A Variational Bayesian Perspective on Massive MIMO Detection

Duy H. N. Nguyen, Italo Atzeni, Antti Tölli, and A. Lee Swindlehurst

***Abstract*—Optimal data detection in massive multiple-input multiple-output (MIMO) communication systems requires prohibitive computational complexity. A variety of detection algorithms have been proposed in the literature, offering different trade-offs between complexity and detection error performance. In this paper, we build upon variational Bayes (VB) inference, a powerful statistical inference framework, to design efficient and low-complexity data detection algorithms for massive MIMO systems. We first examine the massive MIMO detection problem with perfect channel state information at the receiver (CSIR) and show that a conventional VB method with known noise variance yields poor detection error performance. To address this limitation, we devise two new VB algorithms that use the noise variance and covariance matrix postulated by the algorithms themselves. We further develop the VB framework for massive MIMO detection with imperfect CSIR. Simulation results show that the proposed VB methods achieve significantly lower detection errors compared with existing schemes for a wide range of channel models.**

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

I. Introduction

Massive multiple-input multiple-output (MIMO) arrays are a key technology for emerging 5G networks, enabling higher spectral efficiency and improved coverage [1], [2]. A massive MIMO base station (BS) can concurrently serve a large number of users in the same time-frequency resource via space division multiple access and highly directional beamforming. However, the increase in the spatial dimension leads to large channel matrices and makes low-complexity multi-user detection a formidable task.

The subject of massive MIMO detection has attracted significant interest in recent years, with several contributions offering different trade-offs between computational complexity and detection error performance. Conventional detection algorithms based on the maximum a-posteriori (MAP) and maximum likelihood (ML) criteria, which jointly recover all the symbols simultaneously, achieve the optimal detection error performance. However, their complexity increases exponentially with the number of users. Linear detectors, such as the matched filter (MF), zero forcing filter, and linear minimum mean squared error (LMMSE) filter, consist of a simple linear pre-processing step to decorrelate the received signal, enabling separate symbol detection on a per-user basis. However, linear detection simply treats the inter-user interference as noise and can thus be highly sub-optimal compared with the MAP/ML detectors, especially in systems with comparable numbers of transmit and receive antennas. Interference cancellation is an attractive alternative solution in terms of both complexity and performance. This family of nonlinear detectors relies on removing already detected symbols to facilitate detection of the remaining ones. Although interference cancellation is prone to error propagation, this issue can be mitigated using soft detected symbols, resulting in the iterative soft interference cancellation (SIC) method [3], [4]. Iterative SIC, involving multiple iterations of symbol detection and interference cancellation, can approach the performance of MAP/ML with manageable complexity [5], [6].

Approximate message passing (AMP), originally developed as a computationally efficient algorithm for the recovery of sparse signals [7], has also been applied in the context of massive MIMO detection [8]. In a MIMO system with independent and identically distributed (i.i.d.) Gaussian channels, AMP decouples the MIMO channel into a set of parallel additive white Gaussian noise (AWGN) channels, thus enabling separate symbol detection. In addition, AMP achieves the minimum symbol error rate (SER) in the large-system limit and shows a near-optimal performance for finite-dimensional systems. More importantly, the superior SER performance of AMP can be obtained with very low complexity. The convergence of AMP is established through the algorithm's state evolution for i.i.d. Gaussian [9] and i.i.d. sub-Gaussian channel matrices [10]. However, AMP may diverge when the channel matrix is ill-conditioned or has non-zero mean. This issue was partially dealt with by the recent development of AMP-like algorithms, such as orthogonal AMP (OAMP) [11] and vector AMP (VAMP) [12], which can easily be applied to the MIMO detection problem. It is worth mentioning that rigorous proofs of the state evolution in AMP-like algorithms are generally quite technical and rely on specific assumptions about the channel statistics, e.g., i.i.d. sub-Gaussian or unitarily invariant channels.

Recently, the MIMO detection problem has been tackled using variational Bayes (VB) inference [13]. VB inference is a powerful statistical inference framework from machine learning that approximates the intractable posterior distribution of latent variables with a known family of simpler distributions through optimization. Among VB methods, the mean-field approximation enables efficient optimization of the variational distribution over a partition of the latent variables while keeping the variational distributions over the other partitions fixed [14]. In this paper, we present the variational Bayesian perspective on the massive MIMO detection problem and compare it with the AMP-based (i.e., AMP and OAMP/VAMP) and SIC methods. While it is common to assume that the noise variance is known at the receiver, a conventional VB detector

relying on knowledge of the noise variance as studied in [13] may yield poor detection performance. We present an analysis of this behavior by connecting the conventional VB detector to the SIC method and make the recommendation to use the noise variance or covariance matrix that is postulated by the VB framework itself instead. We then develop several new VB algorithms for massive MIMO detection based on closed-form and computationally efficient updates. The resulting iterative algorithms have low complexity that is comparable to that of AMP-based schemes.

The contributions of this work are listed as follows.

- We present a comparison between conventional mean-field VB and SIC methods and provide a new perspective to explain the poor detection error performance of the former. We propose two new mean-field VB algorithms for massive MIMO detection with perfect channel state information at the receiver (CSIR) based on the MF (MF-VB) and LMMSE filter (LMMSE-VB), in which the noise variance and covariance matrix are treated as random variables and are thus postulated by the estimation in the algorithms.
- We further develop MF-VB for the case of imperfect CSIR. The proposed method enables joint estimation of the channel matrix, the symbol vector, and the postulated noise variance.
- We evaluate the performance of the developed VB algorithms by comparing them with the LMMSE detector as well as the AMP-based and SIC schemes in various channel settings. Numerical simulations using i.i.d. Gaussian channels indicate that the detection error performance of the developed VB algorithms is better in finite-dimensional systems and comparable to that of SER-optimal AMP-based algorithms for the large-system limit. The VB algorithms, particularly LMMSE-VB, exhibit superior detection performance in systems with correlated channels, realistic 3GPP channels, and channel estimation mismatch.

Notation. x_{ij} and $[\mathbf{X}]_{ij}$ equivalently denote the element in the ith row and jth column of a matrix $\mathbf{X}$; $\mathbf{x}_i$ is the ith column of a matrix $\mathbf{X}$; $\mathrm{Tr}\{\mathbf{X}\}$ and $|\mathbf{X}|$ stand for the trace and the determinant, respectively, of a square matrix $\mathbf{X}$; $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $\mathcal{CN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = 1/(\pi^K|\boldsymbol{\Sigma}|)\exp\big(-(\mathbf{x}-\boldsymbol{\mu})^H\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big)$ denotes the probability distribution function (PDF) of a length-K random vector $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; $\mathbb{E}_{p(x)}[x]$ and $\mathrm{Var}_{p(x)}[x]$ are the mean and the variance of x with respect to its distribution $p(x)$; $\langle x \rangle$, $\langle |x|^2 \rangle$, and $\sigma_x^2 = \langle |x|^2 \rangle - \big|\langle x \rangle\big|^2$ denote the mean, the second moment, and the variance of x with respect to a variational distribution $q(x)$; $\sim$ and $\propto$ stand for "distributed according to" and "proportional to", respectively.

Outline. The rest of the paper is organized as follows. Section II describes the system model. Section III revisits the MIMO detection problem with perfect CSIR, and Section IV presents background on VB inference. Sections V and VI propose new VB methods for MIMO detection with perfect and imperfect CSIR, respectively. Section VII provides numerical results assessing the performance of the proposed algorithms. Finally, Section VIII summarizes our contributions.

II. System Model

We consider a MIMO system with K inputs and M outputs, in which the received signal vector $\mathbf{y} \in \mathbb{C}^M$ is given by

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

Here, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{C}^{M\times K}$ denotes the channel, $\mathbf{x} = [x_1, \ldots, x_K] \in \mathbb{C}^K$ is the input signal vector, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I}_M)$ models the additive noise. Furthermore, we define $\beta = K/M$ as the *system ratio*. Without loss of generality, we refer to the specific case of uplink transmission with K single-antenna users and a M-antenna BS. We assume that the transmitted symbol x_i from user i is drawn from a complex-valued discrete constellation $\mathcal{S}$, e.g., quadrature amplitude modulation (QAM) or phase-shift keying (PSK), and is normalized such that $\mathbb{E}[x_i] = 0$ and $\mathbb{E}\big[|x_i|^2\big] = 1$. The prior distribution of x_i is given by

$$p(x_i) = \sum_{a\in\mathcal{S}} p_a \delta(x_i - a), \tag{2}$$

where p_a corresponds to the known prior probability of the constellation point $a \in \mathcal{S}$ and $\delta(x_i - a)$ indicates the mass point at a.

Unless otherwise stated, we assume that the channel vector $\mathbf{h}_i \in \mathbb{C}^M$ associated with user i is Gaussian with $p(\mathbf{h}_i) = \mathcal{CN}(\mathbf{h}_i; \mathbf{0}, \mathbf{R}_i)$, where $\mathbf{R}_i = \mathbb{E}[\mathbf{h}_i\mathbf{h}_i^H]$ is the covariance matrix. It is noted that $\mathbf{R}_i$ is generally not a scaled identity matrix and is typically modeled to reflect the spatial correlation and the large-scale fading from user i to the BS [15]. Finally, we assume that $\mathbb{E}[\mathbf{h}_i\mathbf{h}_j^H] = \mathbf{0}$ if $i \neq j$. The objective of this paper is to obtain an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ from the observation $\mathbf{y}$ with minimum mean squared detection error $\mathbb{E}\big[\|\mathbf{x} - \hat{\mathbf{x}}\|^2\big]$.

III. MIMO Detection with Perfect CSIR

This section revisits MIMO detection for systems with perfect CSIR and describes some state-of-the-art methods that will be used to benchmark the proposed VB algorithms in Section VII.

A. Conventional MIMO Detection Schemes

When the distribution of $\mathbf{x}$ is discrete, an optimal detector that minimizes the detection error can be obtained through the MAP criterion:

$$\begin{aligned}\hat{\mathbf{x}}_{\mathrm{MAP}} &= \underset{\mathbf{x}\in\mathcal{S}^K}{\arg\max}\ p(\mathbf{y}|\mathbf{x};\mathbf{H})p(\mathbf{x}) \\ &= \underset{\mathbf{x}\in\mathcal{S}^K}{\arg\max}\ \big[\ln p(\mathbf{x}) - N_0^{-1}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big].\end{aligned} \tag{3}$$

While the MAP detector (or the ML detector for uniform $p(x_i)$) is optimal in terms of SER, its complexity grows exponentially with the number of inputs, making it infeasible for large-scale MIMO detection. Linear detectors with low complexity are practical candidates for massive MIMO systems. The estimated symbol is obtained via a linear combination of the received signal $\mathbf{y}$, which is then projected onto

the nearest symbol in the constellation $\mathcal{S}$. Among the linear detectors, the LMMSE detector achieves the best detection error performance. This detector first obtains the LMMSE estimate of $\mathbf{x}$ as

$$\hat{\mathbf{x}}_{\text{LMMSE}} = \left(\mathbf{H}^H\mathbf{H} + N_0\mathbf{I}_K\right)^{-1}\mathbf{H}^H\mathbf{y}, \tag{4}$$

which is then element-wise projected onto $\mathcal{S}$. We note that the LMMSE detector requires the inverse of a $(K \times K)$-dimensional matrix.

B. MIMO Detection via Approximate Message Passing

The AMP algorithm [7] was proposed as a computationally efficient iterative method to recover a sparse vector $\mathbf{x}$ from measurements in the form of (1). Initializing at iteration $t = 1$ with $\hat{x}_i^1 = \mathbb{E}_{p(x_i)}[x_i]$, $\mathbf{r}^1 = \mathbf{y}$, and $\nu_1^2 = \text{Var}_{p(x_i)}[x_i]$, the ***AMP algorithm*** consists of the following steps:

$$\begin{aligned}
\mathbf{z}^t &= \hat{\mathbf{x}}^t + \mathbf{H}^H\mathbf{r}^t, && \text{(linear estimator)}\\
\sigma_t^2 &= N_0 + \beta\nu_t^2, && \text{(error variance of } \mathbf{z}^t)\\
\hat{\mathbf{x}}^{t+1} &= \eta(\mathbf{z}^t, \sigma_t^2), && \text{(nonlinear denoiser)}\\
\nu_{t+1}^2 &= \sigma_t^2\langle\eta'(\mathbf{z}^t, \sigma_t^2)\rangle, && \text{(error variance of } \mathbf{x}^{t+1})\\
\mathbf{r}^{t+1} &= \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{t+1} + \beta\frac{\nu_{t+1}^2}{\sigma_t^2}\mathbf{r}^t, && \text{(Onsager-corrected residual)}
\end{aligned}$$

which are repeated until convergence or until a certain number of iterations is reached. Here, $\eta(\cdot, \sigma_t^2) : \mathbb{C}^K \to \mathbb{C}^K$ is a nonlinear *denoising* function parameterized by σ_t^2 and $\langle\eta'(\mathbf{z}^t, \sigma_t^2)\rangle = (1/K)\text{Tr}\{\partial\eta(\mathbf{z}^t, \sigma_t^2)/\partial\mathbf{z}^t\}$ is its *divergence* at $\mathbf{z}^t$. When $\mathbf{H}$ is a large i.i.d. sub-Gaussian matrix, the linear estimator applied to the Onsager-corrected residual decouples the system into K parallel AWGN channels $z_i = x_i + \mathcal{CN}(0, \sigma_t^2)$ [7], [12]. Thus, the denoiser is separable and can be applied element-wise to $\mathbf{z}^t$. It is worth noting that the postulated error variance σ_t^2 comprises two terms: one that reflects the true noise variance N_0 and one that accounts for the error variance in the denoising step. The AMP method was originally developed for real-valued system models. The complex Bayesian AMP (cB-AMP) algorithm presented above was recently analyzed for large-scale MIMO detection with complex-valued symbols in [8]. The so-called large MIMO AMP (LAMA) algorithm developed in [8] employs a minimum mean squared error (MMSE) denoiser $\eta(z_i^t, \sigma_t^2) = \mathsf{F}(z_i^t, \sigma_t^2)$, which is defined as the mean of the posterior distribution $p(x_i|z_i^t; \sigma_t^2)$. Given $\mathsf{G}(z_i^t, \sigma_t^2)$ as the variance of the posterior distribution $p(x_i|z_i^t; \sigma_t^2)$, the divergence $\langle\mathsf{F}'(\mathbf{z}^t, \sigma_t^2)\rangle$ is equal to $\left(1/(K\sigma_t^2)\right)\sum_{i=1}^K \mathsf{G}(z_i^t, \sigma_t^2)$ [16]. Effectively, $\nu_{t+1}^2 = \sigma_t^2\langle\mathsf{F}'(\mathbf{z}^t, \sigma_t^2)\rangle = (1/K)\sum_{i=1}^K \mathsf{G}(z_i^t, \sigma_t^2)$ is the empirical error variance of the MMSE denoiser $\mathsf{F}(z_i^t, \sigma_t^2)$.

The convergence of the AMP algorithm is rigorously established through the algorithm's state evolution for i.i.d. Gaussian [9] and i.i.d. sub-Gaussian $\mathbf{H}$ [10]. However, AMP diverges in many practical scenarios, e.g., in the case of ill-conditioned or non-zero-mean $\mathbf{H}$ [12]. In the context of MIMO detection, AMP diverges when the columns of $\mathbf{H}$ exhibit correlated elements or their vector norms are significantly uneven due to users with different large-scale fading coefficients. This limitation prompted the development of AMP-like algorithms that converge for a larger class of matrices than i.i.d. sub-Gaussian, including OAMP [11] and VAMP [12] for unitarily invariant matrices.

While there are subtle differences between OAMP and VAMP in terms of implementation, they are essentially equivalent [17]. The OAMP/VAMP algorithm involves the iterations between a linear estimator and a nonlinear denoiser. Initializing at iteration $t = 1$ with $\hat{x}_1^1 = \mathbb{E}_{p(x_i)}[x_i]$ and $\nu_1^2 = \left(1/\text{Tr}\{\mathbf{H}^H\mathbf{H}\}\right)\left(\|\mathbf{y}\|^2 - MN_0\right)$, the ***OAMP/VAMP algorithm*** consists of the following steps:

$$\begin{aligned}
\hat{\mathbf{A}}^t &= \nu_t^2(\nu_t^2\mathbf{H}^H\mathbf{H} + N_0\mathbf{I}_K)^{-1}\mathbf{H}^H,\\
\mathbf{A}^t &= \frac{K}{\text{Tr}\{\hat{\mathbf{A}}^t\mathbf{H}\}}\hat{\mathbf{A}}^t, && \text{(linear filter)}\\
\mathbf{z}^t &= \hat{\mathbf{x}}^t + \mathbf{A}^t(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^t), && \text{(linear estimator)}\\
\sigma_t^2 &= \frac{N_0\|\mathbf{A}^t\|_F^2 + \nu^t\|\mathbf{I}_K - \mathbf{A}^t\mathbf{H}\|_F^2}{K}, && \text{(error variance of } \mathbf{z}^t)\\
\hat{\mathbf{x}}^{t+1} &= \frac{\eta(\mathbf{z}^t, \sigma_t^2) - \langle\eta'(\mathbf{z}^t, \sigma_t^2)\rangle\mathbf{z}^t}{1 - \langle\eta'(\mathbf{z}^t, \sigma_t^2)\rangle}, && \text{(nonlinear denoiser)}\\
\nu_{t+1}^2 &= \frac{\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{t+1}\|^2 - MN_0}{\text{Tr}\{\mathbf{H}^H\mathbf{H}\}}, && \text{(error variance of } \mathbf{x}^{t+1})
\end{aligned}$$

which are repeated until convergence or until a certain number of iterations is reached. In OAMP/VAMP, the linear filter $\hat{\mathbf{A}}^t$ can be the MF $\mathbf{H}^H$ (as in the AMP scheme), the pseudo-inverse filter $\mathbf{H}^\dagger$, or the LMMSE filter (as in the form presented above). The linear estimate $\mathbf{z}^t$ is passed through a nonlinear denoiser which also removes the divergence $\langle\eta'(\mathbf{z}^t, \sigma_t^2)\rangle$ to obtain a *divergence-free* estimate $\hat{\mathbf{x}}^{t+1}$. Effectively, the OAMP/VAMP algorithm decouples the linear MIMO channel into K parallel AWGN channels $z_i = x_i + \mathcal{CN}(0, \sigma_t^2)$. In its optimal form, OAMP/VAMP adopts the LMMSE filter and the MMSE denoiser, i.e., $\eta(\mathbf{z}^t, \sigma_t^2) = \mathsf{F}(\mathbf{z}^t, \sigma_t^2)$ and $\langle\eta'(\mathbf{z}^t, \sigma_t^2)\rangle = \left(1/(K\sigma_t^2)\right)\sum_{i=1}^K \mathsf{G}(z_i^t, \sigma_t^2)$. Since OAMP/VAMP with the LMMSE filter significantly outperforms its counterparts with MF and the pseudo-inverse filter [11], we will only present numerical results for this version and hereafter refer to it simply as the ***OAMP/VAMP algorithm***. Compared with AMP, OAMP/VAMP requires one matrix inversion per iteration. Interestingly, the matrix inversion in the LMMSE filter can be circumvented in the *economy form* of OAMP/VAMP by first performing the singular value decomposition of the channel. More details on the computation of $p(x_i|z_i^t; \sigma_t^2)$, $\mathsf{F}(z_i^t, \sigma_t^2)$, and $\mathsf{G}(z_i^t, \sigma_t^2)$ with discrete $p(x_i)$ used in the AMP and OAMP/VAMP algorithms are given in Appendix A, which also provides an expression for the final MAP estimate $\hat{x}_i$.

IV. Background on VB Inference

This section presents background on VB inference, which we will exploit to solve the MIMO detection problem at hand. The goal of VB inference is to find a suitable approximation for a computationally intractable posterior distribution $p(\mathbf{x}|\mathbf{y})$ given a probabilistic model that specifies the joint distribution $p(\mathbf{x}, \mathbf{y})$. Here, $\mathbf{y}$ represents the set of all observed variables and $\mathbf{x}$ represents the set of m latent variables and parameters.

The VB method consists of finding a distribution function $q(\mathbf{x})$ within a family $\mathcal{Q}$ of distributions with its own set of variational parameters that make $q(\mathbf{x})$ as close as possible to the posterior distribution of interest $p(\mathbf{x}|\mathbf{y})$. In particular, VB inference amounts to solving the following optimization problem:

$$q(\mathbf{x}) = \underset{q(\mathbf{x})\in\mathcal{Q}}{\arg\min}\ \mathrm{KL}\big(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{y})\big), \tag{5}$$

where

$$\mathrm{KL}\big(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{y})\big) = \mathbb{E}_{q(\mathbf{x})}\big[\ln q(\mathbf{x})\big] - \mathbb{E}_{q(\mathbf{x})}\big[\ln p(\mathbf{x}|\mathbf{y})\big] \tag{6}$$

denotes the Kullback-Leibler (KL) divergence of $q(\mathbf{x})$ from $p(\mathbf{x}|\mathbf{y})$. Expanding $p(\mathbf{x}|\mathbf{y})$ reveals

$$\begin{aligned}\mathrm{KL}\big(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{y})\big) &= \mathbb{E}_{q(\mathbf{x})}\big[\ln q(\mathbf{x})\big] - \mathbb{E}_{q(\mathbf{x})}\big[\ln p(\mathbf{x},\mathbf{y})\big] \\ &\quad + \ln p(\mathbf{y}).\end{aligned} \tag{7}$$

Since $p(\mathbf{y})$ does not depend on $q(\mathbf{x})$, maximizing the evidence lower bound (ELBO),[1] defined as

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\mathbf{x})}\big[\ln p(\mathbf{x},\mathbf{y})\big] - \mathbb{E}_{q(\mathbf{x})}\big[\ln q(\mathbf{x})\big], \tag{8}$$

is equivalent to minimizing the KL divergence. The maximum possible value of $\mathrm{ELBO}(q)$ occurs when $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$.

Since attempting to match the true posterior distribution with an arbitrary $q(\mathbf{x})$ is typically intractable, it is more practical to consider a restricted family of distributions $q(\mathbf{x})$. Here, the *mean-field variational family* is constructed such that

$$q(\mathbf{x}) = \prod_{i=1}^{m} q_i(x_i), \tag{9}$$

where the latent variables are taken to be mutually independent and each is governed by a distinct factor in the variational distribution. Among all distributions $q(\mathbf{x})$ having the form in (9), the general expression for the optimal solution of the variational distribution $q_i(x_i)$ that maximizes the ELBO can be obtained as [14]

$$q_i(x_i) \propto \exp\left\{\big\langle \ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\big\rangle\right\}, \tag{10}$$

where $\langle\cdot\rangle$ denotes the expectation with respect to all latent variables except x_i using the currently fixed variational distribution $q_{-i}(\mathbf{x}_{-i}) = \prod_{j\neq i} q_j(x_j)$. By iterating the update of $q_i(x_i)$ sequentially over all j, the $\mathrm{ELBO}(q)$ objective function can be monotonically improved. This is the basis behind the *coordinate ascent variational inference (CAVI)* algorithm, which guarantees convergence to at least a local optimum of $\mathrm{ELBO}(q)$ [14], [18].

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

Theorem 1. *Let the random matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ and the random vector $\mathbf{x} \in \mathbb{C}^n$ be independent with respect to a variational distribution $q(\mathbf{A},\mathbf{x}) = q(\mathbf{A})q(\mathbf{x})$. Assuming that $\mathbf{A}$ is column-wise independent, let $\langle\mathbf{a}_i\rangle$ and $\mathbf{\Sigma}_{\mathbf{a}_i}$ denote the variational mean and covariance matrix, respectively, of the ith column of $\mathbf{A}$. Furthermore, let $\langle\mathbf{x}\rangle$ and $\mathbf{\Sigma}_{\mathbf{x}} = \mathrm{diag}(\sigma^2_{x_1},\ldots,\sigma^2_{x_n})$ denote the variational mean and covariance matrix, respectively, of $\mathbf{x}$. Considering an arbitrary vector $\mathbf{y}\in\mathbb{C}^m$ and defining the expectation $\big\langle\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2\big\rangle$ with respect to $q(\mathbf{A},\mathbf{x})$, we have*

$$\begin{aligned}\big\langle\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2\big\rangle &= \big\|\mathbf{y}-\langle\mathbf{A}\rangle\langle\mathbf{x}\rangle\big\|^2 + \mathrm{Tr}\big\{\langle\mathbf{A}\rangle\mathbf{\Sigma}_{\mathbf{x}}\langle\mathbf{A}^H\rangle\big\} \\ &\quad + \sum_{i=1}^{n}\big\langle|x_i|\big\rangle^2\mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_i}\}.\end{aligned} \tag{11}$$

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

$$\begin{aligned}\big\langle\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2\big\rangle &= \|\mathbf{y}\|^2 - 2\,\Re\big\{\mathbf{y}^H\langle\mathbf{A}\mathbf{x}\rangle\big\} + \langle\mathbf{x}^H\mathbf{A}^H\mathbf{A}\mathbf{x}\rangle \\ &= \big\|\mathbf{y}-\langle\mathbf{A}\rangle\langle\mathbf{x}\rangle\big\|^2 - \langle\mathbf{x}^H\rangle\langle\mathbf{A}^H\rangle\langle\mathbf{A}\rangle\langle\mathbf{x}\rangle \\ &\quad + \mathrm{Tr}\big\{\langle\mathbf{A}^H\mathbf{A}\rangle\langle\mathbf{x}\mathbf{x}^H\rangle\big\}.\end{aligned} \tag{12}$$

Note that $\langle\mathbf{x}\mathbf{x}^H\rangle = \langle\mathbf{x}\rangle\langle\mathbf{x}^H\rangle + \mathbf{\Sigma}_{\mathbf{x}}$. In addition, we have

$$\begin{aligned}\big[\langle\mathbf{A}^H\mathbf{A}\rangle\big]_{ij} &= \langle\mathbf{a}_i^H\mathbf{a}_j\rangle \\ &= \begin{cases}\langle\mathbf{a}_i^H\rangle\langle\mathbf{a}_i\rangle + \mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_i}\}, & \text{if } i=j \\ \langle\mathbf{a}_i^H\rangle\langle\mathbf{a}_j\rangle, & \text{otherwise.}\end{cases}\end{aligned} \tag{13}$$

Thus, it follows that $\langle\mathbf{A}^H\mathbf{A}\rangle = \langle\mathbf{A}^H\rangle\langle\mathbf{A}\rangle + \mathbf{D}$, where $\mathbf{D} = \mathrm{diag}\big(\mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_1}\},\ldots,\mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_n}\}\big)$ and, as a result, we have

$$\begin{aligned}\mathrm{Tr}\big\{\langle\mathbf{A}^H\mathbf{A}\rangle\langle\mathbf{x}\mathbf{x}^H\rangle\big\} &= \langle\mathbf{x}^H\rangle\langle\mathbf{A}^H\rangle\langle\mathbf{A}\rangle\langle\mathbf{x}\rangle \\ &\quad + \mathrm{Tr}\big\{\langle\mathbf{A}\rangle\mathbf{\Sigma}_{\mathbf{x}}\langle\mathbf{A}^H\rangle\big\} \\ &\quad + \langle\mathbf{x}^H\rangle\mathbf{D}\langle\mathbf{x}\rangle + \mathrm{Tr}\{\mathbf{D}\mathbf{\Sigma}_{\mathbf{x}}\}.\end{aligned} \tag{14}$$

The proof is concluded by removing the duplicated terms in (12) and exploiting the fact that $\langle\mathbf{x}\rangle^H\mathbf{D}\langle\mathbf{x}\rangle + \mathrm{Tr}\{\mathbf{D}\mathbf{\Sigma}_{\mathbf{x}}\} = \sum_{i=1}^n\big|\langle x_i\rangle\big|^2\mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_i}\} + \sum_{i=1}^n\sigma^2_{x_i}\mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_i}\} = \sum_{i=1}^n\big\langle|x_i|\big\rangle^2\mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_i}\}$. ∎

Corollary 1. *If $\mathbf{A}$ is deterministic, we have*

$$\big\langle\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2\big\rangle = \big\|\mathbf{y}-\mathbf{A}\langle\mathbf{x}\rangle\big\|^2 + \mathrm{Tr}\{\mathbf{A}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{A}^H\}. \tag{15}$$

Proof: This is a direct result of Theorem 1 by exploiting the fact that $\mathrm{Tr}\{\mathbf{\Sigma}_{\mathbf{a}_i}\} = \mathbf{0},\ \forall i$. ∎

V. VB Inference for MIMO Detection with Perfect CSIR

In this section, we apply VB inference to the MIMO detection problem with known channel matrix $\mathbf{H}$. We first review the conventional mean-field VB framework with known noise variance [13], which will be referred to in the following as the ***conv-VB algorithm***. We then compare conv-VB with the SIC method [5] and identify its limitation with respect to the latter. Lastly, we develop MF-VB and LMMSE-VB algorithms for MIMO detection.

A. Conventional VB Inference with Known Noise Variance

With known noise statistics, the joint distribution $p(\mathbf{y},\mathbf{x};\mathbf{H},N_0)$ can be factorized as

$$p(\mathbf{y},\mathbf{x};\mathbf{H},N_0) = p(\mathbf{y}|\mathbf{x};\mathbf{H},N_0)p(\mathbf{x}), \tag{16}$$

with $p(\mathbf{y}|\mathbf{x};\mathbf{H},N_0) = \mathcal{CN}(\mathbf{y};\mathbf{H}\mathbf{x},N_0\mathbf{I}_M)$. Given the observation $\mathbf{y}$, the mean-field variational distribution

[1]The negative of the ELBO is commonly referred to as the Gibbs free energy.

$$g_i(x_i) = \exp\Big\{ \ln p(x_i) - N_0^{-1}\Big(\|\mathbf{h}_i\|^2|x_i|^2 - 2\,\Re\Big\{ \mathbf{h}_i^H \Big(\mathbf{y} - \sum_{j\neq i}^{K} \mathbf{h}_j \langle x_j \rangle \Big) x_i^* \Big\} \Big) \Big\}. \tag{18}$$

$q(\mathbf{x}) = \prod_{i=1}^{K} q_i(x_i)$ is derived to approximate the posterior distribution $p(\mathbf{x}|\mathbf{y}; \mathbf{H}, N_0)$. Expanding the conditional $p(\mathbf{y}|\mathbf{x}; \mathbf{H}, N_0)p(\mathbf{x})$, taking the expectation with respect to all latent variables except x_i using the variational distribution $\prod_{j\neq i} q_j(x_j)$, and retaining only the components that are related to x_i, we have

$$\begin{aligned} q_i(x_i) &\propto \exp\big\{ \big\langle \ln p(\mathbf{y}|\mathbf{x}; \mathbf{H}, N_0) + \ln p(\mathbf{x}) \big\rangle \big\} \\ &\propto \exp\big\{ \big\langle \ln p(\mathbf{x}) - N_0^{-1} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \big\rangle \big\} \\ &\propto \exp\Big\{ \ln p(x_i) - N_0^{-1} \Big\langle \Big\| \mathbf{y} - \mathbf{h}_i x_i - \sum_{j\neq i}^{K} \mathbf{h}_j x_j \Big\|^2 \Big\rangle \Big\} \\ &= g_i(x_i), \end{aligned} \tag{17}$$

where $g_i(x_i)$ can be expanded as in (18) at the top of the page.

In the context of MIMO detection, the distribution $q_i(x_i)$ can be normalized such that

$$q_i(a) = \frac{g_i(a)}{\sum_{b\in\mathcal{S}} g_i(b)}, \quad \forall a \in \mathcal{S}. \tag{19}$$

The variational mean $\langle x_i \rangle$ with respect to $q_i(x_i)$ is then given by

$$\langle x_i \rangle = \sum_{a\in\mathcal{S}} a\, q_i(a). \tag{20}$$

Here, $\langle x_i \rangle$ can be interpreted as a *soft* detection of x_i. By iterating the update of $q_i(x_i)$ and $\langle x_i \rangle$ for $i = 1, \ldots, K$, we attain the CAVI algorithm for soft symbol detection. This procedure is summarized in Algorithm 1, where $\hat{x}_i^t$ is used to replace the variational mean $\langle x_i \rangle$ at iteration t. The iterative procedure is repeated until convergence or a until certain number of iterations is reached. We note that Algorithm 1 corresponds to the version of the VB method for MIMO detection developed in [13]. However, we observe in our simulations that Algorithm 1 occasionally yields an NaN error. This happens when the argument inside the exponential function in (18) becomes too large. In the following, we present an equivalent form of the variational distribution $q_i(x_i)$ that helps tackle this issue.

Let the variational mean $\langle x_j \rangle$ be the current *soft* estimate of symbol x_j, $\forall j$, and denote

$$z_i = \frac{\mathbf{h}_i^H}{\|\mathbf{h}_i\|^2} \Big(\mathbf{y} - \sum_{j\neq i}^{K} \mathbf{h}_j \langle x_j \rangle \Big). \tag{21}$$

Noting that z_i is a constant with respect to the variational distribution $q_i(x_i)$, (17)–(18) can be rewritten as

$$\begin{aligned} q_i(x_i) &\propto p(x_i) \exp\big\{ -N_0^{-1} \|\mathbf{h}_i\|^2 \big(|x_i|^2 - 2\,\Re\{x_i^* z_i\} \big) \big\} \\ &\propto p(x_i) \exp\big\{ -N_0^{-1} \|\mathbf{h}_i\|^2 |x_i - z_i|^2 \big\} \\ &\propto p(x_i)\, \mathcal{CN}(z_i; x_i, \zeta_i^2), \end{aligned} \tag{22}$$

where $\zeta_i^2 = N_0/\|\mathbf{h}_i\|^2$. Here, $\mathcal{CN}(x_i; z_i, \zeta_i^2)$ can be interpreted as the likelihood function $p(z_i|x_i; \zeta_i^2)$. In other words, the mean-field VB approximation decouples the linear MIMO system into K parallel AWGN channels $z_i = x_i + \mathcal{CN}\big(0, \zeta_i^2\big)$.

Algorithm 1: ***conv-VB algorithm*** with known noise variance [13]

Input: $\mathbf{y}$, $\mathbf{H}$, N_0, and prior distributions $\{p(x_i)\}$;
Output: $\hat{\mathbf{x}}$;
Initialize $\hat{\mathbf{x}}^1 = \mathbf{0}$;
for $t = 1, 2, \ldots$ **do**
  for $i = 1, 2, \ldots, K$ **do**
    Compute $g_i(x_i)$ as in (17);
    Normalize the distribution $q_i(x_i)$ as in (19);
    Compute $\hat{x}_i^t$ as in (20) with respect to $q_i(x_i)$;
  end
end
MAP estimate: $\hat{x}_i \leftarrow \arg\max_{a\in\mathcal{S}} q_i(a)$.

Remark 1: *The expression for z_i in* (21) *can be expressed as $z_i = \langle x_i \rangle + \big(\mathbf{h}_i^H/\|\mathbf{h}_i\|^2\big)\big(\mathbf{y} - \mathbf{H}\langle \mathbf{x} \rangle\big)$. Interestingly, this expression presents a linear estimator of x_i, which is similar to the first step in the AMP method and in the OAMP scheme with MF, albeit with the subtle change involving use of the MF $\mathbf{h}_i^H/\|\mathbf{h}_i\|^2$. The variational distribution $q_i(x_i)$ and the corresponding variational mean $\langle x_i \rangle$ and variance $\sigma_{x_i}^2$ can be obtained in the same manner as $p(x_i|z_i; \zeta_i^2)$ and the corresponding posterior mean $\mathsf{F}(x_i, \zeta_i^2)$ and variance $\mathsf{G}(x_i, \zeta_i^2)$ presented in Appendix A. Note that ζ_i^2 is now used in place of σ_i^2 in the AMP-based algorithms. Since now the argument inside the exponential function of $q_i(x_i)$ is always negative, the overflow issue with $q_i(x_i)$ is averted.*

B. Comparison of Conventional VB Inference with SIC

The variational distribution $q_i(x_i)$ in the form of (22) is the *exact* posterior distribution of the following linear model:

$$\mathbf{y} = \mathbf{h}_i x_i + \sum_{j\neq i}^{K} \mathbf{h}_j \langle x_j \rangle + \mathbf{n}, \tag{23}$$

where $\sum_{j\neq i}^{K} \mathbf{h}_j \langle x_j \rangle$ is the realized inter-user interference and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, N_0 \mathbf{I}_M)$. Comparing with the system model (1), the variational distribution $q_i(x_i)$ corresponds to the true posterior distribution $p(x_i|\mathbf{y})$ if the estimate $\langle x_j \rangle$ is the same as the true signal x_j, $\forall j \neq i$. In this case, the mean-field VB estimation of x_i is Bayes optimal. However, the system model (1) can also be written as

$$\mathbf{y} = \mathbf{h}_i x_i + \sum_{j\neq i}^{K} \mathbf{h}_j \langle x_j \rangle + \mathbf{n}_i, \tag{24}$$

where $\mathbf{n}_i = \sum_{j\neq i}^{K} \mathbf{h}_j \big(x_j - \langle x_j \rangle \big) + \mathbf{n}$ is the residual interference-plus-noise. With respect to the variational distribution $q_{-i}(\mathbf{x}_{-i})$, $\mathbf{n}_i$ has covariance matrix $\mathbf{C}_i =$

$\sum_{j\neq i}^{K}\sigma_{x_j}^2\mathbf{h}_j\mathbf{h}_j^H + N_0\mathbf{I}_M$. Since $\mathbf{C}_i \succeq N_0\mathbf{I}_M$, the variational distribution $q_i(x_i)$ in (22) may not represent a good approximation of the posterior distribution $p(x_i|\mathbf{y})$, especially when the residual inter-user interference is not negligible. This observation explains the poor performance of conv-VB with respect to the AMP-based algorithms. We note that this issue was previously reported in [19] for the sparse signal recovery problem by comparing the difference between the fixed points of the AMP and the mean-field VB schemes.

We now compare conv-VB with SIC-based MIMO detection [5]. With a slight abuse of notation, we use $\langle x_i\rangle$ to denote the current soft estimate of x_i in SIC. The key idea of SIC is to first approximate the residual interference-plus-noise $\mathbf{n}_i$ as Gaussian while fixing the estimate $\langle x_j\rangle,\ \forall j\neq i$. Then, the likelihood function $p\big(\mathbf{y}|x_i;\langle\mathbf{x}_{-i}\rangle\big)$ is approximated as $\mathcal{CN}\big(\mathbf{y};\mathbf{h}_i x_i+\sum_{j\neq i}^{K}\mathbf{h}_j\langle x_j\rangle,\mathbf{C}_i\big)$, enabling a tractable derivation of the posterior distribution $p\big(x_i|\mathbf{y};\langle\mathbf{x}_{-i}\rangle\big)$ via Bayes' theorem.[2] To this end, we examine two approaches to combine the output signal $\mathbf{y}$, cancel the interference $\sum_{j\neq i}^{K}\mathbf{h}_j\langle x_j\rangle$, and generate the soft estimate $\langle x_i\rangle$.

- SIC with MF (***MF-SIC algorithm***): By applying the MF $\mathbf{h}_i^H/\|\mathbf{h}_i\|^2$ as in (21) to (24), one attains a linear estimate z_i of x_i with interference cancellation as $z_i \approx x_i + \mathcal{CN}\big(0,\mathbf{h}_i^H\mathbf{C}_i\mathbf{h}_i/\|\mathbf{h}_i\|^4\big)$. The soft estimate $\langle x_i\rangle$ can then be obtained as the posterior mean of $p\big(x_i|z_i;\mathbf{h}_i^H\mathbf{C}_i\mathbf{h}_i/\|\mathbf{h}_i\|^4\big)$.
- SIC with LMMSE filter (***LMMSE-SIC algorithm***): We note that

$$\begin{aligned}
&p\big(\mathbf{y}|x_i;\langle\mathbf{x}_{-i}\rangle\big)\\
&\approx \mathcal{CN}\bigg(\mathbf{y};\mathbf{h}_i x_i+\sum_{j\neq i}^{K}\mathbf{h}_j\langle x_j\rangle,\mathbf{C}_i\bigg)\\
&= \mathcal{CN}\bigg(x_i;\frac{\mathbf{h}_i^H\mathbf{C}_i^{-1}\big(\mathbf{y}-\sum_{j\neq i}^{K}\mathbf{h}_j\langle x_j\rangle\big)}{\mathbf{h}_i^H\mathbf{C}_i^{-1}\mathbf{h}_i},\frac{1}{\mathbf{h}_i^H\mathbf{C}_i^{-1}\mathbf{h}_i}\bigg)\\
&= \mathcal{CN}\bigg(z_i;x_i,\frac{1}{\mathbf{h}_i^H\mathbf{C}_i^{-1}\mathbf{h}_i}\bigg),
\end{aligned}\tag{25}$$

  where

$$z_i=\frac{\mathbf{h}_i^H\mathbf{C}_i^{-1}\big(\mathbf{y}-\sum_{j\neq i}^{K}\mathbf{h}_j\langle x_j\rangle\big)}{\mathbf{h}_i^H\mathbf{C}_i^{-1}\mathbf{h}_i}.\tag{26}$$

  Thus, we have $z_i \approx x_i + \mathcal{CN}\big(0,1/(\mathbf{h}_i^H\mathbf{C}_i^{-1}\mathbf{h}_i)\big)$. The soft estimate $\langle x_i\rangle$ can then be obtained as the mean of the posterior distribution $p\big(x_i|z_i;1/(\mathbf{h}_i^H\mathbf{C}_i^{-1}\mathbf{h}_i)\big)$. We observe that z_i in the form of (26) is the LMMSE estimate of x_i after canceling the inter-user interference and whitening with the colored residual interference-plus-noise covariance matrix $\mathbf{C}_i$.

MF-SIC and LMMSE-SIC proceed with the iterative update over $\{x_i\}$ until convergence or a certain number of iterations is reached. Except for the removal of the divergence terms, the implementation of MF-SIC and LMMSE-SIC is similar to that of the AMP and OAMP/VAMP algorithms, respectively. However, while SIC methods yield good estimation performance as shown in several simulation scenarios in Section VII, there are two major shortcomings in the algorithm. First, the K $(M\times M)$-dimensional residual interference-plus-noise covariance matrices $\{\mathbf{C}_i\}$ need to be computed (and inverted in LMMSE-SIC) at each iteration, which leads to prohibitive complexity for large systems. Second, the SIC-based MIMO detection [5] may not be provably convergent. To address the poor performance of conv-VB with known noise variance and the shortcomings of the SIC algorithms, we develop two novel VB schemes that jointly estimate the symbol vector and the postulated noise variance or covariance matrix.

C. Proposed MF-VB for MIMO Detection

In practice, the noise variance N_0 is not known *a priori* and needs to be estimated as well. Moreover, the conventional VB method with known noise variance does not take into account the residual inter-user interference. Here, we consider the residual interference-plus-noise as a random variable $N_0^{\rm post}$, which is postulated by the estimation in the VB framework. For ease of computation, we use $\gamma = 1/N_0^{\rm post}$ to denote the precision of the estimation. We assume a conjugate prior Gamma distribution $\mathrm{Gamma}(a_0,b_0)$ for γ, where a_0 and b_0 are the shape and rate parameters of the distribution, respectively. The PDF of γ is thus given by

$$p(\gamma)=\frac{b_0^{\gamma}}{\Gamma(a_0)}\gamma^{a_0-1}\mathrm{e}^{-b_0\gamma},\tag{27}$$

where $\Gamma(a_0)$ is the Gamma function. Treating the precision γ as a random variable, the joint distribution $p(\mathbf{y},\mathbf{x},\gamma;\mathbf{H})$ can be factorized as

$$p(\mathbf{y},\mathbf{x},\gamma;\mathbf{H})=p(\mathbf{y}|\mathbf{x},\gamma;\mathbf{H})p(\mathbf{x})p(\gamma),\tag{28}$$

where $p(\mathbf{y}|\mathbf{x},\gamma;\mathbf{H}) = \mathcal{CN}(\mathbf{y};\mathbf{H}\mathbf{x},\gamma^{-1}\mathbf{I}_M)$. Given the observation $\mathbf{y}$, we aim at obtaining the mean-field variational distribution $q(\mathbf{x},\gamma)$ such that

$$p(\mathbf{x},\gamma|\mathbf{y};\mathbf{H})\approx q(\mathbf{x},\gamma)=\prod_{i=1}^{K}q_i(x_i)q(\gamma).\tag{29}$$

The optimization of $q(\mathbf{x},\gamma)$ is executed by iteratively updating $\{x_i\}$ and γ as follows.

1) Updating x_i. The variational distribution $q_i(x_i)$ is obtained by expanding the conditional in (28) and taking the expectation with respect to all latent variables except x_i using the variational distribution $\prod_{j\neq i}^{K}q_j(x_j)q(\gamma)$:

$$\begin{aligned}
q_i(x_i) &\propto \exp\big\{\big\langle \ln p(\mathbf{y}|\mathbf{x},\gamma;\mathbf{H})+\ln p(\mathbf{x})\big\rangle\big\}\\
&\propto \exp\big\{\big\langle \ln p(\mathbf{x})-\gamma\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big\rangle\big\}\\
&\propto p(x_i)\exp\big\{-\langle\gamma\rangle\|\mathbf{h}_i\|^2|x_i-z_i|^2\big\}\\
&\propto p(x_i)\,\mathcal{CN}\bigg(z_i;x_i,\frac{1}{\langle\gamma\rangle\|\mathbf{h}_i\|^2}\bigg),
\end{aligned}\tag{30}$$

where z_i is a linear estimate of x_i as defined in (21). The variational distribution $q_i(x_i)$ can be easily realized by normalizing $p(x_i)\,\mathcal{CN}\big(z_i;x_i,1/\big(\langle\gamma\rangle\|\mathbf{h}_i\|^2\big)\big)$. The variational mean and

[2]It can be proved that $\mathbf{n}_i$ is Gaussian for sufficiently large K using Lindeberg's condition for the central limit theorem. Specifically, the condition requires the independence between $x_j-\langle x_j\rangle,\ \forall j\neq i$ and the finite posterior variances $\{\sigma_{x_j}^2\}$ [20]. In addition, the covariance matrix $\mathbf{C}_i$ of $\mathbf{n}_i$ for an i.i.d. channel $\mathbf{H}$ and sufficiently large K tends to a scaled identity matrix, making $\mathbf{n}_i$ i.i.d. as well.

$$q(\mathbf{W}) \propto \exp\big\{ \ln|\mathbf{W}| - \big\langle (\mathbf{y}-\mathbf{H}\mathbf{x})^H \mathbf{W}(\mathbf{y}-\mathbf{H}\mathbf{x})\big\rangle + (n-M)\ln|\mathbf{W}| - \mathrm{Tr}\{\mathbf{W}_0^{-1}\mathbf{W}\}\big\}$$
$$\propto \exp\Big\{ (n-M+1)\ln|\mathbf{W}| - \mathrm{Tr}\Big\{ \big(\mathbf{W}_0^{-1} + \big(\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big)\big(\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big)^H + \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H\big)\mathbf{W}\Big\}\Big\}. \tag{39}$$

$$\langle\mathbf{W}\rangle = (n+1)\big(\mathbf{W}_0^{-1} + \big(\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big)\big(\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big)^H + \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H\big)^{-1}. \tag{40}$$

variance are then computed as $\langle x_i\rangle = \mathsf{F}\big(z_i, 1/\big(\langle\gamma\rangle\|\mathbf{h}_i\|^2\big)\big)$ and $\sigma_{x_i}^2 = \mathsf{G}\big(z_i, 1/\big(\langle\gamma\rangle\|\mathbf{h}_i\|^2\big)\big)$, respectively.

2) Updating γ. The variational distribution $q(\gamma)$ is obtained by taking the expectation of the conditional in (28) with respect to $q(\mathbf{x})$:

$$q(\gamma) \propto \exp\big\{\big\langle \ln p(\mathbf{y}|\mathbf{x},\gamma;\mathbf{H}) + \ln p(\gamma)\big\rangle\big\}$$
$$\propto \exp\big\{M\ln\gamma - \gamma\big\langle\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big\rangle + (a_0-1)\ln\gamma - b_0\gamma\big\}$$
$$\propto \exp\big\{(a_0+M-1)\ln\gamma - \gamma\big(b_0 + \big\|\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big\|^2 + \mathrm{Tr}\{\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H\}\big)\big\}. \tag{31}$$

Note that the last step is obtained as a result of Corollary 1. The variational distribution $q(\gamma)$ is thus Gamma with mean

$$\langle\gamma\rangle = \frac{a_0+M}{b_0 + \big\|\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big\|^2 + \mathrm{Tr}\{\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H\}}. \tag{32}$$

By iteratively optimizing $\big\{q_i(x_i)\big\}$ and $q(\gamma)$, we obtain the CAVI algorithm for estimating $\mathbf{x}$ and the precision γ. We refer to this scheme as the ***MF-VB algorithm*** due to the use of the MF $\mathbf{h}_i^H/\|\mathbf{h}_i\|^2$ to obtain the linear estimate z_i in (21).

Remark 2: *If the improper prior* $\mathrm{Gamma}(0,0)$ *is used,* $1/\langle\gamma\rangle = (1/M)\big(\big\|\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big\|^2 + \mathrm{Tr}\{\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H\}\big)$ *is now the point estimate of the deterministic unknown* N_0^{post}*. Similar to the AMP-based algorithms, the term* $\big\|\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big\|^2$ *reflects the empirical estimate of the true noise variance* N_0*, whereas the term* $\mathrm{Tr}\{\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^H\}$ *reflects the empirical error variance induced by the MMSE denoiser* $\mathsf{F}\big(z_i, 1/\big(\langle\gamma\rangle\|\mathbf{h}_i\|^2\big)\big)$*. This result, corresponding to the M-step in variational Bayesian expectation-maximization [18], coincides with the estimated noise variance proposed in [19]. However, unlike [19], we treat the reciprocal of the postulated noise variance as a random variable under the VB framework.*

Remark 3: *If* N_0 *is known, a weakly informative Gamma prior, e.g.,* $\mathrm{Gamma}(1, N_0)$ *that results in* $\mathbb{E}[\gamma] = 1/N_0$*, can be used instead of* $\mathrm{Gamma}(0,0)$*. Interestingly, we observe through numerous simulations that the MF-VB approach based on the improper prior* $\mathrm{Gamma}(0,0)$ *yields marginally more accurate estimation of* $\mathbf{x}$ *than the one based on a weakly informative prior. Thus, we use the improper prior* $\mathrm{Gamma}(0,0)$ *in the numerical simulation for MF-VB.*

D. Proposed LMMSE-VB for MIMO Detection

We now develop a VB method to estimate the input signal $\mathbf{x}$ using the postulated noise covariance matrix $\mathbf{C}^{\mathrm{post}}$ instead of using the postulated noise variance N_0^{post} as in Section V-C. For ease of computation, we use $\mathbf{W} = (\mathbf{C}^{\mathrm{post}})^{-1}$ to denote the precision matrix and assume a conjugate prior complex Wishart distribution $\mathcal{CW}(\mathbf{W}_0, n)$ for $\mathbf{W}$, where $\mathbf{W}_0 \succeq \mathbf{0}$ is the scale matrix and $n \geq M$ is the number of degrees of freedom. The PDF of $\mathbf{W} \sim \mathcal{CW}(\mathbf{W}_0, n)$ satisfies

$$p(\mathbf{W}) \propto |\mathbf{W}|^{n-M}\exp\big(-\mathrm{Tr}\{\mathbf{W}_0^{-1}\mathbf{W}\}\big). \tag{33}$$

The joint distribution $p(\mathbf{y},\mathbf{x},\mathbf{W};\mathbf{H})$ can be factorized as

$$p(\mathbf{y},\mathbf{x},\mathbf{W};\mathbf{H}) = p(\mathbf{y}|\mathbf{x},\mathbf{W};\mathbf{H})p(\mathbf{x})p(\mathbf{W}), \tag{34}$$

where $p(\mathbf{y}|\mathbf{x},\mathbf{W};\mathbf{H}) = \mathcal{CN}(\mathbf{y};\mathbf{H}\mathbf{x},\mathbf{W}^{-1})$. Given the observation $\mathbf{y}$, we aim at obtaining the mean-field variational distribution $q(\mathbf{x},\mathbf{W})$ such that

$$p(\mathbf{x},\mathbf{W}|\mathbf{y};\mathbf{H}) \approx q(\mathbf{x},\mathbf{W}) = \prod_{i=1}^{K} q_i(x_i)q(\mathbf{W}). \tag{35}$$

The optimization of $q(\mathbf{x},\mathbf{W})$ is executed by iteratively updating $\{x_i\}$ and $\mathbf{W}$ as follows.

1) Updating x_i. The variational distribution $q_i(x_i)$ is obtained by expanding the conditional in (34) and taking the expectation with respect to all latent variables except x_i using the variational distribution $\prod_{j\neq i}^{K} q_j(x_j)q(\mathbf{W})$:

$$q_i(x_i) \propto \exp\big\{\big\langle \ln p(\mathbf{y}|\mathbf{x},\mathbf{W};\mathbf{H}) + \ln p(\mathbf{x})\big\rangle\big\}$$
$$\propto \exp\big\{\big\langle \ln p(\mathbf{x}) - (\mathbf{y}-\mathbf{H}\mathbf{x})^H\mathbf{W}(\mathbf{y}-\mathbf{H}\mathbf{x})\big\rangle\big\}$$
$$\propto p(x_i)\exp\big\{-\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i|x_i-z_i|^2\big\}$$
$$\propto p(x_i)\,\mathcal{CN}\bigg(z_i; x_i, \frac{1}{\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i}\bigg), \tag{36}$$

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

$$z_i = \frac{\mathbf{h}_i^H\langle\mathbf{W}\rangle}{\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i}\bigg(\mathbf{y} - \sum_{j\neq i}^{K}\mathbf{h}_j\langle x_j\rangle\bigg)$$
$$= \langle x_i\rangle + \frac{\mathbf{h}_i^H\langle\mathbf{W}\rangle}{\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i}\big(\mathbf{y}-\mathbf{H}\langle\mathbf{x}\rangle\big). \tag{37}$$

The variational distribution $q_i(x_i)$ can easily be realized by normalizing $p(x_i)\,\mathcal{CN}\big(z_i; x_i, 1/(\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i)\big)$. The variational mean and variance are then computed as $\langle x_i\rangle = \mathsf{F}\big(z_i, 1/(\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i)\big)$ and $\sigma_{x_i}^2 = \mathsf{G}\big(z_i, 1/(\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i)\big)$, respectively.

2) Updating $\mathbf{W}$. The variational distribution $q(\mathbf{W})$ is obtained by taking the expectation of the conditional in (34) with respect to $q(\mathbf{x})$:

$$q(\mathbf{W}) \propto \exp\big\{\big\langle \ln p(\mathbf{y}|\mathbf{x},\mathbf{W};\mathbf{H}) + \ln p(\mathbf{W})\big\rangle\big\}. \tag{38}$$

Note that (38) can be expanded as in (39) at the top of the page by applying Corollary 1. The variational distribution $q(\mathbf{W})$ is thus complex Wishart with $n+1$ degrees of freedom and mean $\langle\mathbf{W}\rangle$ given in (40) at the top of the page.

By iteratively optimizing $\big\{q_i(x_i)\big\}$ and $q(\mathbf{W})$, we obtain the CAVI algorithm for estimating $\mathbf{x}$ and the precision matrix $\mathbf{W}$.

Algorithm 2: ***MF-VB/LMMSE-VB algorithm*** with postulated noise variance/covariance matrix

Input: $\mathbf{y}$, $\mathbf{H}$, and prior distributions $\{p(x_i)\}$;
Output: $\hat{\mathbf{x}}$;
Initialize $\hat{x}_i^1 = 0$ and $\sigma_{x_i,1}^2 = \mathrm{Var}_{p(x_i)}[x_i],\ \forall i$, and $\mathbf{r} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^1$;
for $t = 1, 2, \ldots$ **do**
  Update $\mathbf{\Sigma_x} = \mathrm{diag}(\sigma_{x_1,t}^2, \ldots, \sigma_{x_K,t}^2)$;
  $\gamma_t \leftarrow M/\big(\|\mathbf{r}\|^2 + \mathrm{Tr}\{\mathbf{H}\mathbf{\Sigma_x}\mathbf{H}^H\}\big)$ for MF-VB *or* $\mathbf{W}_t \leftarrow \big((\|\mathbf{r}\|^2/M)\mathbf{I}_M + \mathbf{H}\mathbf{\Sigma_x}\mathbf{H}^H\big)^{-1}$ for LMMSE-VB;
  for $i = 1, 2, \ldots, K$ **do**
    For MF-VB, compute

$$z_i^t \leftarrow \hat{x}_i^t + \mathbf{h}_i^H\mathbf{r}/\|\mathbf{h}_i\|^2$$
$$\hat{x}_i^{t+1} \leftarrow \mathsf{F}\big(z_i^t, 1/(\gamma_t\|\mathbf{h}_i\|^2)\big)$$
$$\sigma_{x_i,t+1}^2 \leftarrow \mathsf{G}\big(z_i^t, 1/(\gamma_t\|\mathbf{h}_i\|^2)\big)$$

    or for LMMSE-VB, compute

$$z_i^t \leftarrow \hat{x}_i^t + \mathbf{h}_i^H\mathbf{W}_t\mathbf{r}/(\mathbf{h}_i^H\mathbf{W}_t\mathbf{h}_i)$$
$$\hat{x}_i^{t+1} \leftarrow \mathsf{F}\big(z_i^t, 1/(\mathbf{h}_i^H\mathbf{W}_t\mathbf{h}_i)\big)$$
$$\sigma_{x_i,t+1}^2 \leftarrow \mathsf{G}\big(z_i^t, 1/(\mathbf{h}_i^H\mathbf{W}_t\mathbf{h}_i)\big)$$

    Update residual: $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{h}_i(\hat{x}_i^t - \hat{x}_i^{t+1})$
  end
end
MAP estimate:
$\hat{x}_i \leftarrow \arg\max_{a\in\mathcal{S}} p_a\mathcal{CN}\big(z_i^t; a, 1/(\gamma_t\|\mathbf{h}_i\|^2)\big)$ for MF-VB *or* $\arg\max_{a\in\mathcal{S}} p_a\mathcal{CN}\big(z_i^t; a, 1/(\mathbf{h}_i^H\mathbf{W}_t\mathbf{h}_i)\big)$ for LMMSE-VB.

We refer to this scheme as the ***LMMSE-VB algorithm*** since z_i resembles an LMMSE estimate of x_i due to the cancellation of the inter-user interference and the whitening of the postulated noise covariance matrix $\mathbf{C}^{\mathrm{post}}$.

Remark 4: *If an improper prior $\mathcal{CW}(0, \mathbf{0})$ is used, the variational mean $\langle\mathbf{W}\rangle$ in* (40) *cannot be computed due to the rank deficiency of $\big(\mathbf{y} - \mathbf{H}\langle\mathbf{x}\rangle\big)\big(\mathbf{y} - \mathbf{H}\langle\mathbf{x}\rangle\big)^H + \mathbf{H}\mathbf{\Sigma_x}\mathbf{H}^H$. In fact, it may not be possible to estimate the covariance matrix $\mathbf{C}^{\mathrm{post}} = \langle\mathbf{W}\rangle^{-1}$ with only one degree of freedom. To circumvent this issue, we propose to use the estimator*

$$\langle\mathbf{W}\rangle \approx \left(\frac{\big\|\mathbf{y} - \mathbf{H}\langle\mathbf{x}\rangle\big\|^2}{M}\mathbf{I}_M + \mathbf{H}\mathbf{\Sigma_x}\mathbf{H}^H\right)^{-1} \tag{41}$$

for the precision matrix $\mathbf{W}$. Similar to the AMP-based algorithms and MF-VB, the term $\big(\big\|\mathbf{y} - \mathbf{H}\langle\mathbf{x}\rangle\big\|^2/M\big)\mathbf{I}_M$ reflects the empirical estimate of the true noise variance N_0 (and also guarantees the existence of the inverse), whereas the term $\mathbf{H}\mathbf{\Sigma_x}\mathbf{H}^H$ reflects the empirical error covariance matrix induced by the MMSE denoiser $\mathsf{F}\big(z_i, 1/\big(\mathbf{h}_i^H\langle\mathbf{W}\rangle\mathbf{h}_i\big)\big)$. Although the convergence of LMMSE-VB using $\langle\mathbf{W}\rangle$ in (41) *is not analytically proved, all the simulations presented in Section VII indicate a robust and fast convergence as well as a remarkable performance.*

MF-VB and LMMSE-VB are summarized side by side in Algorithm 2. Here, we use $\hat{x}_i^t$ to replace the variational mean $\langle x_i\rangle$ at iteration t and each iteration consists of one round of updating $\{x_i\}$ and γ (or $\mathbf{W}$). To reduce the complexity of both algorithms, we also include the residual term $\mathbf{r}$, which is initialized as $\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^1$. After re-estimating x_i, i.e., $\hat{x}_i^t$ into $\hat{x}_i^{t+1}$, the residual $\mathbf{r}$ is updated as in step 9 to reflect the update of $\hat{x}_i^{t+1}$. This step allows the matrix multiplication $\mathbf{H}\hat{\mathbf{x}}^t$ to be bypassed in the linear estimator to obtain z_{i+1}^t.

Remark 5: *MF-VB and LMMSE-VB are analogous to MF-SIC and LMMSE-SIC, respectively, except for a key difference. The reciprocal of the noise variance (or covariance matrix) in the VB algorithms is estimated only once per iteration. This implementation significantly reduces the complexity, especially for LMMSE-VB, compared with their SIC counterparts. In addition, the convergence of MF-VB to at least a local optimal solution can be analytically proved due to the coordinate ascent approach of the algorithm.*

Remark 6: *MF-VB and LMMSE-VB are similar to AMP and OAMP/VAMP implemented with MF and LMMSE filters, respectively, in the linear estimation step. However, the VB algorithms do not compute nor remove the divergence term as do their AMP-based counterparts. Another difference between the two frameworks lies in the updating step. The VB algorithms use successive updates (Gauss-Seidel method), in which each z_i is computed based on the latest $\hat{\mathbf{x}}$ followed by the update of $\hat{x}_i$. On the other hand, the AMP-based schemes allow parallel updates (Jacobi method) of $\{z_i\}$ based on $\hat{\mathbf{x}}$ from the previous iteration followed by the update of $\{\hat{x}_i\}$. Interestingly, by using the residual update in step 9 of Algorithm 2, the complexity of each iteration of the VB methods becomes comparable to that of the AMP-based algorithms.*

E. Computational Complexity Analysis

This section presents a comparative analysis on the computational complexity of LMMSE, AMP-based, SIC-based, and VB-based algorithms. It is assumed that $M \geq K$, as in the case of uplink MIMO. Except for the LMMSE detector, which has the complexity of $\mathcal{O}\big(MK^2 + |\mathcal{S}|K\big)$, the complexity of other algorithms is evaluated on a per-iteration basis. Their Big-$\mathcal{O}$ complexity analyses are summarized in Table I and elaborated in the following.

The AMP algorithm has the complexity of $\mathcal{O}\big(MK\big)$ in the linear estimation step and $\mathcal{O}\big(|\mathcal{S}|K\big)$ in the nonlinear denoiser for all K users. Due to the matrix inversion in the LMMSE filter, the OAMP/VAMP algorithm increases its total complexity to $\mathcal{O}\big(MK^2 + |\mathcal{S}|K\big)$. The MF-SIC algorithm requires the computation of the residual interference-plus-noise covariance matrix $\mathbf{C}_i$, which induces the complexity of $\mathcal{O}\big(M^2\big)$ with proper implementation. Thus, the complexity of MF-SIC for all K users is $\mathcal{O}\big(M^2K + |\mathcal{S}|K\big)$. On the other hand, the LMMSE-SIC algorithm demands the inversion of $\mathbf{C}_i$, which raises the complexity in the LMMSE estimation step to $\mathcal{O}\big(M^3\big)$ per user and the total complexity to $\mathcal{O}\big(M^3K + |\mathcal{S}|K\big)$. In the conv-VB algorithm, the computation of the variational distribution $q_i(x_i)$ using $g_i(x_i)$ in (18) or via the computation of z_i in (21) has the complexity of $\mathcal{O}\big(M^2 + |\mathcal{S}\big)$. However, by defining the residual term $\mathbf{r} = \mathbf{y} - \mathbf{H}\hat{\mathbf{x}}$ and properly updating $\mathbf{r}$, the complexity can be reduced to $\mathcal{O}\big(M + |\mathcal{S}|\big)$ per user, resulting in the total complexity of $\mathcal{O}\big(MK + |\mathcal{S}|K\big)$ for the conv-VB algorithm. Similarly, the MF-VB algorithm also has the complexity of

TABLE I
COMPUTATIONAL COMPLEXITY OF MIMO DETECTION METHODS.

Method	Complexity
LMMSE	$\mathcal{O}\big(MK^2 + \lvert\mathcal{S}\rvert K\big)$
AMP	$\mathcal{O}\big(MK + \lvert\mathcal{S}\rvert K\big)$
OAMP/VAMP	$\mathcal{O}\big(MK^2 + \lvert\mathcal{S}\rvert K\big)$
MF-SIC	$\mathcal{O}\big(M^2K + \lvert\mathcal{S}\rvert K\big)$
LMMSE-SIC	$\mathcal{O}\big(M^3K + \lvert\mathcal{S}\rvert K\big)$
conv-VB	$\mathcal{O}\big(MK + \lvert\mathcal{S}\rvert K\big)$
MF-VB	$\mathcal{O}\big(MK + \lvert\mathcal{S}\rvert K\big)$
LMMSE-VB	$\mathcal{O}\big(M^3 + \lvert\mathcal{S}\rvert K\big)$

$\mathcal{O}\big(MK + |\mathcal{S}|K\big)$ with the use and update of $\mathbf{r}$ in step 9 of its implementation. Note that the update of γ_t in step 6 requires the computation of $\mathrm{Tr}\{\mathbf{H}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}^H\} = \sum_{i=1}^{K} \|\mathbf{h}_i\|^2 \sigma^2_{x_{i,t}}$, which itself has the complexity of $\mathcal{O}\big(MK\big)$. In the implementation of the LMMSE-VB algorithm, the computation of $\mathbf{W}_t$ in step 6 induces the complexity of $\mathcal{O}\big(M^3\big)$, where as the computation in step 9 has the complexity of $\mathcal{O}\big(M^2 + |\mathcal{S}|\big)$ per user. Thus, the total complexity of the LMMSE-VB algorithm is $\mathcal{O}\big(M^3 + |\mathcal{S}|K\big)$.

VI. VB Inference for MIMO Detection with Imperfect CSIR

In this section, we develop a new VB method for MIMO detection in the presence of imperfect CSIR. We assume that there is a mismatch between the estimated channel, denoted by $\hat{\mathbf{H}}$, and the true channel $\mathbf{H}$.

A. Conventional MIMO Detection with Imperfect CSIR

We first examine a conventional approach in which the BS estimates the uplink channel during the pilot transmission phase and uses the estimated channel for data detection. Let $\mathbf{x}_{\mathrm{p},i} \in \mathbb{C}^{T_\mathrm{p}}$ be the pilot sequence transmitted by user i. The received signal during the pilot transmission phase over T_p time slots can be modeled as

$$\mathbf{Y}_\mathrm{p} = \mathbf{H}\mathbf{X}_\mathrm{p} + \mathbf{N}_\mathrm{p}, \tag{42}$$

where $\mathbf{X}_\mathrm{p} = [\mathbf{x}_{\mathrm{p},1}, \ldots, \mathbf{x}_{\mathrm{p},K}]^T \in \mathbb{C}^{K \times T_\mathrm{p}}$ is the pilot matrix and $\mathbf{N}_\mathrm{p}$ is additive Gaussian noise comprised of i.i.d. $\mathcal{CN}(0, N_0)$ random variables. Here, we assume that the pilot sequences from the K users are orthogonal to each other, i.e., $\mathbf{X}_\mathrm{p}\mathbf{X}_\mathrm{p}^H = P_\mathrm{p}T_\mathrm{p}\mathbf{I}_K$, where P_p is the transmit power during the pilot transmission phase. We first correlate the received signal with the associated pilot signal $\mathbf{x}_{\mathrm{p},i}$ from user i to obtain

$$\begin{aligned}
\mathbf{y}_{\mathrm{p},i} &= \frac{1}{\sqrt{P_\mathrm{p}T_\mathrm{p}}}\mathbf{Y}_\mathrm{p}\mathbf{x}^*_{\mathrm{p},i} \\
&= \frac{1}{\sqrt{P_\mathrm{p}T_\mathrm{p}}}\sum_{j=1}^{K}\mathbf{h}_j\mathbf{x}^T_{\mathrm{p},j}\mathbf{x}^*_{\mathrm{p},i} + \frac{1}{\sqrt{P_\mathrm{p}T_\mathrm{p}}}\mathbf{N}_\mathrm{p}\mathbf{x}^*_{\mathrm{p},i} \\
&= \sqrt{P_\mathrm{p}T_\mathrm{p}}\mathbf{h}_i + \mathbf{n}_{\mathrm{p},i},
\end{aligned} \tag{43}$$

where $\mathbf{n}_{\mathrm{p},i} = \big(1/\sqrt{P_\mathrm{p}T_\mathrm{p}}\big)\mathbf{N}_\mathrm{p}\mathbf{x}^*_{\mathrm{p},i} \sim \mathcal{CN}(0, N_0\mathbf{I}_M)$. The optimal MMSE estimate $\hat{\mathbf{h}}_i$ can be obtained as

$$\begin{aligned}
\hat{\mathbf{h}}_i &= \mathbb{E}[\mathbf{h}_i\mathbf{y}^H_{\mathrm{p},i}]\big(\mathbb{E}[\mathbf{y}_{\mathrm{p},i}\mathbf{y}^H_{\mathrm{p},i}]\big)^{-1}\mathbf{y}_{\mathrm{p},i} \\
&= \sqrt{P_\mathrm{p}T_\mathrm{p}}(P_\mathrm{p}T_\mathrm{p}\mathbf{I}_M + N_0\mathbf{R}_i^{-1})^{-1}\mathbf{y}_{\mathrm{p},i} \\
&= (P_\mathrm{p}T_\mathrm{p}\mathbf{I}_M + N_0\mathbf{R}_i^{-1})^{-1}\mathbf{Y}_\mathrm{p}\mathbf{x}^*_{\mathrm{p},i}.
\end{aligned} \tag{44}$$

The estimation errors $\mathbf{e}_i = \mathbf{h}_i - \hat{\mathbf{h}}_i, \forall i$ are independent and each is distributed as $\mathbf{e}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{K}_i)$, with

$$\mathbf{K}_i = (P_\mathrm{p}T_\mathrm{p}N_0^{-1}\mathbf{I}_M + \mathbf{R}_i^{-1})^{-1}. \tag{45}$$

The channel estimation mismatch can thus be modeled as

$$\mathbf{H} = \hat{\mathbf{H}} + \mathbf{E}, \tag{46}$$

where the channel estimation error $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_K]$ is independent of the estimated channel $\hat{\mathbf{H}}$. Hence, the system model (1) can be rewritten as

$$\mathbf{y} = \hat{\mathbf{H}}\mathbf{x} + \mathbf{E}\mathbf{x} + \mathbf{n}. \tag{47}$$

Conditioned on $\mathbf{x}$, the effective noise $\tilde{\mathbf{n}} = \mathbf{E}\mathbf{x} + \mathbf{n}$ is Gaussian with zero mean and covariance matrix

$$\mathbf{C}_{\tilde{\mathbf{n}}|\mathbf{x}} = \sum_{i=1}^{K}|x_i|^2\mathbf{K}_i + N_0\mathbf{I}_M. \tag{48}$$

Treating $p(\mathbf{y}|\mathbf{x}; \hat{\mathbf{H}}) = \mathcal{CN}(\mathbf{y}; \hat{\mathbf{H}}\mathbf{x}, \mathbf{C}_{\tilde{\mathbf{n}}|\mathbf{x}})$ as the likelihood function, one can apply the MAP (or ML) detector as mentioned in Section III to obtain an optimal estimate of $\mathbf{x}$.

Alternatively, an LMMSE detector can be used to estimate $\mathbf{x}$ with reduced complexity compared to the MAP detector. The LMMSE detector in (4) can be readily applied with a small adjustment by replacing the noise covariance matrix $N_0\mathbf{I}_M$ with the approximate covariance matrix of the effective noise $\mathbf{C}_{\tilde{\mathbf{n}}} = \mathbb{E}_{\mathbf{x}}\big[\mathbf{C}_{\tilde{\mathbf{n}}|\mathbf{x}}\big] = \sum_{i=1}^{K}\mathbb{E}\big[|x_i|^2\big]\mathbf{K}_i + N_0\mathbf{I}_M$. We note that conventional MIMO detection methods simply treat the channel estimation error as noise. Hence, we develop a novel VB scheme to jointly estimate the channel, the symbol vector, and the postulated noise variance.

B. Proposed MF-VB-M for MIMO Detection with Imperfect CSIR

Treating the channel $\mathbf{H}$ and the precision $\gamma = 1/N_0^{\mathrm{post}}$ as random variables, the joint distribution $p(\mathbf{y}, \mathbf{x}, \mathbf{H}, \gamma; \hat{\mathbf{H}}, \mathbf{K})$ can be factored as

$$p(\mathbf{y}, \mathbf{x}, \mathbf{H}, \gamma; \hat{\mathbf{H}}, \mathbf{K}) = p(\mathbf{y}|\mathbf{x}, \mathbf{H}, \gamma)p(\mathbf{H}; \hat{\mathbf{H}}, \mathbf{K})p(\mathbf{x})p(\gamma), \tag{49}$$

where $p(\mathbf{y}|\mathbf{x}, \mathbf{H}, \gamma) = \mathcal{CN}(\mathbf{y}; \mathbf{H}\mathbf{x}, \gamma^{-1}\mathbf{I}_M)$ and $p(\mathbf{H}; \hat{\mathbf{H}}, \mathbf{K}) = \prod_{i=1}^{K}\mathcal{CN}(\mathbf{h}_i; \hat{\mathbf{h}}_i, \mathbf{K}_i)$. Given the observation $\mathbf{y}$ and the estimated channel $\hat{\mathbf{H}}$, we aim at obtaining the mean-field variational distribution $q(\mathbf{x}, \mathbf{H}, \gamma)$ such that

$$\begin{aligned}
p(\mathbf{x}, \mathbf{H}, \gamma|\mathbf{y}; \hat{\mathbf{H}}, \mathbf{K}) &\approx q(\mathbf{x}, \mathbf{H}, \gamma) \\
&= \prod_{i=1}^{K}q_i(x_i)\prod_{i=1}^{K}q_i(\mathbf{h}_i)q(\gamma).
\end{aligned} \tag{50}$$

The optimization of $q(\mathbf{x}, \mathbf{H}, \gamma)$ is executed by iteratively updating $\{\mathbf{h}_i\}$, $\{x_i\}$, and γ as follows.

$$
\begin{aligned}
q_i(x_i) &\propto p(x_i)\exp\Big\{ -\langle\gamma\rangle\big(\big\|\langle\mathbf{h}_i\rangle\big\|^2+\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}\big)|x_i|^2+2\,\Re\Big\{\langle\mathbf{h}_i^H\rangle\Big(\mathbf{y}-\sum_{j\neq i}^{K}\langle\mathbf{h}_j\rangle\langle x_j\rangle\Big)x_i^*\Big\}\Big\}\\
&\propto p(x_i)\exp\Big\{ -\langle\gamma\rangle\big(\big\|\langle\mathbf{h}_i\rangle\big\|^2|x_i-z_i|^2+\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}|x_i|^2\big)\Big\}\\
&\propto p(x_i)\mathcal{CN}\Big(x_i;z_i,\frac{1}{\langle\gamma\rangle\big\|\langle\mathbf{h}_i\rangle\big\|^2}\Big)\mathcal{CN}\Big(x_i;0,\frac{1}{\langle\gamma\rangle\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}}\Big)\\
&\propto p(x_i)\mathcal{CN}\Big(x_i;\frac{z_i\big\|\langle\mathbf{h}_i\rangle\big\|^2}{\big\|\langle\mathbf{h}_i\rangle\big\|^2+\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}},\frac{1}{\langle\gamma\rangle\big(\big\|\langle\mathbf{h}_i\rangle\big\|^2+\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}\big)}\Big).
\end{aligned}
\tag{55}
$$

1) Updating $\mathbf{h}_i$. The variational distribution $q_i(\mathbf{h}_i)$ is obtained by expanding the conditional in (49) and taking the expectation with respect to all latent variables except $\mathbf{h}_i$ using the variational distribution $q(\mathbf{x})\prod_{j\neq i}^{K}q_j(\mathbf{h}_j)q(\gamma)$:

$$
\begin{aligned}
&q_i(\mathbf{h}_i)\\
&\propto \exp\big\{\big\langle \ln p(\mathbf{y}|\mathbf{x},\mathbf{H},\gamma)+\ln p(\mathbf{h}_i;\hat{\mathbf{h}}_i,\mathbf{K}_i)\big\rangle\big\}\\
&\propto \exp\big\{ -\big\langle\gamma\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big\rangle-(\mathbf{h}_i-\hat{\mathbf{h}}_i)^H\mathbf{K}_i^{-1}(\mathbf{h}_i-\hat{\mathbf{h}}_i)\big\}\\
&\propto \exp\Big\{ -\mathbf{h}_i^H\big[\langle\gamma\rangle\big\langle|x_i|^2\big\rangle\mathbf{I}_M+\mathbf{K}_i^{-1}\big]\mathbf{h}_i\\
&\quad+2\,\Re\Big\{\mathbf{h}_i^H\Big[\langle\gamma\rangle\Big(\mathbf{y}-\sum_{j\neq i}^{K}\langle\mathbf{h}_j\rangle\langle x_j\rangle\Big)\langle x_i^*\rangle+\mathbf{K}_i^{-1}\hat{\mathbf{h}}_i\Big]\Big\}\Big\}.
\end{aligned}
\tag{51}
$$

The variational distribution $q_i(\mathbf{h}_i)$ is thus Gaussian with mean and covariance matrix

$$
\langle\mathbf{h}_i\rangle=\boldsymbol{\Sigma}_i\Big(\langle\gamma\rangle\Big(\mathbf{y}-\sum_{j\neq i}^{K}\langle\mathbf{h}_j\rangle\langle x_j\rangle\Big)\langle x_i^*\rangle+\mathbf{K}_i^{-1}\hat{\mathbf{h}}_i\Big),
\tag{52}
$$

$$
\boldsymbol{\Sigma}_i=\big[\langle\gamma\rangle\big\langle|x_i|^2\big\rangle\mathbf{I}_M+\mathbf{K}_i^{-1}\big]^{-1},
\tag{53}
$$

respectively.

2) Updating x_i. The variational distribution $q_i(x_i)$ is obtained by expanding the conditional in (49) and taking the expectation with respect to all latent variables except x_i using the variational distribution $\prod_{j\neq i}^{K}q_j(x_j)q(\mathbf{H})q(\gamma)$:

$$
\begin{aligned}
q_i(x_i)&\propto \exp\big\{\big\langle \ln p(\mathbf{y}|\mathbf{x},\mathbf{H},\gamma)+\ln p(\mathbf{x})\big\rangle\big\}\\
&\propto p(x_i)\exp\big\{ -\big\langle\gamma\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big\rangle\big\}.
\end{aligned}
\tag{54}
$$

Note that (54) can be expanded as in (55) at the top of the next page, where z_i is a linear estimate of x_i that is now defined as

$$
\begin{aligned}
z_i&=\frac{\langle\mathbf{h}_i^H\rangle}{\big\|\langle\mathbf{h}_i\rangle\big\|^2}\Big(\mathbf{y}-\sum_{j\neq i}^{K}\langle\mathbf{h}_j\rangle\langle x_j\rangle\Big)\\
&=\langle x_i\rangle+\frac{\langle\mathbf{h}_i^H\rangle}{\big\|\langle\mathbf{h}_i\rangle\big\|^2}\big(\mathbf{y}-\langle\mathbf{H}\rangle\langle\mathbf{x}\rangle\big),
\end{aligned}
\tag{56}
$$

with $\big\langle\|\mathbf{h}_i\|^2\big\rangle=\big\|\langle\mathbf{h}_i\rangle\big\|^2+\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}$. Now, we define

$$
\tilde{z}_i=\frac{z_i\big\|\langle\mathbf{h}_i\rangle\big\|^2}{\big\|\langle\mathbf{h}_i\rangle\big\|^2+\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}},
\tag{57}
$$

$$
\tilde{\zeta}_i^2=\frac{1}{\langle\gamma\rangle\big(\big\|\langle\mathbf{h}_i\rangle\big\|^2+\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}\big)},
\tag{58}
$$

and obtain the variational distribution $q_i(x_i)$ as[3]

$$
q_i(x_i)\propto p(x_i)\mathcal{CN}(x_i;\tilde{z}_i,\tilde{\zeta}_i^2),
\tag{59}
$$

which can be easily normalized. The variational mean $\langle x_i\rangle$ and variance $\sigma_{x_i}^2$ are then computed accordingly.

3) Updating γ. The variational distribution $q(\gamma)$ is obtained by taking the expectation of the conditional in (49) with respect to $q(\mathbf{x})q(\mathbf{H})$:

$$
\begin{aligned}
q(\gamma)&\propto \exp\big\{\big\langle \ln p(\mathbf{y}|\mathbf{x},\mathbf{H},\gamma)+\ln p(\gamma)\big\rangle\big\}\\
&\propto \exp\big\{M\ln\gamma-\gamma\big\langle\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big\rangle\\
&\quad+(a_0-1)\ln\gamma-b_0\gamma\big\}.
\end{aligned}
\tag{60}
$$

The variational distribution $q(\gamma)$ is thus Gamma with mean

$$
\langle\gamma\rangle=\frac{a_0+M}{b_0+\big\langle\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big\rangle},
\tag{61}
$$

to which we apply Theorem 1 to obtain

$$
\begin{aligned}
\big\langle\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2\big\rangle&=\big\|\mathbf{y}-\langle\mathbf{H}\rangle\langle\mathbf{x}\rangle\big\|^2+\mathrm{Tr}\big\{\langle\mathbf{H}\rangle\boldsymbol{\Sigma}_{\mathbf{x}}\langle\mathbf{H}\rangle^H\big\}\\
&\quad+\sum_{i=1}^{K}\big\langle|x_i|^2\big\rangle\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}.
\end{aligned}
\tag{62}
$$

Similar to the AMP-based algorithms and MF-VB/LMMSE-VB, the term $\big\|\mathbf{y}-\langle\mathbf{H}\rangle\langle\mathbf{x}\rangle\big\|^2$ reflects the empirical estimate of the true noise variance N_0, whereas the term $\mathrm{Tr}\big\{\langle\mathbf{H}\rangle\boldsymbol{\Sigma}_{\mathbf{x}}\langle\mathbf{H}\rangle^H\big\}$ reflects the empirical error covariance matrix induced by the MMSE denoiser $\mathsf{F}\big(\tilde{z}_i,\tilde{\zeta}_i^2\big)$. In addition, the term $\sum_{i=1}^{K}\big\langle|x_i|^2\big\rangle\mathrm{Tr}\{\boldsymbol{\Sigma}_i\}$ reflects the empirical error covariance matrix induced by the channel estimator.

By iteratively optimizing $\big\{q_i(\mathbf{h}_i)\big\}$, $\big\{q_i(x_i)\big\}$, and $q(\gamma)$, we obtain the CAVI algorithm for estimating $\mathbf{H}$, $\mathbf{x}$, and the precision γ. We refer to this scheme as the ***MF-VB-M algorithm*** due to the use of the MF $\langle\mathbf{h}_i\rangle^H/\big\|\langle\mathbf{h}_i\rangle\big\|^2$ to obtain the linear estimate z_i in (56) with channel estimation mismatch. MF-VB-M is summarized in Algorithm 3. Here, we use $\hat{x}_i^t$ and $\check{\mathbf{h}}_i^t$ to replace the variational means $\langle x_i\rangle$ and $\langle\mathbf{h}_i\rangle$, respectively, at iteration t and each iteration consists of one round of updating $\{\mathbf{h}_i\}$, $\{x_i\}$, and γ.

[3]To obtain (59), we use the following property of the Gaussian distribution:

$$
\begin{aligned}
\mathcal{CN}(x;a,A)\mathcal{CN}(x;b,B)&=\mathcal{CN}\Big(x;\frac{a/A+b/B}{1/A+1/B},\frac{1}{1/A+1/B}\Big)\\
&\quad\times\mathcal{CN}(0;a-b,A+B)\\
&\propto\mathcal{CN}\Big(x;\frac{a/A+b/B}{1/A+1/B},\frac{1}{1/A+1/B}\Big).
\end{aligned}
$$

Algorithm 3: ***MF-VB-M algorithm*** with postulated noise variance and imperfect CSIR

Input: $\mathbf{y}$, $\hat{\mathbf{H}}$, $\{\mathbf{K}_i\}$, and prior distributions $\{p(x_i)\}$;
Output: $\hat{\mathbf{x}}$ and $\check{\mathbf{H}}$;
Initialize $\hat{x}_i^1 = 0$, $\sigma_{x_i,1}^2 = \mathrm{Var}_{p(x_i)}[x_i]$, and $\mathbf{\Sigma}_i = \mathbf{K}_i,\ \forall i$, and $\check{\mathbf{H}}^1 = \hat{\mathbf{H}}$;
for $t = 1, 2, \ldots$ **do**
  for $i = 1, 2, \ldots, K$ **do**
    Compute $\mathbf{\Sigma}_i$ as in (53) and $\check{\mathbf{h}}_i^t$ as in (52);
  end
  for $i = 1, 2, \ldots, K$ **do**
    Compute z_i^t, $\tilde{z}_i^t$, and $\tilde{\zeta}_{i,t}^2$ as in (56), (57), and (58), respectively;
    Compute and normalize $q_i(x_i)$ as in (59);
    Compute $\hat{x}_i^t$ and $\sigma_{x_i,t}^2$ with respect to $q_i(x_i)$;
  end
  Compute γ_t using (61)–(62);
end
MAP estimate: $\hat{x} \leftarrow \arg\max_{a \in \mathcal{S}} q_i(a)$.

Remark 7: *Algorithm 3 requires K matrix inversions per iteration to compute the variational distribution $q(\mathbf{H})$. However, if the channel matrix $\mathbf{H}$ is i.i.d. Gaussian, i.e., $\mathbf{R}_i = (1/M)\mathbf{I}_M,\ \forall i$, we obtain the statistics of the channel vector $\mathbf{h}_i$ during the pilot transmission phase as*

$$\hat{\mathbf{h}}_i = \frac{1}{P_{\mathrm{p}} T_{\mathrm{p}} + M N_0} \mathbf{Y}_{\mathrm{p}} \mathbf{x}_{\mathrm{p},i}^*, \tag{63}$$

$$\mathbf{K}_i = \frac{1}{P_{\mathrm{p}} T_{\mathrm{p}} N_0^{-1} + M} \mathbf{I}_M. \tag{64}$$

Algorithm 3 can then be executed without any matrix inversion. More specifically, the variational distribution $q(\mathbf{h}_i)$ is Gaussian with mean and covariance matrix

$$\langle \mathbf{h}_i \rangle = \frac{\langle \gamma \rangle \big(\mathbf{y} - \sum_{j \neq i}^K \langle \mathbf{h}_j \rangle \langle x_j \rangle \big) \langle x_i^* \rangle + (P_{\mathrm{p}} T_{\mathrm{p}} N_0^{-1} + M) \hat{\mathbf{h}}_i}{\langle \gamma \rangle \langle |x_i|^2 \rangle + P_{\mathrm{p}} T_{\mathrm{p}} N_0^{-1} + M}, \tag{65}$$

$$\mathbf{\Sigma}_i = \frac{1}{\langle \gamma \rangle \langle |x_i|^2 \rangle + P_{\mathrm{p}} T_{\mathrm{p}} N_0^{-1} + M} \mathbf{I}_M, \tag{66}$$

respectively.

Remark 8: *LMMSE-VB can also be developed for the joint estimation of $\mathbf{H}$, $\mathbf{x}$, and the precision matrix $\mathbf{W} = \big(\mathbf{C}^{\mathrm{post}}\big)^{-1}$. This would call for a few minor adjustments to MF-VB-M to accommodate the estimation of $\mathbf{W}$ in place of γ similar to those necessary to obtain LMMSE-VB from MF-VB. However, the computation of the variational distribution of $\mathbf{h}_i$, which is Gaussian with covariance matrix $\mathbf{\Sigma}_i = \big[\langle |x_i|^2 \rangle \langle \mathbf{W} \rangle + \mathbf{K}_i^{-1}\big]^{-1}$, requires the inversion of an $(M \times M)$-dimensional matrix, even for i.i.d. channels. Hence, the resulting algorithm would be much more computationally intensive than LMMSE-VB. In addition, we observe through numerical simulations that such an algorithm provides negligible performance gains with respect to MF-VB-M and we thus omit its derivations and discussion.*

VII. Simulation Results

This section presents numerical results comparing the SER performance of the AMP-based, SIC, and VB algorithms along with the LMMSE detector. The number of iterations is capped at 50 for each of these iterative algorithms. Unless otherwise stated, the covariance matrices $\{\mathbf{R}_i\}$ are normalized such that their diagonal elements are $1/M$, which implies $\mathbb{E}\big[\|\mathbf{h}_i\|^2\big] = 1,\ \forall i$. The noise variance N_0 is set according to the operating signal-to-noise ratio (SNR), which is defined as

$$\mathrm{SNR} = \frac{\mathbb{E}\big[\|\mathbf{H}\mathbf{x}\|^2\big]}{\mathbb{E}\big[\|\mathbf{n}\|^2\big]} = \frac{\sum_{i=1}^K \mathrm{Tr}\{\mathbf{R}_i\}}{M N_0} = \frac{K}{M N_0}. \tag{67}$$

A. Perfect CSIR with i.i.d. Gaussian Channels

We first examine the case where the channel matrix $\mathbf{H}$ consists of i.i.d. Gaussian coefficients (corresponding to i.i.d. Rayleigh fading) and is perfectly known at the BS.

Fig. 1 illustrates the SER performance for a case with $M = K = 32$ and QPSK signaling. At high SNR, MF-VB outperforms AMP and MF-SIC, whereas LMMSE-VB significantly outperforms OAMP/VAMP and LMMSE-SIC. In this relatively small MIMO system, the algorithms using the LMMSE filter in the linear estimation step significantly outperform their counterparts based on the MF. This gain comes at the expense of increased complexity, especially for LMMSE-SIC. It is noted that conv-VB performs very poorly, even worse than the LMMSE detector, at high SNR. The proposed MF-VB and LMMSE-VB have addressed this limitation.

Fig. 2 depicts the SER performance for a case with $M = K = 128$ and QPSK signaling. In this relatively large MIMO system, AMP, OAMP/VAMP, MF-SIC, MF-VB, and LMMSE-VB obtain similar SER results. Thus, in this case there is no benefit to use more computationally intensive schemes like OAMP/VAMP and LMMSE-VB. Due to the computational burden of LMMSE-SIC, we omit its simulation. However, LMMSE-SIC is expected to achieve a similar performance to the other algorithms (except conv-VB) since AMP and OAMP/VAMP are optimal in the large-system limit. In addition, as the residual inter-user interference becomes i.i.d. for large K and i.i.d. channels, MF-SIC and LMMSE-SIC are thus equivalent. In Figs. 1 and 2, the curve corresponding to AWGN channels is also plotted as a lower bound for i.i.d. Rayleigh fading channels with $\beta = 1$ and $M \to \infty$.

Fig. 3 plots the SER performance for a case with $M = K = 32$ and 16-QAM signaling. Compared with the results in Fig. 1, it is worth noting that a higher modulation scheme requires more than a simple increase in SNR to approach the SER performance as for AWGN channels. In this relatively small MIMO system, AMP, MF-SIC, and MF-VB saturate quite quickly and perform very poorly compared with OAMP/VAMP, LMMSE-SIC, and LMMSE-VB. Fig. 3 also indicates the superior performance of the proposed LMMSE-VB over OAMP/VAMP and LMMSE-SIC, achieving gains of up to 5 dB and 8 dB, respectively, at high SNR.

Fig. 4 displays the convergence behavior of the above algorithms for a case with $M = K = 64$, QPSK signaling, and an SNR of 12 dB. The convergence plots are obtained by averaging over 500 channel realizations. It is observed that the SIC algorithms converge faster than their AMP and VB counterparts. Specifically, LMMSE-SIC converges within

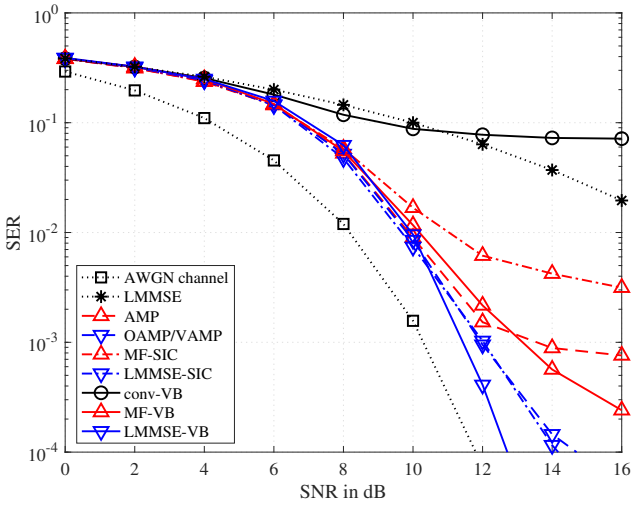

Fig. 1. SER performance of the LMMSE detector, AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming i.i.d. Rayleigh fading channels with $M = K = 32$ and QPSK signaling. LMMSE-VB achieves the lowest SER at high SNR.

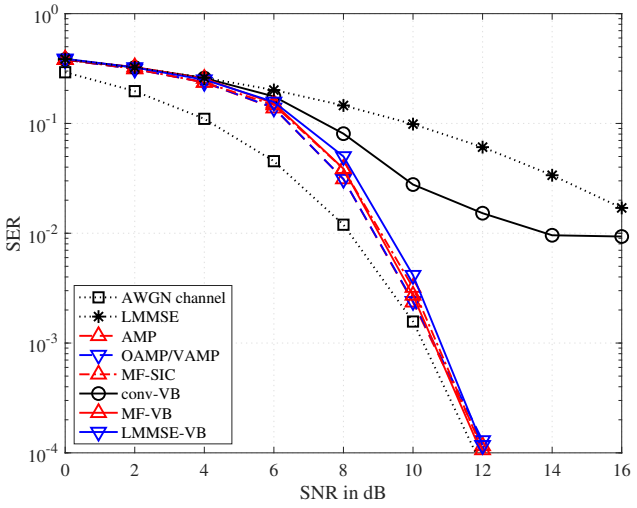

Fig. 2. SER performance of the LMMSE detector, AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming i.i.d. Rayleigh fading channels with $M = K = 32$ and QPSK signaling. All the algorithms achieve comparable SER, except for the LMMSE detector and conv-VB.

5 iterations, whereas OAMP/VAMP requires 10 iterations. However, the quick convergence of LMMSE-SIC comes at the cost of much higher complexity per iteration. Interestingly, MF-VB converges faster and to a lower SER than AMP. Although its convergence is not analytically proved, LMMSE-VB converges fairly quickly in all the considered simulation scenarios, as indicated in Fig. 4.

B. Perfect CSIR with Correlated Channels

We now study the case where the channel matrix $\mathbf{H}$ consists of correlated Gaussian coefficients (corresponding to correlated Rayleigh fading) and is perfectly known at the BS.

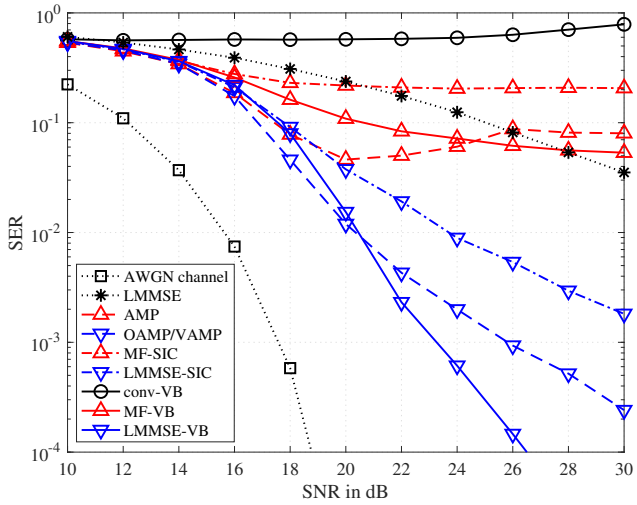

Fig. 3. SER performance of the LMMSE detector, AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming i.i.d. Rayleigh fading channels with $M = K = 32$ and 16-QAM signaling. Only the algorithms using the LMMSE filter in the linear estimation step achieve acceptable SER, and LMMSE-VB shows the lowest SER at high SNR.

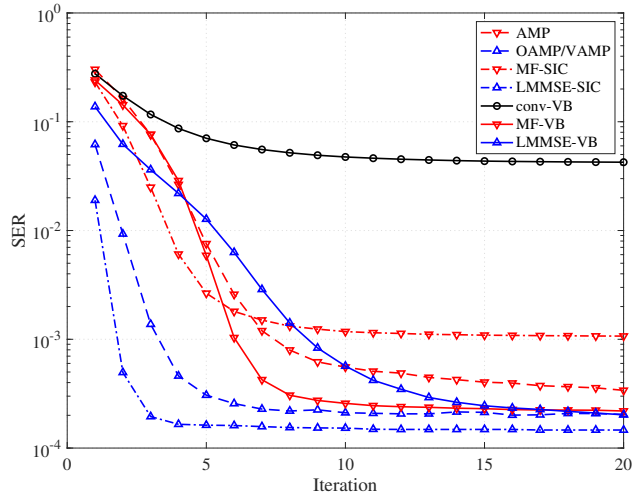

Fig. 4. Convergence of the AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming i.i.d. Rayleigh fading channels with $M = K = 64$, QPSK signaling, and SNR of 12 dB. All the algorithms exhibit very quick convergence (less than 20 iterations).

We first consider the exponential spatial correlation model [21] for each column of $\mathbf{H}$, in which each covariance matrix $\mathbf{R}_i$ is set to

$$[\mathbf{R}_i]_{k\ell} = \begin{cases} (1/M)\alpha^{k-\ell}, & \text{if } k \geq \ell \\ (1/M)(\alpha^{\ell-k})^*, & \text{if } k < \ell, \end{cases} \tag{68}$$

where α is the (complex) correlation coefficient between neighboring receive antennas. Fig. 5 presents the SER performance with $M = K = 64$, QPSK signaling, and using the exponential model with $\alpha = 0.5+j0.5$. It is observed that only the algorithms using the LMMSE filter in the linear estimation step achieve acceptable SER at high SNR. AMP, MF-SIC, VB, and MF-VB are even worse than the LMMSE detector as

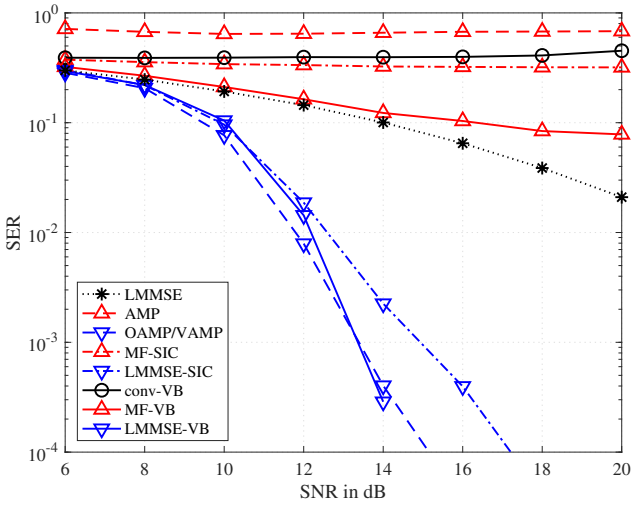

Fig. 5. SER performance of the LMMSE detector, AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming correlated Rayleigh fading channels based on the *exponential model*, $M = K = 64$, and QPSK signaling. Only the algorithms using the LMMSE filter in the linear estimation step achieve acceptable SER, and LMMSE-VB tends to achieve the lowest SER at high SNR.

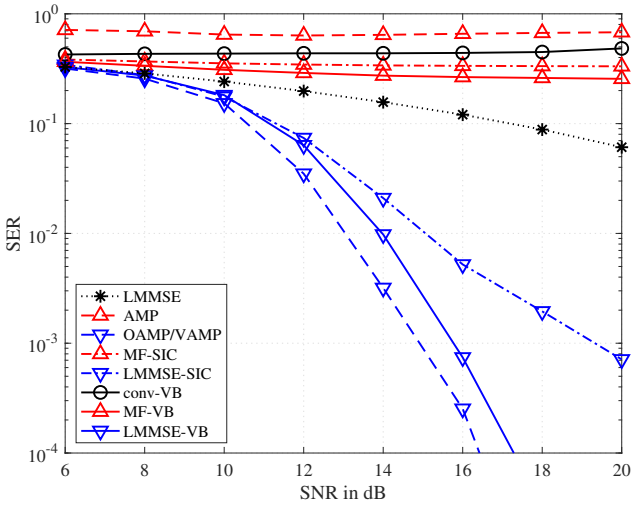

Fig. 6. SER performance of the LMMSE detector, AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming correlated Rayleigh fading channels based on the *one-ring model*, $M = K = 64$, and QPSK signaling. Only the algorithms using the LMMSE filter in the linear estimation step achieve acceptable SER, and the OAMP/VAMP algorithm outperforms LMMSE-VB.

they fail to account for the correlated MIMO channels in the linear estimation step. At very high SNR, LMMSE-VB tends to outperform OAMP/VAMP, and both achieve much lower SER than LMMSE-SIC.

As a further example, we examine the SER performance using the one-ring spatial correlation model [22]. This is characterized by a ring of scatterers around the users and no significant local scattering around the BS. In this context, the multipath components arrive at the BS with a small angular spread and the covariance matrices $\{\mathbf{R}_i\}$ tend to have low rank as M grows large [23]. Fig. 6 presents the SER performance for a case with $M = K = 64$, QPSK signaling, and using the one-ring model with a 15° angular spread. Similar to the results in Fig. 5, only OAMP/VAMP, LMMSE-SIC, and LMMSE-VB achieve acceptable SER at high SNR. However, OAMP/VAMP now outperforms LMMSE-VB by a small margin (i.e., < 1 dB).

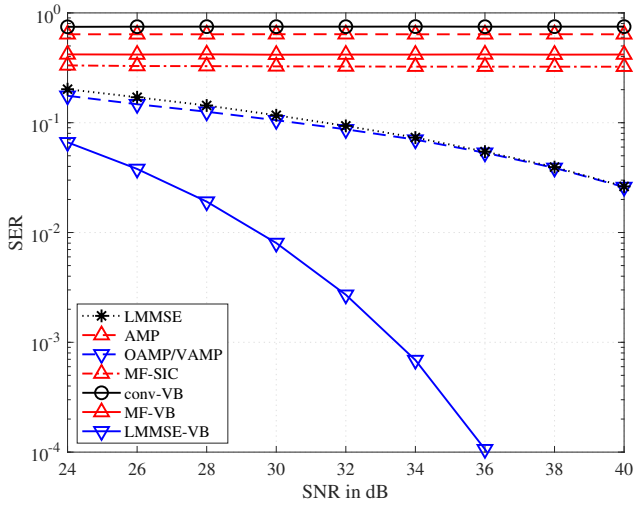

Fig. 7. SER performance of the LMMSE detector, AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming the *QuaDRiGa channel simulator* with $M = 128$, $K = 64$, and QPSK signaling. LMMSE-VB outperforms all the other algorithms by a large margin.

We next consider a realistic channel model used to mimic urban cellular deployments. In particular, we assume 3D MIMO channels generated by the 3GPP QuaDRiGa channel simulator [24]. We consider a BS equipped with a rectangular planar array with 64 dual-polarized antennas (i.e., $M = 128$) installed at a height of 25 m. The BS is assumed to cover a 120° cell sector of radius 500 m within which $K = 64$ users are uniformly distributed. We generate 200 channel realizations by creating 200 independent realizations of the user locations. Since the pathloss can vary dramatically between different users and we assume no power control ($\mathbb{E}[|x_i|^2] = 1,\ \forall i$), the operating SNRs can vary significantly from user to user. For each channel realization, we vary the noise variance N_0 at the BS accordingly to achieve an average operating SNR for all users, which is now defined as

$$\text{SNR} = \frac{\mathbb{E}\big[\|\mathbf{Hx}\|^2\big]}{\mathbb{E}\big[\|\mathbf{n}\|^2\big]} = \frac{\text{Tr}\{\mathbf{HH}^H\}}{MN_0}. \tag{69}$$

Fig. 7 illustrates the SER performance using the QuaDRiGa channel simulator described above, where LMMSE-SIC is omitted due to its prohibitive complexity with $M = 128$. It is observed that OAMP/VAMP performs only slightly better than the LMMSE detector, and both are significantly worse than LMMSE-VB. The large gap between OAMP/VAMP and LMMSE-VB in this non-homogeneous SNR setting is due to the difference in their methods of decoupling the MIMO channel. OAMP/VAMP decouples the MIMO channel

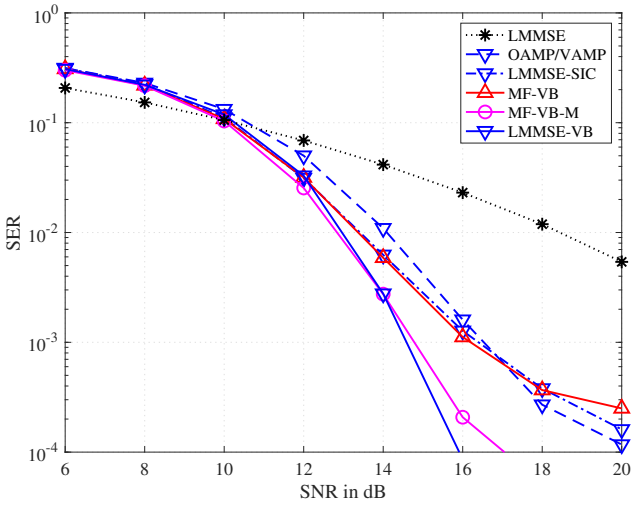

Fig. 8. SER performance of the LMMSE detector, AMP-based algorithms (in *dashed* lines), SIC algorithms (in *dashed-dotted* lines), and VB algorithms (in *solid* lines) assuming i.i.d. Rayleigh fading channels with imperfect CSIR, $M = K = 32$, and QPSK signaling. MF-VB-M reduces the SER with respect to MF-VB by nearly one order of magnitude at high SNR.

into K parallel Gaussian channels with the same SNR, i.e., $z_i = x_i + \mathcal{CN}(0, \sigma_t^2)$. On the contrary, LMMSE-VB decouples it into K parallel channels with possibly different SNRs, i.e., $z_i = x_i + \mathcal{CN}\big(0, 1/(\mathbf{h}_i^H \mathbf{W}_t \mathbf{h}_i)\big)$, enabling the consideration of user-specific channel conditions.

C. Imperfect CSIR

Lastly, we examine the i.i.d. Rayleigh fading case with imperfect CSIR and compare the SER performance of the proposed MF-VB-M with that of MF-VB and of the other algorithms using the LMMSE filter in the linear estimation step. We consider $M = K = 32$, QPSK signaling, pilot transmission time $T_{\rm p} = 32$, and pilot transmit power $P_{\rm p} = 1$. The estimated channel $\hat{\mathbf{H}}$ is obtained via the optimal MMSE channel estimator in (44). In all the algorithms except MF-VB-M, the estimated channel $\hat{\mathbf{H}}$ is treated as the true channel $\mathbf{H}$. The resulting SER performance is illustrated in Fig. 8. Compared with the results in Fig. 1, we see that LMMSE-VB outperforms OAMP/VAMP and LMMSE-SIC by a wider margin in this case with channel estimation mismatch. Furthermore, the SER of MF-VB is close to that of OAMP/VAMP and LMMSE-SIC. The improved performance of MF-VB and LMMSE-VB relative to OAMP/VAMP and LMMSE-SIC is due to the fact that the postulated noise variance/covariance matrix implicitly takes into account the channel estimation error. The proposed MF-VB-M algorithm for MIMO detection with imperfect CSIR is much better than MF-VB and performs similarly to LMMSE-VB at high SNR. This performance gain only requires a few additional simple computation steps to derive the variational distribution of $\mathbf{h}_i$, as detailed in ***Remark 7***.

VIII. Conclusion

This paper presented a study of massive MIMO detection from a variational Bayesian perspective. For the case of perfect CSIR, we developed the MF-VB and LMMSE-VB algorithms that use the noise variance and covariance matrix, respectively, postulated by the VB framework itself. These algorithms address the limitation in the conventional VB method with known noise variance and can approach and outperform their AMP-based and SIC counterparts in numerous channel settings. In addition, they involve closed-form and computationally efficient updates and exhibit very quick convergence. Finally, we proposed the MF-VB-M algorithm for the case of imperfect CSIR. Numerical results confirm the superior performance of the developed VB algorithms over the AMP-based and SIC schemes under various channel models. Future work may consider extensions to nonlinear, time-varying, and/or wideband MIMO channels.

Appendix A
Computation of $p(x_i|z_i,\sigma_t^2)$, $\mathsf{F}(z_i^t,\sigma_t^2)$, and $\mathsf{G}(z_i^t,\sigma_t^2)$

To compute the posterior mean and variance used in AMP and OAMP/VAMP, it is noted that the posterior distribution is given by

$$\begin{aligned} p(x_i|z_i^t;\sigma_t^2) &= \frac{1}{Z} p(z_i^t|x_i;\sigma_t^2) p(x_i) \\ &= \frac{1}{Z} \mathcal{CN}(z_i^t; x_i, \sigma_t^2) p(x_i), \end{aligned} \tag{70}$$

where Z is the normalization factor. In the context of MIMO detection, the posterior distribution is discrete with probability mass function given by

$$p(a|z_i^t;\sigma_t^2) = \frac{1}{Z} \exp\bigg(-\frac{|z_i^t - a|^2}{\sigma_t^2} \bigg) p_a, \tag{71}$$

with $Z = \sum_{b\in\mathcal{S}} \exp\big(-\frac{|z_i^t - b|^2}{\sigma_t^2} \big) p_b$. The corresponding posterior mean $\mathsf{F}(z_i^t,\sigma_t^2)$ and variance $\mathsf{G}(z_i^t,\sigma_t^2)$ can be computed accordingly. The final MAP estimate of x_i can be obtained as

$$\begin{aligned} \hat{x}_i &= \arg\max_{x_i\in\mathcal{S}} \; p(x_i|z_i^t;\sigma_t^2) \\ &= \arg\max_{a\in\mathcal{S}} \bigg(\ln p_a - \frac{|z_i^t - a|^2}{\sigma_t^2} \bigg). \end{aligned} \tag{72}$$

References

[1] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

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

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

[4] P. Alexander, M. Reed, J. Asenstorfer, and C. Schlegel, "Iterative multiuser interference reduction: Turbo CDMA," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1008–1014, Jul. 1999.

[5] W.-J. Choi, K.-W. Cheong, and J. Cioffi, "Iterative soft interference cancellation for multiple antenna systems," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Sep. 2000.

[6] N. Shlezinger, R. Fu, and Y. C. Eldar, "DeepSIC: Deep soft interference cancellation for multiuser MIMO detection," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1349–1362, Feb. 2021.

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

[8] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015.

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

[10] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *Ann. Appl. Probability*, vol. 25, no. 2, pp. 753–822, Apr. 2015.

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

[12] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, May 2019.

[13] 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.

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

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

[16] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2011.

[17] K. Takeuchi, "Bayes-optimal convolutional AMP," *IEEE Trans. Inf. Theory*, vol. 67, no. 7, pp. 4405–4428, May 2021.

[18] 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.

[19] F. Krzakala, A. Manoel, E. W. Tramel, and L. Zdeborová, "Variational free energies for compressed sensing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2014.

[20] W. Feller, *An Introduction to Probability Theorey and Its Applications*. New York, NY, USA: Wiley, 1968.

[21] S. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Letters*, vol. 5, no. 9, pp. 369–371, Sep. 2001.

[22] D.-S. Shiu, G. Foschini, M. Gans, and J. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.

[23] A. Adhikary, J. Nam, J. Y. Ahn, and G. Caire, "Joint spatial division and multiplexing – The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[24] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.