Non-Coherent Multi-User Detection Based on Expectation Propagation

Ngo Khac-Hoang, Maxime Guillaud, Alexis Decurninge, Yang Sheng, Subrata Sarkar, Philip Schniter

Non-Coherent Multi-User Detection Based on Expectation Propagation

Khac-Hoang Ngo*†, Maxime Guillaud*, Alexis Decurninge*, Sheng Yang†, Subrata Sarkar‡, Philip Schniter‡
*Mathematical and Algorithmic Sciences Lab., Huawei Technologies France, 92100 Boulogne-Billancourt, France
†Laboratory of Signals and Systems, CentraleSupélec, 91190 Gif-sur-Yvette, France
‡Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA
Email: {ngo.khac.hoang, maxime.guillaud, alexis.decurninge}@huawei.com,
sheng.yang@centralesupelec.fr, {sarkar.51,schniter.1}@osu.edu

***Abstract*—In this paper, we propose a novel soft-output multi-user detector for non-coherent multiple access with Grassmannian signaling under Rayleigh block fading. Our detector is based on expectation propagation (EP) approximate inference and has polynomial complexity in the number of users. A simplified version of this scheme coincides with a scheme based on soft minimum-mean-square-error (MMSE) estimation and successive interference cancellation (SIC). Both schemes, especially EP, produce accurate approximates of the true posterior. They outperform a baseline decoder based on projecting the received signal onto the subspace orthogonal to the interference in terms of both hard-detected symbol error rate and coded bit error rate.**

***Index Terms*—non-coherent communications, multiple access, expectation propagation, Grassmannian constellations**

I. Introduction

In wireless communications, multiple-input multiple-output (MIMO) technology is capable of improving significantly both the system spectral efficiency and reliability [1], [2]. In practical MIMO systems, the transmitted symbols are normally drawn from a finite discrete constellation. The task of the receiver is to detect these symbols based on the received signal and available channel information. If the *instantaneous* value of the channel matrix is treated as known, such as when it is obtained via channel estimation, the detection problem is said to be *coherent* and has been investigated extensively in the literature [3]. If only *statistical* information about the channel is available, the detection problem is said to be *non-coherent*.

In the non-coherent case, the transmitted symbols are typically structured, e.g., using differential encoding, or such that (s.t.) the matrix of symbols in the space-time domain is orthonormal and isotropically distributed [4]. The latter was proposed for the block fading channel where the channel matrix remains constant for each coherence block of T symbols and varies independently between blocks. There, information is carried in the subspace of the signal matrix, which is invariant to multiplication by the channel matrix. Thus, a non-coherent constellation can be designed as a collection of points in the Grassmann manifold $G(\mathbb{C}^T, K)$, which is the space of K-dimensional subspaces in $\mathbb{C}^T$, where K is the number of transmit antennas. This was shown to be capacity-achieving at high signal-to-noise-ratio (SNR) for the Rayleigh block fading channel [5]. The optimal maximum-likelihood (ML) detector is NP-hard, thus low-complexity sub-optimal detectors have been proposed for Grassmannian constellations with additional structure, e.g., [6], [7].

In this paper, we focus on non-coherent detection in the single-input multiple-output (SIMO) multiple-access channel with K single-antenna users under flat and block Rayleigh fading with coherence time T. The transmitted signals are constructed from *disjoint* Grassmannian constellations in $G(\mathbb{C}^T, 1)$. The receiver is interested not only in the hard detections of the symbols but also in their posterior marginals to, e.g., compute the bit-wise log-likelihood ratios (LLRs) required for channel decoding. Exact posterior marginalization is prohibitive with many users 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 [3], the probabilistic non-coherent MIMO detection has not been well investigated. The detection scheme in [8] decouples the multi-user detection into K single-user detection problems, but it is sub-optimal and compatible only with the constellation structure therein. The list-based soft demapper in [9] 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 multi-user case.

In this work, we propose message-passing algorithms for posterior marginal inference in non-coherent multi-user MIMO channels. Our algorithms are based on expectation propagation (EP) approximate inference [10]. For EP, we build a factor graph whose variable nodes correspond to the noiseless received signal vectors and the Grassmannian symbol indices. The EP algorithm passes messages between these variable nodes and the corresponding factor nodes. We also propose a simplification of this scheme that can be interpreted as soft MMSE estimation and successive interference cancellation (SIC).

We numerically compare the performance of our EP and MMSE-SIC detectors to the optimal ML detector (when possible), a genie-aided detector, the conventional coherent detector, and the state-of-the-art detector from [8]. We find that EP and MMSE-SIC achieve near-optimal symbol error rate and coded bit error rate. To the best of our knowledge, these are the first message-passing schemes for non-coherent multi-user MIMO detection with Grassmannian signaling.

The remainder of this paper is organized as follows. We

present the system model in Section II. A brief review of EP is presented in Section III, and the EP approach to the non-coherent MIMO detection is presented in Section IV. In Section V, a MMSE-SIC detector is presented as a simplification of the EP detector. Numerical results are presented in Section VI, and conclusions are presented in Section VII.

Notation: We denote vectors and matrices with italic bold letters in respectively lowercase and uppercase, e.g., 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 trace, transpose, and conjugated transpose of $\boldsymbol{M}$ are $\mathrm{tr}\{\boldsymbol{M}\}$, $\boldsymbol{M}^{\mathsf{T}}$ and $\boldsymbol{M}^{\mathsf{H}}$, respectively. $\mathbb{1}\{\cdot\}$ is the indicator function. $\mathbf{0}$ denotes the all-zero vectors/matrices. $[n] := \{1, 2, \ldots, n\}$. The Grassmann manifold $G(\mathbb{C}^T, K)$ is defined as the space of K-dimensional subspaces in $\mathbb{C}^T$. In particular, $G(\mathbb{C}^T, 1)$ is the Grassmannian of lines. $D(q\|p)$ denotes the Kullback-Leibler (KL) divergence between two distributions p and q. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian vector distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and thus probability density function (pdf) $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{\exp(-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{\pi^n \det(\boldsymbol{\Sigma})}$, $\boldsymbol{x} \in \mathbb{C}^n$.

II. System Model

We consider a SIMO multiple access channel in which K single-antenna users transmit to a receiver having N antennas. We assume that the channel between the receiver and each user is flat and block fading with an equal-length and synchronous (across the users) coherence interval of T symbols. That is, the channel vectors $\boldsymbol{h}_k \in \mathbb{C}^{N\times 1}$ between the transmit antenna of user k and the N receive antennas remain constant within each coherence block of $T > 1$ symbols, and change independently between blocks. The distribution of $\boldsymbol{h}_k$ is assumed to be known to the receiver, but its realizations are *unknown* to both the receiver and users. We consider independent and identically distributed (i.i.d.) Rayleigh fading, i.e., $\boldsymbol{h}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$.

Within a coherence block, each user k sends a signal vector $\boldsymbol{s}_k \in \mathbb{C}^T$, and the receiver receives a $T \times N$ signal matrix

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

where $\boldsymbol{S} = [\boldsymbol{s}_1 \ldots \boldsymbol{s}_K] \in \mathbb{C}^{T\times K}$ and $\boldsymbol{H} = [\boldsymbol{h}_1 \ldots \boldsymbol{h}_K] \in \mathbb{C}^{N\times K}$ concatenate the transmitted signals and channel vectors, respectively, $\boldsymbol{W}$ is the Gaussian noise with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, and the block index is omitted for simplicity. We assume that the transmitted signals have average unit norm, i.e., $\mathbb{E}\left[\|\boldsymbol{s}_k\|^2\right] = 1$, $\forall k$. Under this normalization, the SNR of each transmitted signal at each receive antenna is $\mathrm{SNR} = 1/(T\sigma^2)$. We assume that the transmitted signals belong to *disjoint* finite individual Grassmannian constellations in $G(\mathbb{C}^T, 1)$. That is, $\boldsymbol{s}_k \in \mathcal{S}_k := \{\boldsymbol{s}_k^{(1)}, \ldots, \boldsymbol{s}_k^{(|\mathcal{S}_k|)}\}$, where each symbol $\boldsymbol{s}_k^{(i)}$ is a unit-norm vector representative of a point in $G(\mathbb{C}^T, 1)$.

Given $\boldsymbol{S}$, the matrix $\boldsymbol{Y}$ is Gaussian with independent columns having the same covariance matrix $\sigma^2\boldsymbol{I}_T + \boldsymbol{S}\boldsymbol{S}^{\mathsf{H}}$. Thus,

$$p(\boldsymbol{Y}|\boldsymbol{S}) = \frac{\exp\left(-\mathrm{tr}\left\{\boldsymbol{Y}^{\mathsf{H}}(\sigma^2\boldsymbol{I}_T + \boldsymbol{S}\boldsymbol{S}^{\mathsf{H}})^{-1}\boldsymbol{Y}\right\}\right)}{\pi^{NT}\det^N(\sigma^2\boldsymbol{I}_T + \boldsymbol{S}\boldsymbol{S}^{\mathsf{H}})}. \tag{2}$$

When a channel code is used, most channel decoders require the LLR of the bits computed from the posteriors $p(\boldsymbol{s}_k|\boldsymbol{Y})$, $k \in [K]$, which are marginalized from

$$p(\boldsymbol{S}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{S})p(\boldsymbol{S})}{p(\boldsymbol{Y})} \propto p(\boldsymbol{Y}|\boldsymbol{S})p(\boldsymbol{S}). \tag{3}$$

Assuming that the transmitted signals are independent and uniformly distributed over the discrete constellations, the prior $p(\boldsymbol{S})$ factorizes as $p(\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K]) = \prod_{k=1}^{K} \frac{1}{|\mathcal{S}_k|}\mathbb{1}\{\boldsymbol{s}_k \in \mathcal{S}_k\}$. On the other hand, the likelihood function $p(\boldsymbol{Y}|\boldsymbol{S})$ involves all the signals $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K$ in a manner that does not easily factorize. Exact marginalization of $p(\boldsymbol{S}|\boldsymbol{Y})$ requires computing

$$p(\boldsymbol{s}_k|\boldsymbol{Y}) = \sum_{\boldsymbol{s}_l \in \mathcal{S}_l, \forall l \neq k} p(\boldsymbol{S}|\boldsymbol{Y}), \quad \text{for } k \in [K]. \tag{4}$$

This becomes formidable in the case of many users or large constellations. Thus, we seek a low-complexity approximation

$$p(\boldsymbol{S}|\boldsymbol{Y}) \approx \hat{p}(\boldsymbol{S}|\boldsymbol{Y}) = \prod_{k=1}^{K} \hat{p}(\boldsymbol{s}_k|\boldsymbol{Y}). \tag{5}$$

In what follows, we design a posterior marginal estimation scheme based on expectation propagation (EP).

III. Expectation Propagation

EP was proposed in [10] for approximate inference in probabilistic graphical models. Let us consider a set of variables contained in a random vector $\boldsymbol{x}$ with posterior of the form

$$p(\boldsymbol{x}) \propto \prod_{\alpha} \psi_\alpha(\boldsymbol{x}_\alpha), \tag{6}$$

where $\boldsymbol{x}_\alpha$ is the subset of variables involved in the factor ψ_α. Let us partition the components of $\boldsymbol{x}$ into some sets $\{\boldsymbol{x}_\beta\}$, where no $\boldsymbol{x}_\beta$ is split across factors (i.e., $\forall\ \alpha, \beta$ either $\boldsymbol{x}_\beta \subset \boldsymbol{x}_\alpha$ or $\boldsymbol{x}_\beta \cap \boldsymbol{x}_\alpha = \emptyset$). We are interested in the posterior marginals with respect to (w.r.t.) the partition $\{\boldsymbol{x}_\beta\}$.

EP approximates the true posterior p from (6) by a distribution $\hat{p}$ that can be expressed in two ways. First, it can be w.r.t. the "target" partition $\{\boldsymbol{x}_\beta\}$ as

$$\hat{p}(\boldsymbol{x}) = \prod_{\beta} \hat{p}_\beta(\boldsymbol{x}_\beta), \tag{7}$$

where $\hat{p}_\beta$ are constrained to be in the exponential family so that $\hat{p}_\beta(\boldsymbol{x}_\beta) = \exp\left(\boldsymbol{\gamma}_\beta^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta) - A_\beta(\boldsymbol{\gamma}_\beta)\right)$, 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 $\{\boldsymbol{x}_\alpha\}$ in accordance with (6) as

$$\hat{p}(\boldsymbol{x}) \propto \prod_{\alpha} m_\alpha(\boldsymbol{x}_\alpha). \tag{8}$$

For (7) and (8) to be consistent, there must exist factors $m_{\alpha,\beta}$ of the form $m_{\alpha,\beta}(\boldsymbol{x}_\beta) = \exp\left(\boldsymbol{\gamma}_{\alpha,\beta}^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)\right)$ such that

$$m_\alpha(\boldsymbol{x}_\alpha) = \prod_{\beta\in\mathfrak{N}_\alpha} m_{\alpha,\beta}(\boldsymbol{x}_\beta) = \exp\Bigg(\sum_{\beta\in\mathfrak{N}_\alpha} \boldsymbol{\gamma}_{\alpha,\beta}^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)\Bigg), \tag{9}$$

$$\hat{p}_\beta(\boldsymbol{x}_\beta) \propto \prod_{\alpha\in\mathfrak{N}_\beta} m_{\alpha,\beta}(\boldsymbol{x}_\beta) = \exp\Bigg(\sum_{\alpha\in\mathfrak{N}_\beta} \boldsymbol{\gamma}_{\alpha,\beta}^{\mathsf{T}}\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)\Bigg), \tag{10}$$

where $\mathfrak{N}_\alpha$ collects the indices β for which $\boldsymbol{x}_\beta \subset \boldsymbol{x}_\alpha$, and $\mathfrak{N}_\beta$ collects the indices α for which $\boldsymbol{x}_\beta \subset \boldsymbol{x}_\alpha$. 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.

EP works by first initializing all $m_\alpha(\boldsymbol{x}_\alpha)$ and $\hat{p}_\beta(\boldsymbol{x}_\beta)$ then iteratively updating each m_α in turn. Let us fix a factor index α. We construct the "tilted" distribution q_α by swapping ψ_α for its approximate m_α in $\hat{p}(\boldsymbol{x})$ as $q_\alpha(\boldsymbol{x}) = \frac{\hat{p}(\boldsymbol{x})\psi_\alpha(\boldsymbol{x}_\alpha)}{m_\alpha(\boldsymbol{x}_\alpha)}$, and then project it back onto the exponential family by solving

$$\hat{p}_\alpha^{\text{new}}(\boldsymbol{x}) = \prod_\beta \hat{p}_{\alpha,\beta}^{\text{new}}(\boldsymbol{x}_\beta) = \arg\min_{\underline{p}\in\mathcal{P}} D\big(q_\alpha(\boldsymbol{x}) \,\big\|\, \underline{p}(\boldsymbol{x})\big), \quad (11)$$

where $\mathcal{P}$ is the set of distributions with the form of $\hat{p}$ in (7). After some manipulations following [10], we deduce that for each $\beta \in \mathfrak{N}_\alpha$, the optimal $\hat{p}_{\alpha,\beta}^{\text{new}}$ is the moment match of $q_{\alpha,\beta}$ in the exponential family with sufficient statistics $\boldsymbol{\phi}_\beta(\boldsymbol{x}_\beta)$, where

$$q_{\alpha,\beta}(\boldsymbol{x}_\beta) := \int \psi_\alpha(\boldsymbol{x}_\alpha) \left[\prod_{\beta\in\mathfrak{N}_\alpha} \prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)\right] \mathrm{d}\boldsymbol{x}_{\alpha\setminus\beta} \quad (12)$$

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 $\boldsymbol{x}_\beta$. For $\beta \notin \mathfrak{N}_\alpha$, the optimal $\hat{p}_{\alpha,\beta}^{\text{new}}$ is simply $\hat{p}_\beta(\boldsymbol{x}_\beta)$. The factor m_α is then updated via

$$m_\alpha^{\text{new}}(\boldsymbol{x}_\alpha) = \frac{\hat{p}_\alpha^{\text{new}}(\boldsymbol{x}) m_\alpha(\boldsymbol{x}_\alpha)}{\hat{p}(\boldsymbol{x})} \propto \prod_{\beta\in\mathfrak{N}_\alpha} m_{\alpha,\beta}^{\text{new}}(\boldsymbol{x}_\beta), \quad (13)$$

$$\text{with} \quad m_{\alpha,\beta}^{\text{new}}(\boldsymbol{x}_\beta) := \frac{\hat{p}_{\alpha,\beta}^{\text{new}}(\boldsymbol{x}_\beta)}{\prod_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha} m_{\alpha',\beta}(\boldsymbol{x}_\beta)}. \quad (14)$$

Observe that the update of m_α only affects the approximate posterior of the variable nodes β in the neighborhood of factor node α. Equation (14) says that the new message $m_{\alpha,\beta}^{\text{new}}$ passed from α to $\beta \in \mathfrak{N}_\alpha$ equals $\hat{p}_{\alpha,\beta}^{\text{new}}$ divided by the message product $\{m_{\alpha',\beta}\}_{\alpha'\in\mathfrak{N}_\beta\setminus\alpha}$, i.e., previous messages to β from all directions except α. After that, the process is repeated with the next α.

IV. Application of EP to Non-Coherent Detection

To apply EP to the problem of non-coherent detection, we express the signal of user k as $\boldsymbol{s}_k = \boldsymbol{s}_k^{(i_k)}$, where i_k are random symbol indices that are independent and uniformly distributed over $[|\mathcal{S}_k|]$. We rewrite (1) in vector form as

$$\boldsymbol{y} = \sum_{k=1}^K \boldsymbol{z}_k + \boldsymbol{w}, \quad (15)$$

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

With $\boldsymbol{z} := [\boldsymbol{z}_1^\mathsf{T}, \ldots, \boldsymbol{z}_K^\mathsf{T}]^\mathsf{T}$ and $\boldsymbol{i} := [i_1, \ldots, i_k]^\mathsf{T}$, we can write

$$\begin{aligned} p(\boldsymbol{i}, \boldsymbol{z}|\boldsymbol{y}) &\propto p(\boldsymbol{i}, \boldsymbol{z}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{i})p(\boldsymbol{i}) \\ &= \psi_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K)\left[\prod_{k=1}^K \psi_{k1}(\boldsymbol{z}_k, i_k)\right]\left[\prod_{k=1}^K \psi_{k2}(i_k)\right], \quad (16) \end{aligned}$$

corresponding to (6), where

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

We will use EP to infer the posterior distribution of the indices $\{i_k\}$. To do so, we choose the partition $\boldsymbol{x} = \{\boldsymbol{z}_k, i_k\}_{k=1}^K$ and illustrate the interaction between these variables and the factors ψ_0, ψ_{k1}, and ψ_{k2} by the bipartite factor graph in Fig. 1. This graph has a tree structure with a root $\boldsymbol{y}$ and K leaves $\{\psi_{k2}\}_{k=1}^K$.

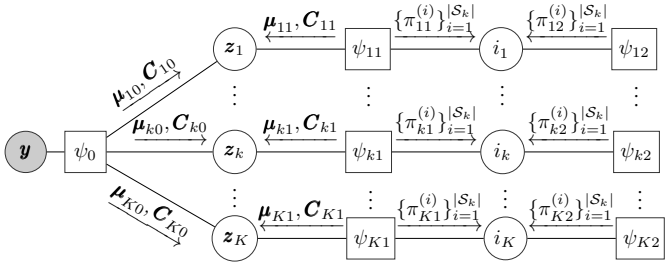

Fig. 1. A factor graph representation of the non-coherent detection problem.

We write the EP approximation according to (7) as

$$\hat{p}(\boldsymbol{x}|\boldsymbol{y}) = \hat{p}(\boldsymbol{i}, \boldsymbol{z}|\boldsymbol{y}) = \prod_{k=1}^K \hat{p}_k(\boldsymbol{z}_k)\hat{p}_k(i_k), \quad (17)$$

where $\hat{p}_k(\boldsymbol{z}_k)$ and $\hat{p}_k(i_k)$ are implicitly conditioned on $\boldsymbol{y}$ and constrained to be a Gaussian vector distribution and a discrete distributions with support $[|\mathcal{S}|]$, respectively, i.e.,

$$\hat{p}_k(\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{z}_k; \hat{\boldsymbol{z}}_k, \boldsymbol{\Sigma}_k) \;\; \text{s.t.} \;\; \boldsymbol{\Sigma}_k \text{ is positive definite}, \quad (18)$$

$$\hat{p}_k(i_k) = \hat{\pi}_k^{(i_k)} \;\; \text{for } i_k \in [|\mathcal{S}_k|] \;\; \text{s.t.} \;\; \sum_{i=1}^{|\mathcal{S}_k|} \hat{\pi}_k^{(i)} = 1. \quad (19)$$

We also write the EP approximation according to (8) as

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

where $m_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K) \propto \prod_{k=1}^K \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})$, $m_{k1}(\boldsymbol{z}_k, i_k) \propto \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})\pi_{k1}^{(i_k)}$, and $m_{k2}(i_k) = \pi_{k2}^{(i_k)}$ for $i_k \in [|\mathcal{S}_k|]$. On the factor graph in Fig. 1, we can interpret $(\boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})$ as the message from factor node ψ_0 to variable node $\boldsymbol{z}_k$, $(\boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})$ as the message from node ψ_{k1} to node $\boldsymbol{z}_k$, $\{\pi_{k1}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ as the message from node ψ_{k1} to node i_k, and $\{\pi_{k2}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ as the message from node ψ_{k2} to node i_k.

1) The EP message updates: Following (12) and (14), we derive the messages as follows.[1] First, the message $\{\pi_{k2}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ from node ψ_{k2} to node i_k is simply $\pi_{k2}^{(i_k)} = \frac{1}{|\mathcal{S}_k|}$ for $i_k \in [|\mathcal{S}_k|]$.

[1] A full derivation can be found in the long version [11].

The message $\{\pi_{k1}^{(i_k)}\}_{i_k=1}^{|\mathcal{S}_k|}$ from node ψ_{k1} to node i_k is given by

$$\pi_{k1}^{(i_k)} = \frac{\mathcal{N}\big(\mathbf{0}; \boldsymbol{\mu}_{k0}, (\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}}) \otimes \boldsymbol{I}_N + \boldsymbol{C}_{k0}\big)}{\sum_{i=1}^{|\mathcal{S}_k|} \mathcal{N}\big(\mathbf{0}; \boldsymbol{\mu}_{k0}, (\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \boldsymbol{I}_N + \boldsymbol{C}_{k0}\big)}, \quad (20)$$

for $i_k \in [|\mathcal{S}_k|]$. The message $(\boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})$ from node ψ_{k1} to nodes $\boldsymbol{z}_k$ is given by

$$\boldsymbol{C}_{k1} = \big(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{C}_{k0}^{-1}\big)^{-1}, \ \boldsymbol{\mu}_{k1} = \boldsymbol{C}_{k1}\big(\boldsymbol{\Sigma}_k^{-1}\hat{\boldsymbol{z}}_k - \boldsymbol{C}_{k0}^{-1}\boldsymbol{\mu}_{k0}\big), \quad (21)$$

where

$$\hat{\boldsymbol{z}}_k = \sum_{i=1}^{|\mathcal{S}_k|} \pi_{k1}^{(i)} \hat{\boldsymbol{z}}_{ki}, \ \boldsymbol{\Sigma}_k = \sum_{i=1}^{|\mathcal{S}_k|} \pi_{k1}^{(i)} (\hat{\boldsymbol{z}}_{ki}\hat{\boldsymbol{z}}_{ki}^{\mathsf{H}} + \boldsymbol{\Sigma}_{ki}) - \hat{\boldsymbol{z}}_k\hat{\boldsymbol{z}}_k^{\mathsf{H}}, \quad (22)$$

with $\boldsymbol{\Sigma}_{ki} = \Big([(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \boldsymbol{I}_N]^{-1} + \boldsymbol{C}_{k0}^{-1}\Big)^{-1}$, and $\hat{\boldsymbol{z}}_{ki} = \boldsymbol{\Sigma}_{ki}\boldsymbol{C}_{k0}^{-1}\boldsymbol{\mu}_{k0}$. Finally, the message $(\boldsymbol{\mu}_{k0}, \boldsymbol{C}_{k0})$ from node ψ_0 to node $\boldsymbol{z}_k$ is given by

$$\boldsymbol{C}_{k0} = \sigma^2\boldsymbol{I}_{NT} + \sum_{j\neq k} \boldsymbol{C}_{j1}, \quad \boldsymbol{\mu}_{k0} = \boldsymbol{y} - \sum_{j\neq k} \boldsymbol{\mu}_{j1}. \quad (23)$$

2) Initialization of the EP messages: We choose the non-informative initialization $\boldsymbol{C}_{k0}^{-1} = \mathbf{0}$ and $\boldsymbol{\mu}_{k0} = \mathbf{0}$, so that, from (20), the initial message from node ψ_{k1} to node i_k coincides with the uniform prior $\pi_{k1}^{(i)} = \frac{1}{|\mathcal{S}_k|}$ for $i \in [|\mathcal{S}_k|]$; the initial parameters $\boldsymbol{\Sigma}_{ki} = (\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \boldsymbol{I}_N$ and $\boldsymbol{z}_{ki} = \mathbf{0}$. This leads to the initial parameters of $\hat{p}_k(\boldsymbol{z}_k)$ from (22) as $\hat{\boldsymbol{z}}_k = \mathbf{0}$, and $\boldsymbol{\Sigma}_k = \frac{1}{|\mathcal{S}_k|}\sum_{i=1}^{|\mathcal{S}_k|}(\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \boldsymbol{I}_N$, and the initial message from node ψ_{k1} to node $\boldsymbol{z}_k$ given in (21) as $\boldsymbol{C}_{k1} = \boldsymbol{\Sigma}_k$, and $\boldsymbol{\mu}_{k1} = \hat{\boldsymbol{z}}_k$. Finally, the initial messages from node ψ_0 to node $\boldsymbol{z}_k$ follows from (23) as $\boldsymbol{C}_{k0} = \sigma^2\boldsymbol{I}_{NT} + \sum_{j\neq k}\frac{1}{|\mathcal{S}_j|}\sum_{i=1}^{|\mathcal{S}_j|}(\boldsymbol{s}_j^{(i)}\boldsymbol{s}_j^{(i)\mathsf{H}}) \otimes \boldsymbol{I}_N$, and $\boldsymbol{\mu}_{k0} = \boldsymbol{y}$.

After the initialization, the EP algorithm proceeds by iteratively updating the messages. In particular, it goes through the branches of the tree graph in Fig. 1 in a round-robin manner, and in each branch, the factor nodes are visited in the order from leaf to root (other message passing schedulings can be implemented). In the end, according to (10) and (19), the estimated pmf of $\hat{p}(\boldsymbol{s}_k|\boldsymbol{Y})$ is $\hat{p}_k(i_k) = \hat{\pi}_k^{(i_k)} \propto \pi_{k1}^{(i_k)}\pi_{k2}^{(i_k)}$, that is $\hat{p}_k(i_k) = \pi_{k1}^{(i_k)}$ since $\pi_{k2}^{(i_k)}$ is constant over $i_k \in [|\mathcal{S}_k|]$.

V. MMSE-SIC: A Simplification of EP

In the EP message updates, if we replace (21) by

$$\boldsymbol{\mu}_{k1} = \mathbf{0} \ \text{ and } \ \boldsymbol{C}_{k1} = \sum_{i=1}^{|\mathcal{S}_k|} \pi_{k1}^{(i)} (\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}) \otimes \boldsymbol{I}_N, \quad (24)$$

which arises by skipping a projection onto the Gaussian family in the derivation of $(\boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k1})$, it follows from (23) that $\boldsymbol{\mu}_{k0} = \boldsymbol{y}$ and $\boldsymbol{C}_{k0} = \sigma^2\boldsymbol{I}_{NT} + \sum_{j\neq k}\sum_{i=1}^{|\mathcal{S}_j|}\pi_{j1}^{(i)}(\boldsymbol{s}_j^{(i)}\boldsymbol{s}_j^{(i)\mathsf{H}}) \otimes \boldsymbol{I}_N$. Let $\boldsymbol{R}_k := \sum_{i=1}^{|\mathcal{S}_k|}\pi_{k1}^{(i)}\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}}$ and $\boldsymbol{Q}_k := \sum_{l\neq k}\boldsymbol{R}_l + \sigma^2\boldsymbol{I}_T$, then $\boldsymbol{C}_{k1} = \boldsymbol{R}_k \otimes \boldsymbol{I}_N$ and $\boldsymbol{C}_{k0} = \boldsymbol{Q}_k \otimes \boldsymbol{I}_N$. It follows that the posterior update (20) of the EP scheme can be written as

$$\pi_{k1}^{(i_k)} = \frac{\mathcal{N}\Big(\mathbf{0}; \boldsymbol{y}, (\boldsymbol{s}_k^{(i_k)}\boldsymbol{s}_k^{(i_k)\mathsf{H}} + \boldsymbol{Q}_k) \otimes \boldsymbol{I}_N\Big)}{\sum_{i=1}^{|\mathcal{S}_k|}\mathcal{N}\Big(\mathbf{0}; \boldsymbol{y}, (\boldsymbol{s}_k^{(i)}\boldsymbol{s}_k^{(i)\mathsf{H}} + \boldsymbol{Q}_k) \otimes \boldsymbol{I}_N\Big)}, i_k \in [|\mathcal{S}_k|]. \quad (25)$$

This simplified scheme can be alternatively constructed as follows. In the channel output (15), the interference from other users while decoding the signal of user k is $\boldsymbol{t}_k := \sum_{l\neq k}\boldsymbol{z}_l$ with mean $\mathbb{E}[\boldsymbol{t}_k] = \mathbf{0}$ and covariance matrix $\mathbb{E}[\boldsymbol{t}_k\boldsymbol{t}_k^{\mathsf{H}}] = \sum_{l\neq k}\mathbb{E}[\boldsymbol{s}_l\boldsymbol{s}_l^{\mathsf{H}}] \otimes \boldsymbol{I}_N = \sum_{l\neq k}\boldsymbol{R}_l \otimes \boldsymbol{I}_N$. If we treat $\boldsymbol{t}_k$ as a Gaussian vector with the same mean and covariance matrix, then $\boldsymbol{t}_k + \boldsymbol{w} \sim \mathcal{N}\big(\mathbf{0}, \boldsymbol{Q}_k \otimes \boldsymbol{I}_N\big)$. Since $\boldsymbol{y} = (\boldsymbol{s}_k \otimes \boldsymbol{I}_N)\boldsymbol{h}_k + \boldsymbol{t}_k + \boldsymbol{w}$, the single-user likelihood under this approximation is

$$\hat{p}(\boldsymbol{y}|\boldsymbol{s}_k) = \mathcal{N}\left(\boldsymbol{y}; \mathbf{0}, \left(\boldsymbol{s}_k\boldsymbol{s}_k^{\mathsf{H}} + \boldsymbol{Q}_k\right) \otimes \boldsymbol{I}_N\right). \quad (26)$$

Then, the update of the approximate posterior $\hat{p}(\boldsymbol{s}_k|\boldsymbol{y}) \propto \hat{p}(\boldsymbol{y}|\boldsymbol{s}_k)$ coincides with (25). $\boldsymbol{R}_k$ is then recalculated with the updated value of $\hat{p}(\boldsymbol{s}_k|\boldsymbol{y})$, and $\boldsymbol{Q}_l$, $l \neq k$, are updated accordingly. This is done for each user $k \in [K]$, and then the next iteration starts.

In short, the derived simplification of the EP scheme iteratively MMSE-estimates the signal $\boldsymbol{z}_k$ 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-SIC. As for the general EP scheme, we can start with the non-informative initialization $\hat{p}(\boldsymbol{s}_k|\boldsymbol{Y}) = \frac{1}{|\mathcal{S}_k|}\mathbb{1}\{\boldsymbol{s}_k \in \mathcal{S}_k\}$.

VI. Performance Evaluation

We evaluate the performance of the proposed schemes for a given set of individual constellations with $|\mathcal{S}_k| = 2^B$, $\forall k \in [K]$. We consider the design in [8], which generates $\mathcal{S}_k$ as $\boldsymbol{s}_k^{(i)} = \frac{\boldsymbol{U}_k\boldsymbol{d}^{(i)}}{\|\boldsymbol{U}_k\boldsymbol{d}^{(i)}\|}, i \in [2^B]$, where $\boldsymbol{U}_k \in \mathbb{C}^{T\times(T-K+1)}$ is a full-rank precoder defined uniquely for user k and $\mathcal{D} = \{\boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(2^B)}\}$ is a Grassmannian constellation in $G(\mathbb{C}^{T-K+1}, 1)$. We consider the precoders $\boldsymbol{U}_k$ defined in [8, Eq.(11)] and the cube-split constellation proposed in [7] for $\mathcal{D}$. This structured constellation has good packing properties, allows for low-complexity single-user decoding and a simple yet effective binary labeling scheme. We take the binary label of $\boldsymbol{d}^{(i)}$ for $\boldsymbol{s}_k^{(i)}$, $\forall k$. Exploiting the precoder structure, [8] introduces a detector [8, Sec.V-B-3] that 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).

We set the number of iterations of EP and MMSE-SIC as 20, and of POCIS as 3 since it quickly converges.[2]

First, in Fig. 2, we plot the hard-detected symbol error rate (SER) of EP, MMSE-SIC, and POCIS for $T = 6$, $K = 3$, $N = 8$, and $B = 8$. For a benchmark, since the optimal joint ML detector is computationally infeasible, we consider a genie-aided detector consisting in giving the receiver, while it decodes $\boldsymbol{s}_k$, the knowledge of the signals $\boldsymbol{s}_l$ (but not the channels $\boldsymbol{h}_l$) of all interfering users $l \neq k$. The performance of EP is very close to this genie-aided detector and better than MMSE-SIC at $\mathrm{SNR} \geq 10$ dB. Both EP and MMSE-SIC are better than POCIS. We also show the SER of a non-coherent time division multiple access (TDMA) scheme where each user transmits from a

[2]To stabilize, we damp the update of $\boldsymbol{C}_{k1}, \boldsymbol{\mu}_{k1}, \boldsymbol{C}_{k0}, \boldsymbol{\mu}_{k0}$ in EP and of $\boldsymbol{R}_k, \boldsymbol{Q}_k$ in MMSE-SIC.

cube-split constellation of size 2^{BK} in $G(\mathbb{C}^T, 1)$ in a round-robin manner. We also show a coherent pilot-based scheme with quadrature amplitude modulation signals, MMSE channel estimation, and MMSE symbol equalization. These latter two schemes are outperformed by the non-coherent multiple-access scheme [8] with EP, MMSE-SIC, and POCIS detectors.

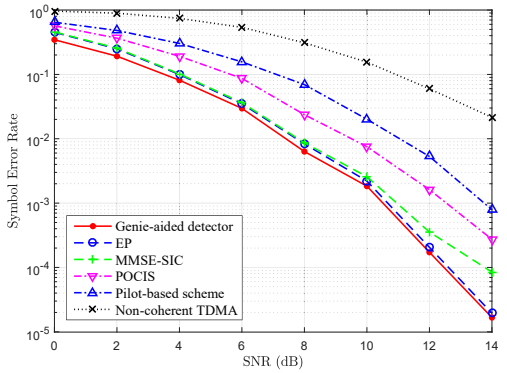

Fig. 2. The symbol error rate of EP, MMSE-SIC, POCIS, and a genie-aided detector for $T = 6$, $K = 3$, $N = 8$ in comparison with a pilot-based scheme and non-coherent TDMA for the same transmission rate of 8 bits/user/block.

Next, we integrate a rate-$1/3$ turbo code. The turbo encoder accepts packets of 1008 bits; the turbo decoder computes the bit-wise LLRs from the detector's soft outputs and performs 10 decoding iterations. In Fig. 3, 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 with exact marginalization (4). The BER of MMSE-SIC vanishes slower with SNR than the other schemes, and becomes better than POCIS as K grows. For $T = 7$ and $K = N = 4$, the power gain of EP w.r.t. MMSE-SIC and POCIS for the same BER of 10^{-3} is about 3 dB and 4 dB, respectively.

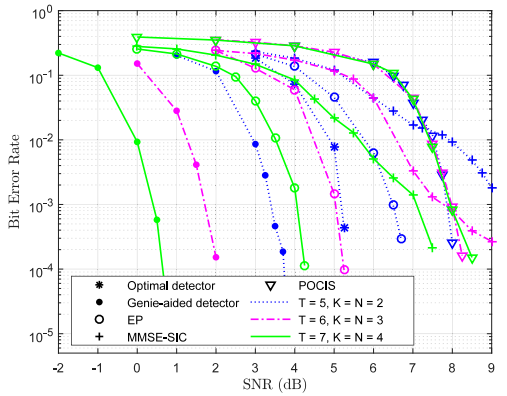

Fig. 3. The bit error rate with turbo codes of EP, MMSE-SIC, POCIS, and the optimal/genie-aided detector for $B = 8$ bits/symbol and $K = N$.

Finally, in Fig. 4, we compare the BER with the same turbo code with different constellation sizes for $T = 6$, $K = 3$, and $N = 4$. For $B = 5$, i.e., small constellations, MMSE-SIC can be slightly better than EP (both have performance close to the optimal detector). This is because it can happen that all the mass of the pmf $\pi_{k1}^{(i_k)}$ is concentrated on a possibly wrong symbol at early iterations, and EP may not be able to refine significantly the pmf 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. 3, the performance of MMSE-SIC is significantly improved as the number of receive antennas N increases from 3 to 4.

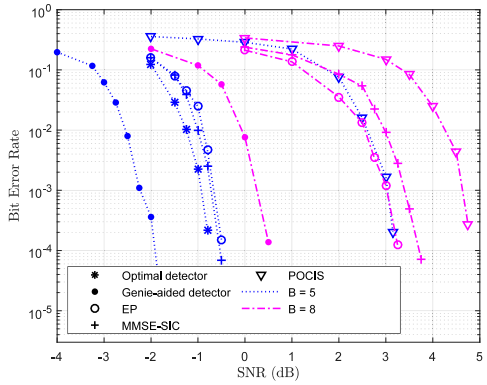

Fig. 4. The bit error rate with turbo codes of EP, MMSE-SIC, POCIS, and the optimal/genie-aided detector for $T = 6$, $K = 3$, and $N = 4$.

VII. Conclusion

We proposed an expectation propagation based scheme and a MMSE-SIC scheme for soft-output multi-user detection in non-coherent SIMO communications. The latter scheme can be interpreted as a simplification of the former. Both schemes are shown to achieve good performance, especially the EP scheme, in terms of symbol error rate when they are used for hard detection, and bit error rate when used for channel decoding.

References

[1] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecommun.*, vol. 10, pp. 585–595, Nov./Dec. 1999.

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

[3] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 1941–1988, Fourthquarter 2015.

[4] 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.

[5] 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.

[6] 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.

[7] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured Grassmannian constellation for non-coherent SIMO communications," *arXiv preprint arXiv:1905.08745*, 2019, submitted to IEEE Trans. Wireless Commun.

[8] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "A multiple access scheme for non-coherent SIMO communications," in *52nd Asilomar Conference on Signals, Systems, and Computers*, CA, USA, Oct. 2018, pp. 1846–1850.

[9] 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.

[10] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, Jan. 2001.

[11] K.-H. Ngo, M. Guillaud, A. Decurninge, S. Yang, and P. Schniter, "Multi-user detection based on expectation propagation for the non-coherent SIMO multiple access channel," *submitted to IEEE Trans. Wireless Commun.*, 2019, (arXiv preprint arXiv:1905.11152).