Multi-User Detection Based on Expectation Propagation for the Non-Coherent SIMO Multiple Access Channel

Khac-Hoang Ngo, *Graduate Student Member, IEEE*, Maxime Guillaud, *Senior Member, IEEE*, Alexis Decurninge, *Member, IEEE*, Sheng Yang, *Member, IEEE*, and Philip Schniter, *Fellow, IEEE*

***Abstract*— We consider the non-coherent single-input multiple-output (SIMO) multiple access channel with general signaling under spatially correlated Rayleigh block fading. We propose a novel soft-output multi-user detector that computes an approximate marginal posterior of each transmitted signal using only the knowledge about the channel distribution. Our detector is based on expectation propagation (EP) approximate inference and has polynomial complexity in the number of users, number of receive antennas and channel coherence time. We also propose two simplifications of this detector with reduced complexity. With Grassmannian signaling, the proposed detectors outperform a state-of-the-art non-coherent detector with projection-based interference mitigation. With pilot-assisted signaling, the EP detector outperforms, in terms of symbol error rate, some conventional coherent pilot-based detectors, including a sphere decoder and a joint channel estimation–data detection scheme. Our EP-based detectors produce accurate approximates of the true posterior leading to high achievable sum-rates. The gains of these detectors are further observed in terms of the bit error rate when using their soft outputs for a turbo channel decoder.**

***Index Terms*— Non-coherent communications, multiple access, detection, expectation propagation, Grassmannian constellations.**

I. Introduction

In WIRELESS communications, multi-antenna based multiple-input multiple-output (MIMO) technology is capable of improving significantly both the system spectral efficiency and reliability due to its multiplexing and diversity gains [2], [3]. MIMO is at the heart of current cellular systems, and large-scale (massive) MIMO [4] is considered as one of the fundamental technologies for the fifth-generation (5G) wireless communications [5]. In practical MIMO systems, the transmitted symbols are normally drawn from a finite discrete constellation to reduce complexity. Due to propagation effects, the symbols sent from different transmit antennas interfere, and the receiver observes a linear superposition of these symbols corrupted by noise. The task of the receiver is to detect these symbols (or rather the underlying bits) based on the received signal and the available knowledge about the channel.

If the *instantaneous* value of the channel matrix is treated as known, the detection is said to be *coherent* and has been investigated extensively in the literature [6]. In this case, the data symbols are normally taken from a scalar constellation such as the quadrature amplitude modulation (QAM). Since the optimal maximum-likelihood (ML) coherent detection problem is known to be non-deterministic polynomial-time hard (NP-hard) [7], many sub-optimal coherent MIMO detection algorithms have been proposed. These range from linear schemes, such as the zero forcing (ZF) and minimum mean square error (MMSE) detectors, to non-linear schemes based on, for example, interference cancellation, tree search, and lattice reduction [6].

If only *statistical* information about the channel is available, the detection problem is said to be *non-coherent*. In the block fading channel where the channel matrix remains constant for each length-T coherence block and varies between blocks, the receiver can estimate (normally imperfectly) the channel based on the transmitted pilot symbols, then perform coherent detection based on the channel estimate. Channel estimation and data detection can also be done iteratively [8], [9], or jointly based on tree search [10], [11]. These schemes requires pilot transmission for an initial channel estimate or to guarantee the identifiability of the data symbols. Another approach not involving pilot transmission is unitary space time modulation, in which the matrix of symbols in the space-time domain is orthonormal and isotropically distributed [12]. There, information is carried by the signal matrix subspace position, which is invariant to multiplication by the channel matrix. Therefore, a constellation over matrix-valued symbols can be designed as a collection of subspaces in $\mathbb{C}^T$. Such constellations belong to the Grassmann manifold $G(\mathbb{C}^T, K)$,

Manuscript received October 3, 2019; revised March 17, 2020 and May 28, 2020; accepted May 29, 2020. Date of publication June 12, 2020; date of current version September 10, 2020. This article was presented in part at the 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2019. The associate editor coordinating the review of this article and approving it for publication was M. Payaró. *(Corresponding author: Khac-Hoang Ngo.)*

Khac-Hoang Ngo is with the Mathematical and Algorithmic Sciences Laboratory, Paris Research Center, Huawei Technologies, 92100 Boulogne-Billancourt, France, and also with the Laboratory of Signals and Systems, CentraleSupélec, University of Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: ngo.khac.hoang@huawei.com).

Maxime Guillaud and Alexis Decurninge are with the Mathematical and Algorithmic Sciences Laboratory, Paris Research Center, Huawei Technologies, 92100 Boulogne-Billancourt, France (e-mail: maxime.guillaud@huawei.com; alexis.decurninge@huawei.com).

Sheng Yang is with the Laboratory of Signals and Systems, CentraleSupélec, University of Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: sheng.yang@centralesupelec.fr).

Philip Schniter is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: schniter.1@osu.edu).

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

Digital Object Identifier 10.1109/TWC.2020.3000419

which is the space of K-dimensional subspaces in $\mathbb{C}^T$, where K is the number of transmit antennas. For the independent and identically distributed (i.i.d.) Rayleigh block fading channel, when the signal-to-noise-ratio (SNR) is large, Grassmannian signaling was shown to achieve a rate within a vanishing gap from the capacity if $T \geq N + \min\{K, N\}$ [13], and within a constant gap if $2K \leq T \leq N + K$ [14], where N is the number of receive antennas. Like with coherent detection, the optimal ML non-coherent detection problem under Grassmannian signaling is NP-hard. Thus, low-complexity sub-optimal detectors have been proposed for constellations with additional structure, e.g., [15]–[17].

In this paper, we focus on the non-coherent detection problem in the Rayleigh flat and block fading single-input multiple-output (SIMO) multiple-access channel (MAC) with coherence time T. There, the communication signals are independently transmitted from K single-antenna users. If the users could cooperate, the high-SNR optimal joint signaling scheme would be a Grassmannian signaling on $G(\mathbb{C}^T, K)$ [13]. However, we assume uncoordinated users, for which the optimal non-coherent transmission scheme is not known, although some approximate optimality design criteria have been proposed in [18]. In this work, we design the detector without assuming any specific structure of the signal transmitted over a coherence block. We consider the case where the receiver is interested not only in the hard detection of the symbols but also in their posterior marginal probability mass functions (PMFs). This "soft" information is needed, for example, when computing the bit-wise log-likelihood ratios (LLRs) required for soft-input soft-output channel decoding. Computing an exact marginal PMF would require enumerating all possible combinations of other-user signals, which is infeasible with many users, many antennas, or large constellations. Thus, we seek sub-optimal schemes with practical complexity.

In contrast to probabilistic coherent MIMO detection, for which many schemes have been proposed (e.g., [19]–[21]), the probabilistic non-coherent MIMO detection under general signaling, and Grassmannian signaling in particular, has not been well investigated. The detection scheme proposed in [22] is sub-optimal and compatible only with the specific constellation structure considered therein. The list-based soft demapper in [23] reduces the number of terms considered in posterior marginalization by including only those symbols at a certain distance from a reference point. However, it was designed for the single-user case only and has no obvious generalization to the MAC. The semi-blind approaches [8]–[11] for the MIMO point-to-point channel can be extended to the MAC. However, these schemes are restricted to transmitted signals with pilots.

In this work, we propose message-passing algorithms for posterior marginal inference of non-coherent multi-user MIMO transmissions over spatially correlated Rayleigh block fading channels. Our algorithms are based on expectation propagation (EP) approximate inference [24], [25]. EP provides an iterative framework for approximating posterior beliefs by parametric distributions in the exponential family [26, Sec. 1.6]. Although there are many possible ways to apply EP to our non-coherent multi-user detection problem, we do so by choosing as variable nodes the indices of the transmitted symbols and the noiseless received signal from each user. The EP algorithm passes messages between the corresponding variable nodes and factor nodes on a bipartite factor graph. In doing so, the approximate posteriors of these variables are iteratively refined. We also address numerical implementation issues.

To measure the accuracy of the approximate posterior generated by the soft detectors, we compute the mismatched sum-rate of the system that uses the approximate posterior as the decoding metric. This mismatched sum-rate approaches the achievable rate of the system as the approximate posterior gets close to the true posterior. We also evaluate the symbol error rate when using the proposed schemes for hard detection, and the bit error rate when using these schemes for turbo equalization with a standard turbo code.

The contributions of this work are summarized as follows:

1) We propose soft and hard multi-user detectors for the non-coherent SIMO MAC using EP approximate inference, and methods to stabilize the EP updates. The proposed detectors work for general vector-valued transmitted symbols within each channel coherence block, i.e., it is general enough to include both the pilot-assisted and pilot-free signaling cases.
2) We propose two simplifications of the EP detector with reduced complexity. The first one, so-called EPAK, is based on approximating the EP messages with Kronecker products. The second one can be interpreted as soft MMSE estimation and successive interference approximation (SIA).
3) We analyze the complexity and numerically evaluate the convergence, running time, and performance of the proposed EP, EPAK, and MMSE-SIA detectors, the optimal ML detector, a genie-aided detector, the state-of-the-art detector from [22], and some conventional coherent pilot-based schemes. Our results suggest that the proposed detectors offer significantly improved mismatched sum-rate, symbol error rate, and coded bit error rate with respect to (w.r.t.) some existing sub-optimal schemes, while having lower complexity than the ML detector.

To the best of our knowledge, our proposed approach is the first message-passing scheme for non-coherent multi-user MIMO detection with general constellations.

The remainder of this paper is organized as follows. The system model is presented in Section II. A brief review of EP is presented in Section III, and the EP approach to non-coherent detection is presented in Section IV. In Section V, two simplifications (MMSE-SIA and EPAK) of the EP detector are presented. Implementation aspects of EP, MMSE-SIA, and EPAK are discussed in Section VI. Numerical results and conclusions are presented in Section VII and Section VIII, respectively. The mathematical preliminaries and proofs are provided in the appendices.

Notations: Random quantities are denoted with non-italic letters with sans-serif fonts, e.g., a scalar x, a vector $\mathbf{v}$, and a matrix $\mathbf{M}$. Deterministic quantities are denoted with italic letters, e.g., a scalar x, a vector $\boldsymbol{v}$, and a matrix $\boldsymbol{M}$. The Euclidean norm is denoted by $\|\boldsymbol{v}\|$ and the Frobenius

norm $\|\boldsymbol{M}\|_F$. The conjugate, transpose, conjugate transpose, trace, and vectorization of $\boldsymbol{M}$ are denoted by $\boldsymbol{M}^*$, $\boldsymbol{M}^{\mathsf{T}}$, $\boldsymbol{M}^{\mathsf{H}}$, $\mathrm{tr}\{\boldsymbol{M}\}$, and $\mathrm{vec}(\boldsymbol{M})$, respectively. $\prod$ denotes the conventional or Cartesian product, depending on the factors. $\otimes$ denotes the Kronecker product. $\mathbb{1}\{A\}$ denotes the indicator function whose value is 1 if A is true and 0 otherwise. $[n] := \{1, 2, \ldots, n\}$. $\propto$ means "proportional to". The Grassmann manifold $G(\mathbb{C}^T, K)$ is the space of K-dimensional subspaces in $\mathbb{C}^T$. In particular, $G(\mathbb{C}^T, 1)$ is the Grassmannian of lines. The Kullback-Leibler divergence of a distribution p from another distribution q of a random vector x with domain $\mathcal{X}$ is defined by $D(q\|p) := \int_{\mathcal{X}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x}$ if $\mathcal{X}$ is continuous and $D(q\|p) := \sum_{\boldsymbol{x}\in\mathcal{X}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$ if $\mathcal{X}$ is discrete. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the *complex* Gaussian vector distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and thus probability density function (PDF)

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{\exp\big(-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)}{\pi^n \det(\boldsymbol{\Sigma})}, \quad \boldsymbol{x} \in \mathbb{C}^n.$$

II. System Model

A. Channel Model

We consider a SIMO MAC in which K single-antenna users transmit to an N-antenna receiver. We assume that the channel is flat and block fading with an equal-length and synchronous (across the users) coherence interval of T channel uses. That is, the channel vectors $\mathsf{h}_k \in \mathbb{C}^{N\times 1}$, which contain the fading coefficients between the transmit antenna of user $k \in [K]$ and the N receive antennas, remain constant within each coherence block of T channel uses and change independently between blocks. Furthermore, the *distribution* of h_k is assumed to be known to the receiver, but its *realizations* are unknown to both ends of the channel. Since the users are not co-located, we assume that the h_k are independent across users. We consider Rayleigh fading with receiver-side correlation, i.e., $\mathsf{h}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Xi}_k)$, where $\boldsymbol{\Xi}_k \in \mathbb{C}^{N\times N}$ is the spatial correlation matrix. We assume that $\frac{1}{N}\mathrm{tr}\{\boldsymbol{\Xi}_k\} =: \xi_k$ where ξ_k is the large-scale average channel gain from one of the receive antennas to user k. We assume that $T > K$ and $N \geq K$.

Within a coherence block, each transmitter k sends a signal vector $\mathsf{s}_k \in \mathbb{C}^T$, and the receiver receives a realization $\boldsymbol{Y}$ of the random matrix

$$\mathsf{Y} = \sum_{k=1}^{K} \mathsf{s}_k \mathsf{h}_k^{\mathsf{T}} + \mathsf{W} = \mathsf{S}\mathsf{H}^{\mathsf{T}} + \mathsf{W}, \tag{1}$$

where $\mathsf{S} = [\mathsf{s}_1 \ \ldots \ \mathsf{s}_K] \in \mathbb{C}^{T\times K}$ and $\mathsf{H} = [\mathsf{h}_1 \ \ldots \ \mathsf{h}_K] \in \mathbb{C}^{N\times K}$ concatenate the transmitted signals and channel vectors, respectively, $\mathsf{W} \in \mathbb{C}^{T\times N}$ is the Gaussian noise with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries independent of H, and the block index is omitted for simplicity.

We assume that the transmitted signals have average unit norm, i.e., $\mathbb{E}\left[\|\mathsf{s}_k\|^2\right] = 1, k \in [K]$. Under this normalization, the signal-to-noise ratio (SNR) of the transmitted signal from user k at each receive antenna is $\mathrm{SNR}_k = \xi_k/(T\sigma^2)$. We assume that the transmitted signals belong to *disjoint* finite discrete individual constellations with vector-valued symbols. That is, $\mathsf{s}_k \in \mathcal{S}_k := \{\boldsymbol{s}_k^{(1)}, \ldots, \boldsymbol{s}_k^{(|\mathcal{S}_k|)}\},\ k \in [K]$. In particular, $\mathcal{S}_k$ can be a Grassmannian constellation on $G(\mathbb{C}^T, 1)$, i.e., each constellation symbol $\boldsymbol{s}_k^{(i)}$ is a unit-norm vector representative of a point in $G(\mathbb{C}^T, 1)$. Another example is when the constellation symbols contain pilots and scalar data symbols.[1] Each symbol in $\mathcal{S}_k$ is labeled with a binary sequence of length $B_k := \log_2 |\mathcal{S}_k|$.

B. Multi-User Detection Problem

Given $\mathsf{S} = \boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K]$, the conditional probability density $p_{\mathsf{Y}|\mathsf{S}}$, also known as likelihood function, is derived similar to [27, Eq.(9)] as

$$p_{\mathsf{Y}|\mathsf{S}}(\boldsymbol{Y}|\boldsymbol{S}) = \frac{\exp\big(-\mathrm{vec}(\boldsymbol{Y}^{\mathsf{T}})^{\mathsf{H}}\big(\sigma^2\boldsymbol{I}_{NT} + \sum_{k=1}^{K}\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}} \otimes \boldsymbol{\Xi}_k\big)^{-1}\mathrm{vec}(\boldsymbol{Y}^{\mathsf{T}})\big)}{\pi^{NT}\det(\sigma^2\boldsymbol{I}_{NT} + \sum_{k=1}^{K}\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}} \otimes \boldsymbol{\Xi}_k)}.$$

Given the received signal $\mathsf{Y} = \boldsymbol{Y}$, the joint multi-user ML symbol decoder is then

$$\hat{\boldsymbol{S}} = \arg\min_{\boldsymbol{S}\in\prod_{k=1}^{K}\mathcal{S}_k} \Bigg(\mathrm{vec}(\boldsymbol{Y}^{\mathsf{T}})^{\mathsf{H}}\Big(\sigma^2\boldsymbol{I}_{NT} + \sum_{k=1}^{K}\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}} \otimes \boldsymbol{\Xi}_k\Big)^{-1} \times \mathrm{vec}(\boldsymbol{Y}^{\mathsf{T}}) + \log\det\Big(\sigma^2\boldsymbol{I}_{NT} + \sum_{k=1}^{K}\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}} \otimes \boldsymbol{\Xi}_k\Big)\Bigg). \tag{2}$$

Since the ML decoding metric depends on $\boldsymbol{S}$ only through $\sum_{k=1}^{K}\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}} \otimes \boldsymbol{\Xi}_k$, for identifiability, it must hold that $\sum_{k=1}^{K}\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}} \otimes \boldsymbol{\Xi}_k \neq \sum_{k=1}^{K}\boldsymbol{s}'_k\boldsymbol{s}'^{\mathsf{H}}_k \otimes \boldsymbol{\Xi}_k$ for any pair of distinct joint symbols $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K]$ and $\boldsymbol{S}' = [\boldsymbol{s}'_1, \ldots, \boldsymbol{s}'_K]$ in $\prod_{k=1}^{K}\mathcal{S}_k$.

When a channel code is used, most channel decoders require the LLRs of the bits. The LLR of the j-th bit of user k, denoted by $\mathsf{b}_{k,j}$, given the observation $\mathsf{Y} = \boldsymbol{Y}$ is defined as

$$\begin{aligned}\mathrm{LLR}_{k,j}(\boldsymbol{Y}) &:= \log\frac{p_{\mathsf{Y}|\mathsf{b}_{k,j}}(\boldsymbol{Y}|1)}{p_{\mathsf{Y}|\mathsf{b}_{k,j}}(\boldsymbol{Y}|0)} \\ &= \log\frac{\sum_{\boldsymbol{\alpha}\in\mathcal{S}_{k,j}^{(1)}} p_{\mathsf{Y}|\mathsf{s}_k}(\boldsymbol{Y}|\boldsymbol{\alpha})}{\sum_{\boldsymbol{\beta}\in\mathcal{S}_{k,j}^{(0)}} p_{\mathsf{Y}|\mathsf{s}_k}(\boldsymbol{Y}|\boldsymbol{\beta})} \\ &= \log\frac{\sum_{\boldsymbol{\alpha}\in\mathcal{S}_{k,j}^{(1)}} p_{\mathsf{s}_k|\mathsf{Y}}(\boldsymbol{\alpha}|\boldsymbol{Y})}{\sum_{\boldsymbol{\beta}\in\mathcal{S}_{k,j}^{(0)}} p_{\mathsf{s}_k|\mathsf{Y}}(\boldsymbol{\beta}|\boldsymbol{Y})}\end{aligned} \tag{3}$$

where $\mathcal{S}_{k,j}^{(b)}$ denotes the set of all possible symbols in $\mathcal{S}_k$ with the j-th bit being equal to b for $j \in [B_k]$ and $b \in \{0, 1\}$. To compute (3), the posteriors $p_{\mathsf{s}_k|\mathsf{Y}}$, $k \in [K]$, are marginalized from

$$p_{\mathsf{S}|\mathsf{Y}}(\boldsymbol{S}|\boldsymbol{Y}) = \frac{p_{\mathsf{Y}|\mathsf{S}}(\boldsymbol{Y}|\boldsymbol{S})p_{\mathsf{S}}(\boldsymbol{S})}{p_{\mathsf{Y}}(\boldsymbol{Y})} \propto p_{\mathsf{Y}|\mathsf{S}}(\boldsymbol{Y}|\boldsymbol{S})p_{\mathsf{S}}(\boldsymbol{S}).$$

Assuming that the transmitted signals are independent and uniformly distributed over the respective constellations, the prior p_{S} factorizes as $\Pr(\mathsf{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K]) = \prod_{k=1}^{K}\frac{1}{|\mathcal{S}_k|}\mathbb{1}\{\boldsymbol{s}_k \in \mathcal{S}_k\}$. On the other hand, the likelihood

[1] In this case, the constellations are disjoint thanks to the fact that pilot sequences are user-specific.

function $p_{\mathbf{Y}|\mathbf{S}}(\boldsymbol{Y}|[\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K])$ involves all $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K$ in such a manner that it does not straightforwardly factorize. Exact marginalization of $p_{\mathbf{S}|\mathbf{Y}}$ requires computing

$$p_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k|\boldsymbol{Y}) = \sum_{\boldsymbol{s}_l\in\mathcal{S}_l,\forall l\neq k} p_{\mathbf{S}|\mathbf{Y}}([\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K]|\boldsymbol{Y}), \quad k\in[K]. \tag{4}$$

That is, it requires computing $p_{\mathbf{Y}|\mathbf{S}}(\boldsymbol{Y}|\boldsymbol{S})$ (which requires the inversion of an $NT\times NT$ matrix) for all $\boldsymbol{S}\in\prod_{k=1}^K\mathcal{S}_k$. Thus, the total complexity of exact marginalization is $O(K^6 2^{KB})$.[2] This is formidable for many users or large constellations. Thus, we seek alternative approaches to estimate

$$\begin{aligned} p_{\mathbf{S}|\mathbf{Y}}([\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K]|\boldsymbol{Y}) &\approx \hat{p}_{\mathbf{S}|\mathbf{Y}}([\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K]|\boldsymbol{Y}) \\ &= \prod_{k=1}^K \hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k|\boldsymbol{Y}). \end{aligned} \tag{5}$$

C. Achievable Rate

According to [28, Sec. II], the highest sum-rate reliably achievable with a given decoding metric $\hat{p}_{\mathbf{S}|\mathbf{Y}}$, so-called the mismatched sum-rate, is lower bounded by the generalized mutual information (GMI) given by

$$\begin{aligned} &R_{\mathrm{GMI}} \\ &= \frac{1}{T}\sup_{s\geq 0}\mathbb{E}\left[\log_2\frac{\hat{p}_{\mathbf{S}|\mathbf{Y}}(\mathbf{S}|\mathbf{Y})^s}{\sum_{\boldsymbol{S}'\in\prod_{k=1}^K\mathcal{S}_k}\Pr(\mathbf{S}=\boldsymbol{S}')\hat{p}_{\mathbf{S}|\mathbf{Y}}(\boldsymbol{S}'|\mathbf{Y})^s}\right] \\ &= \frac{1}{T}\sup_{s\geq 0}\mathbb{E}\left[\sum_{k=1}^K B_k - \log_2\frac{\sum_{\boldsymbol{S}'\in\prod_{k=1}^K\mathcal{S}_k}\hat{p}_{\mathbf{S}|\mathbf{Y}}(\boldsymbol{S}'|\boldsymbol{Y})^s}{\hat{p}_{\mathbf{S}|\mathbf{Y}}(\mathbf{S}|\mathbf{Y})^s}\right] \quad (6) \\ &= \frac{1}{T}\sum_{k=1}^K B_k - \frac{1}{T}\inf_{s\geq 0}\mathbb{E}\left[\sum_{k=1}^K\log_2\frac{\sum_{\boldsymbol{s}'_k\in\mathcal{S}_k}\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}'_k|\mathbf{Y})^s}{\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\mathbf{s}_k|\mathbf{Y})^s}\right] \quad (7) \end{aligned}$$

bits/channel use, where the expectation is over the joint distribution of $\mathbf{S}$ and $\mathbf{Y}$, i.e., $p_{\mathbf{Y}|\mathbf{S}}p_{\mathbf{S}}$, (6) holds because the transmitted symbols are independent and have uniform prior distribution, and (7) follows from the factorization of $\hat{p}_{\mathbf{S}|\mathbf{Y}}$ in (5). The generalized mutual information R_{GMI} is upper bounded by the sum-rate achieved with the optimal decoding metric $p_{\mathbf{S}|\mathbf{Y}}$ given by

$$\begin{aligned} R &= \frac{1}{T}I(\mathbf{S};\mathbf{Y}) \\ &= \frac{1}{T}h(\mathbf{S}) - \frac{1}{T}h(\mathbf{S}|\mathbf{Y}) \\ &= \frac{1}{T}\sum_{k=1}^K B_k - \frac{1}{T}\mathbb{E}\left[\log_2\frac{1}{p_{\mathbf{S}|\mathbf{Y}}(\mathbf{S}|\mathbf{Y})}\right] \\ &= \frac{1}{T}\sum_{k=1}^K B_k - \frac{1}{T}\mathbb{E}\left[\log_2\frac{\sum_{\boldsymbol{S}'\in\prod_{k=1}^K\mathcal{S}_k}p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\boldsymbol{S}')}{p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S})}\right] \end{aligned} \tag{8}$$

bits/channel use, where (8) follows from the Bayes' law and the uniformity of the prior distribution. R_{GMI} approaches R as $\hat{p}_{\mathbf{S}|\mathbf{Y}}$ gets close to $p_{\mathbf{S}|\mathbf{Y}}$. Note that if we fix $s=1$ in place of the infimum in (7), it holds that

$$\begin{aligned} R - R_{\mathrm{GMI}}(s=1) &= \frac{1}{T}\mathbb{E}\left[\log_2\frac{p_{\mathbf{S}|\mathbf{Y}}(\mathbf{S}|\mathbf{Y})}{\hat{p}_{\mathbf{S}|\mathbf{Y}}(\mathbf{S}|\mathbf{Y})}\right] \\ &= \frac{1}{T}\mathbb{E}_{\mathbf{Y}}\big[D(p_{\mathbf{S}|\mathbf{Y}}\|\hat{p}_{\mathbf{S}|\mathbf{Y}})\big], \end{aligned}$$

which converges to zero when the KL divergence between $\hat{p}_{\mathbf{S}|\mathbf{Y}}$ and $p_{\mathbf{S}|\mathbf{Y}}$ vanishes.

The expectations in (7) and (8) cannot be derived in closed form in general. Alternatively, we can evaluate R and R_{GMI} (and also $\mathbb{E}_{\mathbf{Y}}[D(p_{\mathbf{S}|\mathbf{Y}}\big\|\hat{p}_{\mathbf{S}|\mathbf{Y}})]$) numerically with the Monte Carlo method. Note that when K or B_k is large, even a numerical evaluation of R and $\mathbb{E}_{\mathbf{Y}}[D(p_{\mathbf{S}|\mathbf{Y}}\big\|\hat{p}_{\mathbf{S}|\mathbf{Y}})]$ is not possible. Therefore, we choose to use the mismatched sum-rate lower bound R_{GMI} as an information-theoretic metric to evaluate how close $\hat{p}_{\mathbf{S}|\mathbf{Y}}$ is to $p_{\mathbf{S}|\mathbf{Y}}$.

In what follows, we design a posterior marginal estimation scheme based on EP. We start by providing a brief review of EP in the next section.

III. Expectation Propagation

The EP algorithm was first proposed in [24] and summarized in, e.g., [25] for approximate inference in probabilistic graphical models. EP is an iterative framework for approximating posterior beliefs by parametric distributions in the exponential family [26, Sec. 1.6]. Let us consider a set of unknown variables represented by a random vector $\mathbf{x}$ with posterior of the form

$$p_{\mathbf{x}}(\boldsymbol{x}) \propto \prod_{\alpha}\psi_\alpha(\boldsymbol{x}_\alpha), \tag{9}$$

where $\boldsymbol{x}_\alpha$ is the subset of variables involved in the factor ψ_α corresponding to a partition $\{\mathbf{x}_\alpha\}$ of $\mathbf{x}$. Furthermore, let us partition the components of $\mathbf{x}$ into some sets $\{\mathbf{x}_\beta\}$, where no $\mathbf{x}_\beta$ is split across factors (i.e., $\forall\alpha,\beta$ either $\mathbf{x}_\beta\subset\mathbf{x}_\alpha$ or $\mathbf{x}_\beta\cap\mathbf{x}_\alpha=\emptyset$). The partition $\{\mathbf{x}_\alpha\}$ represents the local dependency of the variables given by the intrinsic factorization (9), while the partition $\{\mathbf{x}_\beta\}$ groups the variables that always occur together in a factor. We are interested in the posterior marginals w.r.t. the partition $\{\mathbf{x}_\beta\}$. In the following, we omit $\mathbf{x}$ in the subscripts since it is obvious.

EP approximates the true posterior p from (9) by a distribution $\hat{p}$ that can be expressed in two ways. First, it can be expressed w.r.t. the "target" partition $\{\mathbf{x}_\beta\}$ as

$$\hat{p}(\boldsymbol{x}) = \prod_{\beta}\hat{p}_\beta(\boldsymbol{x}_\beta), \tag{10}$$

where $\hat{p}_\beta$ are constrained to be in the exponential family [26, Sec. 1.6], such that (s.t.)

$$\hat{p}_\beta(\boldsymbol{x}_\beta) = \exp\big(\boldsymbol{\gamma}_\beta^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta) - A_\beta(\boldsymbol{\gamma}_\beta)\big), \tag{11}$$

for sufficient statistics $\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)$, parameters $\boldsymbol{\gamma}_\beta$, and log-partition function $A_\beta(\boldsymbol{\gamma}) := \ln\int e^{\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)}\,\mathrm{d}\boldsymbol{x}_\beta$. Second, $\hat{p}$ can also be expressed w.r.t. the partition $\{\mathbf{x}_\alpha\}$ as

$$\hat{p}(\boldsymbol{x}) \propto \prod_{\alpha} m_\alpha(\boldsymbol{x}_\alpha), \tag{12}$$

[2] Throughout the paper, as far as the complexity analysis is concerned, we assume for notational simplicity that $T=O(K)$, $N=O(K)$, and $|\mathcal{S}_k|=O(2^B)$, $\forall k\in[K]$. If the channels are uncorrelated ($\boldsymbol{\Xi}_k=\boldsymbol{I}_N$), the likelihood function can be simplified as $p_{\mathbf{Y}|\mathbf{S}}(\boldsymbol{Y}|\boldsymbol{S}) = \frac{\exp\left(-\mathrm{tr}\left\{\boldsymbol{Y}^{\mathsf{H}}(\sigma^2\boldsymbol{I}_T+\boldsymbol{S}\boldsymbol{S}^{\mathsf{H}})^{-1}\boldsymbol{Y}\right\}\right)}{\pi^{NT}\det^N(\sigma^2\boldsymbol{I}_T+\boldsymbol{S}\boldsymbol{S}^{\mathsf{H}})}$. Thus, the complexity of exact marginalization is reduced to $O(K^3 2^{KB})$.

in accordance with (9). For (10) and (12) to be consistent, the terms m_α should also factorize over β, i.e., there exist factors $m_{\alpha,\beta}$ of the form $m_{\alpha,\beta}(\boldsymbol{x}_\beta) = \exp\left(\boldsymbol{\gamma}_{\alpha,\beta}^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)\right)$ s.t.

$$m_\alpha(\boldsymbol{x}_\alpha) = \prod_{\beta\in\mathfrak{N}_\alpha} m_{\alpha,\beta}(\boldsymbol{x}_\beta) = \exp\Bigg(\sum_{\beta\in\mathfrak{N}_\alpha} \boldsymbol{\gamma}_{\alpha,\beta}^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)\Bigg),$$
$$\hat{p}_\beta(\boldsymbol{x}_\beta) \propto \prod_{\alpha\in\mathfrak{N}_\beta} m_{\alpha,\beta}(\boldsymbol{x}_\beta) = \exp\Bigg(\sum_{\alpha\in\mathfrak{N}_\beta} \boldsymbol{\gamma}_{\alpha,\beta}^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)\Bigg), \quad (13)$$

where $\mathfrak{N}_\alpha$ collects the indices β for which $\mathsf{x}_\beta \subset \mathsf{x}_\alpha$, and $\mathfrak{N}_\beta$ collects the indices α for which $\mathsf{x}_\beta \subset \mathsf{x}_\alpha$. It turns out that $m_{\alpha,\beta}$ can be interpreted as a message from the factor node α to the variable node β on a bipartite factor graph [29]. In this case, $\hat{p}_\beta(\boldsymbol{x}_\beta)$ is proportional to the product of all messages impinging on variable node β.

EP works by first initializing all $m_\alpha(\boldsymbol{x}_\alpha)$ and $\hat{p}_\beta(\boldsymbol{x}_\beta)$ (typically by the respective priors, which are assumed to also belong to the considered exponential family), then iteratively updating each approximation factor m_α in turn. Let us fix a factor index α. According to [24], the "tilted" distribution q_α is constructed by swapping the true potential ψ_α for its approximate m_α in $\hat{p}(\boldsymbol{x})$ as $q_\alpha(\boldsymbol{x}) = \frac{\hat{p}(\boldsymbol{x})\psi_\alpha(\boldsymbol{x}_\alpha)}{m_\alpha(\boldsymbol{x}_\alpha)}$, where it is assumed that $\int q_\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$. This tilted distribution is projected back onto the exponential family by minimizing the KL divergence:

$$\hat{p}_\alpha^{\mathrm{new}}(\boldsymbol{x}) = \arg\min_{\underline{p}\in\mathcal{P}} D\big(q_\alpha(\boldsymbol{x})\big\|\underline{p}(\boldsymbol{x})\big), \quad (14)$$

where $\mathcal{P}$ is the set of distributions of the form of $\hat{p}$ in (10), i.e., $\underline{p}(\boldsymbol{x}) = \prod_\beta \underline{p}_\beta(\boldsymbol{x}_\beta) = \prod_\beta \exp\left(\underline{\boldsymbol{\gamma}}_\beta^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta) - A_\beta(\underline{\boldsymbol{\gamma}}_\beta)\right)$ for some $\{\underline{\boldsymbol{\gamma}}_\beta\}$. Following [24], the solution to (14) is as follows.

Proposition 1: The solution to (14) *is given by* $\hat{p}_\alpha^{\mathrm{new}}(\boldsymbol{x}) = \prod_\beta \hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta)$ *with* $\hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta) = \hat{p}_\beta(\boldsymbol{x}_\beta)$, $\forall\beta \notin \mathfrak{N}_\alpha$, *and* $\hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta) = \exp\left(\underline{\boldsymbol{\gamma}}_\beta^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta) - A_\beta(\underline{\boldsymbol{\gamma}}_\beta)\right)$ *with* $\underline{\boldsymbol{\gamma}}_\beta$ *s.t.* $\mathbb{E}_{\hat{p}_{\alpha,\beta}^{\mathrm{new}}}[\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)] = \mathbb{E}_{q_\alpha}[\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)]$, $\forall\beta \in \mathfrak{N}_\alpha$, *whenever the expectation* $\mathbb{E}_{q_\alpha}[\cdot]$ *exists.*

Proof: The proof is given in Appendix B. □

The factor m_α is then updated via

$$\begin{aligned} m_\alpha^{\mathrm{new}}(\boldsymbol{x}_\alpha) &= \frac{\hat{p}_\alpha^{\mathrm{new}}(\boldsymbol{x})m_\alpha(\boldsymbol{x}_\alpha)}{\hat{p}(\boldsymbol{x})} \quad (15)\\ &= \Bigg[\prod_{\beta\in\mathfrak{N}_\alpha} m_{\alpha,\beta}(\boldsymbol{x}_\beta)\Bigg]\frac{\prod_{\beta\in\mathfrak{N}_\alpha}\hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta)}{\prod_{\beta\in\mathfrak{N}_\alpha}\hat{p}_\beta(\boldsymbol{x}_\beta)}\\ &\propto \Bigg[\prod_{\beta\in\mathfrak{N}_\alpha} m_{\alpha,\beta}(\boldsymbol{x}_\beta)\Bigg]\\ &\quad\times\frac{\prod_{\beta\in\mathfrak{N}_\alpha}\hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta)}{\prod_{\beta\in\mathfrak{N}_\alpha}\left[m_{\alpha,\beta}(\boldsymbol{x}_\beta)\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)\right]}\\ &= \prod_{\beta\in\mathfrak{N}_\alpha} m_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta), \quad (16)\end{aligned}$$

with

$$m_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta) := \frac{\hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta)}{\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)}. \quad (17)$$

Note that, on the right-hand side (RHS) of (15), all terms dependent on $\{\boldsymbol{x}_\beta\}_{\beta\notin\mathfrak{N}_\alpha}$ cancel, leaving the dependence only on $\{\boldsymbol{x}_\beta\}_{\beta\in\mathfrak{N}_\alpha}$. Thus, the update of m_α only affects the approximate posterior of nodes β in the neighborhood of node α. After that, the process is repeated with the next α.

A message-passing view of Proposition 1 can be seen by expanding $q_\alpha(\boldsymbol{x})$ as

$$\begin{aligned} q_\alpha(\boldsymbol{x}) &= \frac{\psi_\alpha(\boldsymbol{x}_\alpha)}{m_\alpha(\boldsymbol{x}_\alpha)}\Bigg[\prod_{\beta\in\mathfrak{N}_\alpha}\prod_{\alpha'\in\mathfrak{N}_\beta} m_{\alpha',\beta}(\boldsymbol{x}_\beta)\Bigg]\Bigg[\prod_{\beta\notin\mathfrak{N}_\alpha}\hat{p}_\beta(\boldsymbol{x}_\beta)\Bigg]\\ &= \psi_\alpha(\boldsymbol{x}_\alpha)\Bigg[\prod_{\beta\in\mathfrak{N}_\alpha}\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)\Bigg]\Bigg[\prod_{\beta\notin\mathfrak{N}_\alpha}\hat{p}_\beta(\boldsymbol{x}_\beta)\Bigg],\end{aligned}$$

then, using the natural logarithm for the KL divergence, it follows that

$$\begin{aligned} &D\big(q_\alpha(\boldsymbol{x})\big\|\underline{p}(\boldsymbol{x})\big)\\ &= \int q_\alpha(\boldsymbol{x})\ln\frac{q_\alpha(\boldsymbol{x})}{\underline{p}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}\\ &= \int \psi_\alpha(\boldsymbol{x}_\alpha)\Bigg[\prod_{\beta\in\mathfrak{N}_\alpha}\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)\Bigg]\Bigg[\prod_{\beta\notin\mathfrak{N}_\alpha}\hat{p}_\beta(\boldsymbol{x}_\beta)\Bigg]\\ &\quad\times\ln\Bigg(\frac{\psi_\alpha(\boldsymbol{x}_\alpha)\prod_{\beta\in\mathfrak{N}_\alpha}\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)}{\prod_{\beta\in\mathfrak{N}_\alpha}\underline{p}_\beta(\boldsymbol{x}_\beta)}\\ &\quad\times\frac{\prod_{\beta\notin\mathfrak{N}_\alpha}\hat{p}_\beta(\boldsymbol{x}_\beta)}{\prod_{\beta\notin\mathfrak{N}_\alpha}\underline{p}_\beta(\boldsymbol{x}_\beta)}\Bigg)\,\mathrm{d}\boldsymbol{x}\\ &= \int \psi_\alpha(\boldsymbol{x}_\alpha)\Bigg[\prod_{\beta\in\mathfrak{N}_\alpha}\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)\Bigg]\\ &\quad\times\ln\frac{\psi_\alpha(\boldsymbol{x}_\alpha)\prod_{\beta\in\mathfrak{N}_\alpha}\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)}{\prod_{\beta\in\mathfrak{N}_\alpha}\underline{p}_\beta(\boldsymbol{x}_\beta)}\,\mathrm{d}\boldsymbol{x}_\alpha\\ &\quad+\sum_{\beta\notin\mathfrak{N}_\alpha}\int\hat{p}_\beta(\boldsymbol{x}_\beta)\ln\frac{\hat{p}_\beta(\boldsymbol{x}_\beta)}{\underline{p}_\beta(\boldsymbol{x}_\beta)}\,\mathrm{d}\boldsymbol{x}_\beta\\ &= \sum_{\beta\in\mathfrak{N}_\alpha}\int q_{\alpha,\beta}(\boldsymbol{x}_\beta)\ln\frac{q_{\alpha,\beta}(\boldsymbol{x}_\beta)}{\underline{p}_\beta(\boldsymbol{x}_\beta)}\,\mathrm{d}\boldsymbol{x}_\beta + \sum_{\beta\notin\mathfrak{N}_\alpha} D\big(\hat{p}_\beta\big\|\underline{p}_\beta\big) + c_0\\ &= \sum_{\beta\in\mathfrak{N}_\alpha} D\big(q_{\alpha,\beta}\big\|\underline{p}_\beta\big) + \sum_{\beta\notin\mathfrak{N}_\alpha} D\big(\hat{p}_\beta\big\|\underline{p}_\beta\big) + c_0, \quad (18)\end{aligned}$$

where

$$q_{\alpha,\beta}(\boldsymbol{x}_\beta) := \int\psi_\alpha(\boldsymbol{x}_\alpha)\Bigg[\prod_{\beta\in\mathfrak{N}_\alpha}\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)\Bigg]\,\mathrm{d}\boldsymbol{x}_{\alpha\setminus\beta} \quad (19)$$

and c_0 represents a constant w.r.t. the distribution $\underline{p}$ (which we optimize) whose value is irrelevant and may change at each occurrence. Equation (18) says that, for each β in the neighborhood of node α, the optimal $\underline{p}_\beta$ (i.e., $\hat{p}_{\alpha,\beta}^{\mathrm{new}}$) is uniquely identified as the moment match of $q_{\alpha,\beta}$ in the exponential family with sufficient statistics $\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)$, where $q_{\alpha,\beta}$ is formed by taking the product of the true factor ψ_α and all the messages impinging on that factor, and then integrating out all variables except $\boldsymbol{x}_\beta$. Furthermore, (17) says that the new message $m_{\alpha,\beta}^{\mathrm{new}}$ passed from α to $\beta\in\mathfrak{N}_\alpha$ equals $\hat{p}_{\alpha,\beta}^{\mathrm{new}}$ divided by the product of messages $\{m_{\alpha',\beta}\}_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha}$, i.e., previous messages to β from all directions except α. An illustrative example is shown in Fig. 1.

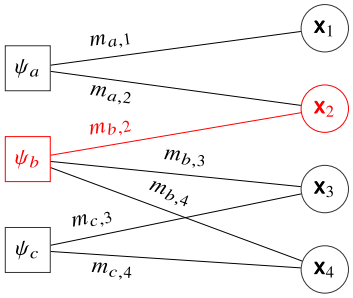


Fig. 1. An example of the factor graph representation of EP for $\alpha \in \{a, b, c\}$ and $\beta \in \{1, 2, 3, 4\}$. For $\alpha = b$ and $\beta = 2$, according to (19) and (17), $q_{b,2}(\mathbf{x}_2) = \int \psi_b(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) m_{a,2}(\mathbf{x}_2) m_{c,3}(\mathbf{x}_3) m_{c,4}(\mathbf{x}_4)\, \mathrm{d}\mathbf{x}_3\, \mathrm{d}\mathbf{x}_4$ and $m_{b,2}^{\text{new}}(\mathbf{x}_2) = \frac{\hat{p}_{b,2}^{\text{new}}(\mathbf{x}_2)}{m_{a,2}(\mathbf{x}_2)}$, respectively.

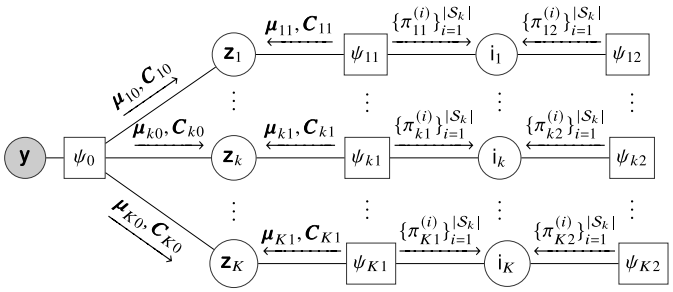


Fig. 2. A factor graph representation of the non-coherent detection problem. The messages are depicted with under-arrows showing their direction from a factor node to a variable node.

IV. Application of EP to Non-Coherent Detection

In order to apply EP to the non-coherent detection problem described in Section II, we express the transmitted signal as $\mathbf{s}_k = \boldsymbol{s}_k^{(\mathsf{i}_k)}$, where $\mathsf{i}_1, \ldots, \mathsf{i}_K$ are independent random indices.[3] With the assumption that the constellation symbols are transmitted with equal probability, i_k are uniformly distributed over $[|\mathcal{S}_k|]$, $k \in [K]$. We rewrite the received signal (1) in vector form as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{z}_k + \mathbf{w}, \tag{20}$$

where $\mathbf{y} := \mathrm{vec}(\mathbf{Y}^{\mathsf{T}})$, $\mathbf{z}_k := (\boldsymbol{s}_k^{(\mathsf{i}_k)} \otimes \boldsymbol{I}_N)\mathbf{h}_k$, and $\mathbf{w} := \mathrm{vec}(\mathbf{W}^{\mathsf{T}}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{NT})$. The problem of estimating $p_{\mathbf{s}_k|\mathbf{Y}}$ is equivalent to estimating $p_{\mathsf{i}_k|\mathbf{Y}}$ since they admit the same PMF.

With $\mathbf{z} := [\mathbf{z}_1^{\mathsf{T}}, \ldots, \mathbf{z}_K^{\mathsf{T}}]^{\mathsf{T}}$ and $\mathbf{i} := [\mathsf{i}_1, \ldots, \mathsf{i}_K]^{\mathsf{T}}$, we can write

$$\begin{aligned} &p_{\mathbf{i},\mathbf{z}|\mathbf{y}}(\boldsymbol{i}, \boldsymbol{z}|\boldsymbol{y}) \\ &\propto p_{\mathbf{i},\mathbf{z},\mathbf{y}}(\boldsymbol{i}, \boldsymbol{z}, \boldsymbol{y}) \\ &= p_{\mathbf{y}|\mathbf{z}}(\boldsymbol{y}|\boldsymbol{z}) p_{\mathbf{z}|\mathbf{i}}(\boldsymbol{z}|\boldsymbol{i}) p_{\mathbf{i}}(\boldsymbol{i}) \\ &= \psi_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K) \left[\prod_{k=1}^{K} \psi_{k1}(\boldsymbol{z}_k, i_k)\right] \left[\prod_{k=1}^{K} \psi_{k2}(i_k)\right], \end{aligned}$$

corresponding to (9), where

$$\begin{aligned} \psi_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K) &:= p_{\mathbf{y}|\mathbf{z}}(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}\Big(\boldsymbol{y}; \sum_{k=1}^{K} \boldsymbol{z}_k, \sigma^2 \boldsymbol{I}_{NT}\Big), \\ \psi_{k1}(\boldsymbol{z}_k, i_k) &:= p_{\mathbf{z}_k|\mathsf{i}_k}(\boldsymbol{z}_k) = \mathcal{N}\big(\boldsymbol{z}_k; \mathbf{0}, (\boldsymbol{s}_k^{(i_k)} \boldsymbol{s}_k^{(i_k)\mathsf{H}}) \otimes \boldsymbol{\Xi}_k\big), \\ \psi_{k2}(i_k) &:= p_{\mathsf{i}_k}(i_k) = \frac{1}{|\mathcal{S}_k|} \quad \text{for} \quad i_k \in [|\mathcal{S}_k|]. \end{aligned} \tag{21}$$

In the following, we consider a realization $\boldsymbol{y}$ of $\mathbf{y}$ and use EP to infer the posterior of the indices $\{\mathsf{i}_k\}$ and, as a by-product, the posterior of $\mathbf{z}_k$, $k \in [K]$. To do so, we choose the partition $\mathbf{x} = \{\mathbf{z}_k, \mathsf{i}_k\}_{k=1}^{K}$ and illustrate the interaction between these variables and the factors $\psi_0, \psi_{k1}, \psi_{k2}$ on the bipartite factor graph in Fig. 2. This graph is a tree with a root $\mathbf{y}$ and K leaves $\{\psi_{k2}\}_{k=1}^{K}$.

[3]The application of EP to non-coherent multi-user detection is non-trivial. Many choices can be made to model and partition the unknowns, but may not result in tractable derivation. Our choice is carefully made to enable closed-form message updates.

We write the EP approximation according to (10) as

$$\hat{p}_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \hat{p}_{\mathbf{i},\mathbf{z}|\mathbf{y}}(\boldsymbol{i}, \boldsymbol{z}|\boldsymbol{y}) = \prod_{k=1}^{K} \hat{p}_{\mathbf{z}_k}(\boldsymbol{z}_k) \hat{p}_{\mathsf{i}_k}(i_k), \tag{22}$$

where $\hat{p}_{\mathbf{z}_k}(\boldsymbol{z}_k)$ and $\hat{p}_{\mathsf{i}_k}(i_k)$ are implicitly conditioned on $\mathbf{y} = \boldsymbol{y}$ and constrained to be a Gaussian vector distribution and a discrete distribution with support $[|\mathcal{S}|]$ (both belong to the exponential family), respectively. Specifically, they are parameterized as

$$\hat{p}_{\mathbf{z}_k}(\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{z}_k; \hat{\boldsymbol{z}}_k, \boldsymbol{\Sigma}_k) \quad \text{s.t.} \quad \boldsymbol{\Sigma}_k \text{ is positive definite}, \tag{23}$$

$$\hat{p}_{\mathsf{i}_k}(i_k) = \hat{\pi}_k^{(i_k)} \quad \text{for} \quad i_k \in [|\mathcal{S}_k|] \quad \text{s.t.} \quad \sum_{i=1}^{|\mathcal{S}_k|} \hat{\pi}_k^{(i)} = 1. \tag{24}$$

We also write the EP approximation according to (12) as

$$\begin{aligned} \hat{p}_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) &= \hat{p}_{\mathbf{i},\mathbf{z}|\mathbf{y}}(\boldsymbol{i}, \boldsymbol{z}|\boldsymbol{y}) \\ &\propto m_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K) \left[\prod_{k=1}^{K} m_{k1}(\boldsymbol{z}_k, i_k)\right] \left[\prod_{k=1}^{K} m_{k2}(i_k)\right], \end{aligned}$$

where we define

$$\begin{aligned} m_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K) &\propto \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0}), \\ m_{k1}(\boldsymbol{z}_k, i_k) &\propto \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1}) \pi_{k1}^{(i_k)}, \\ m_{k2}(i_k) &= \pi_{k2}^{(i_k)} \quad \text{for} \quad i_k \in [|\mathcal{S}_k|]. \end{aligned}$$

On the factor graph in Fig. 2, we can interpret $(\boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})$ as the message from factor node ψ_0 to variable node $\mathbf{z}_k$, $(\boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})$ as the message from factor node ψ_{k1} to variable node $\mathbf{z}_k$, $\{\pi_{k1}^{(\mathsf{i}_k)}\}_{\mathsf{i}_k=1}^{|\mathcal{S}_k|}$ as the message from factor node ψ_{k1} to variable node i_k, and $\{\pi_{k2}^{(\mathsf{i}_k)}\}_{\mathsf{i}_k=1}^{|\mathcal{S}_k|}$ as the message from factor node ψ_{k2} to variable node i_k.

Remark 1: Our choice of Gaussian distribution (within the exponential family) in (23) *is motivated by the fact that when the noise and channel are Gaussian, the symbol posterior takes the form of a Gaussian mixture. It also allows a tractable derivation (using the Gaussian PDF multiplication rule) and closed-form update expressions, as will be shown in the next subsection. If a general (possibly non-Gaussian) channel model is considered, the factor* $\psi_{k1}(\boldsymbol{z}_k, i_k)$ *in* (21) *may be different, but the factor graph in Fig. 2 remains unchanged.*

A. The EP Message Updates

In the following, we derive the message updates from each of the factor nodes ψ_0, ψ_{k1}, and ψ_{k2}, $k \in [K]$, to the corresponding variable nodes. To do so, for each $\alpha \in \{k1, k2, 0\}$, we compute the projected density $\hat{p}_\alpha^{\text{new}} = \prod_{k=1}^K \hat{p}_{\alpha,\mathbf{z}_k}^{\text{new}}(\boldsymbol{z}_k)\hat{p}_{\alpha,\mathsf{i}_k}^{\text{new}}(i_k)$ according to (22) and Proposition 1, and then update the factor m_α according to (16).

1) Message $\{\pi_{k2}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ From Factor Node ψ_{k2} to Variable Node i_k: First, we compute $\hat{p}_{k2,\mathsf{i}_k}^{\text{new}}$ and then the EP message $\{\pi_{k2}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ from node ψ_{k2} to node i_k. From (18) and (24), we know that $\hat{p}_{k2,\mathsf{i}_k}^{\text{new}}$ is the discrete distribution with PMF $\{\hat{\pi}_{k2}^{(i)}\}_{i=1}^{|\mathcal{S}_k|}$ proportional to $\psi_{k2}(i_k)\pi_{k1}^{(i_k)}$, and so

$$\hat{\pi}_{k2}^{(i_k)} = \frac{\psi_{k2}(i_k)\pi_{k1}^{(i_k)}}{\sum_{i=1}^{|\mathcal{S}_k|}\psi_{k2}(i)\pi_{k1}^{(i)}} = \frac{\pi_{k1}^{(i_k)}}{\sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}} \quad \text{for } i_k \in [|\mathcal{S}_k|],$$

since $\psi_{k2}(i_k)$ is constant over these i_k. With $\hat{p}_{k2,\mathsf{i}_k}^{\text{new}}$ computed, (16) implies that the message from node ψ_{k2} to node i_k is the PMF proportional to

$$\frac{\hat{p}_{k2,\mathsf{i}_k}^{\text{new}}(i_k)}{\pi_{k1}^{(i_k)}} = \frac{\hat{\pi}_{k2}^{(i_k)}}{\pi_{k1}^{(i_k)}} = \frac{1}{\sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}} = c_0 \quad \text{for } i_k \in [|\mathcal{S}_k|],$$

and thus $\pi_{k2}^{(i_k)} = \frac{1}{|\mathcal{S}_k|}$ for $i_k \in [|\mathcal{S}_k|]$.

2) Messages From Factor Node ψ_{k1} to Variable Nodes $\mathbf{z}_k$ and i_k: Next, we compute $\hat{p}_{k1}^{\text{new}} = \prod_{k=1}^K \hat{p}_{k1,\mathbf{z}_k}^{\text{new}}(\boldsymbol{z}_k)\hat{p}_{k1,\mathsf{i}_k}^{\text{new}}(i_k)$ and the messages $\{\pi_{k1}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ and $(\boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})$ from node ψ_{k1} to nodes i_k and $\mathbf{z}_k$, respectively.

a) Message $\{\pi_{k1}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ from node ψ_{k1} to node i_k: We first compute $\hat{p}_{k1,\mathsf{i}_k}^{\text{new}}(i_k)$. From (18) and (24), we know that $\hat{p}_{k1,\mathsf{i}_k}^{\text{new}}(i_k)$ is the discrete distribution with support $[|\mathcal{S}_k|]$ and PMF $\hat{\pi}_{k1}^{(i_k)}$ proportional to

$$\begin{aligned}
&\int \psi_{k1}(\boldsymbol{z}_k, i_k)\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})\pi_{k2}^{(i_k)}\, \mathrm{d}\boldsymbol{z}_k \\
&= \frac{1}{|\mathcal{S}_k|}\int \mathcal{N}\big(\boldsymbol{z}_k; \mathbf{0}, (\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}) \otimes \Xi_k\big)\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})\, \mathrm{d}\boldsymbol{z}_k \\
&= \frac{1}{|\mathcal{S}_k|}\int \mathcal{N}\big(\boldsymbol{z}_k; \hat{\boldsymbol{z}}_{ki_k}, \boldsymbol{\Sigma}_{ki_k}\big) \\
&\quad \times \mathcal{N}\big(\mathbf{0}; \boldsymbol{\mu}_{k0}, (\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}) \otimes \Xi_k + \boldsymbol{C}_{k0}\big)\, \mathrm{d}\boldsymbol{z}_k \\
&= \frac{1}{|\mathcal{S}_k|}\mathcal{N}\big(\mathbf{0}; \boldsymbol{\mu}_{k0}, (\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}) \otimes \Xi_k + \boldsymbol{C}_{k0}\big),
\end{aligned}$$

where the second equality follows from the Gaussian PDF multiplication rule in Lemma 1 with

$$\begin{aligned}
\boldsymbol{\Sigma}_{ki} &= \big([(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k]^{-1} + \boldsymbol{C}_{k0}^{-1}\big)^{-1} \\
&= \big[(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k\big]\big((\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k + \boldsymbol{C}_{k0}\big)^{-1}\boldsymbol{C}_{k0}, \qquad (25)
\end{aligned}$$

$$\begin{aligned}
\hat{\boldsymbol{z}}_{ki} &= \boldsymbol{\Sigma}_{ki}\boldsymbol{C}_{k0}^{-1}\boldsymbol{\mu}_{k0} \\
&= \big[(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k\big]\big((\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k + \boldsymbol{C}_{k0}\big)^{-1}\boldsymbol{\mu}_{k0}. \qquad (26)
\end{aligned}$$

Thus

$$\hat{\pi}_{k1}^{(i_k)} = \frac{\mathcal{N}\big(\mathbf{0}; \boldsymbol{\mu}_{k0}, (\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}) \otimes \Xi_k + \boldsymbol{C}_{k0}\big)}{\sum_{i=1}^{|\mathcal{S}_k|}\mathcal{N}\big(\mathbf{0}; \boldsymbol{\mu}_{k0}, (\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k + \boldsymbol{C}_{k0}\big)}, \quad i_k \in [|\mathcal{S}_k|]. \qquad (27)$$

With $\hat{p}_{k1,\mathsf{i}_k}^{\text{new}}(i_k)$ computed, (16) implies that the message $\pi_{k1}^{(i_k)}$ from node ψ_{k1} to node i_k is the PMF proportional to $\frac{\hat{p}_{k1,\mathsf{i}_k}^{\text{new}}(i_k)}{\pi_{k2}^{(i_k)}} = |\mathcal{S}_k|\hat{\pi}_{k1}^{(i_k)}$ for $i_k \in [|\mathcal{S}_k|]$, and thus

$$\pi_{k1}^{(i_k)} = \frac{|\mathcal{S}_k|\hat{\pi}_{k1}^{(i_k)}}{\sum_{i=1}^{|\mathcal{S}_k|}|\mathcal{S}_k|\hat{\pi}_{k1}^{(i)}} = \hat{\pi}_{k1}^{(i_k)} \quad \text{for } i_k \in [|\mathcal{S}_k|]. \qquad (28)$$

b) Message $(\boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})$ from node ψ_{k1} to nodes $\mathbf{z}_k$: We next compute $\hat{p}_{k1,\mathbf{z}_k}^{\text{new}}(\boldsymbol{z}_k)$. From (18) and (23), we know that $\hat{p}_{k1,\mathbf{z}_k}^{\text{new}}(\boldsymbol{z}_k)$ is the Gaussian distribution with mean $\hat{\boldsymbol{z}}_k$ and covariance $\boldsymbol{\Sigma}_k$ matched to that of the PDF proportional to

$$\begin{aligned}
&\sum_{i_k=1}^{|\mathcal{S}_k|}\psi_{k1}(\boldsymbol{z}_k, i_k)\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})\pi_{k2}^{(i_k)} \\
&= \frac{1}{|\mathcal{S}_k|}\sum_{i=1}^{|\mathcal{S}_k|}\mathcal{N}\big(\boldsymbol{z}_k; \mathbf{0}, (\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k\big)\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0}) \\
&= \frac{1}{|\mathcal{S}_k|}\sum_{i=1}^{|\mathcal{S}_k|}\mathcal{N}\big(\boldsymbol{z}_k; \hat{\boldsymbol{z}}_{ki}, \boldsymbol{\Sigma}_{ki}\big)\mathcal{N}\big(\mathbf{0}; \boldsymbol{\mu}_{k0}, (\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \Xi_k + \boldsymbol{C}_{k0}\big) \\
&\propto \sum_{i=1}^{|\mathcal{S}_k|}\mathcal{N}\big(\boldsymbol{z}_k; \hat{\boldsymbol{z}}_{ki}, \boldsymbol{\Sigma}_{ki}\big)\hat{\pi}_{k1}^{(i)}, \qquad (29)
\end{aligned}$$

where the second equality follows from the Gaussian PDF multiplication rule in Lemma 1 with $\boldsymbol{\Sigma}_{ki}$ and $\hat{\boldsymbol{z}}_{ki}$ defined in (25) and (26), respectively. Thus, from (28), we have

$$\hat{\boldsymbol{z}}_k = \sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}\hat{\boldsymbol{z}}_{ki}, \qquad (30)$$

$$\boldsymbol{\Sigma}_k = \sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}(\hat{\boldsymbol{z}}_{ki}\hat{\boldsymbol{z}}_{ki}^{\mathsf{H}} + \boldsymbol{\Sigma}_{ki}) - \hat{\boldsymbol{z}}_k\hat{\boldsymbol{z}}_k^{\mathsf{H}}. \qquad (31)$$

With $\hat{p}_{k1,\mathbf{z}_k}^{\text{new}}(\boldsymbol{z}_k)$ computed, (16) implies that the message from node ψ_{k1} to node $\mathbf{z}_k$ is proportional to

$$\frac{\hat{p}_{k1,\mathbf{z}_k}^{\text{new}}(\boldsymbol{z}_k)}{\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})} = \frac{\mathcal{N}(\boldsymbol{z}_k; \hat{\boldsymbol{z}}_k, \boldsymbol{\Sigma}_k)}{\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})} \propto \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1}), \qquad (32)$$

with

$$\boldsymbol{C}_{k1} = \big(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{C}_{k0}^{-1}\big)^{-1}, \qquad (33)$$

$$\boldsymbol{\mu}_{k1} = \boldsymbol{C}_{k1}\big(\boldsymbol{\Sigma}_k^{-1}\hat{\boldsymbol{z}}_k - \boldsymbol{C}_{k0}^{-1}\boldsymbol{\mu}_{k0}\big). \qquad (34)$$

Equations (33) and (34) can be verified using $\mathcal{N}(\boldsymbol{z}_k; \hat{\boldsymbol{z}}_k, \boldsymbol{\Sigma}_k) \propto \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})$, which follows from (13) and the Gaussian PDF multiplication rule in Lemma 1.

3) Message $(\boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})$ From Node ψ_0 to Node $\mathbf{z}_k$: Finally, we compute $\hat{p}_{0,\mathbf{z}_k}^{\text{new}}$ and the EP message $(\boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})$ from node ψ_0 to node $\mathbf{z}_k$ for each $k \in [K]$. From (18) and (23), we know that $\hat{p}_{0,\mathbf{z}_k}^{\text{new}}$ is the Gaussian distribution with mean $\hat{\boldsymbol{z}}_{k0}$ and

covariance $\boldsymbol{\Sigma}_{k0}$ matched to that of the PDF proportional to

$$\begin{aligned}&\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})\int\psi_0(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_K)\Big[\prod_{j\neq k}\mathcal{N}(\boldsymbol{z}_j;\boldsymbol{\mu}_{j1},\boldsymbol{C}_{j1})\ \mathrm{d}\boldsymbol{z}_j\Big]\\&=\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})\\&\quad\times\int\mathcal{N}\Big(\boldsymbol{y};\boldsymbol{z}_k+\sum_{j\neq k}\boldsymbol{z}_j,\sigma^2\boldsymbol{I}_{NT}\Big)\Big[\prod_{j\neq k}\mathcal{N}(\boldsymbol{z}_j;\boldsymbol{\mu}_{j1},\boldsymbol{C}_{j1})\ \mathrm{d}\boldsymbol{z}_j\Big]\\&=\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})\mathcal{N}\Big(\boldsymbol{z}_k;\boldsymbol{y}-\sum_{j\neq k}\boldsymbol{\mu}_{j1},\sigma^2\boldsymbol{I}_{NT}+\sum_{j\neq k}\boldsymbol{C}_{j1}\Big),\end{aligned}\tag{35}$$

where (35) follows by applying repeatedly Lemma 1. Applying the Gaussian PDF multiplication rule to (35), we obtain

$$\boldsymbol{\Sigma}_{k0}=\Big(\boldsymbol{C}_{k1}^{-1}+\Big[\sigma^2\boldsymbol{I}_{NT}+\sum_{j\neq k}\boldsymbol{C}_{j1}\Big]^{-1}\Big)^{-1},\tag{36}$$

$$\hat{\boldsymbol{z}}_{k0}=\boldsymbol{\Sigma}_{k0}\Big(\boldsymbol{C}_{k1}^{-1}\boldsymbol{\mu}_{k1}+\Big[\sigma^2\boldsymbol{I}_{NT}+\sum_{j\neq k}\boldsymbol{C}_{j1}\Big]^{-1}\Big[\boldsymbol{y}-\sum_{j\neq k}\boldsymbol{\mu}_{j1}\Big]\Big).\tag{37}$$

Given $\hat{p}^{\mathrm{new}}_{0,\mathbf{z}_k}(\boldsymbol{z}_k)=\mathcal{N}(\boldsymbol{z}_k;\hat{\boldsymbol{z}}_{k0},\boldsymbol{\Sigma}_{k0})$, (16) implies that the message from node ψ_0 to node $\mathbf{z}_k$ is proportional to

$$\frac{\hat{p}^{\mathrm{new}}_{0,\mathbf{z}_k}(\boldsymbol{z}_k)}{\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})}=\frac{\mathcal{N}(\boldsymbol{z}_k;\hat{\boldsymbol{z}}_{k0},\boldsymbol{\Sigma}_{k0})}{\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})}\propto\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0}),$$

with $\boldsymbol{C}_{k0}=\big(\boldsymbol{\Sigma}_{k0}^{-1}-\boldsymbol{C}_{k1}^{-1}\big)^{-1}$ and $\boldsymbol{\mu}_{k0}=\boldsymbol{C}_{k0}\big(\boldsymbol{\Sigma}_{k0}^{-1}\boldsymbol{z}_{k0}-\boldsymbol{C}_{k1}^{-1}\boldsymbol{\mu}_{k1}\big)$. This is verified using $\mathcal{N}(\boldsymbol{z}_k;\hat{\boldsymbol{z}}_{k0},\boldsymbol{\Sigma}_{k0})\propto\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0})$, which follows from (13), and the Gaussian PDF multiplication rule in Lemma 1. Plugging in the expressions for $\boldsymbol{\Sigma}_{k0}^{-1}$ and $\hat{\boldsymbol{z}}_{k0}$ from (36) and (37) yields

$$\boldsymbol{C}_{k0}=\sigma^2\boldsymbol{I}_{NT}+\sum_{j\neq k}\boldsymbol{C}_{j1},\tag{38}$$

$$\boldsymbol{\mu}_{k0}=\boldsymbol{y}-\sum_{j\neq k}\boldsymbol{\mu}_{j1}.\tag{39}$$

This concludes the derivation of the EP message updates.

B. Initialization of the EP Messages

We initialize the EP messages as follows. First, we choose the non-informative initialization $\boldsymbol{C}_{k0}^{-1}=\mathbf{0}$ and $\boldsymbol{\mu}_{k0}=\mathbf{0}$, so that, from (27), the initial message from node ψ_{k1} to node i_k coincides with the uniform prior $\pi_{k1}^{(i_k)}=\hat{\pi}_{k1}^{(i_k)}=\frac{1}{|\mathcal{S}_k|}$ for $i_k\in[|\mathcal{S}_k|]$, and, from (25) and (26), the initial parameters $\boldsymbol{\Sigma}_{ki}=(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}})\otimes\boldsymbol{\Xi}_k$ and $\boldsymbol{z}_{ki}=\mathbf{0}$, respectively, for $k\in[K]$ and $i\in[|\mathcal{S}_k|]$. This leads to the initial parameters of $\hat{p}_k(\boldsymbol{z}_k)$ from (30) and (31) as $\hat{\boldsymbol{z}}_k=\mathbf{0}$ and $\boldsymbol{\Sigma}_k=\frac{1}{|\mathcal{S}_k|}\sum_{i=1}^{|\mathcal{S}_k|}(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}})\otimes\boldsymbol{\Xi}_k$, and the initial message from node ψ_{k1} to node $\mathbf{z}_k$ given in (33) and (34) as $\boldsymbol{C}_{k1}=\boldsymbol{\Sigma}_k=\frac{1}{|\mathcal{S}_k|}\sum_{i=1}^{|\mathcal{S}_k|}(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}})\otimes\boldsymbol{\Xi}_k$, and $\boldsymbol{\mu}_{k1}=\hat{\boldsymbol{z}}_k=\mathbf{0}$. Finally, the initial messages from node ψ_0 to node $\mathbf{z}_k$ follows from (38) and (39) as $\boldsymbol{C}_{k0}=\sigma^2\boldsymbol{I}_{NT}+\sum_{j\neq k}\frac{1}{|\mathcal{S}_j|}\sum_{i=1}^{|\mathcal{S}_j|}(\boldsymbol{s}_j^{(i)}\boldsymbol{s}_j^{(i)\mathsf{H}})\otimes\boldsymbol{\Xi}_k$, and $\boldsymbol{\mu}_{k0}=\boldsymbol{y}$.

C. The Algorithm

We summarize the proposed EP scheme for probabilistic non-coherent detection in Algorithm 1. In the end, according to (13) and (24), the estimated PMF $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k^{(i_k)}|\boldsymbol{Y})$ is given by $\hat{p}_k(i_k)=\hat{\pi}_k^{(i_k)}\propto\pi_{k1}^{(i_k)}\pi_{k2}^{(i_k)}$, that is $\hat{p}_k(i_k)=\pi_{k1}^{(i_k)}$ since $\pi_{k2}^{(i_k)}$ is constant. The algorithm goes through the branches of the tree graph in Fig. 2 in a round-robin manner. In each branch, the factor nodes are visited from leaf to root. We note that other message passing schedules can be implemented.

Algorithm 1: EP for Probabilistic Non-Coherent Detection

Input: the observation $\boldsymbol{Y}$; the constellations $\mathcal{S}_1,\ldots,\mathcal{S}_K$;
set the maximal number of iterations $t_{\max}$;
initialize of the messages $\{\pi_{k1}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|},\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1},\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0}$, for $k\in[K]$;
$t\longleftarrow 0$;
repeat
  $t\longleftarrow t+1$;
  for $k\leftarrow 1$ **to** K **do**
    update $\{\pi_{k1}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ according to (28) and (27) ;
    compute $\{\hat{\boldsymbol{z}}_{ki}\}_{i=1}^{|\mathcal{S}_k|}$ and $\{\boldsymbol{\Sigma}_{ki}\}_{i=1}^{|\mathcal{S}_k|}$ according to (26) and (25), respectively ;
    compute $\hat{\boldsymbol{z}}_k$ and $\boldsymbol{\Sigma}_k$ according to (30) and (31), respectively ;
    update $\boldsymbol{\mu}_{k1}$ and $\boldsymbol{C}_{k1}$ according to (34) and (33), respectively ;
    update $\{\boldsymbol{\mu}_{j0}\}_{j\neq k}$ and $\{\boldsymbol{C}_{j0}\}_{j\neq k}$ according to (39) and (38), respectively ;
  end
until convergence or $t=t_{\max}$;
return The PMF $\{\pi_{k1}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ of $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k^{(i_k)}|\boldsymbol{Y})$ for $k\in[K]$

In the EP algorithm, the dominant operation is the update of $\pi_{k1}^{(i_k)}$, $\boldsymbol{\Sigma}_{ki}$, and $\hat{\boldsymbol{z}}_{ki}$, which involves the inverse of the $NT\times NT$ matrix $(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}})\otimes\boldsymbol{\Xi}_k+\boldsymbol{C}_{k0}$ (with complexity $O(K^6)$) *for all* $k\in[K]$ and $i_k\in[|\mathcal{S}_k|]$. The complexity of computing $\hat{\boldsymbol{z}}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1},\{\boldsymbol{\mu}_{j0}\}_{j\neq k}$, and $\{\boldsymbol{C}_{j0}\}_{j\neq k}$ are all of lower order. Therefore, the complexity per iteration is given by $O(K^7 2^B)$. In order to reduce this complexity, we derive two simplifications of the EP scheme in the next section.

V. Simplifications of the EP Detector

In this section, we attempt to simplify EP by avoiding the inverse of $NT\times NT$ matrices.

A. EP With Approximate Kronecker Products (EPAK)

We observe that if $\boldsymbol{C}_{k0}$ could be expressed as a Kronecker product $\bar{\boldsymbol{C}}_{k0}\otimes\boldsymbol{\Xi}_k$ with $\bar{\boldsymbol{C}}_{k0}\in\mathbb{C}^{T\times T}$, we could rewrite $\pi_{k1}^{(i_k)}$ in (27) as

$$\pi_{k1}^{(i_k)}=\frac{\mathcal{N}\big(\mathbf{0};\boldsymbol{\mu}_{k0},(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}+\bar{\boldsymbol{C}}_{k0})\otimes\boldsymbol{\Xi}_k\big)}{\sum_{i=1}^{|\mathcal{S}_k|}\mathcal{N}\big(\mathbf{0};\boldsymbol{\mu}_{k0},(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}+\bar{\boldsymbol{C}}_{k0})\otimes\boldsymbol{\Xi}_k\big)}.\tag{40}$$

Let $\boldsymbol{M}_{k0} \in \mathbb{C}^{T\times N}$ s.t. $\boldsymbol{\mu}_{k0} = \mathrm{vec}\big(\boldsymbol{M}_{k0}^{\mathsf{T}}\big)$, (40) could be computed efficiently using

$$\begin{aligned}\mathcal{N}\Big(\mathbf{0};\boldsymbol{\mu}_{k0},\big(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}+\bar{\boldsymbol{C}}_{k0}\big)\otimes\boldsymbol{\Xi}_k\Big)\\ \propto\big(1+\boldsymbol{s}_k^{(i_k)\mathsf{H}}\bar{\boldsymbol{C}}_{k0}^{-1}\boldsymbol{s}_k^{(i_k)}\big)^{-N}\\ \times\exp\left(\frac{\mathrm{tr}\big\{\bar{\boldsymbol{C}}_{k0}^{-1}\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}\boldsymbol{M}_{k0}(\boldsymbol{\Xi}_k^{-1})^{\mathsf{T}}\boldsymbol{M}_{k0}^{\mathsf{H}}\big\}}{1+\boldsymbol{s}_k^{(i_k)\mathsf{H}}\bar{\boldsymbol{C}}_{k0}^{-1}\boldsymbol{s}_k^{(i_k)}}\right)\end{aligned}$$

since only the $T\times T$ matrix $\bar{\boldsymbol{C}}_{k0}$ needs to be inverted (the inverse of $\boldsymbol{\Xi}_k$ can be precomputed and stored). In general, $\boldsymbol{C}_{k0}$ does not have a Kronecker structure. Thus we propose to fit $\boldsymbol{C}_{k0}$ to the form of a Kronecker product by solving the least squares problem

$$\min_{\bar{\boldsymbol{C}}_{k0}\in\mathbb{C}^{T\times T}}\|\boldsymbol{C}_{k0}-\bar{\boldsymbol{C}}_{k0}\otimes\boldsymbol{\Xi}\|_F^2$$

as formulated in [30, Sec. 4]. Let $\boldsymbol{C}_{k0}\{i,j\}$ be the $N\times N$ sub-matrix containing the elements in rows from $(i-1)N+1$ to iN and columns from $(j-1)N+1$ to jN of $\boldsymbol{C}_{k0}$. Let $\bar{c}_{ij}$ be the element in row i and column j of $\bar{\boldsymbol{C}}_{k0}$. It follows that

$$\begin{aligned}&\|\boldsymbol{C}_{k0}-\bar{\boldsymbol{C}}_{k0}\otimes\Xi_k\|_F^2\\ &=\sum_{i=1}^{T}\sum_{j=1}^{T}\|\boldsymbol{C}_{k0}\{i,j\}-\bar{c}_{ij}\boldsymbol{\Xi}_k\|_F^2\\ &=\sum_{i=1}^{T}\sum_{j=1}^{T}\|\boldsymbol{C}_{k0}\{i,j\}\|_F^2-\bar{c}_{ij}\mathrm{tr}\left\{\boldsymbol{C}_{k0}\{i,j\}^{\mathsf{H}}\boldsymbol{\Xi}_k\right\}\\ &\quad-\bar{c}_{ij}^{*}\mathrm{tr}\left\{\boldsymbol{\Xi}_k\boldsymbol{C}_{k0}\{i,j\}\right\}+|\bar{c}_{ij}|^2\mathrm{tr}\left\{\boldsymbol{\Xi}_k^2\right\}.\end{aligned}$$

Observe that $\|\boldsymbol{C}_{k0}-\bar{\boldsymbol{C}}_{k0}\otimes\boldsymbol{\Xi}_k\|_F^2$ is the sum of convex quadratic functions of $\bar{c}_{ij}$. Setting the partials $\frac{\partial\|\boldsymbol{C}_{k0}-\bar{\boldsymbol{C}}_{k0}\otimes\boldsymbol{\Xi}_k\|_F^2}{\partial\bar{c}_{ij}}$ to zeros, the optimal $\bar{\boldsymbol{C}}_{k0}$ is given by

$$\bar{c}_{ij}=\frac{\mathrm{tr}\left\{\boldsymbol{C}_{k0}\{i,j\}\boldsymbol{\Xi}_k\right\}}{\mathrm{tr}\left\{\boldsymbol{\Xi}_k^2\right\}}.$$

With the approximation $\boldsymbol{C}_{k0}\approx\bar{\boldsymbol{C}}_{k0}\otimes\boldsymbol{\Xi}_k$, we can approximate $\pi_{k1}^{(i_k)}$ by the RHS of (40). Also, it follows from (25) and (26) that

$$\boldsymbol{\Sigma}_{ki}\approx\left[\big(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}\big)\big(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}+\bar{\boldsymbol{C}}_{k0}\big)^{-1}\bar{\boldsymbol{C}}_{k0}\right]\otimes\boldsymbol{\Xi}_k,\quad(41)$$

$$\hat{\boldsymbol{z}}_{ki}\approx\mathrm{vec}\Big(\Big[\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}\big(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}+\bar{\boldsymbol{C}}_{k0}\big)^{-1}\boldsymbol{M}_{k0}\Big]^{\mathsf{T}}\Big).\quad(42)$$

To compute $\boldsymbol{C}_{k1}$ and $\boldsymbol{\mu}_{k1}$ in (34) and (33), the inversion of $\boldsymbol{C}_{k0}$ can be simplified as $\boldsymbol{C}_{k0}^{-1}\approx\bar{\boldsymbol{C}}_{k0}^{-1}\otimes\boldsymbol{\Xi}_k^{-1}$, but the inverse of $NT\times NT$ matrices involving $\boldsymbol{\Sigma}_k$ is still required.

To keep an accurate message update at early iterations,[4] let us fix a threshold $t_0\in[t_{\max}]$ and modify Algorithm 1 as follows. At iteration t, if $t\le t_0$, the messages are updated as in lines 7-11; if $t>t_0$, in line 7, (27) is replaced by (40) for the update of $\pi_{k1}^{(i_k)}$, and in line 8, (26) and (25) are replaced by (42) and (41) for the update of $\boldsymbol{\Sigma}_{ki}$ and $\hat{\boldsymbol{z}}_{ki}$, respectively. We refer to this scheme as EPAK (EP with Approximate Kronecker). It coincides with EP if $t_0=t_{\max}$. At iteration $t>t_0$, the dominant operations in EPAK are the inverse of $\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}+\bar{\boldsymbol{C}}_{k0}$ (with complexity $O(K^3)$) in (41) and (42) for each $k\in[K]$ and $i\in[|\mathcal{S}_k|]$, and the inverse of $NT\times NT$ matrices (with complexity $O(K^6)$) to compute $\boldsymbol{C}_{k1}$ and $\boldsymbol{\mu}_{k1}$ for each $k\in[K]$. Thus the complexity at iteration t of EPAK is $O(K^7 2^B)$ if $t\le t_0$ and $O(K^4 2^B+K^7)$ if $t>t_0$.

B. Minimum Mean Square Error - Successive Interference Approximation (MMSE-SIA)

Another method to simplify EP is as follows. In the EP scheme, as in (29) and (32), the message $\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})$ from node ψ_{k1} to node $\mathbf{z}_k$ is derived by first projecting $\hat{p}_{k1,\mathbf{z}_k}^{\mathrm{new}}(\boldsymbol{z}_k)\propto\sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}\mathcal{N}(\boldsymbol{z}_k;\hat{\boldsymbol{z}}_{ki},\boldsymbol{\Sigma}_{ki})$ onto the Gaussian family, then dividing the projected Gaussian by $\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0})$. If we skip the projection of $\hat{p}_{k1}^{\mathrm{new}}(\boldsymbol{z}_k)$ onto the Gaussian family, i.e., we derive $\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})$ by dividing directly $\hat{p}_{k1,\mathbf{z}_k}^{\mathrm{new}}(\boldsymbol{z}_k)$ to $\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0})$, then the mean $\boldsymbol{\mu}_{k1}$ and covariance matrix $\boldsymbol{C}_{k1}$ are matched to that of the PDF proportional to

$$\begin{aligned}\frac{\hat{p}_{k1,\mathbf{z}_k}^{\mathrm{new}}(\boldsymbol{z}_k)}{\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0})}&=\sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}\frac{\mathcal{N}(\boldsymbol{z}_k;\hat{\boldsymbol{z}}_{ki},\boldsymbol{\Sigma}_{ki})}{\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0})}\\ &\propto\sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}\mathcal{N}\big(\boldsymbol{z}_k;\mathbf{0},(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}})\otimes\boldsymbol{\Xi}_k\big)\\ &=\mathcal{N}\big(\boldsymbol{z}_k;\mathbf{0},\boldsymbol{R}_k\otimes\boldsymbol{\Xi}_k\big).\end{aligned}\quad(43)$$

where $\boldsymbol{R}_k:=\sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}$. (43) can be verified using $\mathcal{N}(\boldsymbol{z}_k;\hat{\boldsymbol{z}}_{ki},\boldsymbol{\Sigma}_{ki})\propto\mathcal{N}\big(\boldsymbol{z}_k;\mathbf{0},(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}})\otimes\boldsymbol{\Xi}_k\big)\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0})$, which follows from the Gaussian PDF multiplication rule with $\hat{\boldsymbol{z}}_{ki}$ and $\boldsymbol{\Sigma}_{ki}$ given in (26) and (25), respectively. It follows that $\boldsymbol{\mu}_{k1}=\mathbf{0}$ and $\boldsymbol{C}_{k1}=\boldsymbol{R}_k\otimes\boldsymbol{\Xi}_k$. As a consequence (see (39) and (38)), $\boldsymbol{\mu}_{k0}=\boldsymbol{y}$ and $\boldsymbol{C}_{k0}=\sigma^2\boldsymbol{I}_{NT}+\sum_{l\ne k}\boldsymbol{R}_l\otimes\boldsymbol{\Xi}_k$.

This scheme can be alternatively interpreted as follows. We expand $\mathbf{y}$ in (20) as

$$\mathbf{y}=(\mathbf{s}_k\otimes\boldsymbol{I}_N)\mathbf{h}_k+\sum_{l\ne k}(\mathbf{s}_l\otimes\boldsymbol{I}_N)\mathbf{h}_l+\mathbf{w}.$$

The second term $\mathbf{t}_k:=\sum_{l\ne k}(\mathbf{s}_l\otimes\boldsymbol{I}_N)\mathbf{h}_l$ is the interference from other users while decoding the signal of user k. Since the signals $\mathbf{s}_l$ are independent of the channels $\mathbf{h}_l$ and the channels $\mathbf{h}_l$ have zero mean, we have that $\mathbb{E}[\mathbf{t}_k]=\mathbf{0}$. The covariance matrix of $\mathbf{t}_k$ is $\mathbb{E}[\mathbf{t}_k\mathbf{t}_k^{\mathsf{H}}]=\sum_{l\ne k}\mathbb{E}[\mathbf{s}_l\mathbf{s}_l^{\mathsf{H}}]\otimes\boldsymbol{\Xi}_k=\sum_{l\ne k}\boldsymbol{R}_l\otimes\boldsymbol{\Xi}_k$. If we treat the interference term $\mathbf{t}_k$ as a Gaussian vector with the same mean and covariance matrix,[5] then $\mathbf{t}_k+\mathbf{w}\sim\mathcal{N}\big(\mathbf{0},\sum_{l\ne k}\boldsymbol{R}_l\otimes\boldsymbol{\Xi}_k+\sigma^2\boldsymbol{I}_{NT}\big)$. The single-user likelihood under this approximation is $\hat{p}_{\mathbf{y}|\mathbf{s}_k}(\boldsymbol{y}|\boldsymbol{s}_k)=\mathcal{N}\big(\boldsymbol{y};\mathbf{0},\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}}\otimes\boldsymbol{\Xi}_k+\sum_{l\ne k}\boldsymbol{R}_l\otimes\boldsymbol{\Xi}_l+\sigma^2\boldsymbol{I}_{NT}\big)$. With this and Lemma 1, the update of the approximate posterior $\hat{p}_{\mathbf{s}_k|\mathbf{y}}(\boldsymbol{s}_k|\boldsymbol{y})\propto\hat{p}_{\mathbf{y}|\mathbf{s}_k}(\boldsymbol{y}|\boldsymbol{s}_k)$ coincides with (27) for $\boldsymbol{\mu}_{k0}=\boldsymbol{y}$ and $\boldsymbol{C}_{k0}=\sigma^2\boldsymbol{I}_{NT}+\sum_{l\ne k}\boldsymbol{R}_l\otimes\boldsymbol{\Xi}_k$. The matrix $\boldsymbol{R}_k$ is then recalculated with the updated value of $\hat{p}_{\mathbf{s}_k|\mathbf{y}}(\boldsymbol{s}_k^{(i_k)}|\boldsymbol{y})$,

[4]In the uncorrelated fading case, i.e. $\boldsymbol{\Xi}_k=\boldsymbol{I}_N$, the approximation of $\boldsymbol{C}_{k0}$ with Kronecker products becomes more accurate when $\hat{\pi}_{k1}$ is closer to a Kronecker-delta distribution, i.e., we have high confidence in one of the symbols. This is likely the case at high SNR after some EP iterations. At early iterations, however, the approximation $\boldsymbol{C}_{k0}\approx\bar{\boldsymbol{C}}_{k0}\otimes\boldsymbol{\Xi}$ can be inaccurate.

[5]Another choice is to treat each $\mathbf{s}_l$, $l\ne k$, as a Gaussian. With this choice, however, the interference term $\mathbf{t}_k$ is a product of Gaussians which makes the approximate single-user likelihood difficult to evaluate.

$i_k \in [|\mathcal{S}_k|]$. The matrices $\boldsymbol{C}_{l0}$ are updated accordingly, and then used to update $\hat{p}_{\mathbf{s}_l|\mathbf{y}}(\boldsymbol{s}_l^{(i_l)}|\boldsymbol{y})$, $i_l \in [|\mathcal{S}_l|]$, $l \neq k$.

In short, the derived simplification of the EP scheme above iteratively MMSE-estimates the signal $\mathbf{z}_k$ of one user at a time while treating the interference as Gaussian. At each iteration, the Gaussian approximation of the interference for each user is successively improved using the estimates of the signals of other users. We refer to this scheme as MMSE-SIA and summarize it in Algorithm 2. In particular, as for the EP scheme, we can start with the non-informative initialization $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}|\boldsymbol{Y}) = \frac{1}{|\mathcal{S}_k|}\mathbb{1}\{\boldsymbol{s} \in \mathcal{S}_k\}$.

Algorithm 2: MMSE-SIA for Probabilistic Non-Coherent Detection

Input: the observation $\boldsymbol{Y}$; the constellations $\mathcal{S}_1, \ldots, \mathcal{S}_K$;
set the maximal number of iterations $t_{\max}$;
initialize of the posteriors $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k|\boldsymbol{Y})$ for $\boldsymbol{s}_k \in \mathcal{S}_k$, and $\boldsymbol{R}_k = \mathbb{E}_{\hat{p}_{\mathbf{s}_k|\mathbf{Y}}}[\mathbf{s}_k\mathbf{s}_k^{\mathsf{H}}]$ for $k \in [K]$;
$t \longleftarrow 0$;
repeat
  $t \longleftarrow t+1$;
  for $k \leftarrow 1$ **to** K **do**
    compute $\boldsymbol{C}_{k0} = \sigma^2\boldsymbol{I}_{NT} + \sum_{l\neq k}\boldsymbol{R}_l \otimes \Xi_k$;
    update $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k|\boldsymbol{Y})$, $\boldsymbol{s}_k \in \mathcal{S}_k$, according to (27) with $\boldsymbol{\mu}_{k0} = \boldsymbol{y}$ and $\boldsymbol{C}_{k0}$ computed ;
    update $\boldsymbol{R}_k = \mathbb{E}_{\hat{p}_{\mathbf{s}_k|\mathbf{Y}}}[\mathbf{s}_k\mathbf{s}_k^{\mathsf{H}}]$;
  end
until convergence or $t = t_{\max}$;
return $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k|\boldsymbol{Y})$ for $\boldsymbol{s}_k \in \mathcal{S}_k$, $k \in [K]$

The complexity order of Algorithm 2 is the same as EP due to the $NT \times NT$ matrix inversion in (27). However, MMSE-SIA still has complexity advantage over EP since no other matrix inversion is required, and there is no need to compute $\{\hat{\boldsymbol{z}}_{ki}\}$, $\{\boldsymbol{\Sigma}_{ki}\}$, $\hat{\boldsymbol{z}}_k$, $\boldsymbol{\Sigma}_k$, or update $\boldsymbol{\mu}_{k1}$. If the channel is uncorrelated ($\Xi_k = \boldsymbol{I}_N$), the complexity order of MMSE-SIA can be reduced. In this case, $\boldsymbol{C}_{k0}$ is the Kronecker product $\boldsymbol{Q}_k \otimes \boldsymbol{I}_N$ with $\boldsymbol{Q}_k := \sum_{l=1,l\neq k}^{K}\boldsymbol{R}_l + \sigma^2\boldsymbol{I}_T$, and thus in (27),

$$\begin{aligned}&\mathcal{N}\big(\mathbf{0};\boldsymbol{\mu}_{k0},(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}})\otimes\Xi_k+\boldsymbol{C}_{k0}\big)\\&=\mathcal{N}\left(\mathbf{0};\mathbf{y},\left(\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}+\boldsymbol{Q}_k\right)\otimes\boldsymbol{I}_N\right)\\&\propto\left(1+\boldsymbol{s}_k^{(i_k)\mathsf{H}}\boldsymbol{Q}_k^{-1}\boldsymbol{s}_k^{(i_k)}\right)^{-N}\exp\left(\frac{\left\|\mathbf{Y}^{\mathsf{H}}\boldsymbol{Q}_k^{-1}\boldsymbol{s}_k^{(i_k)}\right\|^2}{1+\boldsymbol{s}_k^{(i_k)\mathsf{H}}\boldsymbol{Q}_k^{-1}\boldsymbol{s}_k^{(i_k)}}\right).\end{aligned}\quad(44)$$

Then, only the inverse of $\boldsymbol{Q}_k$ is computed, which requires $O(K^3)$ operations. Given $\boldsymbol{Q}_k^{-1}$, the complexity of computing the RHS of (44) is then $O(K^2)$ for each $i_k \in [|\mathcal{S}_k|]$. Therefore, the complexity of computing $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}(\boldsymbol{s}_k|\boldsymbol{Y})$ is $O(K^3 + K^2 2^B)$ for $k \in [K]$. Finally, the complexity per iteration of the MMSE-SIA algorithm for uncorrelated fading is given by $O(K^4 + K^3 2^B)$.

VI. Implementation Aspects

A. Complexity

We summarize the computational complexity of the considered schemes in Table I.

TABLE I
Complexity Order of Different Non-Coherent Detectors With $T = O(K), N = O(K)$, and $|\mathcal{S}_k| = O(2^B)$, $k \in [K]$

Detector	Complexity order	
	Correlated fading	Uncorrelated fading $\Xi_k = \boldsymbol{I}_N, \forall k$
Optimal (exact marginalization)	$O(K^6 2^{BK})$	$O(K^3 2^{BK})$
EP	$O(K^7 2^B t_{\max})$	
EPAK	$O\big(K^7 2^B t_0 + (K^4 2^B + K^7)(t_{\max} - t_0)\big)$	
MMSE-SIA	$O(K^7 2^B t_{\max})$	$O(K^4 t_{\max} + K^3 2^B t_{\max})$

$t_{\max}$ denotes the number of iterations. $t_0 \in [t_{\max}]$.

B. Stabilization

We discuss some possible numerical problems in the EP algorithm and our solutions.

1) Singularity of $\boldsymbol{\Sigma}_k$: First, in (31), since the $NT \times NT$ matrix $\boldsymbol{\Sigma}_k$ is the weighted sum of the terms of rank less than NT, it can be close to singular if at a certain iteration, only few of the weights $\pi_{k1}^{(i)}$ are sufficiently larger than zero. The singularity of $\boldsymbol{\Sigma}_k$ can also arise from the constellation structure. For example, the constellations proposed in [22] are precoded versions of a constellation in $G(\mathbb{C}^{T-K+1}, 1)$ and the maximal rank of $\boldsymbol{\Sigma}_k$ is $N(T-K+1) \leq NT$. To avoid the inverse of $\boldsymbol{\Sigma}_k$, we express $\boldsymbol{C}_{k1}$ in (33) and $\boldsymbol{\mu}_{k1}$ in (34) respectively as

$$\begin{aligned}\boldsymbol{C}_{k1} &= -\boldsymbol{C}_{k0}\big(\boldsymbol{\Sigma}_k - \boldsymbol{C}_{k0}\big)^{-1}\boldsymbol{\Sigma}_k,\\ \boldsymbol{\mu}_{k1} &= \boldsymbol{C}_{k0}\big(\boldsymbol{\Sigma}_k - \boldsymbol{C}_{k0}\big)^{-1}\left(\boldsymbol{\Sigma}_k - \sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}\boldsymbol{\Sigma}_{ki}\right)\boldsymbol{C}_{k0}^{-1}\boldsymbol{\mu}_{k0}.\end{aligned}\quad(45)$$

2) "Negative Variance": Another problem is that $\boldsymbol{C}_{k1}$ is not guaranteed to be positive definite even if both $\boldsymbol{C}_{k0}$ and $\boldsymbol{\Sigma}_k$ are. When $\boldsymbol{C}_{k1}$ is not positive definite, from (38), $\boldsymbol{C}_{k0}$ can have negative eigenvalues, which, through (27), can make $\hat{\pi}_{k1}^{(i_k)}$ become close to a Kronecker-delta distribution (even at low SNR) where the position of the mode can be arbitrary, and the algorithm may diverge. Note that this "negative variance" problem is common in EP (see, e.g., [24, Sec. 3.2.1], [31, Sec. 5.3]). There has been no generally accepted solution and one normally resorts to various heuristics adapted to each problem. In our problem, to control the eigenvalues of $\boldsymbol{C}_{k1}$, we modify (45) by first computing the eigendecomposition $-\boldsymbol{C}_{k0}\big(\boldsymbol{\Sigma}_k - \boldsymbol{C}_{k0}\big)^{-1}\boldsymbol{\Sigma}_k = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$, then computing $\boldsymbol{C}_{k1}$ as $\boldsymbol{C}_{k1} = \boldsymbol{V}|\boldsymbol{\Lambda}|\boldsymbol{V}^{-1}$, where $|\boldsymbol{\Lambda}|$ is the element-wise absolute value of $\boldsymbol{\Lambda}$. This manipulation of replacing the variance parameters by their absolute values was also used in [32].

3) Overconfidence at Early Iterations: Finally, due to the nature of the message passing between continuous and discrete distribution, it can happen that all the mass of the PMF $\hat{\pi}_{k1}^{(i_k)}$ is concentrated on a small region of a potentially large constellation $\mathcal{S}_k$. For example, if $\pi_{k1}^{(i_k)}$ is close to a Kronecker-delta distribution with a single mode at i_0, then (26) and (25) implies that $\boldsymbol{\Sigma}_k$ is approximately $\boldsymbol{\Sigma}_{ki_0}$, and then from (33), $\boldsymbol{C}_{k1} \approx (\boldsymbol{s}_k^{(i_0)}\boldsymbol{s}_k^{(i_0)\mathsf{H}}) \otimes \Xi_k$. In this case, almost absolute certainty is placed on the symbol $\boldsymbol{s}_k^{(i_0)}$, and the algorithm will not be able significantly update its belief in the subsequent iterations. This can be problematic when the mode of $\pi_{k1}^{(i_k)}$ is placed on the wrong symbol at early iterations. To smooth the updates,

we apply damping on the update of the parameters of the continuous distributions $\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k1},\boldsymbol{C}_{k1})$ and $\mathcal{N}(\boldsymbol{z}_k;\boldsymbol{\mu}_{k0},\boldsymbol{C}_{k0})$. That is, with a damping factor $\eta \in [0;1]$, at iteration t and for each user k, we update

$$\boldsymbol{C}_{k1}(t) = \eta \boldsymbol{V}(t)|\boldsymbol{\Lambda}(t)|\boldsymbol{V}^{-1}(t) + (1-\eta)\boldsymbol{C}_{k1}(t-1), \tag{46}$$

$$\begin{aligned}\boldsymbol{\mu}_{k1}(t) = {} & \eta \boldsymbol{C}_{k0}(t-1)\big(\boldsymbol{\Sigma}_k(t) - \boldsymbol{C}_{k0}(t-1)\big)^{-1} \\ & \times\left(\boldsymbol{\Sigma}_k(t) - \sum_{i=1}^{|\mathcal{S}_k|} \pi_{k1}^{(i)}(t)\boldsymbol{\Sigma}_{ki}(t)\right)\boldsymbol{C}_{k0}^{-1}(t-1)\boldsymbol{\mu}_{k0}(t-1) \\ & + (1-\eta)\boldsymbol{\mu}_{k1}(t-1),\end{aligned} \tag{47}$$

$$\boldsymbol{C}_{l0}(t) = \eta\left(\sigma^2 \boldsymbol{I}_{NT} + \sum_{j\neq l}\boldsymbol{C}_{j1}(t)\right) + (1-\eta)\boldsymbol{C}_{l0}(t-1), \quad \forall l \neq k, \tag{48}$$

$$\boldsymbol{\mu}_{l0}(t) = \eta\left(\boldsymbol{y} - \sum_{j\neq l}\boldsymbol{\mu}_{j1}(t)\right) + (1-\eta)\boldsymbol{\mu}_{l0}(t-1), \quad \forall l \neq k. \tag{49}$$

In short, we stabilize the EP message updates by replacing (46), (47), (48), and (49) for (33), (34), (38), and (39), respectively. This technique also applies to EPAK. For MMSE-SIA, we damp the update of $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$ in a similar manner as $\boldsymbol{Q}_k(t) = \eta\big(\sum_{l\neq k}\boldsymbol{R}_l(t-1) + \sigma^2\boldsymbol{I}_T\big) + (1-\eta)\boldsymbol{Q}_k(t-1)$ and $\boldsymbol{R}_k(t) = \eta\sum_{i_k=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i_k)}(t)\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}} + (1-\eta)\boldsymbol{R}_k(t-1)$. Note that damping does not change the complexity order of these schemes. The approaches described in this subsection were implemented for the numerical results in the next section.

VII. Performance Evaluation

In this section, we evaluate the performance of our proposed schemes for a given set of individual constellations. We assume that $B_1 = \ldots B_K =: B$. We consider the local scattering model [4, Sec. 2.6] for the correlation matrices $\boldsymbol{\Xi}_k$. Specifically, the (l,m)-th element of $\boldsymbol{\Xi}_k$ is generated as $[\boldsymbol{\Xi}_k]_{l,m} = \xi_k \mathbb{E}_{\delta_k}[\exp(2\pi d_H(l-m)\sin(\varphi_k+\delta_k))]$, where d_H is the antenna spacing in the receiver array (measured in number of wavelengths), φ_k is a deterministic nominal angle, and δ_k is a random deviation. We consider $d_H = \frac{1}{2}$, φ_k generated uniformly in $[-\pi,\pi]$, and δ_k uniformly distributed in $[-\sqrt{3}\sigma_\varphi, \sqrt{3}\sigma_\varphi]$ with angular standard deviation $\sigma_\varphi = 10^\circ$. We also consider $\xi_k = 1, \forall k$. We set a damping factor $\eta = 0.9$ for EP, EPAK, and MMSE-SIA.

A. Test Constellations, State-of-the-Art Detectors, and Benchmarks

1) Precoding-Based Grassmannian Constellations: We consider the constellation design in [22], which imposes a geometric separation between the individual constellations through a set of precoders $\boldsymbol{U}_k$, $k \in [K]$. Specifically, starting with a Grassmannian constellation $\mathcal{D} = \{\boldsymbol{d}^{(1)},\ldots,\boldsymbol{d}^{(2^B)}\}$ in $G(\mathbb{C}^{T-K+1},1)$, the individual constellation $\mathcal{S}_k$ is generated as

$$\boldsymbol{s}_k^{(i)} = \frac{\boldsymbol{U}_k\boldsymbol{d}^{(i)}}{\|\boldsymbol{U}_k\boldsymbol{d}^{(i)}\|}, \quad i \in [2^B].$$

We consider the precoders $\boldsymbol{U}_k$ defined in [22, Eq.(11)] and two candidates for $\mathcal{D}$:

- A numerically optimized constellation generated by solving the max-min distance criteria

$$\max_{\boldsymbol{d}^{(i)}\in G(\mathbb{C}^{T-K+1},1), i=1,\ldots,2^B} \ \min_{1\leq i<j\leq 2^B} d(\boldsymbol{d}^{(i)},\boldsymbol{d}^{(j)}), \tag{50}$$

  where $d(\boldsymbol{d}^{(i)},\boldsymbol{d}^{(j)}) := \sqrt{1-|\boldsymbol{d}^{(i)\mathsf{H}}\boldsymbol{d}^{(j)}|^2}$ is the chordal distance between two Grassmannian points represented by $\boldsymbol{d}^{(i)}$ and $\boldsymbol{d}^{(j)}$. A constellation with maximal minimum pairwise distance leads to low symbol error rate in the absence of the interference. In our simulation, we approximate (50) by $\min_{\mathcal{D}} \log\sum_{1\leq i<j\leq 2^B}\exp\big(\frac{|\boldsymbol{d}^{(i)\mathsf{H}}\boldsymbol{d}^{(j)}|}{\epsilon}\big)$ with a small ϵ for smoothness, then solve it using gradient descent on the Grassmann manifold using the Manopt toolbox [33].
- The cube-split constellation proposed in [17], [34]. This structured constellation has good distance properties and allows for low-complexity single-user decoding and a simple yet effective binary labeling scheme.

Exploiting the precoder structure, [22] introduced a detector [22, Sec. V-B-3] that iteratively mitigates interference by projecting the received signal onto the subspace orthogonal to the interference subspace. We refer to it as POCIS (Projection onto the Orthogonal Complement of the Interference Subspace). For each user k, POCIS first estimates the row space of the interference $\sum_{l\neq k}\mathbf{s}_l\mathbf{h}_l^{\mathsf{T}}$ based on the precoders and projects the received signal onto the orthogonal complement of this space. It then performs single-user detections to obtain point estimates of the transmitted symbols. From these estimates, POCIS estimates the column space of the interference and projects the received signal onto its orthogonal complement. This process is repeated in the next iteration. The complexity order of POCIS is equivalent to the MMSE-SIA scheme. Note that only the indices of the estimated symbols are passed in POCIS, as opposed to the soft information on the symbols as in EP, MMSE-SIA, and EPAK.

2) Pilot-Based Constellations: We also consider the pilot-based constellations in which the symbols are generated as $\boldsymbol{s}_k^{(i)} = \left[\sqrt{\frac{K}{T}}\boldsymbol{e}_k^{\mathsf{T}}\ \sqrt{\frac{T-K}{TP_{\mathrm{avg}}}}\tilde{\boldsymbol{s}}_k^{(i)\mathsf{T}}\right]^{\mathsf{T}}$ where $\boldsymbol{e}_k$ is the k-th column of $\boldsymbol{I}_K$, $\tilde{\boldsymbol{s}}_k^{(i)}$ is a vector of data symbols taken from a scalar constellation, such as QAM, and P_{avg} is the average symbol power of the considered scalar constellation. Note that this corresponds to the scenario where the K users transmit mutually orthogonal pilot sequences, followed by spatially multiplexed parallel data transmission. Many MIMO detectors have been proposed specifically for these constellations. We consider some representatives as follows.

- The receiver MMSE-estimates the channel based on the first K rows of $\mathbf{Y}$, then MMSE-equalizes the received data symbols in the remaining $T-K$ rows of $\mathbf{Y}$, and performs a scalar demapper on the equalized symbols.
- The receiver MMSE-estimates the channel, then decodes the data symbols using the Schnorr-Euchner sphere decoder [35], referred to as SESD.
- The receiver performs the semi-blind joint ML channel estimation and data detection scheme in [9] with repeated weighted boosting search (RWBS) for channel estimation

and the Schnorr-Euchner sphere decoder for data detection, referred to as RWBS-SESD.

We note that the sphere decoder has near optimal performance given the channel knowledge, but its complexity is non-deterministic and can be exponential in the channel dimension if the channel matrix is ill-conditioned.

3) Benchmarks: We consider the optimal ML detector, whenever it is feasible, as a benchmark. When the optimal detector is computationally infeasible, we resort to another benchmark consisting in giving the receiver, while it decodes the signal $\mathbf{s}_k$ of user k, the knowledge of the signals $\mathbf{s}_l$ (but not the channel $\mathbf{h}_l$) of all the interfering users $l \neq k$. With this genie-aided information, optimal ML decoding (2) can be performed by keeping $\mathbf{s}_l$ fixed for all $l \neq k$ and searching for the best $\mathbf{s}_k$ in $\mathcal{S}_k$, thus reducing the total search space size from 2^{BK} to $K2^B$. The posterior marginals are computed separately for each user accordingly. This genie-aided detector gives an upper bound on the performance of EP, MMSE-SIA, EPAK, and POCIS.

B. Convergence and Running Time

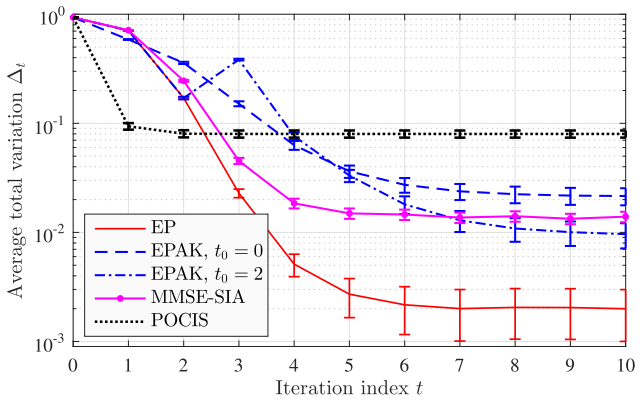
To assess the convergence of the algorithms, we evaluate the total variation distance between the estimated marginal posteriors $\hat{p}_{\mathbf{s}_k|\mathbf{Y}}$ at each iteration and the exact marginal posteriors $p_{\mathbf{s}_k|\mathbf{Y}}$ when exact marginalization (4) is possible. The total variation distance between two probability measures P and Q on $\mathcal{X}$ is defined as $\mathcal{TV}(P,Q) := \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|$. At iteration t where the estimated posteriors are $\hat{p}^{(t)}_{\mathbf{s}_k|\mathbf{Y}}$, $k \in [K]$, we evaluate the average total variation distance as

$$\Delta_t = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{Y}}[\mathcal{TV}(\hat{p}^{(t)}_{\mathbf{s}_k|\mathbf{Y}}, p_{\mathbf{s}_k|\mathbf{Y}})].$$

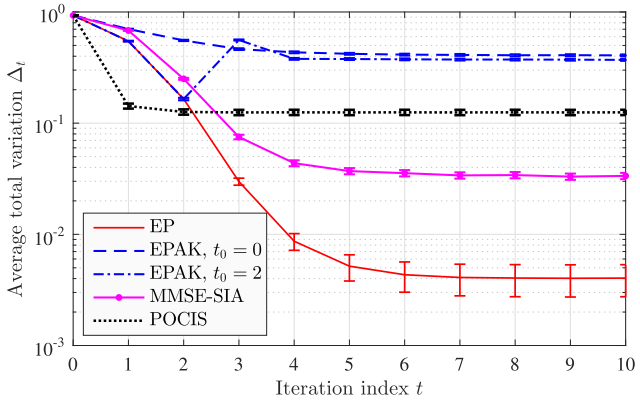
We consider the precoding-based Grassmannian constellations. Fig. 3 shows the empirical average total variation Δ_t for $T = 6$, $K = 3$, $N = 4$, and $B = 4$ at SNR $= 8$ dB. As can be seen, at convergence, EP provides the most accurate estimates of the marginal posteriors although it is less stable than other schemes. EP converges after 6 iterations while MMSE-SIA converges after 5 iterations. For uncorrelated fading, EPAK with $t_0 = 2$ can be eventually better than MMSE-SIA, but converges slower. For correlated fading, EPAK totally fails because of the inaccuracy of the approximation with Kronecker products. POCIS converges very quickly after 2 iterations but achieves a relatively low accuracy of the posterior estimation.

Fig. 4 depicts the average running time (on a local server) of exact marginalization compared with 6 iterations of EP, EPAK, MMSE-SIA, and POCIS at SNR $= 8$ dB. These schemes have significantly lower computation time than exact marginalization. The running time saving of EPAK w.r.t. EP is not significant, even with $t_0 = 0$. For uncorrelated fading, MMSE-SIA has much shorter running time than all other schemes.

From these convergence behaviors, hereafter, we fix the number of iterations of EP, MMSE-SIA, and EPAK as 6 and of POCIS as 3. Furthermore, we consider EPAK only for uncorrelated fading. For correlated fading, we generate the correlation matrices once and fix them over the simulation.



(a) Uncorrelated fading $\Xi_k = \boldsymbol{I}_N$



(b) Correlated Fading

Fig. 3. The empirical average total variation Δ_t over 1000 realizations of the transmitted signal, channel, and noise versus iteration for different non-coherent soft detection schemes for $T = 6$, $K = 3$, $B = 4$, and $N = 4$ at SNR $= 8$ dB. The error bars show the standard error, which is the standard deviation normalized by the square root of the number of samples. For correlated fading, these figures are further averaged over 10 realizations of the correlation matrices.

C. Achievable Rate

We first plot the achievable mismatched sum-rate R_{GMI} of the system calculated as in (7) for $T = 6$, $K = 3$, $N = 4$ and $B \in \{4, 8\}$ in Fig. 5. We consider the precoding-based Grassmannian constellations. For $\mathcal{D}$, we use the numerically optimized constellation if $B = 4$ and the cube-split constellation if $B = 8$. For uncorrelated fading (Fig. 5(a), the rates achieved with EP and MMSE-SIA detectors are very close to the achievable rate of the system (with the optimal detector) and not far from that of the genie-aided detector. EPAK (with $t_0 = 2$) achieves a very low rate, especially in the low SNR regime where the Kronecker approximation is not accurate. For correlated fading, (Fig. 5(b)), the rates achieved with EP and MMSE-SIA are only marginally lower than that of the optimal detector and genie-aided detector. In both cases, the rate achieved with POCIS is lower than that of EP and MMSE-SIA in the lower SNR regime and converges slowly with SNR to the limit $\frac{BK}{T}$ bits/channel use.

D. Symbol Error Rates of Hard Detection

Next, we use the outputs of EP, EPAK, MMSE-SIA and POCIS for a maximum-a-posteriori (MAP) hard detection.

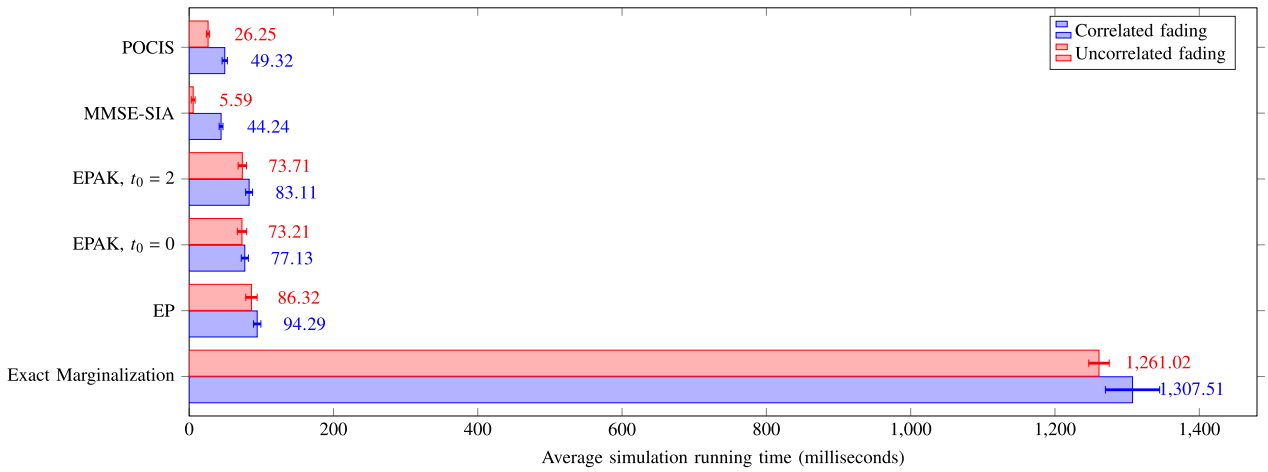
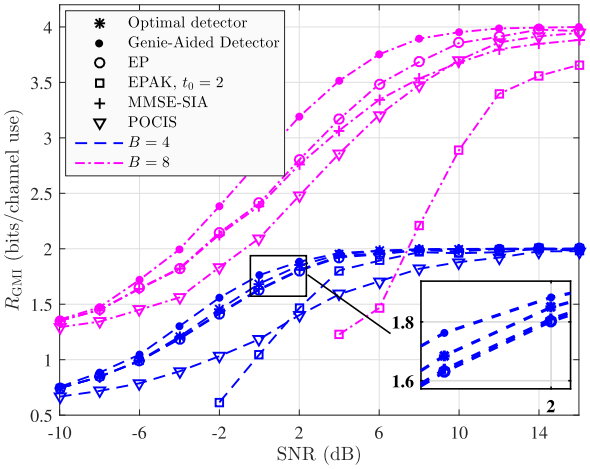
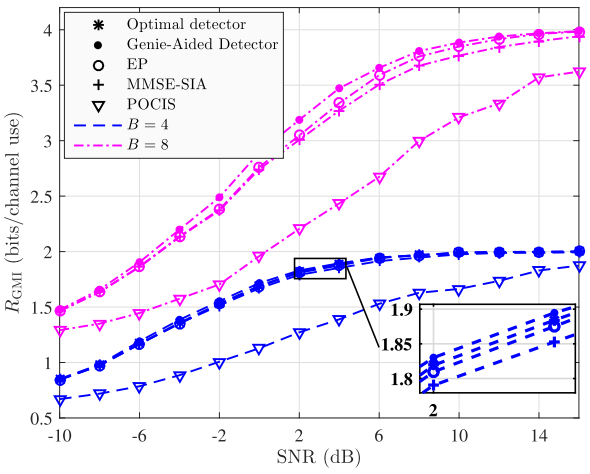


Fig. 4. The average running time over 1000 realizations of the transmitted signal, channel, and noise of exact marginalization vs. 6 iterations of the considered detection schemes for $T = 6$, $K = 3$, $B = 4$, and $N = 4$ at SNR $= 8$ dB. The error bars show the standard deviation. For correlated fading, the running time is further averaged over 10 realizations of the correlation matrices.

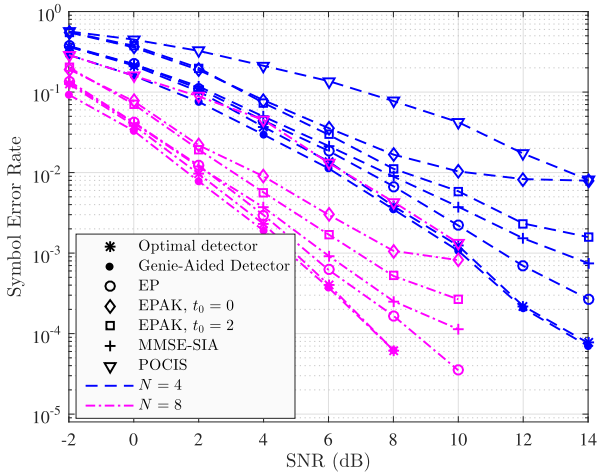


(a) Uncorrelated fading

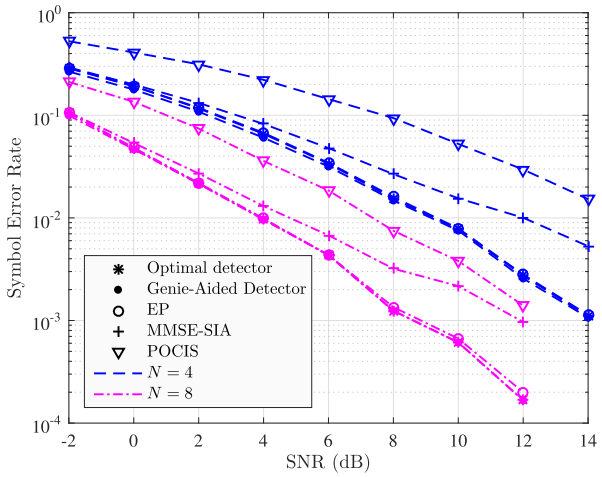


(b) Correlated fading

Fig. 5. The mismatched rate of the system with EP, EPAK (with $t_0 = 2$), MMSE-SIA, and POCIS detectors in comparison with the optimal detector and/or the genie-aided detector for $T = 6$, $K = 3$, $N = 4$, and $B \in \{4, 8\}$.



(a) Uncorrelated fading

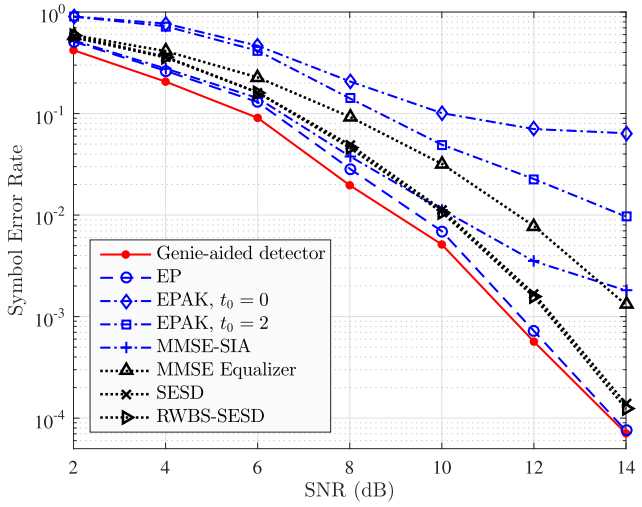


(b) Correlated fading

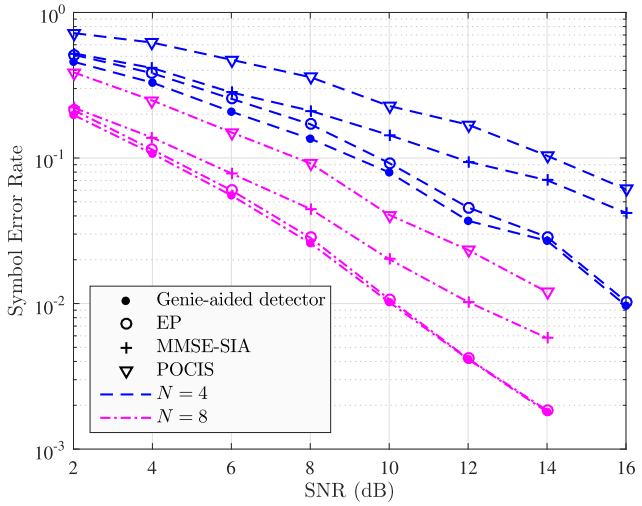
Fig. 6. The symbol error rate of the system with EP, EPAK (with $t_0 \in \{0, 2\}$), MMSE-SIA, and POCIS detectors in comparison with the optimal detector and the genie-aided detector for $T = 6$, $K = 3$, $N \in \{4, 8\}$ and $B = 4$.

We evaluate the performance in terms of symbol error rate (SER).

In Fig. 6, we consider the precoding-based constellations with $T = 6$, $K = 3$, $N \in \{4, 8\}$, and $B = 4$, for which the optimal ML detector (2) is computationally feasible. We observe that the SER of the EP and MMSE-SIA detectors are not much higher than that of the optimal detector,

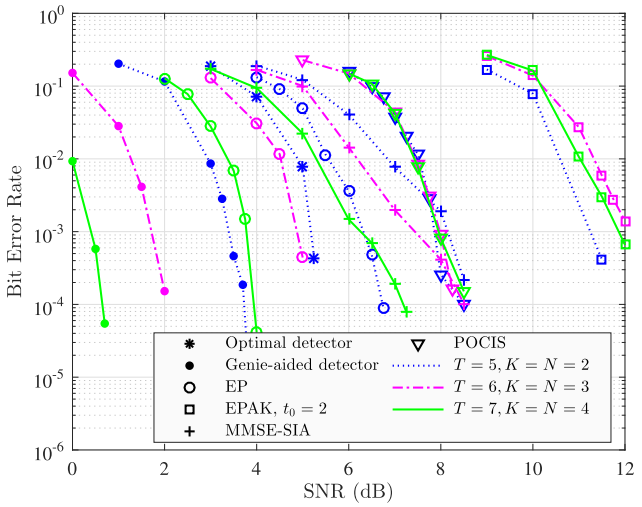


(a) Uncorrelated fading, pilot-based constellations

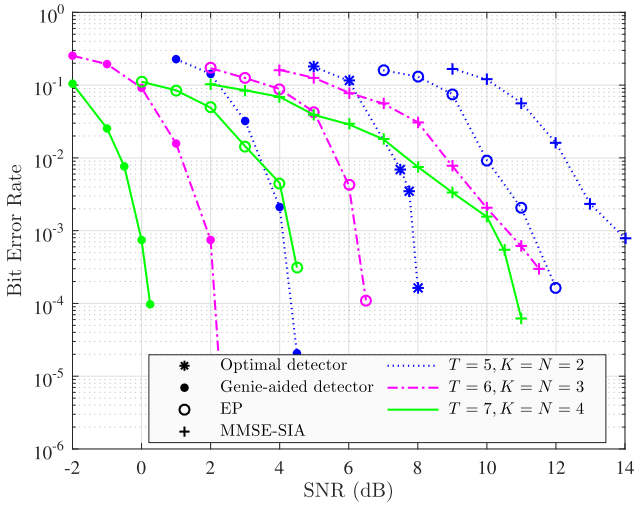


(b) Correlated fading, precoding-based constellations

Fig. 7. The symbol error rate of the system with EP, EPAK ($t_0 \in \{0, 2\}$), MMSE-SIA, POCIS vs. the genie-aided detector for $T = 6$, $K = 3$, $N = 8$, and $B = 9$. For uncorrelated fading, these schemes are compared with three pilot-based detectors using respectively MMSE equalizer, sphere decoding [35], and joint channel estimation–data detection [9].



(a) Uncorrelated fading



(b) Correlated fading

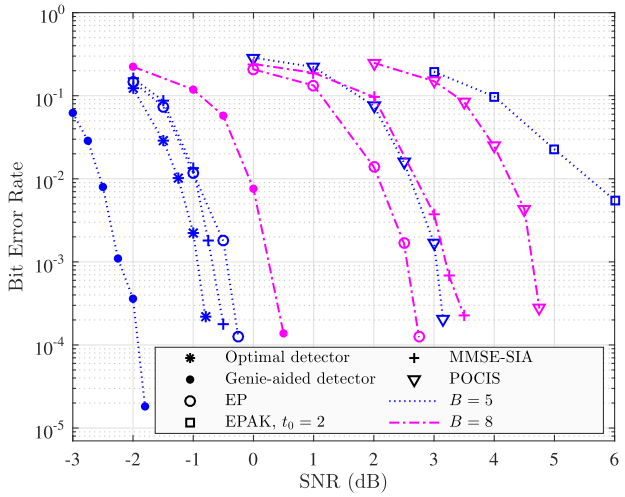
Fig. 8. The bit error rate with turbo codes of EP, EPAK (with $t_0 = 2$), MMSE-SIA, POCIS, and the optimal/genie-aided detector for $B = 8$ bits/symbol and $K = N$.

especially in the lower SNR regime. The SER of EPAK is significantly higher than that of EP and MMSE-SIA for $t_0 = 0$. This is greatly improved by setting $t_0 = 2$, i.e., keeping the first two iterations of EP. The gain of EP w.r.t. EPAK and MMSE-SIA is more pronounced when the SNR increases. For correlated fading, EP performs almost as good as the optimal detector, whose SER performance is closely approximated by the genie-aided detector.

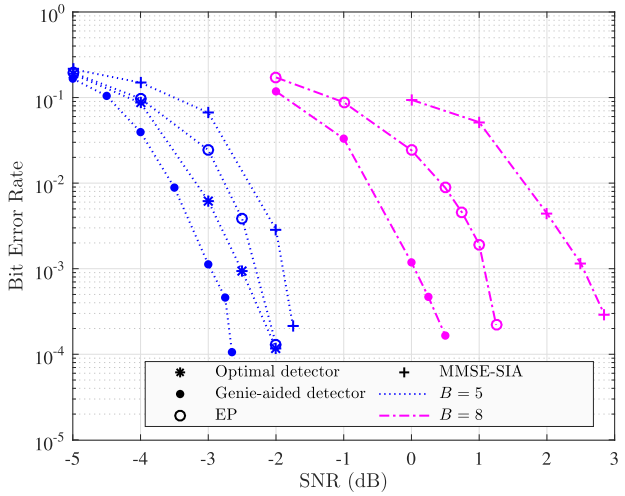
In Fig. 7, we consider $T = 6$, $K = 3$, $N = 8$, and $B = 9$ and use the genie-aided detector as a benchmark. In Fig. 7(a), we consider uncorrelated fading and use the pilot-based constellations with 8-QAM data symbols. The performance of EP is very close to that of the genie-aided detector. The performance of MMSE-SIA is close to EP in the low SNR regime (SNR $\leq$ 8 dB). We also depict the SER of the three pilot-based detectors in Section VII-A.2, namely, 1) MMSE channel estimation, MMSE equalizer, and QAM demapper, 2) SESD, and 3) RWBS-SESD. These three schemes are outperformed by the EP detectors. In Fig. 7(b), we consider correlated fading and use the precoding-based Grassmannian constellations with $\mathcal{D}$ numerically optimized. We observe again that EP achieves almost the same SER performance as the genie-aided detector.

E. Bit Error Rates With a Channel Code

In this subsection, we use the output of the soft detectors for channel decoding. We consider the precoding-based Grassmannian constellations with the cube-split constellation for $\mathcal{D}$ since it admits an effective and simple binary labeling [17]. We take the binary labels of the symbols in $\mathcal{D}$ for the corresponding symbols in $\mathcal{S}_k$. We integrate a standard symmetric parallel concatenated rate-1/3 turbo code [36]. The turbo encoder accepts packets of 1008 bits; the turbo decoder computes the



(a) Uncorrelated fading



(b) Correlated fading

Fig. 9. The bit error rate with turbo codes of EP, EPAK (with $t_0 = 2$), MMSE-SIA, POCIS, and the optimal/genie-aided detector for $T = 6$, $K = 3$, and $N = 4$.

bit-wise LLR from the soft outputs of the detection scheme as in (3) and performs 10 decoding iterations for each packet.

In Fig. 8, we show the bit error rate (BER) with this turbo code using $B = 8$ bits/symbol and different values of T and $K = N$. EP achieves the closest performance to the genie-aided detector and the optimal detector (4). The BER of MMSE-SIA vanishes slower with the SNR than the other schemes, and becomes better than POCIS as K and N increase. The BER of EPAK with $t_0 = 2$ is higher than all other schemes. Under uncorrelated fading, for $T = 7$ and $K = N = 4$, the power gain of EP w.r.t. MMSE-SIA, POCIS, and EPAK for the same BER of 10^{-3} is about 3 dB, 4 dB, and 8 dB, respectively. We also observe that the genie-aided detector gives very optimistic BER performance results compared to the optimal detector.

Finally, in Fig. 9, we consider $T = 6$, $K = 3$, $N = 4$, and compare the BER with the same turbo code for different B. For $B = 5$, both EP and MMSE-SIA have performance close to the optimal detector. Under uncorrelated fading, MMSE-SIA can be slightly better than EP. This is due to the residual effect (after damping) of the phenomenon that all the mass of $\pi_{k1}^{(i_k)}$ is concentrated on a possibly wrong symbol at early iterations, and EP may not be able to refine significantly the PMF in the subsequent iterations if the constellation is sparse. This situation is not observed for $B = 8$, i.e., larger constellations. Also, as compared to the case $T = 6, K = 3, B = 8$ in Fig. 8, the performance of MMSE-SIA is significantly improved as the number of receive antennas increases from $N = 3$ to $N = 4$. As in the previous case, EPAK does not perform well.

VIII. Conclusion

We proposed an expectation propagation (EP) based scheme and two simplifications (EPAK and MMSE-SIA) of this scheme for multi-user detection in non-coherent SIMO multiple access channel with spatially correlated Rayleigh fading. EP and MMSE-SIA are shown to achieve good performance in terms of mismatched sum-rate, symbol error rate when they are used for hard detection, and bit error rate when they are used for soft-input soft-output channel decoding. EPAK has acceptable performance with uncorrelated fading. It performs well for hard symbol detection but inadequately for soft-output detection. While MMSE-SIA and EPAK have lower complexity than EP, the performance gain of EP with respect to MMSE-SIA and EPAK is more significant when the number of users and/or the constellation size increase. Possible extensions of this work include considering more complicated fading models and analyzing theoretically the performance of EP for non-coherent reception.

Appendix A
Properties of the Gaussian PDF

Lemma 1: Let $\mathbf{x}$ *be an* n*-dimensional complex Gaussian vector. It holds that*

1) $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}+\boldsymbol{y};\boldsymbol{\mu}-\boldsymbol{y},\boldsymbol{\Sigma})$ *for* $\boldsymbol{y} \in \mathbb{C}^n$*;*

2) Gaussian PDF multiplication rule:

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_{\text{new}},\boldsymbol{\Sigma}_{\text{new}}) \times \mathcal{N}(\mathbf{0};\boldsymbol{\mu}_1-\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2),$$

where $\boldsymbol{\Sigma}_{\text{new}} := \left(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}\right)^{-1}$ *and* $\boldsymbol{\mu}_{\text{new}} := \boldsymbol{\Sigma}_{\text{new}}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right)$.

Proof: The first part follows readily from the definition of $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$. The complex Gaussian PDF multiplication rule is a straightforward generalization of the real counterpart [37]. □

Appendix B
Proof of Proposition 1

Using the natural logarithm for the KL divergence, we derive

$$D\big(q_\alpha(\boldsymbol{x})\big\|\underline{p}(\boldsymbol{x})\big) = \int q_\alpha(\boldsymbol{x}) \ln \frac{q_\alpha(\boldsymbol{x})}{\prod_\beta \underline{p}_\beta(\boldsymbol{x}_\beta)}\, \mathrm{d}\boldsymbol{x}$$

$$= \sum_\beta \int q_\alpha(\boldsymbol{x}) \ln \frac{1}{\underline{p}_\beta(\boldsymbol{x}_\beta)}\, \mathrm{d}\boldsymbol{x} + c_0$$

$$= \sum_{\beta\in\mathfrak{N}_\alpha} \int q_\alpha(\boldsymbol{x}) \ln \frac{1}{\underline{p}_\beta(\boldsymbol{x}_\beta)}\, \mathrm{d}\boldsymbol{x} + \sum_{\beta\notin\mathfrak{N}_\alpha} \int q_\alpha(\boldsymbol{x}) \ln \frac{1}{\underline{p}_\beta(\boldsymbol{x}_\beta)}\, \mathrm{d}\boldsymbol{x}$$

$$
\begin{aligned}
&+ c_0 \\
&= \sum_{\beta \in \mathfrak{N}_\alpha} \int q_\alpha(\boldsymbol{x}) \ln \frac{1}{\underline{p}_\beta(\boldsymbol{x}_\beta)} \, \mathrm{d}\boldsymbol{x} \\
&\quad + \sum_{\beta \notin \mathfrak{N}_\alpha} \int \hat{p}_\beta(\boldsymbol{x}_\beta) \ln \frac{1}{\underline{p}_\beta(\boldsymbol{x}_\beta)} \, \mathrm{d}\boldsymbol{x}_\beta + c_0 \qquad (51) \\
&= -\sum_{\beta \in \mathfrak{N}_\alpha} \int q_\alpha(\boldsymbol{x}) \left[\underline{\boldsymbol{\gamma}}_\beta^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_\beta) - A_\beta(\underline{\boldsymbol{\gamma}}_\beta) \right] \mathrm{d}\boldsymbol{x} + \sum_{\beta \notin \mathfrak{N}_\alpha} D\big(\hat{p}_\beta \big\| \underline{p}_\beta\big) \\
&\quad + c_0 \qquad (52) \\
&= \sum_{\beta \in \mathfrak{N}_\alpha} \left[A_\beta(\underline{\boldsymbol{\gamma}}_\beta) - \underline{\boldsymbol{\gamma}}_\beta^{\mathsf{T}} \mathbb{E}_{q_\alpha}\left[\boldsymbol{\phi}(\boldsymbol{x}_\beta)\right] \right] + \sum_{\beta \notin \mathfrak{N}_\alpha} D\big(\hat{p}_\beta \big\| \underline{p}_\beta\big) + c_0, \qquad (53)
\end{aligned}
$$

where (51) follows from $q_\alpha(\boldsymbol{x}) = \frac{\psi_\alpha(\boldsymbol{x}_\alpha)}{m_\alpha(\boldsymbol{x}_\alpha)} \Big[\prod_{\beta \in \mathfrak{N}_\alpha} \hat{p}_\beta(\boldsymbol{x}_\beta) \Big] \Big[\prod_{\beta \notin \mathfrak{N}_\alpha} \hat{p}_\beta(\boldsymbol{x}_\beta) \Big]$, and (52) follows from (11). From (53), we can see that the optimization (14) of $\underline{p}$ decouples over $\underline{p}_\beta$, and the optimal distribution can be expressed as $\hat{p}_\alpha^{\mathrm{new}}(\boldsymbol{x}) = \prod_\beta \hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta)$. For $\beta \notin \mathfrak{N}_\alpha$, the minimum of $D\big(\hat{p}_\beta \big\| \underline{p}_\beta\big)$ is simply 0 and achieved with $\hat{p}_{\alpha,\beta}^{\mathrm{new}}(\boldsymbol{x}_\beta) = \hat{p}_\beta(\boldsymbol{x}_\beta)$. For $\beta \in \mathfrak{N}_\alpha$, since the log-partition function $A_\beta(\underline{\boldsymbol{\gamma}}_\beta)$ is convex in $\underline{\boldsymbol{\gamma}}_\beta$ (see, e.g., [38, Lemma 1]), the minimum of $A_\beta(\underline{\boldsymbol{\gamma}}_\beta) - \underline{\boldsymbol{\gamma}}_\beta^{\mathsf{T}} \mathbb{E}_{q_\alpha}\left[\boldsymbol{\phi}(\boldsymbol{x}_\beta)\right]$ is achieved at the value of $\underline{\boldsymbol{\gamma}}_\beta$ where its gradient is zero. Using the well-known property of the log-partition function, $\nabla_{\boldsymbol{\gamma}} A_\beta(\boldsymbol{\gamma}) = \mathbb{E}_{\hat{p}_\beta}[\boldsymbol{\phi}_\beta(\boldsymbol{\gamma})]$, we get that the zero-gradient equation is equivalent to the moment matching criterion $\mathbb{E}_{\hat{p}_{\alpha,\beta}^{\mathrm{new}}}[\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta) = \mathbb{E}_{q_\alpha}[\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)]$.

REFERENCES

[1] K.-H. Ngo, M. Guillaud, A. Decurninge, S. Yang, S. Sarkar, and P. Schniter, "Non-coherent multi-user detection based on expectation propagation," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 2092–2096.

[2] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

[3] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[4] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[5] E. G. Larsson, "Massive MIMO for 5G: Overview and the road ahead," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2017, p. 1.

[6] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1941–1988, 4th Quart., 2015.

[7] S. Verdú, "Computational complexity of optimum multiuser detection," *Algorithmica*, vol. 4, nos. 1–4, pp. 303–312, Jun. 1989.

[8] S. Buzzi, M. Lops, and S. Sardellitti, "Performance of iterative data detection and channel estimation for single-antenna and multiple-antennas wireless communications," *IEEE Trans. Veh. Technol.*, vol. 53, no. 4, pp. 1085–1104, Jul. 2004.

[9] M. Abuthinien, S. Chen, and L. Hanzo, "Semi-blind joint maximum likelihood channel estimation and data detection for MIMO systems," *IEEE Signal Process. Lett.*, vol. 15, pp. 202–205, 2008.

[10] W. Xu, M. Stojnic, and B. Hassibi, "ON exact maximum-likelihood detection for non-coherent MIMO wireless systems: A branch-estimate-bound optimization framework," in *Proc. IEEE Int. Symp. Inf. Theory*, Toronto, ON, Canada, Jul. 2008, pp. 2017–2021.

[11] H. A. J. Alshamary and W. Xu, "Efficient optimal joint channel estimation and data detection for massive MIMO systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 875–879.

[12] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 543–564, Mar. 2000.

[13] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 359–383, Feb. 2002.

[14] W. Yang, G. Durisi, and E. Riegler, "On the capacity of large-MIMO block-fading channels," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 117–132, Feb. 2013.

[15] B. M. Hochwald, T. L. Marzetta, T. J. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 1962–1973, Sep. 2000.

[16] I. Kammoun, A. M. Cipriano, and J.-C. Belfiore, "Non-coherent codes over the Grassmannian," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3657–3667, Oct. 2007.

[17] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured Grassmannian constellation for non-coherent SIMO communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1948–1964, Mar. 2020.

[18] K.-H. Ngo, S. Yang, M. Guillaud, and A. Decurninge, "Joint constellation design for the two-user non-coherent multiple-access channel," 2020, *arXiv:2001.04970*. [Online]. Available: http://arxiv.org/abs/2001.04970

[19] G. Caire, R. R. Muller, and T. Tanaka, "Iterative multiuser joint decoding: Optimal power allocation and low-complexity implementation," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1950–1973, Sep. 2004.

[20] J. Goldberger and A. Leshem, "MIMO detection for high-order QAM based on a Gaussian tree approximation," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4973–4982, Aug. 2011.

[21] J. Cespedes, P. M. Olmos, M. Sanchez-Fernandez, and F. Perez-Cruz, "Probabilistic MIMO symbol detection with expectation consistency approximate inference," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3481–3494, Apr. 2018.

[22] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "A multiple access scheme for non-coherent SIMO communications," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 1846–1850.

[23] M. A. El-Azizy, R. H. Gohary, and T. N. Davidson, "A BICM-IDD scheme for non-coherent MIMO communication," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 541–546, Feb. 2009.

[24] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, Jan. 2001.

[25] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, "Approximate inference techniques with expectation constraints," *J. Stat. Mech., Theory Exp.*, vol. 2005, Nov. 2005, Art. no. 11015.

[26] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco, CA, USA: Holden-Day, 1977.

[27] S. A. Jafar and A. Goldsmith, "Multiple-antenna capacity in correlated Rayleigh fading with channel covariance information," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 990–997, May 2005.

[28] I. E. Telatar, A. Lapidoth, and A. Ganti, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[29] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[30] C. F. Van Loan and N. Pitsianis, "Approximation with Kronecker products," in *Linear Algebra for Large Scale and Real-Time Applications*. Dordrecht, The Netherlands: Springer, 1993, pp. 293–314.

[31] A. Vehtari *et al.*, "Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data," 2014, *arXiv:1412.4869*. [Online]. Available: http://arxiv.org/abs/1412.4869

[32] P. Sun, C. Zhang, Z. Wang, C. N. Manchon, and B. H. Fleury, "Iterative receiver design for ISI channels using combined Belief- and expectation-propagation," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1733–1737, Oct. 2015.

[33] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a MATLAB toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, pp. 1455–1459, Aug. 2014. [Online]. Available: http://www.manopt.org

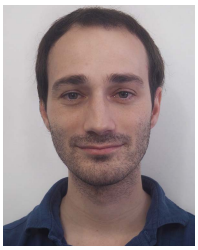
[34] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Design and analysis of a practical codebook for non-coherent communications," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2017, pp. 1237–1241.

[35] C. P. Schnorr and M. Euchner, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," *Math. Program.*, vol. 66, nos. 1–3, pp. 181–199, Aug. 1994.
[36] *Multiplexing and Channel Coding, Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA)*, 3GPP document TS 36.212 V8.0.0, 2007.
[37] P. Bromiley, "Products and convolutions of Gaussian probability density functions," *Tina-Vis. Memo*, vol. 3, no. 4, p. 1, Aug. 2003.
[38] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "A new class of upper bounds on the log partition function," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2313–2335, Jul. 2005.

Khac-Hoang Ngo (Graduate Student Member, IEEE) received the B.E. degree (Hons.) in electronics and telecommunications from the University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam, in 2014, and the M.Sc. degree (Hons.) in advanced wireless communication systems (Master SAR) from the CentraleSupélec, Université Paris-Saclay, France, in 2016. He is currently pursuing the Ph.D. degree in non-coherent wireless communications with the CentraleSupélec and Mathematical and Algorithmic Sciences Laboratory, Huawei Technologies France. His research interests include multiple-antenna wireless communications and information theory. He received the Honda Award for Young Engineers and Scientists (Honda Y-E-S Award) in Vietnam, in 2013.

Maxime Guillaud (Senior Member, IEEE) received the M.Sc. degree in electrical engineering from ENSEA, Cergy, France, in 2000, and the Ph.D. degree in electrical engineering and communications from Telecom ParisTech, Paris, France, in 2005.

From 2000 to 2001, he was a Research Engineer with Lucent Bell Laboratories (now Nokia), Holmdel, NJ, USA. From 2006 to 2010, he was a Senior Researcher with FTW, Vienna, Austria. From 2010 to 2014, he was a Researcher with the Vienna University of Technology, Vienna. Since 2014, he has been a Principal Researcher with the Huawei Technologies France, where he is currently the Head of the Signal and Information Processing Team. He has authored more than 60 research articles and patents. He was a recipient of the SPAWC 2005 Student Paper Award, and a co-recipient of the Mario Boella Business Idea Prize of the NEWCOM NoE in 2005. He was a contributor to the FP6 Newcom and FP7 Newcom++ NoEs, of the FP6 MASCOT Project, and a Key Member of the EU FP7 Project HIATUS. He worked on the transceiver architecture of multiuser cellular systems, and on various aspects of wireless channel modeling, including sparse representations and channel state inference methods. He introduced the principle of relative calibration for the exploitation of channel reciprocity. His recent interests revolve around the physical layer of fifth-generation cellular networks (5G). He is an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.



Alexis Decurninge (Member, IEEE) received the Ph.D. degree in statistics from the Université Pierre et Marie Curie, Paris, France, in 2015. His Ph.D. thesis on statistical methods for radar signal processing was made in collaboration with Thales Air Systems. Since 2015, he has been a Research Engineer with the Mathematical and Algorithmic Sciences Laboratory, Huawei Technologies France, Paris. His research interests focus on statistical signal processing, robust statistical methods, Riemannian geometry, and multiple-antenna wireless.

Sheng Yang (Member, IEEE) received the B.E. degree in electrical engineering from Jiaotong University, Shanghai, China, in 2001, and the Engineer and M.Sc. degrees in electrical engineering from Telecom ParisTech, Paris, France, in 2004, respectively, and the Ph.D. degree from the Université de Pierre et Marie Curie (Paris VI), in 2007. From 2007 to 2008, he was a Senior Staff Research Engineer with the Motorola Research Center, Gif-sur-Yvette, France. Since 2008, he has been with the CentraleSupélec, where he is currently a Full Professor. Since 2015, he has held the Honorary Associate Professorship with the Department of Electrical and Electronic Engineering, The University of Hong Kong (HKU). He received the 2015 IEEE ComSoc Young Researcher Award for the Europe, Middle East, and Africa Region (EMEA). He is an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Philip Schniter (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1992 and 1993, respectively, and the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 2000.

From 1993 to 1996, he was a Systems Engineer with Tektronix Inc., Beaverton, OR, USA. After receiving the Ph.D. degree, he joined the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA, where he is currently a Professor. From 2008 to 2009, he was a Visiting Professor with Eurecom, Sophia Antipolis, France, and with Supélec, Gif-sur-Yvette, France. From 2016 to 2017, he was a Visiting Professor with Duke University, Durham, NC, USA. His areas of interests currently include signal processing, wireless communications, and machine learning. In 2002, he was a recipient of the NSF CAREER Award, in 2016, the IEEE Signal Processing Society Best Paper Award, and in 2018, the Qualcomm Faculty Award. He also serves on the IEEE Computational Imaging Technical Committee, and as an Associate Editor for the *SIAM Journal on Imaging Sciences*.