\mathbf{X}_\mathrm{p} + \mathbf{N}_\mathrm{p}$$
$$\mathbf{R}_\mathrm{d} = \mathbf{H}\mathbf{X}_\mathrm{d} + \mathbf{N}_\mathrm{d}, \tag{3}$$

where $\mathbf{X}_\mathrm{p} = [\mathbf{x}_{\mathrm{p},1}, \ldots, \mathbf{x}_{\mathrm{p},T_\mathrm{p}}] \in \mathbb{C}^{K \times T_\mathrm{P}}$ with $\mathbf{x}_{\mathrm{p},t} = [x_{\mathrm{p},1t}, \ldots, x_{\mathrm{p},Kt}]^T$ and $\mathbf{X}_\mathrm{d} = [\mathbf{x}_{\mathrm{d},1}, \ldots, \mathbf{x}_{\mathrm{d},T_\mathrm{d}}] \in \mathbb{C}^{K \times T_\mathrm{d}}$ with $\mathbf{x}_{\mathrm{d},t} = [x_{\mathrm{d},1t}, \ldots, x_{\mathrm{d},Kt}]^T$. We assume that user-$i$ transmits with power ρ_i during the pilot transmission phase.

Each received analog signal is then quantized by a pair of b-bit ADCs to produce the quantized received signal:

$$\Re\{\mathbf{Y}_\mathrm{p}\} = \mathcal{Q}_b(\Re\{\mathbf{R}_\mathrm{p}\}), \quad \Im\{\mathbf{Y}_\mathrm{p}\} = \mathcal{Q}_b(\Im\{\mathbf{R}_\mathrm{p}\});$$
$$\Re\{\mathbf{Y}_\mathrm{d}\} = \mathcal{Q}_b(\Re\{\mathbf{R}_\mathrm{d}\}), \quad \Im\{\mathbf{Y}_\mathrm{d}\} = \mathcal{Q}_b(\Im\{\mathbf{R}_\mathrm{d}\}), \tag{4}$$

where $\mathcal{Q}_b(\cdot)$ denotes the b-bit ADC operation which is applied separately to every element of its matrix or vector argument. It is assumed that $\mathcal{Q}_b(\cdot)$ performs b-bit uniform scalar quantization, which is characterized by a set of $2^b - 1$ thresholds denoted as $\{d_1, \ldots, d_{2^b-1}\}$. Without loss of generality, we assume $-\infty = d_0 < d_1 < \ldots < d_{2^b-1} < d_{2^b} = \infty$. For a quantization step size of Δ, the quantization thresholds are given by

$$d_k = (-2^{b-1} + k)\Delta, \text{ for } k \in \mathcal{K} = \{1, \ldots, 2^b - 1\}. \tag{5}$$

The quantized output q is then defined as

$$q = \mathcal{Q}_b(r) = \begin{cases} d_k - \frac{\Delta}{2}, & \text{if } r \in (d_{k-1}, d_k] \text{ with } k \in \mathcal{K} \\ (2^b - 1)\frac{\Delta}{2}, & \text{if } r \in (d_{2^b-1}, d_{2^b}]. \end{cases} \tag{6}$$

We also define $q^{\mathrm{low}} = d_{k-1}$ and $q^{\mathrm{up}} = d_k$ as lower and upper thresholds of the quantization bin to which q belongs.

In this paper, we first study the data detection problem with known CSI, i.e., where the problem of interest is to detect the data matrix $\mathbf{X}_\mathrm{d}$ using the received signal matrix $\mathbf{Y}_\mathrm{d}$ and the channel matrix $\mathbf{H}$. Then, in the following section we study the problem of joint channel estimation and data detection where $\mathbf{H}$ is estimated and $\mathbf{X}_\mathrm{d}$ detected using knowledge of the pilot matrix $\mathbf{X}_\mathrm{p}$ and the received signal matrices $\mathbf{Y}_\mathrm{p}$ and $\mathbf{Y}_\mathrm{d}$.

III. BACKGROUND ON VARIATIONAL BAYES INFERENCE

This section presents a brief background on the VB method for approximate inference that will be developed for solving the problems of interest in this paper. In variational inference, the posterior distribution $p(\mathbf{x}|\mathbf{y})$ over a set of latent variables $\mathbf{x}$ given some observed data $\mathbf{y}$ is approximated by a variational distribution $q(\mathbf{x})$. A set of variational parameters describing $q(\mathbf{x})$ within a family of densities $\mathcal{Q}$ are determined to minimize the Kullback-Leibler (KL) divergence from $q(\mathbf{x})$ to $p(\mathbf{x}|\mathbf{y})$ [29], [30], i.e.,

$$q(\mathbf{x}) = \arg \min_{q(\mathbf{x}) \in \mathcal{Q}} \mathrm{KL}(q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y})). \tag{7}$$

The KL divergence is defined as

$$\mathrm{KL}(q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y})) = \mathbb{E}_{q(\mathbf{x})}[\ln q(\mathbf{x})] - \mathbb{E}_{q(\mathbf{x})}[\ln p(\mathbf{x}|\mathbf{y})]$$
$$= \mathbb{E}_{q(\mathbf{x})}[\ln q(\mathbf{x})] - \mathbb{E}_{q(\mathbf{x})}[\ln p(\mathbf{x}, \mathbf{y})]$$
$$+ \ln p(\mathbf{y}). \tag{8}$$

Since $p(\mathbf{y})$ is a constant w.r.t. $q(\mathbf{x})$, maximizing the evidence lower bound (ELBO), defined as

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\mathbf{x})}[\ln p(\mathbf{x}, \mathbf{y})] - \mathbb{E}_{q(\mathbf{x})}[\ln q(\mathbf{x})], \tag{9}$$

is equivalent to minimizing the KL divergence.

The maximum of $\mathrm{ELBO}(q)$ occurs when $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$. Since calculating the true posterior is often intractable, it is more convenient to consider a restricted family of distributions for $q(\mathbf{x})$. Here, the VB method assumes the *mean field variational family*, such that

$$q(\mathbf{x}) = \prod_{i=1}^{m} q_i(x_i). \tag{10}$$

In this mean field family, the latent variables are mutually independent and each is governed by a distinct factor in the variational distribution. The general expression for the optimal solution $q_i(x_i)$ can be obtained as [29]

$$q_i(x_i) \propto \exp\left\{\langle \ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x}) \rangle_{-x_i}\right\}. \tag{11}$$

Here, $\langle \cdot \rangle_{-x_i}$ denotes the expectation w.r.t. all latent variables except x_i using the variational distribution $q_{-i}(\mathbf{x}_{-i}) = \prod_{j \neq i} q_j(x_j)$. In the following, if $\langle \cdot \rangle$ is used, it means the variational expectation is taken w.r.t. all the latent variables in the argument. By iterating the update of $q_i(x_i)$ sequentially over all i, the $\mathrm{ELBO}(q)$ objective function can be monotonically improved. Thus, convergence to at least a local optimum of $\mathrm{ELBO}(q)$ is guaranteed [29], [30].

In the following, we present a theorem on the variational posterior mean of multiple random variables that will be applied repeatedly later in the paper.

Theorem 1. *Let $\mathbf{A}$, $\mathbf{y}$, and $\mathbf{x}$ of size $m \times n$, $m \times 1$, and $n \times 1$ be three independent random matrices (vectors) w.r.t. a variational distribution $q_{\mathbf{A},\mathbf{y},\mathbf{x}}(\mathbf{A}, \mathbf{y}, \mathbf{x}) = q(\mathbf{A})q(\mathbf{y})q(\mathbf{x})$. It is assumed that $\mathbf{A}$ is column-wise independent and let $\langle \mathbf{a}_i \rangle$ and $\mathbf{\Sigma}_{\mathbf{a}_i}$ be the variational mean and covariance matrix of the ith column of $\mathbf{A}$. Let $\langle \mathbf{x} \rangle$ and $\mathbf{\Sigma}_{\mathbf{x}}$ (and $\langle \mathbf{y} \rangle$ and $\mathbf{\Sigma}_{\mathbf{y}}$) be the variational mean and covariance matrix of $\mathbf{x}$ (and $\mathbf{y}$), respectively. For an arbitrary Hermitian matrix $\mathbf{B}$, let $\langle (\mathbf{y} - \mathbf{A}\mathbf{x})^H \mathbf{B}(\mathbf{y} - \mathbf{A}\mathbf{x}) \rangle$ be the expectation of $(\mathbf{y} - \mathbf{A}\mathbf{x})^H \mathbf{B}(\mathbf{y} - \mathbf{A}\mathbf{x})$ w.r.t. $q_{\mathbf{A},\mathbf{y},\mathbf{x}}(\mathbf{A}, \mathbf{y}, \mathbf{x})$. We have*

$$\langle (\mathbf{y} - \mathbf{A}\mathbf{x})^H \mathbf{B}(\mathbf{y} - \mathbf{A}\mathbf{x}) \rangle$$
$$= (\langle \mathbf{y} \rangle - \langle \mathbf{A} \rangle \langle \mathbf{x} \rangle)^H \mathbf{B}(\langle \mathbf{y} \rangle - \langle \mathbf{A} \rangle \langle \mathbf{x} \rangle) + \mathrm{Tr}\{\mathbf{B}\mathbf{\Sigma}_{\mathbf{y}}\}$$
$$+ \langle \mathbf{x} \rangle^H \mathbf{D} \langle \mathbf{x} \rangle + \mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{x}}\mathbf{D}\} + \mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{x}}\langle \mathbf{A}^H \rangle \mathbf{B} \langle \mathbf{A} \rangle\}, \tag{12}$$

where $\mathbf{D} = \mathrm{diag}(\mathrm{Tr}\{\mathbf{B}\mathbf{\Sigma}_{\mathbf{a}_1}\}, \ldots, \mathrm{Tr}\{\mathbf{B}\mathbf{\Sigma}_{\mathbf{a}_n}\})$.

Proof: The proof of this theorem is similar to the proof of Theorem 1 in [27], except that $\mathbf{y}$ is now a random vector. Details of the proof are given in Appendix A. ∎

We note that if any of $\mathbf{A}$, $\mathbf{y}$, and $\mathbf{x}$ is deterministic, the corresponding covariance matrices $\{\mathbf{\Sigma}_{\mathbf{a}_i}\}$, $\mathbf{\Sigma}_{\mathbf{y}}$, and $\mathbf{\Sigma}_{\mathbf{x}}$ will be set to $\mathbf{0}$ and the expectation of $(\mathbf{y} - \mathbf{A}\mathbf{x})^H \mathbf{B}(\mathbf{y} - \mathbf{A}\mathbf{x})$ given in (12) can be simplified accordingly.

IV. VB FOR DATA DETECTION IN FEW-BIT MIMO SYSTEMS WITH CSIR

In this section, we develop new VB-based algorithms for solving the data detection problem in few-bit MIMO systems with known channel $\mathbf{H}$. For ease of presentation, we drop the subscripts d, t indicating the data transmission at time index t.

A. Proposed MF-QVB For Few-Bit MIMO Detection

The VB-based methods proposed in [26] assume prior information about the noise variance N_0. However, in practice, N_0 is not known *a priori* and may need to be estimated. Furthermore, using the known noise variance, the conventional VB methods in [26] do not take into account the residual inter-user interference. Here, we consider the residual interference-plus-noise as an unknown parameter N_0^{post}, which is postulated by the estimation in the VB framework [27]. For ease of computation, we use $\gamma = 1/N_0^{\mathrm{post}}$ to denote the precision to be estimated.

The joint distribution $p(\mathbf{y}, \mathbf{r}, \mathbf{x}; \gamma, \mathbf{H})$ of the observed variable $\mathbf{y}$ and the latent variables $\mathbf{r}$ and $\mathbf{x}$ can be factored as

$$
\begin{aligned}
p(\mathbf{y}, \mathbf{r}, \mathbf{x}; \gamma, \mathbf{H}) &= p(\mathbf{y}|\mathbf{r})p(\mathbf{r}|\mathbf{x}; \gamma, \mathbf{H})p(\mathbf{x}) \\
&= \left[\prod_{m=1}^{M} p(y_m|r_m)\right] p(\mathbf{r}|\mathbf{x}; \gamma, \mathbf{H}) \left[\prod_{i=1}^{K} p(x_i)\right],
\end{aligned}
$$
(13)

where $p(y_m|r_m) = \mathbb{1}\big(r_m \in [y_m^{\mathrm{low}}, y_m^{\mathrm{up}}]\big)$ and $p(\mathbf{r}|\mathbf{x}; \gamma, \mathbf{H}) = \mathcal{CN}(\mathbf{r}; \mathbf{Hx}, \gamma^{-1}\mathbf{I}_M)$. We note that the random vector $\mathbf{r}$ is comprised of conditional independent elements due to the same noise variance being imposed on the M receive antennas.

In the *E-step*, for a currently fixed estimate $\hat{\gamma}$ of γ, we aim to derive the mean field variational distribution $q(\mathbf{r}, \mathbf{x})$ of $\mathbf{r}$ and $\mathbf{x}$ given $\mathbf{y}$ such that

$$
p(\mathbf{r}, \mathbf{x}|\mathbf{y}; \hat{\gamma}, \mathbf{H}) \approx q(\mathbf{r}, \mathbf{x}) = q(\mathbf{r})\left[\prod_{i=1}^{K} q(x_i)\right]. \tag{14}
$$

1) Updating $\mathbf{r}$. The variational distribution $q(\mathbf{r})$ is obtained by taking the expectation of the conditional in (13) w.r.t. $q(\mathbf{x})$:

$$
\begin{aligned}
q(\mathbf{r}) &\propto \exp\left\{\big\langle \ln p(\mathbf{y}|\mathbf{r}) + \ln p(\mathbf{r}|\mathbf{x}; \hat{\gamma}, \mathbf{H})\big\rangle_{-\mathbf{r}}\right\} \\
&\propto \exp\left\{\big\langle \ln \mathbb{1}\big(\mathbf{r} \in [\mathbf{y}^{\mathrm{low}}, \mathbf{y}^{\mathrm{up}}]\big) - \hat{\gamma}\|\mathbf{r} - \mathbf{Hx}\|^2\big\rangle_{-\mathbf{r}}\right\} \\
&\propto \mathbb{1}\big(\mathbf{r} \in [\mathbf{y}^{\mathrm{low}}, \mathbf{y}^{\mathrm{up}}]\big) \times \exp\left\{-\hat{\gamma}\|\mathbf{r} - \mathbf{H}\langle\mathbf{x}\rangle\|^2\right\} \\
&\propto \mathbb{1}\big(\mathbf{r} \in [\mathbf{y}^{\mathrm{low}}, \mathbf{y}^{\mathrm{up}}]\big) \times \mathcal{CN}\big(\mathbf{r}; \mathbf{H}\langle\mathbf{x}\rangle, \hat{\gamma}^{-1}\mathbf{I}_M\big). \tag{15}
\end{aligned}
$$

We note that variational distribution $q(\mathbf{r})$ is inherently separable as $\prod_{m=1}^{M} q(r_m)$ without enforcing the mean field approximation on $q(\mathbf{r})$. Thus, the variational mean and variance can be obtained concurrently for all the elements of $\mathbf{r}$. We see in (15) that $q(r_m)$ is the truncated complex normal distribution obtained from bounding $r_m \sim \mathcal{CN}\big(s_m, \hat{\gamma}^{-1}\big)$, where $s_m = \mathbf{H}_{m,:}\langle\mathbf{x}\rangle$, to the interval $(y_m^{\mathrm{low}}, y_m^{\mathrm{up}})$. Thus, its mean $\langle r_m \rangle$ and variance τ_{r_m} are given by $\mathsf{F}_r\big(s_m, \hat{\gamma}, y_m^{\mathrm{low}}, y_m^{\mathrm{up}}\big)$ and $\mathsf{G}_r\big(s_m, \hat{\gamma}, y_m^{\mathrm{low}}, y_m^{\mathrm{up}}\big)$, respectively.[1]

---

[1] The computations of the mean $\mathsf{F}_r(\mu, \gamma, a, b)$ and variance $\mathsf{G}_r(\mu, \gamma, a, b)$ of an arbitrary complex normal distribution $\mathcal{CN}(\mu, \gamma^{-1})$ truncated to an interval (a, b) are presented in Appendix B.

Algorithm 1 – MF-QVB for Few-Bit MIMO Detection

1: **Input: y**, **H**
2: **Output:** $\hat{\mathbf{x}}$
3: Initialize $\hat{x}_i^1 = 0$ and $\tau_{x_i}^1 = \mathrm{Var}_{p(x_i)}[x_i], \forall i$, $\hat{r}_m^1 = y_m$ and $\tau_{r_m}^1 = 0, \forall m$, and $\mathbf{e} = \hat{\mathbf{r}}^1 - \mathbf{H}\hat{\mathbf{x}}^1$
4: **for** $\ell = 1, 2, \ldots$ **do**
5:    $\hat{\gamma}^\ell \leftarrow M/\big(\|\mathbf{e}\|^2 + \sum_{m=1}^{M} \tau_{r_m}^\ell + \sum_{i=1}^{M} \tau_{x_i}^\ell \|\mathbf{h}_i\|^2\big)$
6:    $\mathbf{s}^\ell \leftarrow \hat{\mathbf{r}}^\ell - \mathbf{e}$
7:    $\hat{\mathbf{r}}^{\ell+1} \leftarrow \mathsf{F}_r\big(\mathbf{s}^\ell, \hat{\gamma}^\ell, \mathbf{y}^{\mathrm{low}}, \mathbf{y}^{\mathrm{up}}\big)$ $\triangleright$ Update **r**
8:    $\boldsymbol{\tau}_{\mathbf{r}}^{\ell+1} \leftarrow \mathsf{G}_r\big(\mathbf{s}^\ell, \hat{\gamma}^\ell, \mathbf{y}^{\mathrm{low}}, \mathbf{y}^{\mathrm{up}}\big)$
9:    $\mathbf{e} \leftarrow \mathbf{e} - \hat{\mathbf{r}}^\ell + \hat{\mathbf{r}}^{\ell+1}$ $\triangleright$ Update residual
10:    **for** $i = 1, \ldots, K$ **do** $\triangleright$ Update **x**
11:       $z_i^\ell \leftarrow \hat{x}_i^\ell + \mathbf{h}_i^H \mathbf{e}/\|\mathbf{h}_i\|^2$ $\triangleright$ Matched filter
12:       $\hat{x}_i^{\ell+1} \leftarrow \mathsf{F}_x\big(z_i^\ell, \hat{\gamma}^\ell \|\mathbf{h}_i\|^2\big)$
13:       $\tau_{x_i}^{\ell+1} \leftarrow \mathsf{G}_x\big(z_i^\ell, \hat{\gamma}^\ell \|\mathbf{h}_i\|^2\big)$
14:       $\mathbf{e} \leftarrow \mathbf{e} + \mathbf{h}_i(\hat{x}_i^\ell - \hat{x}_i^{\ell+1})$ $\triangleright$ Update residual
15: $\forall i : \hat{x}_i \leftarrow \arg\max_{a \in \mathcal{S}} p_a \mathcal{CN}\big(z_i^\ell; a, 1/(\hat{\gamma}^\ell \|\mathbf{h}_i\|^2)\big).$

---

2) Updating x_i. The variational distribution $q(x_i)$ is obtained by taking the expectation of the conditional in (13) w.r.t. $q(\mathbf{r}) \prod_{j \neq i} q(x_j)$:

$$
\begin{aligned}
q(x_i) &\propto \exp\left\{\big\langle \ln p(\mathbf{r}|\mathbf{x}; \hat{\gamma}, \mathbf{H}) + \ln p(x_i)\big\rangle_{-x_i}\right\} \\
&\propto p(x_i) \exp\left\{-\hat{\gamma}\big\langle\|\mathbf{r} - \mathbf{Hx}\|^2\big\rangle_{-x_i}\right\} \\
&\propto p(x_i) \exp\left\{-\hat{\gamma}\bigg[\|\mathbf{h}_i\|^2 |x_i|^2 \right. \\
&\qquad\qquad \left. - 2\Re\bigg\{\mathbf{h}_i^H\bigg(\langle\mathbf{r}\rangle - \sum_{j \neq i}^{K} \mathbf{h}_j\langle x_j\rangle\bigg)x_i^*\bigg\}\bigg]\right\} \\
&\propto p(x_i) \exp\left\{-\hat{\gamma}\|\mathbf{h}_i\|^2\big(|x_i|^2 - 2\Re\{x_i^* z_i\}\big)\right\} \\
&\propto p(x_i) \exp\left\{-\hat{\gamma}\|\mathbf{h}_i\|^2 |x_i - z_i|^2\right\} \\
&\propto p(x_i) \mathcal{CN}\big(z_i; x_i, 1/(\hat{\gamma}\|\mathbf{h}\|^2)\big), \tag{16}
\end{aligned}
$$

where we define

$$
\begin{aligned}
z_i &= \frac{\mathbf{h}_i^H}{\|\mathbf{h}_i\|^2}\bigg(\langle\mathbf{r}\rangle - \sum_{j \neq i}^{K} \mathbf{h}_j\langle x_j\rangle\bigg) \\
&= \langle x_i \rangle + \frac{\mathbf{h}_i^H}{\|\mathbf{h}_i\|^2}\big(\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle\big) \tag{17}
\end{aligned}
$$

with $\langle x_i \rangle$ being the *currently fixed* nonlinear estimate of $x_i, \forall i$. We can see in (16) that the mean field VB approximation decouples the few-bit MIMO system into an AWGN channel $z_i = x_i + \mathcal{CN}\big(0, 1/(\hat{\gamma}\|\mathbf{h}\|^2)\big)$ for user-i. The variational distribution $q(x_i)$ can be realized by normalizing $p(x_i) \mathcal{CN}\big(z_i; x_i, 1/(\hat{\gamma}\|\mathbf{h}\|^2)\big)$. The variational mean $\langle x_i \rangle$ and variance τ_{x_i} are now updated as $\mathsf{F}_x\big(z_i, \hat{\gamma}\|\mathbf{h}_i\|^2\big)$ and $\mathsf{G}_x\big(z_i, \hat{\gamma}\|\mathbf{h}_i\|^2\big)$, respectively.[2]

---

[2] The computations of the mean $\mathsf{F}_x(z, \gamma)$ and variance $\mathsf{G}_x(z, \gamma)$ of a discrete random variable x given a prior distribution $p(x)$ and the observation $z = x + \mathcal{CN}(0, \gamma^{-1})$ are presented in Appendix C.

In the **M-step**, the estimate of γ is updated to maximize $\ln p(\mathbf{y}, \mathbf{r}, \mathbf{x}; \gamma, \mathbf{H})$ w.r.t. $q(\mathbf{r}, \mathbf{x})$, i.e.,

$$\hat{\gamma} = \arg \max_{\gamma} \; \langle \ln p(\mathbf{r}|\mathbf{x}; \gamma; \mathbf{H}) \rangle$$
$$= \arg \max_{\gamma} \; M \ln \gamma - \gamma \langle \|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 \rangle. \quad (18)$$

Applying Theorem 1 to evaluate the expectation $\langle \|\mathbf{r} - \mathbf{H}\mathbf{x}\|^2 \rangle$, the new estimate of γ is given by

$$\hat{\gamma} = \frac{M}{\|\langle \mathbf{r} \rangle - \mathbf{H}\langle \mathbf{x} \rangle\|^2 + \sum_{m=1}^{M} \tau_{r_m} + \sum_{i=1}^{K} \tau_{x_i} \|\mathbf{h}_i\|^2}. \quad (19)$$

By iteratively optimizing $q(\mathbf{r})$, $\{q(x_i)\}$, and updating $\hat{\gamma}$, we obtain the variational Bayes expectation-maximization (VBEM) algorithm for estimating $\mathbf{r}$, $\mathbf{x}$, and γ. Similar to our previous work [27], we refer this scheme to as the **MF-QVB algorithm** due to the use of the matched-filter $\mathbf{h}_i^H / \|\mathbf{h}_i\|^2$ to obtain the linear estimate z_i of x_i in (17). If γ is fixed to N_0^{-1}, the **MF-QVB algorithm** will be referred to as the **conv-QVB algorithm**, that was investigated as the QVB-CSIR algorithm in [26]. The MF-QVB approach is summarized in Algorithm 1. Here, we use $\hat{\mathbf{r}}^{\ell}$, $\hat{x}_i^{\ell}$, and $\hat{\gamma}^{\ell}$ to replace $\langle \mathbf{r} \rangle$, $\langle x_i \rangle$, and $\hat{\gamma}$ at iteration ℓ and each iteration consists of one round of updating the estimation of $\mathbf{r}$, $\mathbf{x}$, and γ. To reduce the complexity of the algorithm, we include the residual term $\mathbf{e}$, which is initialized as $\hat{\mathbf{r}}^1 - \mathbf{H}\hat{\mathbf{x}}^1$. We also define $\mathbf{s}^{\ell} = \hat{\mathbf{r}}^{\ell} - \mathbf{e}$ as an efficient way to compute $\mathbf{H}\hat{\mathbf{x}}^{\ell}$ at iteration-ℓ. The computation of $\mathbf{r}^{\ell+1}$ (and $\boldsymbol{\tau}_{\mathbf{r}}^{\ell+1}$) can be carried out element-wise in parallel.

B. Proposed LMMSE-QVB For Few-Bit MIMO Detection

We now develop the LMMSE-QVB method for few-bit MIMO detection that uses a postulated noise covariance matrix $\mathbf{C}^{\text{post}}$ instead of the postulated noise variance N_0^{post} in the MF-QVB method. The idea of using a postulated noise covariance matrix $\mathbf{C}^{\text{post}}$ was proposed in [27] but for infinite-resolution ADCs. For ease of computation, we use $\boldsymbol{\Gamma} = (\mathbf{C}^{\text{post}})^{-1}$ as the precision matrix to be estimated.

The joint distribution $p(\mathbf{y}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Gamma}, \mathbf{H})$ of the observed variable $\mathbf{y}$ and the latent variables $\mathbf{r}$ and $\mathbf{x}$ at time slot t can be factored as

$$p(\mathbf{y}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Gamma}, \mathbf{H})$$
$$= p(\mathbf{y}|\mathbf{r}) p(\mathbf{r}|\mathbf{x}; \boldsymbol{\Gamma}, \mathbf{H}) p(\mathbf{x})$$
$$= \left[\prod_{m=1}^{M} p(y_m|r_m) \right] p(\mathbf{r}|\mathbf{x}; \boldsymbol{\Gamma}, \mathbf{H}) \left[\prod_{i=1}^{K} p(x_i) \right], \quad (20)$$

where $p(\mathbf{r}|\mathbf{x}, \boldsymbol{\Gamma}; \mathbf{H}) = \mathcal{CN}(\mathbf{r}; \mathbf{H}\mathbf{x}, \boldsymbol{\Gamma}^{-1})$. We note that the random vector $\mathbf{r}$ is no longer comprised of conditional independent elements, since the noise covariance matrix $\boldsymbol{\Gamma}^{-1}$ is in general non-diagonal.

In the **E-step**, for a currently fixed estimate $\hat{\boldsymbol{\Gamma}}$ of $\boldsymbol{\Gamma}$, we aim to derive the mean field variational distribution $q(\mathbf{r}, \mathbf{x})$ of $\mathbf{r}$ and $\mathbf{x}$ given $\mathbf{y}$ such that

$$p(\mathbf{r}, \mathbf{x}|\mathbf{y}; \hat{\boldsymbol{\Gamma}}, \mathbf{H}) \approx q(\mathbf{r}, \mathbf{x}) = \left[\prod_{m=1}^{M} q(r_m) \right] \left[\prod_{i=1}^{K} q(x_i) \right]. \quad (21)$$

1) Updating r_m. The variational distribution $q(r_m)$ is obtained by taking the expectation of the conditional in (20) w.r.t. $q(\mathbf{x}) \prod_{n \neq m} q(r_n)$:

$$q(r_m) \propto \exp \left\{ \left\langle \ln p(y_m|r_m) + \ln p(\mathbf{r}|\mathbf{x}; \hat{\boldsymbol{\Gamma}}, \mathbf{H}) \right\rangle_{-r_m} \right\}$$
$$\propto \exp \left\{ \ln \mathbb{1}\left(r_m \in [y_m^{\text{low}}, y_m^{\text{up}}] \right) \right.$$
$$\left. - \left\langle (\mathbf{r} - \mathbf{H}\mathbf{x})^H \hat{\boldsymbol{\Gamma}} (\mathbf{r} - \mathbf{H}\mathbf{x}) \right\rangle_{-r_m} \right\}$$
$$\propto \mathbb{1}\left(r_m \in [y_m^{\text{low}}, y_m^{\text{up}}] \right)$$
$$\times \exp \left\{ - \hat{\gamma}_{mm} \left(|r_m|^2 - 2 \Re\{ r_m^* \mathbf{H}_{m,:} \langle \mathbf{x} \rangle \} \right)^2 \right.$$
$$\left. - 2 \sum_{n \neq m}^{M} \Re\{ r_m^* \hat{\gamma}_{mn} (\langle r_n \rangle - \mathbf{H}_{n,:} \langle \mathbf{x} \rangle) \} \right\}$$
$$\propto \mathbb{1}\left(r_m \in [y_m^{\text{low}}, y_m^{\text{up}}] \right) \times \exp \left\{ - \hat{\gamma}_{mm} |r_m - s_m|^2 \right\}$$
$$\propto \mathbb{1}\left(r_m \in [y_m^{\text{low}}, y_m^{\text{up}}] \right) \times \mathcal{CN}\left(r_m; s_m, \hat{\gamma}_{mm}^{-1} \right), \quad (22)$$

where s_m is now defined as

$$s_m = \mathbf{H}_{m,:} \langle \mathbf{x} \rangle - \hat{\gamma}_{mm}^{-1} \sum_{n \neq m}^{M} \hat{\gamma}_{mn} (\langle r_n \rangle - \mathbf{H}_{n,:} \langle \mathbf{x} \rangle)$$
$$= \langle r_m \rangle - \frac{\hat{\boldsymbol{\Gamma}}_m^H}{\hat{\gamma}_{mm}} (\langle \mathbf{r} \rangle - \mathbf{H}\langle \mathbf{x} \rangle), \quad (23)$$

$\langle r_m \rangle$ is the *currently fixed* nonlinear estimate of r_m and $\hat{\boldsymbol{\Gamma}}_m$ is the mth column of the Hermitian matrix $\hat{\boldsymbol{\Gamma}}$. We can see in (22) that the variational distribution $q(r_m)$ is the truncated complex normal distribution obtained from bounding $r_m \sim \mathcal{CN}(s_m, \hat{\gamma}_{mm}^{-1})$ to the interval $(y_m^{\text{low}}, y_m^{\text{up}})$. Thus, its mean $\langle r_m \rangle$ and variance τ_{r_m} are updated as $\mathsf{F}_r(s_m, \hat{\gamma}_{mm}, y_m^{\text{low}}, y_m^{\text{up}})$ and $\mathsf{G}_r(s_m, \hat{\gamma}_{mm}, y_m^{\text{low}}, y_m^{\text{up}})$, respectively.

2) Updating x_i. The variational distribution $q(x_i)$ is obtained by taking the expectation of the conditional in (13) w.r.t. $q(\mathbf{r}) \prod_{j \neq i} q(x_j)$:

$$q(x_i) \propto \exp \left\{ \left\langle \ln p(\mathbf{r}|\mathbf{x}; \hat{\boldsymbol{\Gamma}}, \mathbf{H}) + \ln p(x_i) \right\rangle_{-x_i} \right\}$$
$$\propto p(x_i) \exp \left\{ - \left\langle (\mathbf{r} - \mathbf{H}\mathbf{x})^H \hat{\boldsymbol{\Gamma}} (\mathbf{r} - \mathbf{H}\mathbf{x}) \right\rangle_{-x_i} \right\}$$
$$\propto p(x_i) \exp \left\{ - \mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i |x_i - z_i|^2 \right\}$$
$$\propto p(x_i) \mathcal{CN}\left(z_i; x_i, 1/(\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i) \right), \quad (24)$$

where z_i is a linear estimate of x_i that is now defined as

$$z_i = \frac{\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}}}{\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i} \left(\langle \mathbf{r} \rangle - \sum_{j \neq i}^{K} \mathbf{h}_j \langle x_j \rangle \right)$$
$$= \langle x_i \rangle + \frac{\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}}}{\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i} \left(\langle \mathbf{r} \rangle - \mathbf{H}\langle \mathbf{x} \rangle \right), \quad (25)$$

and $\langle x_i \rangle$ is the *current* nonlinear estimate of x_i. Here, z_i is the LMMSE estimate of x_i using the LMMSE filter $\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} / (\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i)$. The variational distribution $q(x_i)$ can be realized by normalizing $p(x_i) \mathcal{CN}(z_i; x_i, 1/(\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i))$. The variational mean $\langle x_i \rangle$ and variance τ_{x_i} are updated as $\mathsf{F}_x(z_i, \mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i)$ and $\mathsf{G}_x(z_i, \mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i)$, respectively.

In the **M-step**, the estimate of $\boldsymbol{\Gamma}$ is updated to maximize $\ln p(\mathbf{y}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Gamma}, \mathbf{H})$ w.r.t. $q(\mathbf{r}, \mathbf{x})$, i.e.,

$$\hat{\boldsymbol{\Gamma}} = \arg\max_{\boldsymbol{\Gamma}} \ \big\langle \ln p(\mathbf{r}|\mathbf{x}; \boldsymbol{\Gamma}, \mathbf{H}) \big\rangle$$
$$= \arg\max_{\boldsymbol{\Gamma}} \ \ln|\boldsymbol{\Gamma}| - \big\langle (\mathbf{r} - \mathbf{Hx})^H \boldsymbol{\Gamma} (\mathbf{r} - \mathbf{Hx}) \big\rangle. \quad (26)$$

By applying Theorem 1, we have

$$\big\langle (\mathbf{r} - \mathbf{Hx})^H \boldsymbol{\Gamma} (\mathbf{r} - \mathbf{Hx}) \big\rangle$$
$$= \mathrm{Tr}\Big\{ \big[(\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle)(\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle)^H + \boldsymbol{\Sigma}_{\mathbf{r}} + \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H \big] \boldsymbol{\Gamma} \Big\}, \quad (27)$$

where $\boldsymbol{\Sigma}_{\mathbf{r}} = \mathrm{diag}(\tau_{r_1}, \ldots, \tau_{r_M})$ and $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathrm{diag}(\tau_{x_1}, \ldots, \tau_{x_K})$. Thus, a new estimate of $\boldsymbol{\Gamma}$ is given by

$$\hat{\boldsymbol{\Gamma}} = \Big((\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle)(\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle)^H + \boldsymbol{\Sigma}_{\mathbf{r}} + \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H \Big)^{-1}. \quad (28)$$

We note that the matrix inversion in (28) often results in numerical errors due to the rank deficiency of $(\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle)(\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle)^H + \boldsymbol{\Sigma}_{\mathbf{r}} + \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H$. Similar to the approach in [27], we propose to use the following estimator

$$\hat{\boldsymbol{\Gamma}} = \left(\frac{\|\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle\|^2}{M} \mathbf{I}_M + \boldsymbol{\Sigma}_{\mathbf{r}} + \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H \right)^{-1}. \quad (29)$$

for the precision matrix $\boldsymbol{\Gamma}$.

By iteratively optimizing $\{q(r_m)\}$, $\{q(x_i)\}$, and $\hat{\boldsymbol{\Gamma}}$, we obtain the VBEM algorithm for estimating $\mathbf{r}$, $\mathbf{x}$, and $\boldsymbol{\Gamma}$. We refer to this scheme as the **LMMSE-QVB algorithm** due to the use of the LMMSE filter $\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} / (\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}} \mathbf{h}_i)$ to obtain the linear estimate z_i of x_i in (25). The LMMSE-QVB approach is summarized in Algorithm 2. As before, we use $\hat{r}_m^\ell$, $\hat{x}_i^\ell$, and $\hat{\boldsymbol{\Gamma}}^\ell$ to replace $\langle r_m\rangle$, $\langle x_i\rangle$, and $\hat{\boldsymbol{\Gamma}}$ at iteration ℓ and each iteration consists of one round of updating the estimation of $\mathbf{r}$, $\mathbf{x}$, and $\boldsymbol{\Gamma}$. Unlike MF-QVB, the LMMSE-QVB algorithm requires sequential updates over $\{\hat{r}_m^\ell\}$. Note that LMMSE-QVB is equivalent to MF-QVB in Algorithm 1 if $\hat{\boldsymbol{\Gamma}}^\ell$ is set as $\hat{\gamma}^\ell \mathbf{I}_M$.

C. Practical Aspects of Implementing MF/LMMSE-QVB

1) Computing $\langle r_m\rangle$ and τ_{r_m}: In MF-QVB, $\langle r_m\rangle$ and τ_{r_m} are updated as $\mathsf{F}_r(s_m, \hat{\gamma}, y_m^{\mathrm{low}}, y_m^{\mathrm{up}})$ and $\mathsf{G}_r(s_m, \hat{\gamma}, y_m^{\mathrm{low}}, y_m^{\mathrm{up}})$, respectively. The computations of these two terms presented in Appendix B can result in catastrophic cancellation when $\Phi(\beta) = \Phi(\alpha)$, even if $\beta = \sqrt{2\hat{\gamma}}(y_m^{\mathrm{low}} - s_m)$ and $\alpha = \sqrt{2\hat{\gamma}}(y_m^{\mathrm{up}} - s_m)$ are different. This numerical error often occurs when s_m is not inside the interval $(y_m^{\mathrm{low}}, y_m^{\mathrm{up}})$ and $\langle\gamma\rangle$ is large, i.e., for high SNR. To improve the robustness of the computation, we use the logistic CDF $F(x; c) = \frac{1}{1 + e^{-cx}}$ and logistic PDF $p(x; c) = cF(x; c)(1 - F(x; c))$ in lieu of the normal CDF $\Phi(x)$ and normal PDF $\phi(x)$. We choose $c = 3/\sqrt{\pi}$ to impose a unit variance on the logistic distribution. Note that $F(x)$ is much easier to compute than $\Phi(x)$. We have observed through numerous simulations that this modification eliminates numerical errors due to the heavier tails in the logistic distribution compared to the normal distribution. Interestingly, the

---

Algorithm 2 – LMMSE-QVB for Few-Bit MIMO Detection

1: **Input:** $\mathbf{y}$, $\mathbf{H}$
2: **Output:** $\hat{\mathbf{x}}$
3: Initialize $\hat{x}_i^1 = 0$ and $\tau_{x_i}^1 = \mathrm{Var}_{p(x_i)}[x_i], \forall i$, $\hat{r}_m^1 = y_m$ and $\tau_{r_m}^1 = 0, \forall m$, and $\mathbf{e} = \hat{\mathbf{r}}^1 - \mathbf{H}\hat{\mathbf{x}}^1$
4: **for** $\ell = 1, 2, \ldots$ **do**
5: $\quad \boldsymbol{\Sigma}_{\mathbf{r}} \leftarrow \mathrm{diag}(\tau_{r_1}^\ell, \ldots, \tau_{r_M}^\ell)$
6: $\quad \boldsymbol{\Sigma}_{\mathbf{x}} \leftarrow \mathrm{diag}(\tau_{x_1}^\ell, \ldots, \tau_{x_K}^\ell)$
7: $\quad \hat{\boldsymbol{\Gamma}}^\ell \leftarrow \big((\|\mathbf{e}\|^2/M)\mathbf{I}_M + \boldsymbol{\Sigma}_{\mathbf{r}} + \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H \big)^{-1}$
8: $\quad$ **for** $m = 1, \ldots, M$ **do** $\qquad\qquad \triangleright$ Update $\mathbf{r}$
9: $\quad\quad s_m^\ell \leftarrow \hat{r}_m^\ell - (\hat{\boldsymbol{\Gamma}}_m^\ell)^H \mathbf{e}/w_{mm}^\ell$
10: $\quad\quad \hat{r}_m^{\ell+1} \leftarrow \mathsf{F}_r(s_m^\ell, \hat{w}_{mm}^\ell, y_m^{\mathrm{low}}, y_m^{\mathrm{up}})$
11: $\quad\quad \tau_{r_m}^{\ell+1} \leftarrow \mathsf{G}_r(s_m^\ell, \hat{w}_{mm}^\ell, y_m^{\mathrm{low}}, y_m^{\mathrm{up}})$
12: $\quad\quad e_m \leftarrow e_m - \hat{r}_m^\ell + \hat{r}_m^{\ell+1}$ $\quad\quad \triangleright$ Update residual
13: $\quad$ **for** $i = 1, \ldots, K$ **do** $\qquad\qquad \triangleright$ Update $\mathbf{x}$
14: $\quad\quad z_i^\ell \leftarrow \hat{x}_i^\ell + \mathbf{h}_i^H \hat{\boldsymbol{\Gamma}}^\ell \mathbf{e}/(\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}}^\ell \mathbf{h}_i)$ $\quad \triangleright$ LMMSE
15: $\quad\quad \hat{x}_i^{\ell+1} \leftarrow \mathsf{F}_x(z_i^\ell, \mathbf{h}_i^H \hat{\boldsymbol{\Gamma}}^\ell \mathbf{h}_i)$
16: $\quad\quad \tau_{x_i}^{\ell+1} \leftarrow \mathsf{G}_x(z_i^\ell, \mathbf{h}_i^H \hat{\boldsymbol{\Gamma}}^\ell \mathbf{h}_i)$
17: $\quad\quad \mathbf{e} \leftarrow \mathbf{e} + \mathbf{h}_i(\hat{x}_i^\ell - \hat{x}_i^{\ell+1})$ $\quad\quad \triangleright$ Update residual
18: $\forall i : \hat{x}_i \leftarrow \arg\max_{a \in \mathcal{S}} p_a \mathcal{CN}(z_i^\ell; a, 1/(\mathbf{h}_i^H \hat{\boldsymbol{\Gamma}}^\ell \mathbf{h}_i))$.

---

detection accuracy is slightly better when using $F(x; c)$ and $p(x; c)$ rather than $\Phi(x)$ and $\phi(x)$, even when no numerical errors occur using the latter approach. We also use $F(x; c)$ and $p(x; c)$ in LMMSE-QVB and observe the same effect.

2) Using $\mathbf{e}$: The residual term $\mathbf{e}$ is included in MF/LMMSE-QVB reduce the computational complexity of these algorithms. Due to the sequential nature of VB, s_m and z_i are computed using the latest updated values of $\langle\mathbf{r}\rangle$ and $\langle\mathbf{x}\rangle$. Instead of computing $\langle\mathbf{r}\rangle - \mathbf{H}\langle\mathbf{x}\rangle$ each iteration, which induces a complexity of $\mathcal{O}(MK)$, we use the current value of the residual term $\mathbf{e}$. The update of $\mathbf{e}$ reflecting any update on estimation of $\mathbf{r}$ or x_i only induces a complexity of $\mathcal{O}(M)$.

V. VB for Joint Channel Estimation and Data Detection in Few-Bit MIMO Systems

The algorithms in the previous section assumed that the CSI was already obtained by some other method prior to data detection. In this section, we generalize the MF-QVB and LMMSE-QVB approaches to perform joint channel estimation and data detection.

A. Proposed MF-QVB-JED Algorithm

We denote γ_{p} and $\boldsymbol{\gamma}_{\mathrm{d}} = [\gamma_{\mathrm{d},1}, \ldots, \gamma_{\mathrm{d},T_{\mathrm{d}}}]^T$ as the precision of the noise in the pilot transmission phase and the T_{d} data transmission time slots, respectively. The factorization of the joint distribution of all the observed and latent variables in the system model (3)–(4) is given in (30), where $p(\mathbf{r}_{\mathrm{p},t}|\mathbf{H}; \gamma_{\mathrm{p}}, \mathbf{x}_{\mathrm{p},t}) = \mathcal{CN}(\mathbf{r}_{\mathrm{p},t}; \mathbf{Hx}_{\mathrm{p},t}, \gamma_{\mathrm{p}}^{-1}\mathbf{I}_M)$ and $p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \gamma_{\mathrm{d},t}) = \mathcal{CN}(\mathbf{r}_{\mathrm{d},t}; \mathbf{Hx}_{\mathrm{d},t}, \gamma_{\mathrm{d},t}^{-1}\mathbf{I}_M)$.

In the **E-step**, for currently fixed estimates $\hat{\gamma}_{\mathrm{p}}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{d}}$ of γ_{p} and $\boldsymbol{\gamma}_{\mathrm{d}}$, respectively, we aim to obtain the mean field

$$p(\mathbf{Y}_\mathrm{p}, \mathbf{Y}_\mathrm{d}, \mathbf{R}_\mathrm{p}, \mathbf{R}_\mathrm{d}, \mathbf{X}_\mathrm{d}, \mathbf{H}; \gamma_\mathrm{p}, \boldsymbol{\gamma}_\mathrm{d}, \mathbf{X}_\mathrm{p}, \{\mathbf{C}_i\})$$
$$= p(\mathbf{Y}_\mathrm{p}|\mathbf{R}_\mathrm{p})p(\mathbf{R}_\mathrm{p}|\mathbf{H}; \gamma_\mathrm{p}, \mathbf{X}_\mathrm{p})p(\mathbf{Y}_\mathrm{d}|\mathbf{R}_\mathrm{d})p(\mathbf{R}_\mathrm{d}|\mathbf{X}_\mathrm{d}, \mathbf{H}; \boldsymbol{\gamma}_\mathrm{d})p(\mathbf{X}_\mathrm{d})p(\mathbf{H}|\{\mathbf{C}_i\})$$
$$= \left[\prod_{t=1}^{T_\mathrm{p}} p(\mathbf{y}_{\mathrm{p},t}|\mathbf{r}_{\mathrm{p},t})p(\mathbf{r}_{\mathrm{p},t}|\mathbf{H}; \gamma_\mathrm{p}, \mathbf{x}_{\mathrm{p},t}) \right] \left[\prod_{t=1}^{T_\mathrm{d}} p(\mathbf{y}_{\mathrm{d},t}|\mathbf{r}_{\mathrm{d},t})p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \gamma_{\mathrm{d},t})p(\mathbf{x}_{\mathrm{d},t}) \right] \prod_{i=1}^{K} p(\mathbf{h}_i|\mathbf{C}_i). \tag{30}$$

$$q(\mathbf{h}_i) \propto \exp\left\{ -\left\langle \hat{\gamma}_\mathrm{p} \sum_{t=1}^{T_\mathrm{p}} \|\mathbf{r}_{\mathrm{p},t} - \mathbf{H}\mathbf{x}_{\mathrm{p},t}\|^2 + \sum_{t=1}^{T_\mathrm{d}} \hat{\gamma}_{\mathrm{d},t} \|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2 + \mathbf{h}_i^H \mathbf{C}_i^{-1} \mathbf{h}_i \right\rangle_{-\mathbf{h}_i} \right\}$$
$$\propto \exp\left\{ -\mathbf{h}_i^H \left[\left(\hat{\gamma}_\mathrm{p} \sum_{t=1}^{T_\mathrm{p}} |x_{\mathrm{p},it}|^2 + \sum_{t=1}^{T_\mathrm{d}} \hat{\gamma}_{\mathrm{d},t} \langle |x_{\mathrm{d},it}|^2 \rangle \right) \mathbf{I}_M + \mathbf{C}_i^{-1} \right] \mathbf{h}_i \right.$$
$$\left. + 2\Re\left\{ \mathbf{h}_i^H \left[\hat{\gamma}_\mathrm{p} \sum_{t=1}^{T_\mathrm{p}} \left(\langle \mathbf{r}_{\mathrm{p},t} \rangle - \sum_{j \neq i}^{K} \langle \mathbf{h}_j \rangle x_{\mathrm{p},jt} \right) x_{\mathrm{p},it}^* + \sum_{t=1}^{T_\mathrm{d}} \hat{\gamma}_{\mathrm{d},t} \left(\langle \mathbf{r}_{\mathrm{d},t} \rangle - \sum_{j \neq i}^{K} \langle \mathbf{h}_j \rangle \langle x_{\mathrm{d},jt} \rangle \right) \langle x_{\mathrm{d},it}^* \rangle \right] \right\} \right\}. \tag{35}$$

$$\boldsymbol{\Sigma}_{\mathbf{h}_i} = \left[\left(\hat{\gamma}_\mathrm{p} \sum_{t=1}^{T_\mathrm{p}} |x_{\mathrm{p},it}|^2 + \sum_{t=1}^{T_\mathrm{d}} \hat{\gamma}_{\mathrm{d},t} \langle |x_{\mathrm{d},it}|^2 \rangle \right) \mathbf{I}_M + \mathbf{C}_i^{-1} \right]^{-1} \tag{36}$$

$$\langle \mathbf{h}_i \rangle = \boldsymbol{\Sigma}_{\mathbf{h}_i} \left[\hat{\gamma}_\mathrm{p} \sum_{t=1}^{T_\mathrm{p}} \left(\langle \mathbf{r}_{\mathrm{p},t} \rangle - \sum_{j \neq i}^{K} \langle \mathbf{h}_j \rangle x_{\mathrm{p},jt} \right) x_{\mathrm{p},it}^* + \sum_{t=1}^{T_\mathrm{d}} \hat{\gamma}_{\mathrm{d},t} \left(\langle \mathbf{r}_{\mathrm{d},t} \rangle - \sum_{j \neq i}^{K} \langle \mathbf{h}_j \rangle \langle x_{\mathrm{d},jt} \rangle \right) \langle x_{\mathrm{d},it}^* \rangle \right]. \tag{37}$$

variational distribution $q(\mathbf{R}_\mathrm{p}, \mathbf{R}_\mathrm{d}, \mathbf{H}, \mathbf{X}_\mathrm{d})$ of $\mathbf{R}_\mathrm{p}$, $\mathbf{R}_\mathrm{d}$, $\mathbf{H}$, and $\mathbf{X}_\mathrm{d}$ given $\mathbf{Y}_\mathrm{p}$ and $\mathbf{Y}_\mathrm{d}$ such that

$$p(\mathbf{R}_\mathrm{p}, \mathbf{R}_\mathrm{d}, \mathbf{X}_\mathrm{d}, \mathbf{H}|\mathbf{Y}_\mathrm{p}, \mathbf{Y}_\mathrm{d}; \hat{\gamma}_\mathrm{p}, \hat{\boldsymbol{\gamma}}_\mathrm{d}, \mathbf{X}_\mathrm{p}, \{\mathbf{C}_i\})$$
$$\approx q(\mathbf{R}_\mathrm{p}, \mathbf{R}_\mathrm{d}, \mathbf{X}_\mathrm{d}, \mathbf{H})$$
$$= \left[\prod_{t=1}^{T_\mathrm{p}} q(\mathbf{r}_{\mathrm{p},t}) \right] \left[\prod_{t=1}^{T_\mathrm{d}} q(\mathbf{r}_{\mathrm{d},t}) \right] \left[\prod_{i=1}^{K} q(\mathbf{h}_i) \right] \left[\prod_{i=1}^{K} \prod_{t=1}^{T_\mathrm{d}} q(x_{\mathrm{d},it}) \right]. \tag{31}$$

1) Updating $\mathbf{r}_{\mathrm{p},t}$. Taking the expectation of the conditional (30) w.r.t. all latent variables except $\mathbf{r}_{\mathrm{p},t}$, the variational distribution $q(\mathbf{r}_{\mathrm{p},t})$ is given by

$$q(\mathbf{r}_{\mathrm{p},t})$$
$$\propto \exp\left\{ \left\langle \ln p(\mathbf{y}_{\mathrm{p},t}|\mathbf{r}_{\mathrm{p},t}) + \ln p(\mathbf{r}_{\mathrm{p},t}|\mathbf{H}; \hat{\gamma}_\mathrm{p}, \mathbf{x}_{\mathrm{p},t}) \right\rangle_{-\mathbf{r}_{\mathrm{p},t}} \right\}$$
$$\propto \exp\left\{ \left\langle \ln \mathbb{1}(\mathbf{r}_{\mathrm{p},t} \in [\mathbf{y}_{\mathrm{p},t}^\mathrm{low}, \mathbf{y}_{\mathrm{p},t}^\mathrm{up}]) - \right.\right.$$
$$\left.\left. \hat{\gamma}_\mathrm{p} \|\mathbf{r}_{\mathrm{p},t} - \mathbf{H}\mathbf{x}_{\mathrm{p},t}\|^2 \right\rangle_{-\mathbf{r}_{\mathrm{p},t}} \right\}$$
$$\propto \mathbb{1}(\mathbf{r}_{\mathrm{p},t} \in [\mathbf{y}_{\mathrm{p},t}^\mathrm{low}, \mathbf{y}_{\mathrm{p},t}^\mathrm{up}]) \times \exp\left\{ -\hat{\gamma}_\mathrm{p} \|\mathbf{r}_{\mathrm{p},t} - \langle \mathbf{H} \rangle \mathbf{x}_{\mathrm{p},t}\|^2 \right\}. \tag{32}$$

We note that the variational distribution $q(\mathbf{r}_{\mathrm{p},t})$ is inherently separable as $\prod_{m=1}^{M} q(r_{\mathrm{p},mt})$ and the variational distribution $q(r_{\mathrm{p},mt})$ is the complex complex normal distribution obtained from bounding $r_{\mathrm{p},mt} \sim \mathcal{CN}(\langle \mathbf{H}_{m,:} \rangle \mathbf{x}_{\mathrm{p},t}, \hat{\gamma}_\mathrm{p}^{-1})$ to the interval $(y_{\mathrm{p},mt}^\mathrm{low}, y_{\mathrm{p},mt}^\mathrm{up})$. The variational mean $\langle r_{\mathrm{p},mt} \rangle$ and variance $\tau_{r_{\mathrm{p},mt}}$ are given by $\mathsf{F}_r(\langle \mathbf{H}_{m,:} \rangle \mathbf{x}_{\mathrm{p},t}, \hat{\gamma}_\mathrm{p}^{-1}, y_{\mathrm{p},mt}^\mathrm{low}, y_{\mathrm{p},mt}^\mathrm{up})$ and $\mathsf{G}_r(\langle \mathbf{H}_{m,:} \rangle \mathbf{x}_{\mathrm{p},t}, \hat{\gamma}_\mathrm{p}^{-1}, y_{\mathrm{p},mt}^\mathrm{low}, y_{\mathrm{p},mt}^\mathrm{up})$, respectively.

2) Updating $\mathbf{r}_{\mathrm{d},t}$. Taking the expectation of the conditional (30) w.r.t. all latent variables except $\mathbf{r}_{\mathrm{d},t}$, the variational

distribution $q(\mathbf{r}_{\mathrm{p},t})$ is given by

$$q(\mathbf{r}_{\mathrm{d},t})$$
$$\propto \exp\left\{ \left\langle \ln p(\mathbf{y}_{\mathrm{d},t}|\mathbf{r}_{\mathrm{d},t}) + \ln p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \hat{\gamma}_{\mathrm{d},t}) \right\rangle_{-\mathbf{r}_{\mathrm{d},t}} \right\}$$
$$\propto \exp\left\{ \left\langle \ln \mathbb{1}(\mathbf{r}_{\mathrm{d},t} \in [\mathbf{y}_{\mathrm{d},t}^\mathrm{low}, \mathbf{y}_{\mathrm{d},t}^\mathrm{up}]) - \right.\right.$$
$$\left.\left. \hat{\gamma}_{\mathrm{d},t} \|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2 \right\rangle_{-\mathbf{r}_{\mathrm{d},t}} \right\}$$
$$\propto \mathbb{1}(\mathbf{r}_{\mathrm{d},t} \in [\mathbf{y}_{\mathrm{d},t}^\mathrm{low}, \mathbf{y}_{\mathrm{d},t}^\mathrm{up}]) \times \exp\left\{ -\hat{\gamma}_{\mathrm{d},t} \|\mathbf{r}_{\mathrm{p},t} - \langle \mathbf{H} \rangle \langle \mathbf{x}_{\mathrm{d},t} \rangle\|^2 \right\}. \tag{33}$$

The update of $\mathbf{r}_{\mathrm{d},t}$ is similar to that of $\mathbf{r}_{\mathrm{p},t}$. Due to the inherent decoupling of $q(\mathbf{r}_{\mathrm{d},t})$, the variational mean $\langle r_{\mathrm{d},mt} \rangle$ and variance $\tau_{r_{\mathrm{d},mt}}$ are given by $\mathsf{F}_r(\langle \mathbf{H}_{m,:} \rangle \langle \mathbf{x}_{\mathrm{d},t} \rangle, \hat{\gamma}_{\mathrm{d},t}, y_{\mathrm{d},mt}^\mathrm{low}, y_{\mathrm{d},mt}^\mathrm{up})$ and $\mathsf{G}_r(\langle \mathbf{H}_{m,:} \rangle \langle \mathbf{x}_{\mathrm{d},t} \rangle, \hat{\gamma}_{\mathrm{d},t}, y_{\mathrm{d},mt}^\mathrm{low}, y_{\mathrm{d},mt}^\mathrm{up})$, respectively.

3) Updating $\mathbf{h}_i$. Taking the expectation of the conditional (30) w.r.t. all latent variables except $\mathbf{h}_i$, the variational distribution $q(\mathbf{h}_i)$ is given by

$$q(\mathbf{h}_i) \propto \exp\left\{ \left\langle \ln p(\mathbf{R}_\mathrm{p}|\mathbf{H}; \hat{\gamma}_\mathrm{p}, \mathbf{X}_\mathrm{p}) + \ln p(\mathbf{R}_\mathrm{d}|\mathbf{X}_\mathrm{d}, \mathbf{H}; \hat{\boldsymbol{\gamma}}_\mathrm{d}) \right.\right.$$
$$\left.\left. + \ln p(\mathbf{h}_i; \mathbf{C}_i) \right\rangle_{-\mathbf{h}_i} \right\}, \tag{34}$$

which is expanded into (35). Thus, the variational distribution $q_{\mathbf{h}_i}(\mathbf{h}_i)$ is the pdf of a Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}_{\mathbf{h}_i}$ given in (36) and mean $\langle \mathbf{h}_i \rangle$ given in (37).

4) Updating $x_{\mathrm{d},it}$. Taking the expectation of the conditional (30) w.r.t. all latent variables except $x_{\mathrm{d},it}$, the variational distribution $q(x_{\mathrm{d},it})$ is given by

$$q(x_{\mathrm{d},it}) \propto \exp\left\{ \left\langle \ln p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \hat{\gamma}_{\mathrm{d},t}) + \ln p(x_{\mathrm{d},it}) \right\rangle_{-x_{\mathrm{d},it}} \right\}$$
$$\propto p(x_{\mathrm{d},it}) \exp\left\{ -\hat{\gamma}_{\mathrm{d},t} \left\langle \|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2 \right\rangle_{-x_{\mathrm{d},it}} \right\}. \tag{38}$$

$$q(x_{\mathrm{d},it}) \propto p(x_{\mathrm{d},it}) \exp\left\{ - \hat{\gamma}_{\mathrm{d},t} \left[\langle \|\mathbf{h}_i\|^2 \rangle |x_{\mathrm{d},it}|^2 - 2\Re\left\{ \langle \mathbf{h}_i^H \rangle \left(\langle \mathbf{r}_{\mathrm{d},t} \rangle - \sum_{j\neq i}^{K} \langle \mathbf{h}_j \rangle \langle x_{\mathrm{d},jt} \rangle \right) x_{\mathrm{d},it}^* \right\} \right] \right\}$$

$$\propto p(x_{\mathrm{d},it}) \exp\left\{ - \hat{\gamma}_{\mathrm{d},t} \langle \|\mathbf{h}_i\|^2 \rangle |x_{\mathrm{d},it} - z_{\mathrm{d},it}|^2 \right\}$$

$$\propto p(x_{\mathrm{d},it}) \, \mathcal{CN}\left(z_{\mathrm{d},it}; x_{\mathrm{d},it}, 1/\left(\hat{\gamma}_{\mathrm{d},t} \langle \|\mathbf{h}_i\|^2 \rangle \right) \right). \tag{39}$$

Note that (38) can be expanded into (39), in which we define

$$z_{\mathrm{d},it} = \frac{\langle \mathbf{h}_i^H \rangle}{\langle \|\mathbf{h}_i\|^2 \rangle} \left(\langle \mathbf{r}_{\mathrm{d},t} \rangle - \sum_{j\neq i}^{K} \langle \mathbf{h}_j \rangle \langle x_{\mathrm{d},jt} \rangle \right)$$

$$= \frac{\|\langle \mathbf{h}_i \rangle\|^2 \langle x_{\mathrm{d},it} \rangle + \langle \mathbf{h}_i^H \rangle \left(\langle \mathbf{r}_{\mathrm{d},t} \rangle - \langle \mathbf{H} \rangle \langle \mathbf{x}_{\mathrm{d},t} \rangle \right)}{\langle \|\mathbf{h}_i\|^2 \rangle} \tag{40}$$

as a linear estimate of $x_{\mathrm{d},it}$. We note that $\langle \|\mathbf{h}_i\|^2 \rangle = \|\langle \mathbf{h}_i \rangle\|^2 + \mathrm{Tr}\{\boldsymbol{\Sigma}_{\mathbf{h}_i}\}$. The variational mean and variance of $x_{\mathrm{d},it}$ are given by $\mathsf{F}_x\left(z_{\mathrm{d},it}, \hat{\gamma}_{\mathrm{d},t} \langle \|\mathbf{h}_i\|^2 \rangle \right)$ and $\mathsf{G}_x\left(z_{\mathrm{d},it}, \hat{\gamma}_{\mathrm{d},t} \langle \|\mathbf{h}_i\|^2 \rangle \right)$, respectively.

==In the **_M-step_**, the estimates of γ_{p} and $\gamma_{\mathrm{d},t}$ are updated to maximize== $\ln p(\mathbf{Y}_{\mathrm{p}}, \mathbf{Y}_{\mathrm{d}}, \mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{X}_{\mathrm{d}}, \mathbf{H}; \gamma_{\mathrm{p}}, \boldsymbol{\gamma}_{\mathrm{d}}, \mathbf{X}_{\mathrm{p}}, \{\mathbf{C}_i\})$ w.r.t. the variational distribution $q(\mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{X}_{\mathrm{d}}, \mathbf{H})$, i.e.,

$$\hat{\gamma}_{\mathrm{p}} = \arg\max_{\gamma_{\mathrm{p}}} \ \big\langle \ln p(\mathbf{R}_{\mathrm{p}} | \mathbf{H}; \gamma_{\mathrm{p}}, \mathbf{X}_{\mathrm{p}}) \big\rangle$$

$$= \arg\max_{\gamma_{\mathrm{p}}} \ MT_{\mathrm{p}} - \gamma_{\mathrm{p}} \langle \|\mathbf{R}_{\mathrm{p}} - \mathbf{H}\mathbf{X}_{\mathrm{p}}\|^2 \rangle$$

$$= \frac{MT_{\mathrm{p}}}{\sum_{t=1}^{T_{\mathrm{p}}} \langle \|\mathbf{r}_{\mathrm{p},t} - \mathbf{H}\mathbf{x}_{\mathrm{p},t}\|^2 \rangle} \tag{41}$$

and for $t = 1, \ldots, T_{\mathrm{d}}$

$$\hat{\gamma}_{\mathrm{d},t} = \arg\max_{\gamma_{\mathrm{d},t}} \ \big\langle \ln p(\mathbf{r}_{\mathrm{d},t} | \mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \gamma_{\mathrm{d},t} \big\rangle$$

$$= \arg\max_{\gamma_{\mathrm{d},t}} \ T_{\mathrm{p}} - \gamma_{\mathrm{d},t} \langle \|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2 \rangle$$

$$= \frac{T_{\mathrm{p}}}{\langle \|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2 \rangle}. \tag{42}$$

By applying Theorem 1, we have

$$\langle \|\mathbf{r}_{\mathrm{p},t} - \mathbf{H}\mathbf{x}_{\mathrm{p},t}\|^2 \rangle = \|\langle \mathbf{r}_{\mathrm{p},t} \rangle - \langle \mathbf{H} \rangle \mathbf{x}_{\mathrm{p},t}\|^2 + \sum_{m=1}^{M} \tau_{r_{\mathrm{p},mt}}$$

$$+ \sum_{i=1}^{K} |x_{\mathrm{p},it}|^2 \mathrm{Tr}\{\boldsymbol{\Sigma}_{\mathbf{h}_i}\} \tag{43}$$

and

$$\langle \|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2 \rangle = \|\langle \mathbf{r}_{\mathrm{d},t} \rangle - \langle \mathbf{H} \rangle \langle \mathbf{x}_{\mathrm{d},t} \rangle\|^2 + \sum_{m=1}^{M} \tau_{r_{\mathrm{d},mt}}$$

$$+ \sum_{i=1}^{K} \left[\langle |x_{\mathrm{d},it}|^2 \rangle \mathrm{Tr}\{\boldsymbol{\Sigma}_{\mathbf{h}_i}\} + \tau_{x_i} \|\langle \mathbf{h}_i \rangle\|^2 \right]. \tag{44}$$

By iteratively optimizing $\{q(\mathbf{r}_{\mathrm{p},t})\}$, $\{q(\mathbf{r}_{\mathrm{d},t})\}$, $\{q(\mathbf{h}_i)\}$, $\{q(x_{\mathrm{d},it})\}$, $\hat{\gamma}_{\mathrm{p}}$, and $\{\hat{\gamma}_{\mathrm{d},t}\}$, we obtain the VBEM algorithm for estimating $\mathbf{R}_{\mathrm{p}}$, $\mathbf{R}_{\mathrm{d}}$, $\mathbf{H}$, $\mathbf{X}_{\mathrm{d}}$, γ_{p}, and $\{\gamma_{\mathrm{d},t}\}$. We refer to this scheme as the **_MF-QVB-JED algorithm_** for joint channel estimation and data detection.

Remark 1: _If γ_{p} and $\{\gamma_{\mathrm{d},t}\}$ are all set to N_0^{-1}, the MF-QVB-JED algorithm is equivalent to the VB-based joint channel estimation and data detection approach in [26]. We will refer to the scheme in [26] as the **conv-QVB-JED algorithm**. In conv-QVB-JED, the variational covariance matrix of $\mathbf{h}_i$ given by_

$$\boldsymbol{\Sigma}_{\mathbf{h}_i} = \left[N_0^{-1} \left(\sum_{t=1}^{T_{\mathrm{p}}} |x_{\mathrm{p},it}|^2 + \sum_{t=1}^{T_{\mathrm{d}}} \langle |x_{\mathrm{d},it}|^2 \rangle \right) \mathbf{I}_M + \mathbf{C}_i^{-1} \right]^{-1} \tag{45}$$

becomes smaller with increasing T_{d} or $\langle |x_{\mathrm{d},it}|^2 \rangle, \forall t$. This result, however, implies that the estimation of $\mathbf{h}_i$ becomes more accurate with a longer transmission phase or even with an unreliable estimate of $x_{\mathrm{d},it}$ reflected through large $\tau_{x_{\mathrm{d},it}}$ (and $\langle |x_{\mathrm{d},it}|^2 \rangle$). In MF-QVB-JED, an unreliable estimation of $x_{\mathrm{d},it}$ will decrease $\hat{\gamma}_{\mathrm{d},t}$ in (42). Evidently, the effect of $\langle |x_{\mathrm{d},it}|^2 \rangle$ on the variational covariance matrix $\boldsymbol{\Sigma}_{\mathbf{h}_i}$ of $\mathbf{h}_i$ in (36) is less important than its effect on $\boldsymbol{\Sigma}_{\mathbf{h}_i}$ in (45). Therefore, in the MF-QVB-JED algorithm, an unreliable estimate of $x_{\mathrm{d},it}$ will not increase the accuracy of estimating $\mathbf{h}_i$. This is one of explanations for the superior performance of MF-QVB-JED compared with conv-QVB-JED.

Remark 2: _By denoting_

$$\gamma_i = \hat{\gamma}_{\mathrm{p}} \sum_{t=1}^{T_{\mathrm{p}}} |x_{\mathrm{p},it}|^2 + \sum_{t=1}^{T_{\mathrm{d}}} \hat{\gamma}_{\mathrm{d},t} \langle |x_{\mathrm{d},it}|^2 \rangle \tag{46}$$

and

$$\mathbf{k}_i = \gamma_i^{-1} \left[\hat{\gamma}_{\mathrm{p}} \sum_{t=1}^{T_{\mathrm{p}}} \left(\langle \mathbf{r}_{\mathrm{p},t} \rangle - \sum_{j\neq i}^{K} \langle \mathbf{h}_j \rangle x_{\mathrm{p},jt} \right) x_{\mathrm{p},it}^* \right.$$

$$\left. + \sum_{t=1}^{T_{\mathrm{d}}} \hat{\gamma}_{\mathrm{d},t} \left(\langle \mathbf{r}_{\mathrm{d},t} \rangle - \sum_{j\neq i}^{K} \langle \mathbf{h}_j \rangle \langle x_{\mathrm{d},jt} \rangle \right) \langle x_{\mathrm{d},it}^* \rangle \right], \tag{47}$$

we note that the variational distribution $q(\mathbf{h}_i)$ in (35) can also be expressed as

$$q(\mathbf{h}_i) \propto \mathcal{CN}(\mathbf{h}_i; \mathbf{k}_i, \gamma_i^{-1}\mathbf{I}_M) \, \mathcal{CN}(\mathbf{h}_i; \mathbf{0}, \mathbf{C}_i)$$

$$= \mathcal{CN}\left(\mathbf{h}_i; \gamma_i \left(\gamma_i \mathbf{I}_M + \mathbf{C}_i^{-1} \right)^{-1} \mathbf{k}_i, \left(\gamma_i \mathbf{I}_M + \mathbf{C}_i^{-1} \right)^{-1} \right),$$

which then explains the results $\boldsymbol{\Sigma}_{\mathbf{h}_i} = \left(\gamma_i \mathbf{I}_M + \mathbf{C}_i^{-1} \right)^{-1}$ and $\langle \mathbf{h}_i \rangle = \gamma_i \boldsymbol{\Sigma}_{\mathbf{h}_i} \mathbf{k}_i$ in (36) and (37), respectively. We also note that $\mathbf{k}_i$ can be written as

$$\mathbf{k}_i = \left(1 - \frac{\sum_{t=1}^{T_{\mathrm{d}}} \hat{\gamma}_{\mathrm{d},t} \tau_{x_{\mathrm{d},it}}}{\gamma_i} \right) \langle \mathbf{h}_i \rangle$$

$$+ \gamma_i^{-1} \left[\hat{\gamma}_{\mathrm{p}} \mathbf{E}_{\mathrm{p}}[\mathbf{X}_{\mathrm{p}}]_{i,:}^H + \mathbf{E}_{\mathrm{d}}\left([\langle \mathbf{X}_{\mathrm{p}} \rangle]_{i,:}^H \odot \hat{\boldsymbol{\gamma}}_{\mathrm{d}} \right) \right],$$

enabling its efficient computation using the residual matrices $\mathbf{E}_{\mathrm{p}} = \langle \mathbf{R}_{\mathrm{p}} \rangle - \langle \mathbf{H} \rangle \mathbf{X}_{\mathrm{p}}$ and $\mathbf{E}_{\mathrm{d}} = \langle \mathbf{R}_{\mathrm{d}} \rangle - \langle \mathbf{H} \rangle \langle \mathbf{X}_{\mathrm{d}} \rangle$.

Algorithm 3 – MF-QVB-JED for Few-Bit MIMO Joint Channel Estimation and Data Detection

1: **Input:** $\mathbf{Y}_{\mathrm{p}}, \mathbf{Y}_{\mathrm{d}}, \mathbf{X}_{\mathrm{p}}, \mathbf{C}_i, \forall i$
2: **Output:** $\hat{\mathbf{H}}, \hat{\mathbf{X}}_{\mathrm{d}}$
3: Initialize $\hat{\mathbf{H}}^1 = \mathbf{0}$, $\hat{\mathbf{X}}_{\mathrm{d}}^1 = \mathbf{0}$, $\tau_{x_{\mathrm{d},it}}^1 = \mathrm{Var}_{p(x_i)}[x_i], \forall i, \forall t$,
   $\hat{\mathbf{R}}_{\mathrm{p}}^1 = \mathbf{Y}_{\mathrm{p}}$, $\hat{\mathbf{R}}_{\mathrm{d}}^1 = \mathbf{Y}_{\mathrm{d}}$, $\tau_{r_{\mathrm{p},mt}}^1 = 0$, $\tau_{r_{\mathrm{d},mt}}^1 = 0, \forall m, \forall t$,
   $\mathbf{E}_{\mathrm{p}} = \hat{\mathbf{R}}_{\mathrm{p}}^1 - \hat{\mathbf{H}}^1 \mathbf{X}_{\mathrm{p}}$, and $\mathbf{E}_{\mathrm{d}} = \hat{\mathbf{R}}_{\mathrm{d}}^1 - \hat{\mathbf{H}}^1 \hat{\mathbf{X}}_{\mathrm{d}}^1$.
4: **for** $\ell = 1, 2, \dots$ **do**
5:      Update $\hat{\gamma}_{\mathrm{p}}^\ell$ using (41) and (43)
6:      **for** $t = 1, \dots, T_{\mathrm{p}}$ **do**   ▷ Update $\mathbf{R}_{\mathrm{p}}$
7:          $\mathbf{s}_{\mathrm{p},t}^\ell \leftarrow \hat{\mathbf{r}}_{\mathrm{p},t}^\ell - [\mathbf{E}_{\mathrm{p}}]_{:,t}$
8:          $\hat{\mathbf{r}}_{\mathrm{p},t}^{\ell+1} \leftarrow \mathsf{F}_r(\mathbf{s}_{\mathrm{p},t}^\ell, \hat{\gamma}_{\mathrm{p}}^\ell, \mathbf{y}_{\mathrm{p},t}^{\mathrm{low}}, \mathbf{y}_{\mathrm{p},t}^{\mathrm{up}})$
9:          $\boldsymbol{\tau}_{\mathbf{r}_{\mathrm{p},t}}^{\ell+1} \leftarrow \mathsf{G}_r(\mathbf{s}_{\mathrm{p},t}^\ell, \hat{\gamma}_{\mathrm{p}}^\ell, \mathbf{y}_{\mathrm{p},t}^{\mathrm{low}}, \mathbf{y}_{\mathrm{p},t}^{\mathrm{up}})$
10:      $\mathbf{E}_{\mathrm{p}} \leftarrow \mathbf{E}_{\mathrm{p}} - \hat{\mathbf{R}}_{\mathrm{p}}^\ell + \hat{\mathbf{R}}_{\mathrm{p}}^{\ell+1}$
11:      **for** $t = 1, \dots, T_{\mathrm{d}}$ **do**   ▷ Update $\mathbf{R}_{\mathrm{d}}$
12:          Update $\hat{\gamma}_{\mathrm{d},t}^\ell$ using (42) and (44)
13:          $\mathbf{s}_{\mathrm{d},t}^\ell \leftarrow \hat{\mathbf{r}}_{\mathrm{d},t}^\ell - [\mathbf{E}_{\mathrm{d}}]_{t,:}$
14:          $\hat{\mathbf{r}}_{\mathrm{d},t}^{\ell+1} \leftarrow \mathsf{F}_r(\mathbf{s}_{\mathrm{d},t}^\ell, \hat{\gamma}_{\mathrm{d},t}^\ell, \mathbf{y}_{\mathrm{d},t}^{\mathrm{low}}, \mathbf{y}_{\mathrm{d},t}^{\mathrm{up}})$
15:          $\boldsymbol{\tau}_{\mathbf{r}_{\mathrm{d},t}}^{\ell+1} \leftarrow \mathsf{G}_r(\mathbf{s}_{\mathrm{d},t}^\ell, \hat{\gamma}_{\mathrm{d},t}^\ell, \mathbf{y}_{\mathrm{d},t}^{\mathrm{low}}, \mathbf{y}_{\mathrm{d},t}^{\mathrm{up}})$
16:      $\mathbf{E}_{\mathrm{d}} \leftarrow \mathbf{E}_{\mathrm{d}} - \hat{\mathbf{R}}_{\mathrm{d}}^\ell + \hat{\mathbf{R}}_{\mathrm{d}}^{\ell+1}$
17:      **for** $i = 1, \dots, K$ **do**   ▷ Update $\mathbf{H}$
18:          $\gamma_i^\ell \leftarrow \hat{\gamma}_{\mathrm{p}}^\ell \sum_{t=1}^{T_{\mathrm{p}}} |x_{\mathrm{p},it}|^2 + \sum_{t=1}^{T_{\mathrm{d}}} \hat{\gamma}_{\mathrm{d},t}^\ell \langle |x_{\mathrm{d},it}^\ell|^2 \rangle$
19:          $\mathbf{k}_i^\ell \leftarrow \left(1 - \frac{\sum_{t=1}^{T_{\mathrm{d}}} \hat{\gamma}_{\mathrm{d},t}^\ell \tau_{x_{\mathrm{d},it}}^\ell}{\gamma_i^\ell}\right)\hat{\mathbf{h}}_i^\ell$
              $+ (\gamma_i^\ell)^{-1}\left[\hat{\gamma}_{\mathrm{p}}^\ell \mathbf{E}_{\mathrm{p}}[\mathbf{X}_{\mathrm{p}}]_{i,:}^H + \mathbf{E}_{\mathrm{d}}([\mathbf{X}_{\mathrm{p}}^\ell]_{i,:}^H \odot \hat{\boldsymbol{\gamma}}_{\mathrm{d}}^\ell)\right]$
20:          $\boldsymbol{\Sigma}_{\mathbf{h}_i}^{\ell+1} \leftarrow (\gamma_i^\ell \mathbf{I}_M + \mathbf{C}_i^{-1})^{-1}$
21:          $\hat{\mathbf{h}}_i^{\ell+1} \leftarrow \gamma_i^\ell \boldsymbol{\Sigma}_{\mathbf{h}_i}^{\ell+1} \mathbf{k}_i^\ell$
22:          $\mathbf{E}_{\mathrm{p}} \leftarrow \mathbf{E}_{\mathrm{p}} + (\hat{\mathbf{h}}_i^\ell - \hat{\mathbf{h}}_i^{\ell+1})[\mathbf{X}_{\mathrm{p}}]_{i,:}$
23:          $\mathbf{E}_{\mathrm{d}} \leftarrow \mathbf{E}_{\mathrm{d}} + (\hat{\mathbf{h}}_i^\ell - \hat{\mathbf{h}}_i^{\ell+1})[\hat{\mathbf{X}}_{\mathrm{d}}^\ell]_{i,:}$
24:      **for** $t = 1, \dots, T_{\mathrm{d}}$ **do**   ▷ Update $\mathbf{X}_{\mathrm{d}}$
25:          **for** $i = 1, \dots, K$ **do**
26:             $z_{it}^\ell \leftarrow \frac{\|\hat{\mathbf{h}}_i^\ell\|^2 \hat{x}_{\mathrm{d},it}^\ell + (\hat{\mathbf{h}}_i^\ell)^H [\mathbf{E}_{\mathrm{d}}]_{:,t}}{\|\hat{\mathbf{h}}_i^\ell\|^2 + \mathrm{Tr}\{\boldsymbol{\Sigma}_{\mathbf{h}_i}^\ell\}}$
27:             $\hat{x}_{\mathrm{d},it}^{\ell+1} \leftarrow \mathsf{F}_x(z_{it}^\ell, \hat{\gamma}_{\mathrm{d},t}^\ell(\|\hat{\mathbf{h}}_i^\ell\|^2 + \mathrm{Tr}\{\boldsymbol{\Sigma}_{\mathbf{h}_i}^\ell\}))$
28:             $\tau_{x_{\mathrm{d},it}}^{\ell+1} \leftarrow \mathsf{G}_x(z_{it}^\ell, \hat{\gamma}_{\mathrm{d},t}^\ell(\|\hat{\mathbf{h}}_i^\ell\|^2 + \mathrm{Tr}\{\boldsymbol{\Sigma}_{\hat{\mathbf{h}}_i}^\ell\}))$
29:             $[\mathbf{E}_{\mathrm{d}}]_{:,t} \leftarrow [\mathbf{E}_{\mathrm{d}}]_{:,t} + \hat{\mathbf{h}}_i^\ell(\hat{x}_{\mathrm{d},it}^\ell - \hat{x}_{\mathrm{d},it}^{\ell+1})$
30: $\forall t, \forall i: \hat{x}_{\mathrm{d},it} \leftarrow \arg\max_{a \in \mathcal{S}} p_a \mathcal{CN}(z_{it}^\ell; a, 1/[\hat{\gamma}_{\mathrm{d},t}^\ell(\|\hat{\mathbf{h}}_i^\ell\|^2 + \mathrm{Tr}\{\boldsymbol{\Sigma}_{\mathbf{h}_i}^\ell\})])$.

The proposed MF-QVB-JED algorithm is summarized Algorithm 3. Here, we use $\hat{\mathbf{r}}_{\mathrm{p},t}^\ell$, $\hat{\mathbf{r}}_{\mathrm{d},t}^\ell$, $\hat{\mathbf{h}}_i^\ell$, $\hat{x}_i^\ell$, $\hat{\gamma}_{\mathrm{p}}^\ell$, and $\hat{\gamma}_{\mathrm{d},t}^\ell$ to replace $\langle \mathbf{r}_{\mathrm{p},t} \rangle$, $\langle \mathbf{r}_{\mathrm{d},t} \rangle$, $\langle \mathbf{h}_i \rangle$, $\langle x_i \rangle$, $\hat{\gamma}_{\mathrm{p}}$, and $\hat{\gamma}_{\mathrm{d},t}$ at iteration ℓ. We also include in the algorithm the residual terms $\mathbf{E}_{\mathrm{p}}$ and $\mathbf{E}_{\mathrm{d}}$, which are adjusted to reflect any update to the estimates of $\mathbf{R}_{\mathrm{p}}$, $\mathbf{R}_{\mathrm{d}}$, $\mathbf{H}$, and $\mathbf{X}_{\mathrm{d}}$.

B. Proposed LMMSE-QVB-JED Algorithm

This section extends the LMMSE-QVB algorithm to the case of joint channel estimation and data detection. We denote γ_{p} and $\{\boldsymbol{\Gamma}_t\} = \{\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_{T_{\mathrm{d}}}\}$ as the precision of the noise during the pilot transmission phase and the T_{d} data transmission time slots, respectively. The joint distribution of all the observations and latent

variables in (30) are now factored as given in (48), where $p(\mathbf{r}_{\mathrm{p},t}|\mathbf{H}; \gamma_{\mathrm{p}}, \mathbf{x}_{\mathrm{p},t}) = \mathcal{CN}(\mathbf{r}_{\mathrm{p},t}; \mathbf{H}\mathbf{x}_{\mathrm{p},t}, \gamma_{\mathrm{p}}^{-1}\mathbf{I}_M)$ and $p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \boldsymbol{\Gamma}_t) = \mathcal{CN}(\mathbf{r}_{\mathrm{d},t}; \mathbf{H}\mathbf{x}_{\mathrm{d},t}, \boldsymbol{\Gamma}_t^{-1})$.

In the **E-step**, for currently fixed estimates $\hat{\gamma}_{\mathrm{p}}$ and $\{\hat{\boldsymbol{\Gamma}}_t\}$ of γ_{p} and $\{\boldsymbol{\Gamma}_t\}$, respectively, we aim to obtain the mean field variational distribution $q(\mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{H}, \mathbf{X}_{\mathrm{d}})$ of $\mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{H}$, and $\mathbf{X}_{\mathrm{d}}$ given $\mathbf{Y}_{\mathrm{p}}$ and $\mathbf{Y}_{\mathrm{d}}$ such that

$$
\begin{aligned}
&p(\mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{X}_{\mathrm{d}}, \mathbf{H}|\mathbf{Y}_{\mathrm{p}}, \mathbf{Y}_{\mathrm{d}}; \hat{\gamma}_{\mathrm{p}}, \{\hat{\boldsymbol{\Gamma}}_t\}, \mathbf{X}_{\mathrm{p}}, \{\mathbf{C}_i\}) \\
&\approx q(\mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{X}_{\mathrm{d}}, \mathbf{H}) \\
&= \left[\prod_{t=1}^{T_{\mathrm{p}}} q(\mathbf{r}_{\mathrm{p},t})\right]\left[\prod_{t=1}^{T_{\mathrm{d}}} q(\mathbf{r}_{\mathrm{d},t})\right]\left[\prod_{i=1}^{K} q(\mathbf{h}_i)\right]\left[\prod_{i=1}^{K}\prod_{t=1}^{T_{\mathrm{d}}} q(x_{\mathrm{d},it})\right].
\end{aligned}
\tag{49}
$$

1) Updating $\mathbf{r}_{\mathrm{p},t}$. Similar to the MF-QVB-JED algorithm.

2) Updating $r_{\mathrm{d},mt}$. Similar to the LMMSE-QVB algorithm, the variational mean $\langle r_{\mathrm{d},mt} \rangle$ and variance $\tau_{r_{\mathrm{d},mt}}$ are determined by $\mathsf{F}_r(s_{\mathrm{d},mt}, \hat{\gamma}_{t,mm}, y_{\mathrm{d},mt}^{\mathrm{low}}, y_{\mathrm{d},mt}^{\mathrm{up}})$ and $\mathsf{G}_r(s_{\mathrm{d},mt}, \hat{\gamma}_{t,mm}, y_{\mathrm{d},mt}^{\mathrm{low}}, y_{\mathrm{d},mt}^{\mathrm{up}})$, where

$$
s_{\mathrm{d},mt} = \langle r_{\mathrm{d},mt} \rangle - \frac{\hat{\boldsymbol{\Gamma}}_{t,m}^H}{\hat{\gamma}_{t,mm}}(\langle \mathbf{r}_{\mathrm{d},t} \rangle - \langle \mathbf{H} \rangle \langle \mathbf{x}_{\mathrm{d},t} \rangle)
\tag{50}
$$

and $\hat{\boldsymbol{\Gamma}}_{t,m}$ and $\hat{\gamma}_{t,mm}$ are the mth column and the (m,m)-element of $\hat{\boldsymbol{\Gamma}}_t$, respectively.

3) Updating $\mathbf{h}_i$. Taking the expectation of the conditional (48) w.r.t. all latent variables except $\mathbf{h}_i$, the variational distribution $q(\mathbf{h}_i)$ is given by

$$
\begin{aligned}
q(\mathbf{h}_i) \propto \exp\Big\{&\big\langle \ln p(\mathbf{R}_{\mathrm{p}}|\mathbf{H}; \hat{\gamma}_{\mathrm{p}}, \mathbf{X}_{\mathrm{p}}) + \ln p(\mathbf{R}_{\mathrm{d}}|\mathbf{X}_{\mathrm{d}}, \mathbf{H}; \{\hat{\boldsymbol{\Gamma}}_t\}) \\
&+ \ln p(\mathbf{h}_i; \mathbf{C}_i)\big\rangle_{-\mathbf{h}_i}\Big\}.
\end{aligned}
\tag{51}
$$

Following the same procedure to obtain $q(\mathbf{h}_i)$ as in the MF-QVB-JED algorithm, we have

$$
\begin{aligned}
q(\mathbf{h}_i) &\propto \mathcal{CN}(\mathbf{h}_i; \mathbf{k}_i, \boldsymbol{\Gamma}^{-1})\mathcal{CN}(\mathbf{h}_i; \mathbf{0}, \mathbf{C}_i) \\
&= \mathcal{CN}(\mathbf{h}_i; (\boldsymbol{\Gamma}_i + \mathbf{C}_i^{-1})^{-1}\boldsymbol{\Gamma}_i\mathbf{k}_i, (\boldsymbol{\Gamma}_i + \mathbf{C}_i^{-1})^{-1}),
\end{aligned}
\tag{52}
$$

where $\boldsymbol{\Gamma}_i$ and $\mathbf{k}_i$ are defined as

$$
\boldsymbol{\Gamma}_i = \hat{\gamma}_{\mathrm{p}} \sum_{t=1}^{T_{\mathrm{p}}} |x_{\mathrm{p},it}|^2 \mathbf{I}_M + \sum_{t=1}^{T_{\mathrm{d}}} \langle |x_{\mathrm{d},it}|^2 \rangle \hat{\boldsymbol{\Gamma}}_t
\tag{53}
$$

$$
\begin{aligned}
\mathbf{k}_i = &\left(\mathbf{I}_M - \boldsymbol{\Gamma}_i^{-1}\sum_{t=1}^{T_{\mathrm{d}}} \hat{\boldsymbol{\Gamma}}_t \tau_{x_{\mathrm{d},it}}\right)\langle \mathbf{h}_i \rangle \\
&+ \boldsymbol{\Gamma}_i^{-1}\left[\hat{\gamma}_{\mathrm{p}} \mathbf{E}_{\mathrm{p}} \mathbf{x}_{\mathrm{p},i}^* + \sum_{t=1}^{T_{\mathrm{d}}} \hat{\boldsymbol{\Gamma}}_t \mathbf{e}_{\mathrm{d},t} x_{\mathrm{d},it}^*\right]
\end{aligned}
\tag{54}
$$

and where $\mathbf{E}_{\mathrm{p}} = \langle \mathbf{R}_{\mathrm{p}} \rangle - \langle \mathbf{H} \rangle \mathbf{X}_{\mathrm{p}}$ and $\mathbf{e}_{\mathrm{d},t} = \langle \mathbf{r}_{\mathrm{d},t} \rangle - \langle \mathbf{H} \rangle \langle \mathbf{x}_{\mathrm{d},t} \rangle$ are the residual terms. The variational covariance matrix and mean of $\mathbf{h}_i$ are now given by $\boldsymbol{\Sigma}_{\mathbf{h}_i} = (\boldsymbol{\Gamma}_i + \mathbf{C}_i^{-1})^{-1}$ and $\langle \mathbf{h}_i \rangle = \boldsymbol{\Sigma}_{\mathbf{h}_i} \boldsymbol{\Gamma}_i \mathbf{k}_i$.

4) Updating $x_{\mathrm{d},it}$. Taking the expectation of the conditional (30) w.r.t. all latent variables except $x_{\mathrm{d},it}$, the variational distribution $q(x_{\mathrm{d},it})$ is given by

$$
q(x_{\mathrm{d},it}) \propto \exp\Big\{\big\langle \ln p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \hat{\boldsymbol{\Gamma}}_t) + \ln p(x_{\mathrm{d},it})\big\rangle_{-x_{\mathrm{d},it}}\Big\}.
\tag{55}
$$

$$p(\mathbf{Y}_{\mathrm{p}}, \mathbf{Y}_{\mathrm{d}}, \mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{X}_{\mathrm{d}}, \mathbf{H}; \gamma_{\mathrm{p}}, \{\mathbf{\Gamma}_t\}, \mathbf{X}_{\mathrm{p}}, \{\mathbf{C}_i\})$$
$$= p(\mathbf{Y}_{\mathrm{p}}|\mathbf{R}_{\mathrm{p}})p(\mathbf{R}_{\mathrm{p}}|\mathbf{H}; \gamma_{\mathrm{p}}, \mathbf{X}_{\mathrm{p}})p(\mathbf{Y}_{\mathrm{d}}|\mathbf{R}_{\mathrm{d}})p(\mathbf{R}_{\mathrm{d}}|\mathbf{X}_{\mathrm{d}}, \mathbf{H}; \{\mathbf{\Gamma}_t\})p(\mathbf{X}_{\mathrm{d}})p(\mathbf{H}|\{\mathbf{C}_i\})$$
$$= \left[\prod_{t=1}^{T_{\mathrm{p}}} p(\mathbf{y}_{\mathrm{p},t}|\mathbf{r}_{\mathrm{p},t})p(\mathbf{r}_{\mathrm{p},t}|\mathbf{H}; \gamma_{\mathrm{p}}, \mathbf{x}_{\mathrm{p},t})\right]\left[\prod_{t=1}^{T_{\mathrm{d}}} p(\mathbf{y}_{\mathrm{d},t}|\mathbf{r}_{\mathrm{d},t})p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \mathbf{\Gamma}_t)p(\mathbf{x}_{\mathrm{d},t})\right]\prod_{i=1}^{K} p(\mathbf{h}_i|\mathbf{C}_i) \qquad (48)$$

Similar to the procedure in the LMMSE-QVB and MF-QVB-JED algorithms, we obtain

$$q(x_{\mathrm{d},it}) \propto p(x_{\mathrm{d},it})\mathcal{CN}\left(z_{\mathrm{d},it}; x_{\mathrm{d},it}, 1/\langle\mathbf{h}_i^H\hat{\mathbf{\Gamma}}_t\mathbf{h}_i\rangle\right), \qquad (56)$$

where

$$z_{\mathrm{d},it} = \frac{\langle\mathbf{h}_i^H\rangle\hat{\mathbf{\Gamma}}_t}{\langle\mathbf{h}_i^H\hat{\mathbf{\Gamma}}_t\mathbf{h}_i\rangle}\left(\langle\mathbf{r}_{\mathrm{d},t}\rangle - \sum_{j\neq i}^{K}\langle\mathbf{h}_j\rangle\langle x_{\mathrm{d},jt}\rangle\right)$$
$$= \frac{\langle\mathbf{h}_i^H\rangle\hat{\mathbf{\Gamma}}_t\langle\mathbf{h}_i\rangle\langle x_{\mathrm{d},it}\rangle + \langle\mathbf{h}_i^H\rangle\hat{\mathbf{\Gamma}}_t\mathbf{e}_{\mathrm{d},t}}{\langle\mathbf{h}_i^H\hat{\mathbf{\Gamma}}_t\mathbf{h}_i\rangle} \qquad (57)$$

is a linear estimate of $x_{\mathrm{d},it}$. We note that $\langle\mathbf{h}_i^H\hat{\mathbf{\Gamma}}_t\mathbf{h}_i\rangle = \langle\mathbf{h}_i\rangle\hat{\mathbf{\Gamma}}_t\langle\mathbf{h}_i\rangle + \mathrm{Tr}\{\hat{\mathbf{\Gamma}}_t\mathbf{\Sigma}_{\mathbf{h}_i}\}$.

In the *M-step*, γ_{p} and $\{\mathbf{\Gamma}_t\}$ are estimated to maximize $\ln p(\mathbf{Y}_{\mathrm{p}}, \mathbf{Y}_{\mathrm{d}}, \mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{X}_{\mathrm{d}}, \mathbf{H}; \gamma_{\mathrm{p}}, \{\mathbf{\Gamma}_t\}, \mathbf{X}_{\mathrm{p}}, \{\mathbf{C}_i\})$ w.r.t. the variational distribution $q(\mathbf{R}_{\mathrm{p}}, \mathbf{R}_{\mathrm{d}}, \mathbf{X}_{\mathrm{d}}, \mathbf{H})$. The update of $\hat{\gamma}_{\mathrm{p}}$ is similar to the procedure in the MF-QVB-JED algorithm and is given in (41) and (43). The update of $\hat{\mathbf{\Gamma}}_t$ is given by

$$\hat{\mathbf{\Gamma}}_t = \arg\max_{\mathbf{\Gamma}_t}\ \big\langle\ln p(\mathbf{r}_{\mathrm{d},t}|\mathbf{H}, \mathbf{x}_{\mathrm{d},t}; \mathbf{\Gamma}_t)\big\rangle$$
$$= \arg\max_{\mathbf{\Gamma}_t}\ \Big[\ln|\mathbf{\Gamma}_t|$$
$$- \big\langle(\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t})^H\mathbf{\Gamma}_t(\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t})\big\rangle\Big].$$

Applying Theorem 1, we have

$$\big\langle(\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t})^H\mathbf{\Gamma}_t(\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t})\big\rangle$$
$$= \mathrm{Tr}\Big\{\Big[\big((\langle\mathbf{r}_{\mathrm{d},t}\rangle - \langle\mathbf{H}\rangle\langle\mathbf{x}_{\mathrm{d},t}\rangle)(\langle\mathbf{r}_{\mathrm{d},t}\rangle - \langle\mathbf{H}\rangle\langle\mathbf{x}_{\mathrm{d},t}\rangle)^H + \mathbf{\Sigma}_{\mathbf{r}_{\mathrm{d},t}}$$
$$+ \sum_{i=1}^{K}\langle|x_{\mathrm{d},it}|^2\rangle\mathbf{\Sigma}_{\mathbf{h}_i} + \langle\mathbf{H}\rangle\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{d},t}}\langle\mathbf{H}\rangle^H\Big]\mathbf{\Gamma}_t\Big\}. \qquad (58)$$

Due to the rank deficiency of $\big\langle(\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t})^H\mathbf{\Gamma}_t(\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t})\big\rangle_{-\mathbf{\Gamma}_t}$, we propose to use the following estimator for $\mathbf{\Gamma}_t$:

$$\hat{\mathbf{\Gamma}}_t = \left(\frac{\|\langle\mathbf{r}_{\mathrm{d},t}\rangle - \langle\mathbf{H}\rangle\langle\mathbf{x}_{\mathrm{d},t}\rangle\|^2}{M}\mathbf{I}_M + \mathbf{\Sigma}_{\mathbf{r}_{\mathrm{d},t}}\right.$$
$$\left.+ \sum_{i=1}^{K}\langle|x_{\mathrm{d},it}|^2\rangle\mathbf{\Sigma}_{\mathbf{h}_i} + \langle\mathbf{H}\rangle\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{d},t}}\langle\mathbf{H}\rangle^H\right)^{-1}. \qquad (59)$$

By iteratively optimizing $\{q(\mathbf{r}_{\mathrm{p},t})\}$, $\{q(\mathbf{r}_{\mathrm{d},t})\}$, $\{q(\mathbf{h}_i)\}$, $\{q(x_{\mathrm{d},it})\}$, $\hat{\gamma}_{\mathrm{p}}$, and $\{\hat{\mathbf{\Gamma}}_t\}$, we obtain the VBEM algorithm for estimating $\mathbf{R}_{\mathrm{p}}$, $\mathbf{R}_{\mathrm{d}}$, $\mathbf{H}$, $\mathbf{X}_{\mathrm{d}}$, γ_{p}, and $\{\mathbf{\Gamma}_t\}$. We refer to this scheme as the ***LMMSE-QVB-JED algorithm*** for joint channel estimation and data detection. The implementation of the LMMSE-QVB-JED algorithm is similar to that of MF-QVB-JED presented in Algorithm 3. We skip the summary of the LMMSE-QVB-JED algorithm for brevity.

C. Practical Aspects of Implementing MF/LMMSE-QVB-JED

1) Computing $\langle|x_{\mathrm{d},it}|^2\rangle$: For PSK signaling, the variational second moment $\langle|x_{\mathrm{d},it}|^2\rangle$ is constant and need not be updated in each iteration of the algorithms. We present the proof for this observation in Appendix C.

2) Computing $\mathbf{\Sigma}_{\mathbf{h}_i}$ with uncorrelated channels: When $\mathbf{C}_i$ is a diagonal matrix, the variational covariance matrix $\mathbf{\Sigma}_{\mathbf{h}_i}$ in (45) is also a diagonal matrix and its computation does not require matrix inversion. Thus, the MF-QVB-JED algorithm can be implemented without any matrix inversion. This property does not hold for the LMMSE-QVB-JED algorithm, since $\mathbf{\Gamma}_i$ in (53) is not in general a diagonal matrix.

3) Lite implementation of MF-QVB-JED: Instead of using the latent variable $\gamma_{\mathrm{d},t}$ as the precision at time slot t, we can impose a single latent variable γ_{d} as the precision for all time slots. A lite version of MF-QVB-JED can be devised using the same procedure as in Section V-A where $\{\gamma_{\mathrm{d},t}\}$ is replaced by γ_{d}. In the *M-step*, the estimate of γ_{d} can be found as

$$\hat{\gamma}_{\mathrm{d}} = \frac{MT_{\mathrm{d}}}{\sum_{t=1}^{T_{\mathrm{d}}}\langle\|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2\rangle}, \qquad (60)$$

where $\langle\|\mathbf{r}_{\mathrm{d},t} - \mathbf{H}\mathbf{x}_{\mathrm{d},t}\|^2\rangle$ is given in (44).

5) Lite implementation of LMMSE-QVB-JED: Instead of using the latent variable $\mathbf{\Gamma}_t$ as the precision matrix at the time slot t, we could use the same precision matrix $\mathbf{\Gamma}$ for all data time slots. A lite version of LMMSE-QVB-JED can be devised using the same procedure as in Section V-B where $\{\mathbf{\Gamma}_t\}$ is replaced by $\mathbf{\Gamma}$. In the *M-step*, we propose to use the following estimator for $\mathbf{\Gamma}$:

$$\hat{\mathbf{\Gamma}} = T_{\mathrm{d}}\left(\frac{\|\langle\mathbf{R}_{\mathrm{d}}\rangle - \langle\mathbf{H}\rangle\langle\mathbf{X}_{\mathrm{d}}\rangle\|_F^2}{M}\mathbf{I}_M + \mathbf{\Sigma}_{\mathbf{r}_{\mathrm{d}}}\right.$$
$$\left.+ \sum_{i=1}^{K}\langle\|\mathbf{x}_{\mathrm{d},i}\|^2\rangle\mathbf{\Sigma}_{\mathbf{h}_i} + \langle\mathbf{H}\rangle\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{d}}}\langle\mathbf{H}\rangle^H\right)^{-1}, \qquad (61)$$

where we denote $\mathbf{\Sigma}_{\mathbf{r}_{\mathrm{d}}} = \sum_{t=1}^{T_{\mathrm{d}}}\mathbf{\Sigma}_{\mathbf{r}_{\mathrm{d},t}}$, $\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{d}}} = \sum_{t=1}^{T_{\mathrm{d}}}\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{d},t}}$ and $\langle\|\mathbf{x}_{\mathrm{d},i}\|^2\rangle = \sum_{t=1}^{T_{\mathrm{d}}}\langle|x_{\mathrm{d},it}|^2\rangle$.

We observe in our simulations that the lite version of MF/LMMSE-QVB-JED slightly increases the detection error compared to the original version. However, the lite version can significantly reduce the computational complexity, especially for the LMMSE-QVB-JED algorithm. LMMSE-QVB-JED requires one matrix inversion in (61) for computing $\hat{\mathbf{\Gamma}}$ in the lite version, while requiring T_{d} matrix inversions to compute $\{\hat{\mathbf{\Gamma}}_t\}$ in the original version. In the numerical results, we will use the lite version of these algorithms.
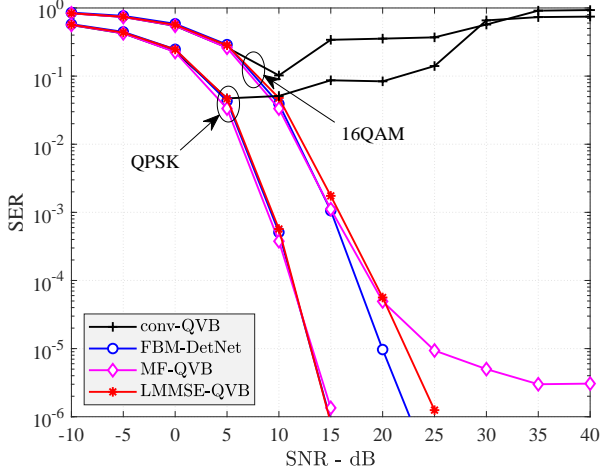
Fig. 1: Data detection performance comparison for i.i.d. channels with perfect CSI, $b = 3$ bits, $K = 16$, $M = 32$ and $M = 64$ for QPSK and 16QAM signaling, respectively.

Fig. 2: Data detection performance comparison for spatially correlated channels with perfect CSI, $b = 3$ bits, $K = 16$, $M = 32$ and $M = 64$ for QPSK and 16QAM signaling, respectively.

VI. Numerical Results

This section presents numerical results comparing the performance of the proposed VB-based methods with the conventional quantized VB-based method, denoted as conv-QVB, in [26] and FBM-DetNet in [20], which are the most recent and related methods to the work in this paper. The maximum number of iterations is set to 50 for all the iterative algorithms. The covariance matrices $\mathbf{C}_i$ are normalized such that their diagonal elements are 1, which implies $\mathbb{E}[\|\mathbf{h}_i\|^2] = M$, $\forall i$. The noise variance N_0 is set according to the operating SNR, which is defined as

$$\text{SNR} = \frac{\mathbb{E}[\|\mathbf{Hx}\|^2]}{\mathbb{E}[\|\mathbf{n}\|^2]} = \frac{\sum_{i=1}^{K} \text{Tr}\{\mathbf{C}_i\}}{MN_0} = \frac{K}{N_0}. \quad (62)$$

For i.i.d. channels, we set $\mathbf{C}_i = \mathbf{I}$, $\forall i$. For spatially correlated channels, we use the typical urban channel model in [8] where the power angle spectrum of the channel model follows a Laplacian distribution with an angle spread of $10°$. The covariance matrix $\mathbf{C}_i$ is obtained according to [31, Eq. (2)]. Unless otherwise stated, we set the training length $T_{\text{p}} = 2K$ and the data transmission length $T_{\text{d}} = 100$.

First, we examine data detection for the case of perfect CSI. Results for i.i.d. and spatially correlated channels are shown in Fig. 1 and Fig. 2, respectively. It can be seen that, for both i.i.d. and correlated channels, the conv-QVB method in [26] is outperformed by all other methods and its performance is severely degraded at high SNRs. This is because conv-QVB does not take into account the residual inter-user interference and often encounters the catastrophic cancellation issue at high SNR. For i.i.d. channels, FBM-DetNet, MF-QVB, and LMMSE-QVB all yield the same performance for QPSK signals, while for 16QAM FBM-DetNet and LMMSE-QVB are similar and both outperform MF-QVB. For spatially correlated channels, LMMSE-QVB provides a significantly lower SER than FBM-DetNet and MF-QVB due to its estimation of the precision matrix $\mathbf{\Gamma}$ which can better represent the effect of the residual inter-user interference.
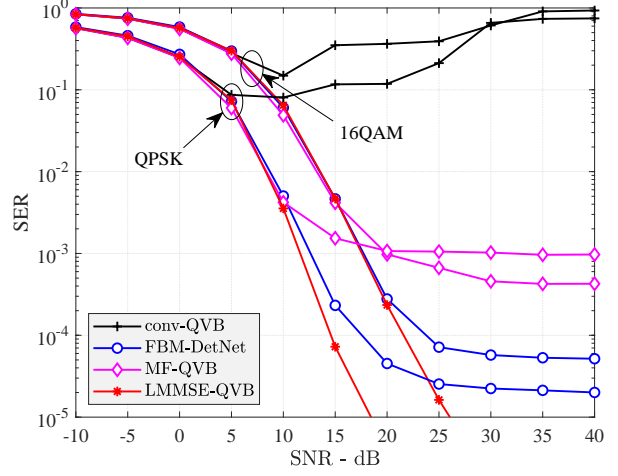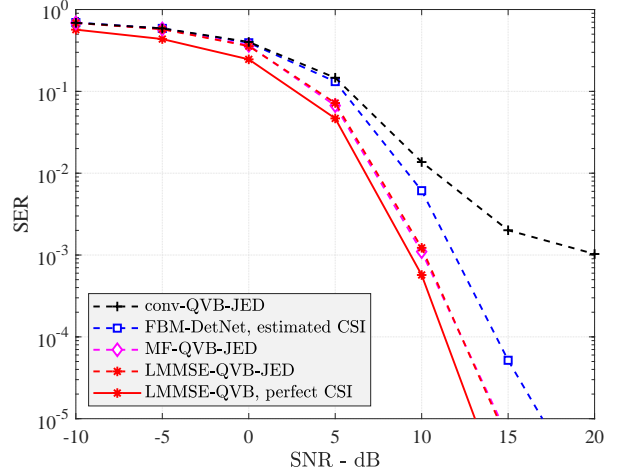

Fig. 3: Data detection comparison between the proposed MF-QVB-JED, LMMSE-QVB-JED, and other existing methods for i.i.d. channels with $K = 16$, $M = 32$, $b = 3$ bits, and QPSK signaling.

Fig. 3 presents results for data detection with estimated CSI and i.i.d. channels. Both MF-QVB-JED and LMMSE-QVB-JED outperform the conventional QVB-JED method as well as the DNN-based detection network FBM-DetNet. Note that FBM-DetNet uses estimated CSI provided by FBM-CENet, a channel estimation network also proposed in [20] and designed to estimate the CSI using only the pilot sequence. MF-QVB-JED and LMMSE-QVB-JED both yield the same SER, which is about 2-3dB better than FBM-DetNet at an SER of 10^{-3} and 10^{-5}, respectively. The performance of MF-QVB-JED and LMMSE-QVB-JED is also quite close to that of LMMSE-QVB with perfect CSI.

Results for data detection with estimated CSI and spatially correlated channels are given in Fig. 4, where we see that the proposed MF-QVB-JED and LMMSE-QVB-JED methods outperform conv-QVB-JED and FBM-DetNet since the effects of both inter-user interference and spatial channel correlation are taken into account. However, unlike the case of i.i.d. chan-
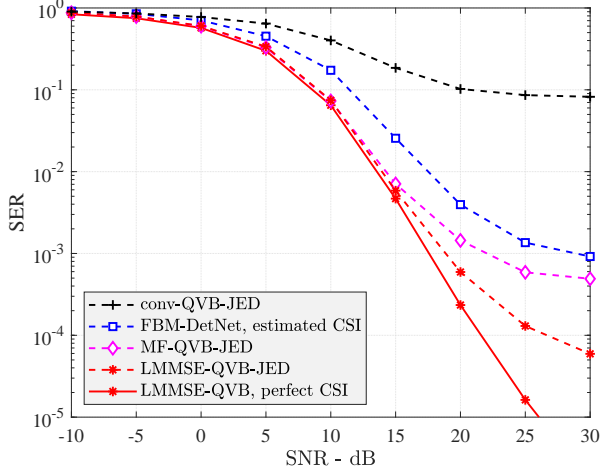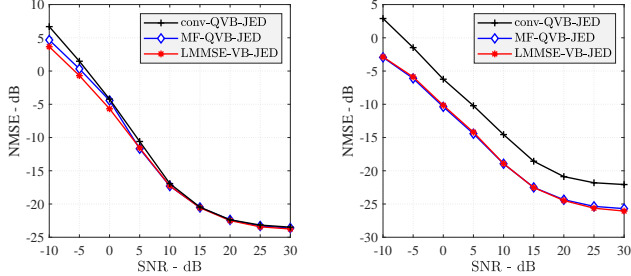
Fig. 4: Data detection performance comparison between the proposed MF-QVB-JED, LMMSE-QVB-JED, and other existing methods for spatially correlated channels with $K = 16$, $M = 64$, $b = 3$ bits, and 16QAM signaling.

(a) i.i.d. channels, $K = 16$, $M = 32$.

(b) spatially correlated channels, $K = 16$, $M = 64$.

Fig. 5: Channel estimation performance comparison.

Fig. 6: Detection performance of MF-QVB-JED and LMMSE-QVB-JED versus T_d with $K = 16$, $M = 64$, $b = 3$ bits, and 16QAM signaling.

Fig. 7: Detection performance of MF-QVB-JED and LMMSE-QVB-JED versus the ADC bit resolution b with $K = 16$, $M = 64$, and 16QAM signaling.

nels where MF-QVB-JED and LMMSE-QVB-JED give the same performance, the LMMSE-QVB-JED method provides a significantly lower SER than MF-QVB-JED at high SNRs for spatially correlated channels. For example, at 30dB, the SER of LMMSE-QVB-JED is about 10 times lower than that of MF-QVB-JED, which is already better than FBM-DetNet.

We provide a channel estimation comparison in Fig. 5 where i.i.d. channels are considered in Fig. 5a and spatially correlated channels are considered in Fig. 5b. The normalized mean squared error (NMSE) in these figures is defined as $\mathrm{NMSE} = \mathbb{E}\left[\|\mathbf{H} - \hat{\mathbf{H}}\|_F^2 / \|\mathbf{H}\|_F^2\right]$. For i.i.d. channels, all three VB-based methods conv-QVB-JED, MF-QVB-JED, and LMMSE-QVB-JED give similar performance but for spatially correlated channels, the proposed MF-QVB-JED and LMMSE-QVB-JED methods are seen to provide lower NMSEs compared to the conv-QVB-JED method.

Fig. 6 presents the SER performance of the proposed MF-QVB-JED and LMMSE-QVB-JED methods w.r.t. the data transmission length T_d. We observe that the SER performance improves with increasing T_d since more received signals are combined to achieve a more accurate channel estimate. Consequently, the data detection phase can result in a lower detection error.
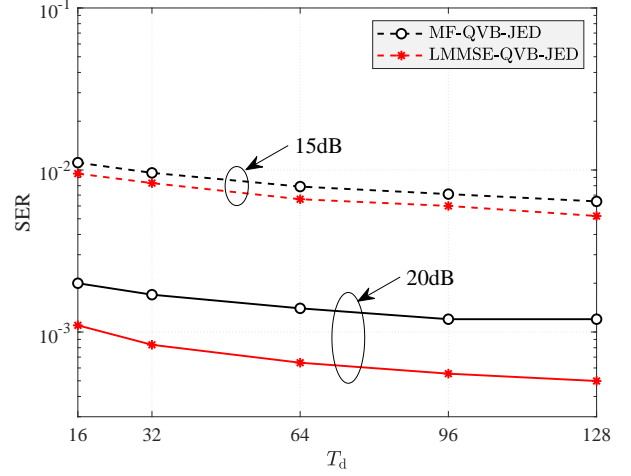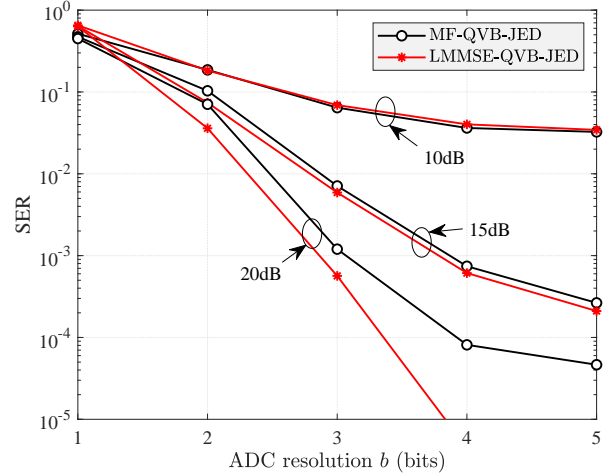
In Fig. 7, we evaluate the data detection performance of the proposed MF-QVB-JED and LMMSE-QVB-JED methods for different ADC bit resolutions. As expected, increasing the resolution b significantly helps improve the detection performance. It is observed that lower SNRs require a lower bit resolution for the best performance, e.g., 4-bit ADCs are sufficient to obtain the lowest SER at 10dB. Increasing the ADC bit resolution to values higher than 4 does not result in a lower SER. It is also interesting to note that at high SNRs, LMMSE-QVB-JED can provide much lower SERs compared to MF-QVB-JED as the bit resolution increases.

VII. CONCLUSION

In this paper, we exploited the VB inference framework to propose different channel estimation and data detection methods for massive MIMO systems with low-resolution ADCs. In particular, we proposed new VB-based algorithms referred to as MF-QVB and LMMSE-QVB for data detection

with known CSI, and MF-QVB-JED and LMMSE-QVB-JED for joint channel estimation and data detection. In the proposed QVB framework, we proposed to float the noise variance/covariance matrix as an unknown random variable which also allows the algorithms to take into account the residual inter-user interference. Numerous practical aspects of the QVB framework were studied to improve the implementation stability. It was also shown via a number of simulation studies that the proposed methods provide robust performance and significantly outperform existing methods, particularly when the channels are spatially correlated.

APPENDIX A
PROOF OF THEOREM 1

Expanding $\langle \|\mathbf{y} - \mathbf{Ax}\|^2 \rangle$ and taking into account the independence between $\mathbf{A}$, $\mathbf{y}$, and $\mathbf{x}$, we have

$$
\begin{aligned}
&\langle (\mathbf{y} - \mathbf{Ax})^H \mathbf{B}(\mathbf{y} - \mathbf{Ax}) \rangle \\
&= \langle \mathbf{y}^H \mathbf{B} \mathbf{y} \rangle - 2 \Re\{ \langle \mathbf{y}^H \mathbf{B} \mathbf{Ax} \rangle \} + \langle \mathbf{x}^H \mathbf{A}^H \mathbf{B} \mathbf{Ax} \rangle \\
&= (\langle \mathbf{y} \rangle - \langle \mathbf{A} \rangle \langle \mathbf{x} \rangle)^H \mathbf{B}(\langle \mathbf{y} \rangle - \langle \mathbf{A} \rangle \langle \mathbf{x} \rangle) + \langle \mathbf{y}^H \mathbf{B} \mathbf{y} \rangle \\
&\quad - \langle \mathbf{y}^H \rangle \mathbf{B} \langle \mathbf{y} \rangle + \langle \mathbf{x}^H \mathbf{A}^H \mathbf{B} \mathbf{Ax} \rangle - \langle \mathbf{x}^H \rangle \langle \mathbf{A}^H \rangle \mathbf{B} \langle \mathbf{A} \rangle \langle \mathbf{x} \rangle.
\end{aligned}
\tag{63}
$$

Note that $\langle \mathbf{xx}^H \rangle = \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^H + \boldsymbol{\Sigma}_\mathbf{x}$ and $\langle \mathbf{y}^H \mathbf{B} \mathbf{y} \rangle = \mathrm{Tr}\{\mathbf{B} \langle \mathbf{yy}^H \rangle\} = \langle \mathbf{y}^H \rangle \mathbf{B} \langle \mathbf{y} \rangle + \mathrm{Tr}\{\mathbf{B} \boldsymbol{\Sigma}_\mathbf{y}\}$. In addition, we have

$$
\begin{aligned}
\left[\langle \mathbf{A}^H \mathbf{B} \mathbf{A} \rangle \right]_{ij} &= \langle \mathbf{a}_i^H \mathbf{B} \mathbf{a}_j \rangle \\
&= \begin{cases} \langle \mathbf{a}_i^H \rangle \mathbf{B} \langle \mathbf{a}_i \rangle + \mathrm{Tr}\{\mathbf{B} \boldsymbol{\Sigma}_{\mathbf{a}_i}\}, & \text{if } i = j \\ \langle \mathbf{a}_i^H \rangle \mathbf{B} \langle \mathbf{a}_j \rangle, & \text{otherwise.} \end{cases}
\end{aligned}
\tag{64}
$$

It thus follows that $\langle \mathbf{A}^H \mathbf{B} \mathbf{A} \rangle = \langle \mathbf{A}^H \rangle \mathbf{B} \langle \mathbf{A} \rangle + \mathbf{D}$, and

$$
\begin{aligned}
\langle \mathbf{x}^H \mathbf{A}^H \mathbf{B} \mathbf{Ax} \rangle &= \mathrm{Tr}\{ \langle \mathbf{A}^H \mathbf{B} \mathbf{A} \rangle \langle \mathbf{xx}^H \rangle \} \\
&= \mathrm{Tr}\{ \langle \mathbf{A}^H \rangle \mathbf{B} \langle \mathbf{A} \rangle \langle \mathbf{x} \rangle \langle \mathbf{x}^H \rangle \} + \langle \mathbf{x} \rangle^H \mathbf{D} \langle \mathbf{x} \rangle \\
&\quad + \mathrm{Tr}\{ \boldsymbol{\Sigma}_\mathbf{x} \mathbf{D} \} + \mathrm{Tr}\{ \boldsymbol{\Sigma}_\mathbf{x} \langle \mathbf{A}^H \rangle \mathbf{B} \langle \mathbf{A} \rangle \}.
\end{aligned}
$$

The statement (12) thus follows by removing the duplicated terms in (63). Note that $\langle \mathbf{x} \rangle^H \mathbf{D} \langle \mathbf{x} \rangle + \mathrm{Tr}\{ \boldsymbol{\Sigma}_\mathbf{x} \mathbf{D} \}$ can also be written as $\langle \mathbf{x} \rangle^H \mathbf{D} \langle \mathbf{x} \rangle + \mathrm{Tr}\{ \boldsymbol{\Sigma}_\mathbf{x} \mathbf{D} \} = \sum_{i=1}^{n} \langle |x_i|^2 \rangle \mathrm{Tr}\{ \mathbf{B} \boldsymbol{\Sigma}_{\mathbf{a}_i} \}$.

APPENDIX B
COMPUTATION OF $\mathsf{F}_r(\mu, \gamma, a, b)$ AND $\mathsf{G}_r(\mu, \gamma, a, b)$

For ease of presentation, we denote

$$
\alpha = \sqrt{2\gamma}(a - \mu), \quad \beta = \sqrt{2\gamma}(b - \mu).
\tag{65}
$$

For an arbitrary complex random variable $\mathcal{CN}(\mu, \gamma^{-1})$ whose real and imaginary parts are both truncated on the interval (a, b), the mean $\mathsf{F}_r(\mu, \gamma, a, b)$ and variance $\mathsf{G}_r(\mu, \gamma, a, b)$ are computed as

$$
\mathsf{F}_r(\mu, \gamma, a, b) = \mu - \frac{1}{\sqrt{2\langle \gamma \rangle}} \frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}
\tag{66}
$$

$$
\mathsf{G}_r(\mu, \gamma, a, b) = \frac{1}{2\langle \gamma \rangle} \left[1 - \frac{\beta_m \phi(\beta) - \alpha \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} \right)^2 \right],
\tag{67}
$$

where the PDF and CDF operators $\phi(\cdot)$ and $\Phi(\cdot)$, as well as the multiplication, division, and square operations are applied individually on the real and imaginary components. The variance $\mathsf{G}_r(\mu, \gamma, a, b)$ is computed by adding the variances of the two components.

APPENDIX C
COMPUTATION OF $\mathsf{F}_x(z, \gamma)$ AND $\mathsf{G}_x(z, \gamma)$

Given $z = x + \mathcal{CN}(0, \gamma^{-1})$, the posterior distribution of x given z is

$$
p(x|z; \gamma) \propto p(x) \, \mathcal{CN}(z; x, \gamma^{-1})
$$

For $a \in \mathcal{S}$, we have

$$
p(x = a|z; \gamma) = \frac{1}{Z} p_a \exp\big(-\gamma |z - a|^2 \big),
$$

where $Z = \sum_{b \in \mathcal{S}} p_b \exp\big(-\gamma |z - b|^2 \big)$ is a normalization factor. The corresponding posterior mean $\mathsf{F}_x(z, \gamma)$ and variance $\mathsf{G}_x(z, \gamma)$ are computed as

$$
\mathsf{F}_x(z, \gamma) = \sum_{a \in \mathcal{S}} a \times p(x = a|z, \gamma)
$$

$$
\mathsf{G}_x(z, \gamma) = \sum_{a \in \mathcal{S}} |a|^2 \times p(x = a|z, \gamma) - |\mathsf{F}_x(z, \gamma)|^2.
$$

We note that $\mathbb{E}\big[|x|^2 | z; \gamma \big]$ is equal to $|a|^2$ for PSK signaling with transmit energy $|a|^2$, as shown below:

$$
\begin{aligned}
\mathbb{E}\big[|x|^2 | z; \gamma \big] &= \sum_{a \in \mathcal{S}} |a|^2 \frac{1}{Z} p_a \exp\big(-\gamma |z - a|^2 \big) \\
&= |a|^2 \frac{\sum_{a \in \mathcal{S}} p_a \exp\big(-\gamma |z - a|^2 \big)}{Z} \\
&= |a|^2.
\end{aligned}
$$

REFERENCES

[1] I. F. Akyildiz, J. M. Jornet, and C. Han, "Terahertz band: Next frontier for wireless communications," *Physical Commun.*, vol. 12, pp. 16–32, 2014.

[2] ——, "TeraNets: Ultra-broadband communication networks in the terahertz band," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 130–135, Aug. 2014.

[3] N. Rajatheva, I. Atzeni, E. Bjornson, A. Bourdoux, S. Buzzi, J.-B. Dore, S. Erkucuk, M. Fuentes, K. Guan, Y. Hu *et al.*, "White paper on broadband connectivity in 6G," *arXiv preprint arXiv:2004.14247*, 2020.

[4] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Select. Topics in Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[5] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4624–4634, Sept. 2017.

[6] K. Roth, H. Pirzadeh, A. L. Swindlehurst, and J. A. Nossek, "A comparison of hybrid beamforming and digital beamforming with low-resolution ADCs for multiple users and imperfect CSI," *IEEE J. Select. Topics in Signal Process.*, vol. 12, no. 3, pp. 484–498, June 2018.

[7] J. Choi, J. Mo, and R. W. Heath, "Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2005–2018, May 2016.

[8] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[9] N. Kolomvakis, T. Eriksson, M. Coldrey, and M. Viberg, "Quantized uplink massive MIMO systems with linear receivers," in *Proc. IEEE Int. Conf. Commun.*, Dublin, Ireland, June 2020.

[10] A. S. Lan, M. Chiang, and C. Studer, "Linearized binary regression," in *Proc. Annual Conf. on Inform. Sciences and Systems*, Princeton, NJ, USA, Mar. 2018.

[11] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "Linear and deep neural network-based receivers for massive MIMO systems with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7333–7345, Nov. 2021.

[12] Y.-S. Jeon, N. Lee, and H. V. Poor, "Robust data detection for MIMO systems with one-bit ADCs: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1663–1676, Mar. 2020.

[13] S. H. Song, S. Lim, G. Kwon, and H. Park, "CRC-aided soft-output detection for uplink multi-user MIMO systems with one-bit ADCs," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Marrakesh, Morocco, Apr. 2019.

[14] Y. Cho and S. Hong, "One-bit Successive-cancellation Soft-output (OSS) detector for uplink MU-MIMO systems with one-bit ADCs," *IEEE Access*, vol. 7, pp. 27 172–27 182, Feb. 2019.

[15] Z. Shao, R. C. de Lamare, and L. T. N. Landau, "Iterative detection and decoding for large-scale multiple-antenna systems with 1-bit ADCs," *IEEE Wireless Commun. Letters*, vol. 7, no. 3, pp. 476–479, June 2018.

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

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

[18] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "SVM-based channel estimation and data detection for one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2086–2099, 2021.

[19] D. H. N. Nguyen, "Neural network-optimized channel estimator and training signal design for MIMO systems with few-bit ADCs," *IEEE Signal Process. Letters*, vol. 27, pp. 1370–1374, 2020.

[20] L. V. Nguyen, D. H. Nguyen, and A. L. Swindlehurst, "Deep learning for estimation and pilot signal design in few-bit massive MIMO systems," *IEEE Trans. Wireless Commun. (Early Access)*, 2022.

[21] S. Khobahi, N. Shlezinger, M. Soltanalian, and Y. C. Eldar, "LoRD-Net: Unfolded deep detection network with low-resolution receivers," *IEEE Trans. Signal Process.*, vol. 69, pp. 5651–5664, 2021.

[22] Y. Jeon, S. Hong, and N. Lee, "Supervised-learning-aided communication framework for MIMO systems with low-resolution ADCs," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7299–7313, Aug. 2018.

[23] L. V. Nguyen, D. T. Ngo, N. H. Tran, A. L. Swindlehurst, and D. H. N. Nguyen, "Supervised and semi-supervised learning for MIMO blind detection with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2427–2442, Apr. 2020.

[24] S. Kim, J. Chae, and S.-N. Hong, "Machine learning detectors for MU-MIMO systems with one-bit ADCs," *IEEE Access*, vol. 8, pp. 86 608–86 616, Apr. 2020.

[25] Y. Xiang, K. Xu, B. Xia, and X. Cheng, "Bayesian joint channel-and-data estimation for quantized OFDM over doubly selective channels," *IEEE Trans. Wireless Commun. (Early Access)*, 2022.

[26] S. S. Thoota and C. R. Murthy, "Variational Bayes' joint channel estimation and soft symbol decoding for uplink massive MIMO systems with low resolution ADCs," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3467–3481, May 2021.

[27] D. H. Nguyen, I. Atzeni, A. Tölli, and A. L. Swindlehurst, "A variational Bayesian perspective on massive MIMO detection," *arXiv preprint arXiv:2205.11649*, 2022.

[28] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "A variational bayesian perspective on MIMO detection with low-resolution ADCs," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, USA, 2022.

[29] C. M. Bishop, *Pattern recognition and machine learning*. New York, NY, USA: Springer, 2006.

[30] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. and Trends® Mach. Learn.*, vol. 1, no. 1–2, pp. 1–305, Jan. 2008.

[31] L. You, X. Gao, X.-G. Xia, N. Ma, and Y. Peng, "Pilot reuse for massive MIMO transmission over spatially correlated rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3352–3366, June 2015.