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

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

**Abstract**

We consider the non-coherent single-input multiple-output (SIMO) multiple access channel with Grassmannian signaling under Rayleigh block fading. We propose a novel soft-output multi-user detector that computes the marginalized 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. A simplification of this detector, with reduced complexity, coincides with soft minimum-mean-square-error (MMSE) estimation and successive interference cancellation (SIC). Both detectors produce accurate approximates of the true posterior which lead to a high mismatched sum-rate of the system that uses the approximate posterior as the decoding metric. The proposed detectors also outperform, in terms of symbol error rate, a conventional coherent pilot-based detector and a state-of-the-art detector based on projecting the received signal onto the subspace orthogonal to the interference. The gain of our EP and MMSE-SIC detectors are further observed in terms of the bit error rate when using their soft outputs for a turbo channel decoder.

Khac-Hoang Ngo is with Mathematical and Algorithmic Sciences Laboratory, Paris Research Center, Huawei Technologies, 92100 Boulogne-Billancourt, France, and 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 Mathematical and Algorithmic Sciences Laboratory, Paris Research Center, Huawei Technologies, 92100 Boulogne-Billancourt, France (e-mail: maxime.guillaud,alexis.decurninge@huawei.com).

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

Subrata Sarkar and Philip Schniter are with Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA (e-mail: {sarkar.51,schniter.1}@osu.edu).

The material in this paper has been submitted in part to the 53rd Asilomar Conference on Signals, Systems, and Computers, CA, USA, 2019.
Index Terms

non-coherent communications, multiple access, 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 [1], [2]. MIMO is at the heart of the technologies incorporated into current cellular systems, and large-scale (massive) MIMO is considered as one of the fundamental technologies for the fifth-generation (5G) wireless communications [3]. 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, such as when it is obtained via channel estimation based on pilot symbols, the detection problem is said to be coherent and has been investigated extensively in the literature [4]. In this case, the data symbols are normally taken from a scalar constellation such as quadrature amplitude modulation (QAM). Since the optimal maximum-likelihood (ML) coherent detection problem is known to be non-deterministic polynomial-time hard (NP-hard) due to the discrete domain of the symbols [5], 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 [4].

If only statistical information about the channel is available, the detection problem is said to be non-coherent. In this case, the transmitted symbols must be structured in a way that permits accurate detection in the face of channel uncertainty. This structure may manifest as differential encoding [6], or codings that constrain the matrix of symbols in the space-time domain to be orthonormal and isotropically distributed [7]. The latter structure was shown to be capacity-achieving in the high-SNR regime for flat Rayleigh block fading channels, where the channel matrix remains constant for each coherence block of $T$ symbols and changes independently between blocks [8], [9]. In this case, information is carried in the subspace of the
transmitted symbol matrix, which is invariant to multiplication by the channel matrix. Therefore, a constellation over matrix-valued symbols can be designed by choosing a collection of subspaces in $\mathbb{C}^T$. Such constellations belong to 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. When the transmitted signal contains pilots, we can also consider pilot-aided channel estimation followed by coherent symbol detection as one form of non-coherent detection. But this latter approach is sub-optimal relative to the Grassmannian signaling [8, Sec. V]. 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 Grassmannian constellations with additional structure, e.g., [10], [11], [12].

In this paper, we focus on the non-coherent MIMO detection problem in the Rayleigh flat and block fading 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 optimal joint signaling scheme would be the Grassmannian signaling on $G(\mathbb{C}^T, K)$. However, we assume uncoordinated users, for which the optimal non-coherent transmission scheme is not known. In this paper, we assume that the transmitted signals are constructed from disjoint Grassmannian constellations in $G(\mathbb{C}^T, 1)$. Furthermore, 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 symbol 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., [13], [14], [15]), the probabilistic non-coherent MIMO detection under Grassmannian signaling has not been well investigated. The detection scheme proposed in [16] decouples the multi-user detection problem into $K$ single-user detection problems by projecting the received signal onto the orthogonal complement of the estimated interference subspace, but it is sub-optimal and compatible only with the constellation structure therein. The list-based soft demapper in [17] 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.
In this work, we propose message-passing algorithms for posterior marginal inference of non-coherent multi-user MIMO transmissions over block Rayleigh fading channels. Our algorithms are based on expectation propagation (EP) approximate inference [18], [19]. EP provides an iterative framework for approximating posterior beliefs by parametric distributions in the exponential family [20, Sec.1.6]. Although there are many possible ways to apply EP to our non-coherent multi-user detection problem, we do so by modeling the unknowns as 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. By updating these messages, the approximate posteriors of these variables are iteratively refined. We also propose a simplification of this EP scheme that can be interpreted as soft MMSE estimation and successive interference cancellation (SIC). Furthermore, we address numerical implementation issues and propose solutions to stabilize the EP updates.

To measure the accuracy of the approximate posterior generated by the EP and MMSE-SIC 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 these schemes for hard detection, and the bit error rate when using these schemes for turbo equalization with a standard turbo code [21].

The contributions of this work are summarized as follows:

1) We propose soft and hard multi-user detectors for the non-coherent single-input multiple-output (SIMO) MAC using EP approximate inference, and we propose methods to stabilize the EP updates.

2) We propose a novel MMSE-SIC detector as a simplification of the EP detector.

3) We analyze the complexity and numerically evaluate the performance of the proposed EP and MMSE-SIC detectors, the optimal ML detector, a genie-aided detector, the existing state-of-the-art detector from [16], and the conventional coherent approach. Our results suggest that the proposed detectors are more complex than existing sub-optimal schemes but offer significantly improved mismatched sum-rate, symbol error rate, and coded bit error rate.

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

The remainder of this paper is organized as follows. The system model and problem formulation
are presented in Section II. A brief review of expectation propagation 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. Implementation aspects of EP and MMSE-SIC are discussed in Section VI. Numerical results are presented in Section VII, and conclusions are presented in Section VIII. The proofs are provided in the appendices.

**Notation:** Random quantities are denoted with non-italic letters with sans-serif fonts, e.g., a scalar $x$, a vector $v$, and a matrix $M$. Deterministic quantities are denoted with italic letters, e.g., a scalar $x$, a vector $v$, and a matrix $M$. The Euclidean norm is denoted by $\|v\|$ and the Frobenius norm $\|M\|_F$. The trace, conjugate, transpose, and conjugated transpose of $M$ are $\text{tr}\{M\}$, $M^*$, $M^T$ and $M^H$, respectively. $\prod$ denotes the conventional or Cartesian product, depending on the factors. $\otimes$ denotes the Kronecker product. $\mathbb{1}\{\cdot\}$ denotes the indicator function. $1$ and $0$ denote respectively the all-one and all-zero vectors/matrices. We use $c_0$ to represent a constant w.r.t. the distributions of interest whose value may change at each occurrence. $[n] := \{1, 2, \ldots, n\}$. $\propto$ means “proportional to”. 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. The Kullback-Leibler divergence of a distribution $p$ from another distribution $q$ of a random vector $x$ with domain $\mathcal{X}$ is defined by $D(q\|p) := \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \, dx$ if $\mathcal{X}$ is continuous and $D(q\|p) := \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)}$ if $\mathcal{X}$ is discrete. $\mathcal{N}(\mu, \Sigma)$ denotes the complex Gaussian vector distribution with mean $\mu$, covariance matrix $\Sigma$, and thus probability density function (pdf) $\mathcal{N}(x; \mu, \Sigma) := \frac{\exp\left(-\frac{(x-\mu)^H\Sigma^{-1}(x-\mu)}{\pi \det(\Sigma)}\right)}{\pi^n \det(\Sigma)}$, $x \in \mathbb{C}^n$.

**II. System Model**

**A. Channel Model**

We consider a SIMO multiple access channel in which $K$ single-antenna users transmit to a receiver equipped with $N$ 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 $h_k \in \mathbb{C}^{N \times 1}$, 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 $T > 1$ symbols 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 the receiver and users, thus the communication is non-coherent. We consider independent and identically distributed (i.i.d.) Rayleigh fading, i.e., $h_k \sim \mathcal{N}(0, I_N)$ for tractability, although the problem can be formulated similarly for other channel statistics.
Within a coherence block, each user $k$ sends a signal vector $s_k \in \mathbb{C}^T$, and the receiver receives

$$ Y = \sum_{k=1}^{K} s_k h_k^T + W = SH^T + W, $$

where $S = [s_1 \ldots s_K] \in \mathbb{C}^{T \times K}$ and $H = [h_1 \ldots h_K] \in \mathbb{C}^{N \times K}$ concatenate the transmitted signals and channel vectors, respectively, $W \in \mathbb{C}^{T \times N}$ 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., $\mathbb{E}[\|s_k\|^2] = 1, k \in [K]$. Under this normalization, the signal-to-noise ratio (SNR) of each transmitted signal at each receive antenna is $SNR = 1/(T\sigma^2)$.

We assume that the transmitted signals belong to disjoint finite discrete individual Grassmannian constellations (a.k.a. codebooks) on $G(\mathbb{C}^T, 1)$. That is, $s_k \in S_k := \{s_k^{(1)}, \ldots, s_k^{(|S_k|)}\}$, where each constellation symbol (a.k.a. codeword) $s_k^{(i)}$ is a unit-norm vector representative of a point in $G(\mathbb{C}^T, 1)$. Let $B_k = \log_2 |S_k|$ be the number of bits per codeword for user $k$.

### B. Multi-user Detection Problem

Given $S$, the received signal $Y$ is a Gaussian matrix with independent columns having the same covariance matrix $\sigma^2 I_T + SS^H$. Therefore, the likelihood function $p(Y|S)$ can be derived as

$$ p(Y|S) = \frac{\exp\left(-\text{tr}\left\{Y^H(\sigma^2 I_T + SS^H)^{-1}Y\right\}\right)}{\pi^T \det^N(\sigma^2 I_T + SS^H)} $$

$$ = \frac{\exp\left(-\frac{1}{\sigma^2}\|Y\|^2_F + \frac{1}{\sigma^2}\|Y^H S(\sigma^2 I_K + S^H S)^{-\frac{1}{2}}\|^2_F\right)}{\pi^T \det^N(\sigma^2 I_K + S^H S)}. $$

The maximum-likelihood (ML) symbol decoder is then

$$ \{s_1, \ldots, \hat{s}_K\} = \arg\max_{s_k \in S_k, k = 1, \ldots, K} \left(\|Y^H S(\sigma^2 I_K + S^H S)^{-\frac{1}{2}}\|^2_F - N\sigma^2 \log \det(\sigma^2 I_K + S^H S)\right). $$

When a channel code is used, most channel decoders require the log-likelihood ratios of the bits, which is computed from the posteriors $p(s_k|Y), k \in [K]$, which are marginalized from

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

Assuming that the transmitted signals are independent and uniformly distributed over the discrete constellations, the prior $p(S)$ factorizes as $p(S = [s_1, \ldots, s_K]) = \prod_{k=1}^{K} \frac{1}{|S_k|} \mathbb{I}\{s_k \in S_k\}$. On the other hand, the likelihood function $p(Y|S)$ involves all the signals $s_1, \ldots, s_K$ in a manner that does not straightforwardly factorize. Exact marginalization of $p(S|Y)$ requires computing

$$ p(s_k|Y) = \sum_{s_{l} \in S_l \forall l \neq k} p(S|Y) \quad \text{for} \quad k \in [K]. $$
This has complexity exponential in $\sum_{i \neq k} B_k$, which is formidable in the case of many users or large codebooks. Thus, we look for alternative approaches to estimate the true posterior as
\[
p(S|Y) \approx \hat{p}(S|Y) = \prod_{k=1}^{K} \hat{p}(s_k|Y). \tag{7}
\]

### C. Achievable Rate

According to [22, Sec.II], the highest sum-rate of the system for reliable communication with a given decoding metric $\hat{p}(S|Y)$, so-called the mismatched sum-rate, is lower bounded by the generalized mutual information (GMI) given by
\[
R_{\text{GMI}} = \frac{1}{T} \sup_{s \geq 0} \mathbb{E}\left[ \log_2 \frac{\hat{p}(S|Y)^s}{\sum_{s' \in \Pi_{k=1}^{K}s_k} p(S)p(S|Y)^s} \right] \tag{8}
\]
\[
= \frac{1}{T} \sup_{s \geq 0} \mathbb{E}\left[ \sum_{k=1}^{K} B_k - \log_2 \frac{\sum_{s' \in \Pi_{k=1}^{K}s_k} \hat{p}(S'|Y)^s}{\hat{p}(S|Y)^s} \right] \tag{9}
\]
\[
= \frac{1}{T} \sum_{k=1}^{K} B_k - \frac{1}{T} \inf_{s \geq 0} \mathbb{E}\left[ \sum_{k=1}^{K} \log_2 \frac{\sum_{s' \in \Pi_{k=1}^{K}s_k} \hat{p}(S'|Y)^s}{\hat{p}(s_k|Y)^s} \right] \tag{10}
\]

where the expectation is over the joint distribution of $S$ and $Y$, i.e., $p(Y|S)p(S)$, (9) holds because the transmitted symbols are independent and have uniform prior distribution, and (10) follows from (7). The generalized mutual information $R_{\text{GMI}}$ is upper bounded by the achievable sum-rate achieved with the optimal decoding metric $p(S|Y)$ given by
\[
R = \frac{1}{T} I(S;Y) = \frac{1}{T} \mathbb{E}\left[ \log_2 \frac{p(S|Y)}{\sum_{s' \in \Pi_{k=1}^{K} s_k} p(S)p(S|Y)} \right] \tag{11}
\]
\[
= \frac{1}{T} \sum_{k=1}^{K} B_k - \frac{1}{T} \mathbb{E}\left[ \log_2 \frac{\sum_{s' \in \Pi_{k=1}^{K} s_k} p(Y|S)}{p(Y|S)} \right] \tag{12}
\]

$R_{\text{GMI}}$ approaches $R$ as $\hat{p}(S|Y)$ gets close to $p(S|Y)$. Note that when $s$ is fixed to 1, it holds that
\[
R - R_{\text{GMI}}(s = 1) = \frac{1}{T} \mathbb{E}\left[ \log_2 \frac{p(S|Y)}{\hat{p}(S|Y)} \right] = \frac{1}{T} \mathbb{E}_Y \left[ D(p(S|Y)\|\hat{p}(S|Y)) \right], \tag{13}
\]

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

The expectations in (10) and (12) cannot be derived in closed form in general. Alternatively, we can estimate $R$ and $R_{\text{GMI}}$ (and also $\mathbb{E}_Y[D(p(S|Y)\|\hat{p}(S|Y))]$) numerically with the Monte Carlo method. Note that when $K$ or $B_k$ is large, even a numerical evaluation of $R$ and $\mathbb{E}_Y[D(p(S|Y)\|\hat{p}(S|Y))]$ is not possible. Therefore, we choose to use the mismatched sum-rate lower bound $R_{\text{GMI}}$ as a information-theoretic metric to evaluate how close $\hat{p}(S|Y)$ is to $p(S|Y)$.

In what follows, we design a posterior marginal estimation scheme based on expectation propagation (EP). We start by providing a brief review of EP in the next section.
III. Expectation Propagation

The EP algorithm was first proposed in [18] and summarized in, e.g., [19] for approximate inference in probabilistic graphical models. EP is an iterative framework for approximating posterior beliefs by parametric distributions in the exponential family [20, Sec.1.6]. Let us consider a set of unknown variables represented by a random vector \( \mathbf{x} \) with posterior of the form

\[
p(\mathbf{x}) \propto \prod_{\alpha} \psi_{\alpha}(\mathbf{x}_{\alpha}), \tag{14}
\]

where \( \mathbf{x}_{\alpha} \) is the subset of variables involved in the factor \( \psi_{\alpha} \). 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 \( \mathbf{x}_{\beta} \subseteq \mathbf{x}_{\alpha} \) or \( \mathbf{x}_{\beta} \cap \mathbf{x}_{\alpha} = \emptyset \)). We are interested in the posterior marginals w.r.t. the partition \( \{ \mathbf{x}_{\beta} \} \).

EP approximates the true posterior \( p \) from (14) 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

\[
\hat{p}(\mathbf{x}) = \prod_{\beta} \hat{p}_{\beta}(\mathbf{x}_{\beta}), \tag{15}
\]

where \( \hat{p}_{\beta} \) are constrained to be in the exponential family [20, Sec.1.6], so that

\[
\hat{p}_{\beta}(\mathbf{x}_{\beta}) = \exp \left( \gamma_{\beta}^T \phi_{\beta}(\mathbf{x}_{\beta}) - A_{\beta}(\gamma_{\beta}) \right), \tag{16}
\]

for sufficient statistics \( \phi_{\beta}(\mathbf{x}_{\beta}) \), parameters \( \gamma_{\beta} \), and log-partition function \( A_{\beta}(\gamma) := \ln \int e^{\gamma^T \phi_{\beta}(\mathbf{x}_{\beta})} d\mathbf{x}_{\beta} \).

Second, the approximate posterior can also be expressed w.r.t. the partition \( \{ \mathbf{x}_{\alpha} \} \) as

\[
\hat{p}(\mathbf{x}) \propto \prod_{\alpha} m_{\alpha}(\mathbf{x}_{\alpha}), \tag{17}
\]

in accordance with (14). For (15) and (17) to be consistent, the terms \( m_{\alpha} \) should factorize over \( \beta \) as well, i.e., there exist factors \( m_{\alpha,\beta} \) of the form \( m_{\alpha,\beta}(\mathbf{x}_{\beta}) = \exp \left( \gamma_{\alpha,\beta}^T \phi_{\beta}(\mathbf{x}_{\beta}) \right) \) such that (s.t.)

\[
m_{\alpha}(\mathbf{x}_{\alpha}) = \prod_{\beta \in \Omega_{\alpha}} m_{\alpha,\beta}(\mathbf{x}_{\beta}) = \exp \left( \sum_{\beta \in \Omega_{\alpha}} \gamma_{\alpha,\beta}^T \phi_{\beta}(\mathbf{x}_{\beta}) \right), \tag{18}
\]

\[
\hat{p}_{\beta}(\mathbf{x}_{\beta}) \propto \prod_{\alpha \in \Omega_{\beta}} m_{\alpha,\beta}(\mathbf{x}_{\beta}) = \exp \left( \sum_{\alpha \in \Omega_{\beta}} \gamma_{\alpha,\beta}^T \phi_{\beta}(\mathbf{x}_{\beta}) \right), \tag{19}
\]

where \( \Omega_{\alpha} \) collects the indices \( \beta \) for which \( \mathbf{x}_{\beta} \subseteq \mathbf{x}_{\alpha} \), and \( \Omega_{\beta} \) collects the indices \( \alpha \) for which \( \mathbf{x}_{\beta} \subseteq \mathbf{x}_{\alpha} \). It turns out that \( m_{\alpha,\beta} \) can be interpreted as a message from the factor node \( \alpha \) to the variable node \( \beta \) on a bipartite factor graph [23]. In this case, \( \hat{p}_{\beta}(\mathbf{x}_{\beta}) \) is proportional to the product of all messages impinging on variable node \( \beta \).

EP works by first initializing all \( m_{\alpha}(\mathbf{x}_{\alpha}) \) and \( \hat{p}_{\beta}(\mathbf{x}_{\beta}) \) (typically by the priors, which is assumed to also belong to the considered exponential family), then iteratively updating each approximation
factor $m_\alpha$ in turn. Let us fix a factor index $\alpha$. According to \cite{18}, the “tilted” distribution $q_\alpha$ is constructed by swapping the true potential $\psi_\alpha$ for its approximate $m_\alpha$ in $\hat{p}(\mathbf{x})$ as

$$q_\alpha(\mathbf{x}) = \frac{\hat{p}(\mathbf{x})\psi_\alpha(\mathbf{x}_\alpha)}{m_\alpha(\mathbf{x}_\alpha)},$$

where it is assumed that $\int q_\alpha(\mathbf{x}) d\mathbf{x} < \infty$. This tilted distribution is projected back onto the exponential family by minimizing the KL divergence:

$$\hat{p}^\text{new}_\alpha(\mathbf{x}) = \arg \min_{\hat{p} \in \mathcal{P}} D(q_\alpha(\mathbf{x}) \parallel \hat{p}(\mathbf{x})),$$

where $\mathcal{P}$ is the set of distributions with the form of $\hat{p}$ from \cite{15}, i.e., $\hat{p}(\mathbf{x}) = \prod_\beta \hat{p}_\beta(\mathbf{x}_\beta) = \prod_\beta \exp(\gamma^T_\beta \phi_\beta(\mathbf{x}_\beta) - A_\beta(\gamma_\beta))$ for some $\{\gamma_\beta\}$.

**Proposition 1.** The solution to (21) is given by $\hat{p}^\text{new}_\alpha(\mathbf{x}) = \prod_\beta \hat{p}_\alpha,\beta(\mathbf{x}_\beta)$ with $\hat{p}_\alpha,\beta(\mathbf{x}_\beta) = \hat{p}_\beta(\mathbf{x}_\beta)$, $\forall \beta \not\in \mathcal{N}_\alpha$, and $\hat{p}^\text{new}_\alpha(\mathbf{x}_\beta) = \exp(\gamma^T_\beta \phi_\beta(\mathbf{x}_\beta) - A_\beta(\gamma_\beta))$ with $\gamma_\beta$ s.t. $\mathbb{E}_{\hat{p}^\text{new}_\alpha,\beta}[\phi_\beta(\mathbf{x}_\beta)] = \mathbb{E}_{q_\alpha}[\phi_\beta(\mathbf{x}_\beta)]$, $\forall \beta \in \mathcal{N}_\alpha$.

**Proof.** The proof is given in Appendix B. \hfill \Box

The factor $m_\alpha$ is then updated via

$$m^\text{new}_\alpha(\mathbf{x}_\alpha) = \frac{\hat{p}^\text{new}_\alpha(\mathbf{x})m_\alpha(\mathbf{x}_\alpha)}{\hat{p}(\mathbf{x})}$$

$$= \left[ \prod_{\beta \in \mathcal{N}_\alpha} m_\alpha,\beta(\mathbf{x}_\beta) \right] \frac{\prod_{\beta \in \mathcal{N}_\alpha} \hat{p}^\text{new}_\alpha(\mathbf{x}_\beta)}{\prod_{\beta \in \mathcal{N}_\alpha} \hat{p}_\beta(\mathbf{x}_\beta)}$$

$$\propto \left[ \prod_{\beta \in \mathcal{N}_\alpha} m_\alpha,\beta(\mathbf{x}_\beta) \right] \frac{\prod_{\beta \in \mathcal{N}_\alpha} \hat{p}^\text{new}_\alpha(\mathbf{x}_\beta)}{\prod_{\beta \in \mathcal{N}_\alpha} \left[ m_\alpha,\beta(\mathbf{x}_\beta) \prod_{\alpha' \in \mathcal{N}_\alpha \setminus \alpha} m_{\alpha',\beta}(\mathbf{x}_\beta) \right]}$$

$$= \prod_{\beta \in \mathcal{N}_\alpha} m_{\alpha,\beta}^\text{new}(\mathbf{x}_\beta),$$

with

$$m_{\alpha,\beta}^\text{new}(\mathbf{x}_\beta) := \frac{\hat{p}^\text{new}_\alpha(\mathbf{x}_\beta)}{\prod_{\alpha' \in \mathcal{N}_\beta \setminus \alpha} m_{\alpha',\beta}(\mathbf{x}_\beta)}.$$  

Note that, on the right-hand side of (22), all terms dependent on $\{\mathbf{x}_\beta\}_{\beta \not\in \mathcal{N}_\alpha}$ cancel, leaving the dependence only on $\{\mathbf{x}_\beta\}_{\beta \in \mathcal{N}_\alpha}$. Thus, the update of $m_\alpha$ only affects the approximate posterior of the variable nodes $\beta$ in the neighborhood of the factor node $\alpha$. After that, the process is repeated with the next $\alpha$. 

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

\[
q_\alpha(x) = \frac{\psi_\alpha(x_\alpha)}{m_\alpha(x_\alpha)} \left[ \prod_{\beta \in \mathcal{V}_\alpha} \prod_{\alpha' \in \mathcal{V}_\beta \setminus \alpha} m_{\alpha',\beta}(x_\beta) \right] \left[ \prod_{\beta \notin \mathcal{V}_\alpha} \hat{p}_\beta(x_\beta) \right]
\]

(27)

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

\[
D(q_\alpha(x) \| p(x)) = \int \psi_\alpha(x_\alpha) \left[ \prod_{\beta \in \mathcal{V}_\alpha} \prod_{\alpha' \in \mathcal{V}_\beta \setminus \alpha} m_{\alpha',\beta}(x_\beta) \right] \left[ \prod_{\beta \notin \mathcal{V}_\alpha} \hat{p}_\beta(x_\beta) \right] \ln \frac{\psi_\alpha(x_\alpha) \left[ \prod_{\beta \in \mathcal{V}_\alpha} \prod_{\alpha' \in \mathcal{V}_\beta \setminus \alpha} m_{\alpha',\beta}(x_\beta) \right] \left[ \prod_{\beta \notin \mathcal{V}_\alpha} \hat{p}_\beta(x_\beta) \right]}{\prod_{\beta \in \mathcal{V}_\alpha} p_\beta(x_\beta) \prod_{\beta \notin \mathcal{V}_\alpha} p_\beta(x_\beta)} \, dx_\alpha
\]

(29)

\[
= \int \psi_\alpha(x_\alpha) \left[ \prod_{\beta \in \mathcal{V}_\alpha} \prod_{\alpha' \in \mathcal{V}_\beta \setminus \alpha} m_{\alpha',\beta}(x_\beta) \right] \ln \frac{\psi_\alpha(x_\alpha) \left[ \prod_{\beta \in \mathcal{V}_\alpha} \prod_{\alpha' \in \mathcal{V}_\beta \setminus \alpha} m_{\alpha',\beta}(x_\beta) \right]}{\prod_{\beta \in \mathcal{V}_\alpha} p_\beta(x_\beta)} \, dx_\alpha
\]

(30)

\[
= \sum_{\beta \in \mathcal{V}_\alpha} \int q_{\alpha,\beta}(x_\beta) \ln \frac{q_{\alpha,\beta}(x_\beta)}{p_\beta(x_\beta)} \, dx_\beta + \sum_{\beta \notin \mathcal{V}_\alpha} D(\hat{p}_\beta \| p_\beta) + c_0
\]

(31)

\[
= \sum_{\beta \in \mathcal{V}_\alpha} D(q_{\alpha,\beta} \| p_\beta) + \sum_{\beta \notin \mathcal{V}_\alpha} D(\hat{p}_\beta \| p_\beta) + c_0,
\]

(32)

where

\[
q_{\alpha,\beta}(x_\beta) := \int \psi_\alpha(x_\alpha) \left[ \prod_{\beta \in \mathcal{V}_\alpha} \prod_{\alpha' \in \mathcal{V}_\beta \setminus \alpha} m_{\alpha',\beta}(x_\beta) \right] \, dx_\alpha \setminus \beta.
\]

Equation (32) says that, for each \( \beta \) in the neighborhood of node \( \alpha \), the optimal \( p_\beta \) (i.e., \( \hat{p}_{\alpha,\beta}^{\text{new}} \)) is the moment match of \( q_{\alpha,\beta} \) in the exponential family with sufficient statistics \( \phi_\beta(x_\beta) \), where \( q_{\alpha,\beta} \) is formed by taking the product of the true factor \( \psi_\alpha \) and all the messages impinging on that factor, and then integrating out all variables except \( x_\beta \). Furthermore, (26) says that the new message \( m_{\alpha,\beta}^{\text{new}} \) passed from \( \alpha \) to \( \beta \in \mathcal{V}_\alpha \) equals \( \hat{p}_{\alpha,\beta}^{\text{new}} \) divided by the message product \( \{m_{\alpha',\beta}\}_{\alpha' \in \mathcal{V}_\beta \setminus \alpha} \), i.e., previous messages to \( \beta \) from all directions except \( \alpha \).
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 \( s_k = s_k^{(i_k)} \), where \( i_k \) are random symbol indices that are independent and uniformly distributed over \([|S_k|]\). We rewrite the received signal (1) in vector form as

\[
y = \sum_{k=1}^{K} z_k + w,
\]

where \( y := \text{vec}(Y) \), \( z_k := (s_k^{(i_k)} \otimes I_N) h_k \), and \( w := \text{vec}(W^T) \sim \mathcal{N}(0, \sigma^2 I_{NT}) \). The problem of estimating \( p(s_k|Y) \) is equivalent to estimating \( p(i_k|Y) \) since they admit the same pmf.

With \( z := [z_1^T, \ldots, z_K^T]^T \) and \( i := [i_1, \ldots, i_K]^T \), we can write the posterior density in factorized form as

\[
p(i, z|y) \propto p(i, z, y) = p(y|z)p(z|i)p(i)
\]

\[
= \psi_0(z_1, \ldots, z_K) \left[ \prod_{k=1}^{K} \psi_{k1}(z_k, i_k) \right] \left[ \prod_{k=1}^{K} \psi_{k2}(i_k) \right],
\]

corresponding to (14), where

\[
\psi_0(z_1, \ldots, z_K) := p(y|z) = \mathcal{N} \left( y; \sum_{k=1}^{K} z_k, \sigma^2 I_{NT} \right),
\]

\[
\psi_{k1}(z_k, i_k) := p(z_k|i_k) = \mathcal{N} \left( z_k; 0, (s_k^{(i_k)} s_k^{(i_k)\dagger}) \otimes I_N \right),
\]

\[
\psi_{k2}(i_k) := p(i_k) = \frac{1}{|S_k|} \quad \text{for} \quad i_k \in [|S_k|].
\]

We will use EP to infer the posterior distribution of the indices \( \{i_k\} \) and, as a by-product, the posterior of \( \{z_k\}, k \in [K] \). To do so, we choose the partition \( x = \{z_k, i_k\}_{k=1}^{K} \) and illustrate the interaction between these variables and the factors \( \psi_0, \psi_{k1}, \text{and} \psi_{k2} \) on the bipartite factor graph in Fig. II. This graph has a tree structure with a root \( y \) and \( K \) leaves \( \psi_{k2}^{(K)} \).

We write the EP approximation according to (15) as

\[
\hat{p}(x|y) = \hat{p}(i, z|y) = \prod_{k=1}^{K} \hat{p}_{z_k}(z_k) \hat{p}_{i_k}(i_k),
\]

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

\[
\hat{p}_{z_k}(z) = \mathcal{N}(z_k; \hat{z}_k, \Sigma_k) \quad \text{s.t.} \quad \Sigma_k \text{ is positive definite},
\]

\[
\hat{p}_{i_k}(i_k) = \pi_k^{(i_k)} \quad \text{for} \quad i_k \in [|S_k|] \quad \text{s.t.} \quad \sum_{i=1}^{|S_k|} \pi_k^{(i)} = 1.
\]
We also write the EP approximation according to (17) as

\[ \hat{p}(x|y) \propto m_0(z_1, \ldots, z_K) \left[ \prod_{k=1}^{K} m_{k1}(z_k, i_k) \right] \left[ \prod_{k=1}^{K} m_{k2}(i_k) \right] \]

(43)

where we define \( m_0(z_1, \ldots, z_K) \propto \prod_{k=1}^{K} \mathcal{N}(z_k; \mu_{k0}, C_{k0}) \), \( m_{k1}(z_k, i_k) \propto \mathcal{N}(z_k; \mu_{k1}, C_{k1}) \pi_{k1}^{(i_k)} \), and \( m_{k2}(i_k) = \pi_{k2}^{(i_k)} \) for \( i_k \in [S_k] \). On the factor graph in Fig. 1, we can interpret \( (\mu_{k0}, C_{k0}) \) as the message from factor node \( \psi_0 \) to variable node \( z_k \), \( (\mu_{k1}, C_{k1}) \) as the message from factor node \( \psi_{k1} \) to variable node \( z_k \), \( \{\pi_{k1}^{(i_k)}\}_{i_k=1}^{S_k} \) as the message from factor node \( \psi_{k1} \) to variable node \( i_k \), and \( \{\pi_{k2}^{(i_k)}\}_{i_k=1}^{S_k} \) as the message from factor node \( \psi_{k2} \) to variable node \( i_k \).

A. The EP message updates

In the following, we derive the message updates from each of the factor nodes \( \psi_0, \psi_{k1}, \) and \( \psi_{k2} \), \( k \in [K] \), to the corresponding variable nodes. To do so, we consider each factor node \( \alpha \in \{k1, k2, 0\} \), compute the projected density \( \hat{p}_{\alpha}^{\text{new}} \) according to Proposition \[ \] and then update the factor \( m_{\alpha} \) according to (25). From (40) and Proposition \[ \] for each factor node \( \alpha \in \{k1, k2, 0\} \), the projected density \( \hat{p}_{\alpha}^{\text{new}} \) is given by

\[ \hat{p}_{\alpha}^{\text{new}} = \prod_{k=1}^{K} \hat{p}_{\alpha, z_k}^{\text{new}}(z_k) \hat{p}_{\alpha, i_k}^{\text{new}}(i_k). \]

1) Message \( \{\pi_{k2}^{(i_k)}\}_{i_k=1}^{S_k} \) from factor node \( \psi_{k2} \) to variable node \( i_k \): First, we compute \( \hat{p}_{k2, i_k}^{\text{new}} \) and then the EP message \( \{\pi_{k2}^{(i_k)}\}_{i_k=1}^{S_k} \) from node \( \psi_{k2} \) to node \( i_k \). From (32) and (42), we know that \( \hat{p}_{k2, i_k}^{\text{new}} \) is the discrete distribution with pmf \( \{\pi_{k2}^{(i_k)}\}_{i_k=1}^{S_k} \) proportional to \( \psi_{k2}(i_k)\pi_{k1}^{(i_k)} \) , and so

\[ \pi_{k2}^{(i_k)} = \frac{\psi_{k2}(i_k)\pi_{k1}^{(i_k)}}{\sum_{i=1}^{S_k} \psi_{k2}(i)\pi_{k1}^{(i)}} = \frac{\pi_{k1}^{(i_k)}}{\sum_{i=1}^{S_k} \pi_{k1}^{(i)}} \text{ for } i_k \in [S_k]. \]
since \( \psi_2(i_k) \) is constant over these \( i_k \). With \( \hat{p}_{k_2,i_k}^\text{new} \) computed, (25) implies that the message from node \( \psi_2 \) to node \( i_k \) is the pmf proportional to
\[
\hat{p}_{k_2,i_k}^\text{new}(i_k) = \frac{\pi_{k_2}^{(i_k)}}{\pi_{k_2}^{(i_k)}} = \frac{1}{\sum_{i=1}^{||S_k||} \pi_{k_1}^{(i)}} = c_0 \text{ for } i_k \in [\|S_k\|],
\]
and thus \( \pi_{k_2}^{(i_k)} = \frac{1}{||S_k||} \) for \( i_k \in [\|S_k\|] \).

2) Messages from factor node \( \psi_{k_1} \) to variable nodes \( z_k \) and \( i_k \): Next, we compute \( \hat{p}_{k_1}^\text{new} = \prod_{k=1}^{K} \hat{p}_{k_1}^\text{new}(z_k) \hat{p}_{k_1,i_k}^\text{new}(i_k) \) and the messages \( \{\pi_{k_1}^{(i_k)}\}_{i_k=1}^{||S_k||} \) and \( (\mu_{k_1}, C_{k_1}) \) from node \( \psi_{k_1} \) to nodes \( i_k \) and \( z_k \), respectively.

Message \( \{\pi_{k_1}^{(i_k)}\}_{i_k=1}^{||S_k||} \) from node \( \psi_{k_1} \) to node \( i_k \): We first compute \( \hat{p}_{k_1,i_k}^\text{new}(i_k) \). From (32) and (42), we know that \( \hat{p}_{k_1,i_k}^\text{new}(i_k) \) is the discrete distribution with support \([\|S_k\|]\) and pmf \( \pi_{k_1}^{(i_k)} \) proportional to
\[
\int \psi_{k_1}(z_k, i_k) \mathcal{N}(z_k; \mu_{k_0}, C_{k_0}) \pi_{k_2}^{(i_k)} \, dz_k
= \frac{1}{||S_k||} \int \mathcal{N}(z_k; \mathbf{0}, (s_k^{(i_k)} s_k^{(i_k)H}) \otimes I_N) \mathcal{N}(z_k; \mu_{k_0}, C_{k_0}) \, dz_k
= \frac{1}{||S_k||} \int \mathcal{N}(z_k; \Sigma_{ki}, \Sigma_{k_0}) \mathcal{N}(\mathbf{0}; \mu_{k_0}, (s_k^{(i_k)} s_k^{(i_k)H}) \otimes I_N + C_{k_0}) \, dz_k
= \frac{1}{||S_k||} \mathcal{N}(\mathbf{0}; \mu_{k_0}, (s_k^{(i_k)} s_k^{(i_k)H}) \otimes I_N + C_{k_0}),
\]
where the second equality follows from the Gaussian pdf multiplication rule in Lemma 1 with
\[
\Sigma_{ki} = \left( [(s_k^{(i)} s_k^{(i)H}) \otimes I_N]^{-1} + C_{k_0}^{-1} \right)^{-1}
= \left( [(s_k^{(i)} s_k^{(i)H}) \otimes I_N] \left( (s_k^{(i)} s_k^{(i)H}) \otimes I_N + C_{k_0} \right)^{-1} \right)^{-1} C_{k_0},
\]
\[
\tilde{z}_{ki} = \Sigma_{ki} C_{k_0}^{-1} \mu_{k_0}
= \left( [(s_k^{(i)} s_k^{(i)H}) \otimes I_N] \left( (s_k^{(i)} s_k^{(i)H}) \otimes I_N + C_{k_0} \right)^{-1} \right) \mu_{k_0},
\]
Thus
\[
\hat{\pi}_{k_1}^{(i_k)} = \frac{\mathcal{N}(\mathbf{0}; \mu_{k_0}, (s_k^{(i_k)} s_k^{(i_k)H}) \otimes I_N + C_{k_0})}{\sum_{i=1}^{||S_k||} \mathcal{N}(\mathbf{0}; \mu_{k_0}, (s_k^{(i_k)} s_k^{(i_k)H}) \otimes I_N + C_{k_0})} \text{ for } i_k \in [\|S_k\|].
\]
With \( \hat{p}_{k_1,i_k}^\text{new}(i_k) \) computed, (25) implies that the message \( \pi_{k_1}^{(i_k)} \) from node \( \psi_{k_1} \) to node \( i_k \) is the pmf proportional to
\[
\frac{\hat{p}_{k_1,i_k}^\text{new}(i_k)}{\pi_{k_2}^{(i_k)}} = |S_k| \hat{\pi}_{k_1}^{(i_k)} \text{ for } i_k \in [\|S_k\|], \text{ and thus}
\]
\[
\pi_{k_1}^{(i_k)} = \frac{|S_k| \hat{\pi}_{k_1}^{(i_k)}}{\sum_{i=1}^{||S_k||} |S_k| \hat{\pi}_{k_1}^{(i_k)}} = \hat{\pi}_{k_1}^{(i_k)} \text{ for } i_k \in [\|S_k\|].
\]
Message \((\mu_{k1}, C_{k1})\) from node \(\psi_{k1}\) to nodes \(z_k\): We next compute \(\hat{p}^{new}_{k1, z_k}(z_k)\). From (32) and (41), we know that \(\hat{p}^{new}_{k1, z_k}(z_k)\) is the Gaussian with mean \(\hat{z}_k\) and covariance \(\Sigma_k\) matched to that of the pdf proportional to
\[
\sum_{i_k=1}^{\vert S_k \vert} \psi^i_{k1}(z_k; i_k) \mathcal{N}(z_k; \mu_{k0}, C_{k0}) \pi^{(i_k)}_{k2}
\]
\[
= \frac{1}{\vert S_k \vert} \sum_{i_k=1}^{\vert S_k \vert} \mathcal{N}(z_k; 0, (s_k^{(i)} s_k^{(i)H}) \otimes I_N) \mathcal{N}(z_k; \mu_{k0}, C_{k0})
\]
\[
= \frac{1}{\vert S_k \vert} \sum_{i_k=1}^{\vert S_k \vert} \mathcal{N}(z_k; \hat{z}_{ki}, \Sigma_{ki}) \mathcal{N}(0; \mu_{k0}, (s_k^{(i)} s_k^{(i)H}) \otimes I_N + C_{k0})
\]
\[
\propto \sum_{i_k=1}^{\vert S_k \vert} \mathcal{N}(z_k; \hat{z}_{ki}, \Sigma_{ki}) \hat{\pi}^{(i)}_{k1},
\]
where the second equality follows from the Gaussian pdf multiplication rule in Lemma 1 with \(\Sigma_{ki}\) and \(\hat{z}_{ki}\) defined in (50) and (52), respectively. Thus, from (54), we have
\[
\hat{z}_k = \sum_{i_k=1}^{\vert S_k \vert} \hat{\pi}^{(i)}_{k1} \hat{z}_{ki},
\]
\[
\Sigma_k = \sum_{i_k=1}^{\vert S_k \vert} \hat{\pi}^{(i)}_{k1} (\hat{z}_{ki} \hat{z}_{ki}^H + \Sigma_{ki}) - \hat{z}_k \hat{z}_k^H.
\]
With \(\hat{p}^{new}_{k1, z_k}(z_k)\) computed, (25) implies that the message from node \(\psi_{k1}\) to \(z_k\) is proportional to
\[
\frac{\hat{p}^{new}_{k1, z_k}(z_k)}{\mathcal{N}(z_k; \mu_{k0}, C_{k0})} = \frac{\mathcal{N}(z_k; \hat{z}_k, \Sigma_k)}{\mathcal{N}(z_k; \mu_{k0}, C_{k0})} \propto \mathcal{N}(z_k; \mu_{k1}, C_{k1}),
\]
with
\[
C_{k1} = (\Sigma_k^{-1} - C_{k0}^{-1})^{-1},
\]
\[
\mu_{k1} = C_{k1} (\Sigma_k^{-1} \hat{z}_k - C_{k0}^{-1} \mu_{k0}).
\]
Equations (61) and (62) can be verified using \(\mathcal{N}(z_k; \hat{z}_k, \Sigma_k) \propto \mathcal{N}(z_k; \mu_{k1}, C_{k1}) \mathcal{N}(z_k; \mu_{k0}, C_{k0})\), which follows from (19) and the Gaussian pdf multiplication rule in Lemma 1.

3) Message \((\mu_{k0}, C_{k0})\) from node \(\psi_0\) to node \(z_k\): Finally, we compute \(\hat{p}^{new}_{0, z_k}\) and the EP message \((\mu_{k0}, C_{k0})\) from node \(\psi_0\) to node \(z_k\) for each \(k \in [K]\). From (32) and (41), we know that \(\hat{p}^{new}_{0, z_k}\)
is the Gaussian distribution with mean \( \hat{z}_{k0} \) and covariance \( \Sigma_{k0} \) matched to that of the pdf proportional to

\[
\mathcal{N}(z_k; \mu_{k1}, C_{k1}) \int \psi_0(z_1, \ldots, z_K) \left[ \prod_{j \neq k} \mathcal{N}(z_j; \mu_{j1}, C_{j1}) \right] dz_j
\]

\[
= \mathcal{N}(z_k; \mu_{k1}, C_{k1}) \int \mathcal{N}(y; z_k + \sum_{j \neq k} z_j, \sigma^2 I_{NT}) \left[ \prod_{j \neq k} \mathcal{N}(z_j; \mu_{j1}, C_{j1}) \right] dz_j
\]

\[
= \mathcal{N}(z_k; \mu_{k1}, C_{k1}) \mathcal{N}(z_k; y - \sum_{j \neq k} \mu_{j1}, \sigma^2 I_{NT} + \sum_{j \neq k} C_{j1}),
\]

where (64) follows by applying repeatedly the two properties in Lemma 1. Applying the Gaussian pdf multiplication rule to (64), we obtain

\[
\Sigma_{k0} = \left( C_{k1}^{-1} + \left[ \sigma^2 I_{NT} + \sum_{j \neq k} C_{j1} \right] \right)^{-1},
\]

(65)

\[
\hat{z}_{k0} = \Sigma_{k0} \left( C_{k1}^{-1} \mu_{k1} + \left[ \sigma^2 I_{NT} + \sum_{j \neq k} C_{j1} \right]^{-1} \left( y - \sum_{j \neq k} \mu_{j1} \right) \right).
\]

(66)

Given \( \hat{p}^{new}_0(z_k) = \mathcal{N}(z_k; \hat{z}_{k0}, \Sigma_{k0}) \), (25) implies that the message from node \( \psi_0 \) to node \( z_k \) is proportional to

\[
\frac{\hat{p}^{new}_0(z_k)}{\mathcal{N}(z_k; \mu_{k1}, C_{k1})} = \frac{\mathcal{N}(z_k; \hat{z}_{k0}, \Sigma_{k0})}{\mathcal{N}(z_k; \mu_{k1}, C_{k1})} \propto \mathcal{N}(z_k; \mu_{k0}, C_{k0}),
\]

(67)

with \( C_{k0} = (\Sigma_{k0}^{-1} - C_{k1}^{-1})^{-1} \) and \( \mu_{k0} = C_{k0} (\Sigma_{k0}^{-1} z_{k0} - C_{k1}^{-1} \mu_{k1}) \). This is verified using \( \mathcal{N}(z_k; \hat{z}_{k0}, \Sigma_{k0}) \propto \mathcal{N}(z_k; \mu_{k0}, C_{k0}) \), which follows from (19), and the Gaussian pdf multiplication rule in Lemma 1. Plugging in the expressions for \( \Sigma_{k0}^{-1} \) and \( z_{k0} \) from (65) and (66) yields

\[
C_{k0} = \sigma^2 I_{NT} + \sum_{j=1, j \neq k}^K C_{j1}
\]

(68)

\[
\mu_{k0} = C_{k0} \left( \sigma^2 I_{NT} + \sum_{j=1, j \neq k}^K C_{j1} \right)^{-1} \left( y - \sum_{j=1, j \neq k}^K \mu_{j1} \right) = y - \sum_{j=1, j \neq k}^K \mu_{j1}.
\]

(69)

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 \( C_{k0}^{-1} = 0 \) and \( \mu_{k0} = 0 \), so that, from (53), the initial message from node \( \psi_{k1} \) to node \( i_k \) coincides with the uniform prior \( \pi_k^{(i_k)} = \frac{1}{|S_k|} \) for \( i_k \in [|S_k|] \), and, from (50) and (52), the initial parameters \( \Sigma_{ki} = (s_k^{(i)})^T \otimes I_N \) and \( z_{ki} = 0 \), respectively, for \( k \in [K] \) and \( i \in [|S_k|] \). This leads to
Algorithm 1: EP for probabilistic non-coherent symbol detection

```plaintext
set the maximal number of iterations \( t_{\text{max}} \);
initialize the messages \( \{ \pi_{k1}^{(i)} \}_{i=1}^{\left| S_k \right|}, \mu_{k1}, C_{k1}, \mu_{k0}, C_{k0} \) for \( k \in [K] \);
\( t \leftarrow 0 \);
repeat
\( t \leftarrow t + 1 \);
for \( k \leftarrow 1 \) to \( K \) do
    update \( \{ \pi_{k1}^{(i)} \}_{i=1}^{\left| S_k \right|} \) according to (54) and (53);
    compute \( \{ \tilde{z}_{ki} \}_{i=1}^{\left| S_k \right|} \) and \( \{ \Sigma_{ki} \}_{i=1}^{\left| S_k \right|} \) according to (52) and (50), respectively;
    compute \( \hat{z}_k \) and \( \Sigma_k \) according to (58) and (59), respectively;
    update \( \mu_{k1} \) and \( C_{k1} \) according to (62) and (61), respectively;
    update \( \{ \mu_{j0} \}_{j=1,j \neq k}^{K} \) and \( \{ C_{j0} \}_{j=1,j \neq k}^{K} \) according to (69) and (68), respectively;
end
until convergence or \( t = t_{\text{max}} \);
return The pmf \( \{ \hat{\pi}_{k}^{(i)} \}_{i=1}^{\left| S_k \right|} = \{ \pi_{k1}^{(i)} \}_{i=1}^{\left| S_k \right|} \) of \( \hat{p}(s_k | Y) \) for \( k \in \{ 1, \ldots, K \} \)
```

the initial parameters of \( \hat{p}_k(z_k) \) from (58) and (59) as \( \hat{z}_k = 0 \) and \( \Sigma_k = \frac{1}{\left| S_k \right|} \sum_{i=1}^{\left| S_k \right|} (s_k^{(i)} s_k^{(i)^H}) \otimes I_N \), and the initial message from node \( \psi_{k1} \) to node \( z_k \) given in (61) and (62) as \( C_{k1} = \Sigma_k = \frac{1}{\left| S_k \right|} \sum_{i=1}^{\left| S_k \right|} (s_k^{(i)} s_k^{(i)^H}) \otimes I_N \), and \( \mu_{k1} = \hat{z}_k = 0 \). Finally, the initial messages from node \( \psi_0 \) to node \( z_k \) follows from (68) and (69) as \( C_{k0} = \sigma^2 I_{NT} + \sum_{j \neq k} \frac{1}{\left| S_j \right|} \sum_{i=1}^{\left| S_j \right|} (s_j^{(i)} s_j^{(i)^H}) \otimes I_N \), and \( \mu_{k0} = y \).

We summarize the proposed EP scheme for probabilistic non-coherent symbol detection in Algorithm 1. Since \( \pi_{k2}^{(i)} \) is constant, it is not required to be updated at each iteration. In the end, according to (19) and (42), the estimated pmf of \( \hat{p}(s_k | Y) \) is \( \hat{p}_k(i_k) = \hat{\pi}_{k}^{(i_k)} \propto \pi_{k1}^{(i_k)} \pi_{k2}^{(i_k)} \), that is \( \hat{p}_k(i_k) = \pi_{k1}^{(i_k)} \). The algorithm 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. We note that other variants of the message passing scheduling can be implemented.

V. MMSE-SIC: A SIMPLIFICATION OF EP

In this section, we derive a simplified version of the EP scheme. In the EP scheme, as in (57) and (60), the message \( \mathcal{N}(z_k; \mu_{k1}, C_{k1}) \) from node \( \psi_{k1} \) to node \( z_k \) is derived by first projecting \( \hat{p}^{\text{new}}_{k1, z_k}(z_k) \propto \sum_{i=1}^{\left| S_k \right|} \pi_{k1}^{(i)} \mathcal{N}(z_k; \tilde{z}_{ki}, \Sigma_{ki}) \) onto the Gaussian family, then dividing the
projected Gaussian by \( \mathcal{N}(z_k; \mu_{k0}, C_{k0}) \). If we skip the projection of \( \hat{p}_{k1}^{\text{new}}(z_k) \) onto the Gaussian family, i.e., we derive \( \mathcal{N}(z_k; \mu_{k1}, C_{k1}) \) by dividing directly \( \hat{p}_{k1}^{\text{new}}(z_k) \) to \( \mathcal{N}(z_k; \mu_{k0}, C_{k0}) \), then the mean \( \mu_{k1} \) and covariance matrix \( C_{k1} \) are matched to that of the pdf proportional to

\[
\frac{\hat{p}_{k1}^{\text{new}}(z_k)}{\mathcal{N}(z_k; \mu_{k0}, C_{k0})} = \sum_{i=1}^{|S_k|} \pi_{k1}^{(i)} \mathcal{N}(z_k; \hat{z}_{ki}, \Sigma_{ki}) \mathcal{N}(z_k; \mu_{k0}, C_{k0})
\]

which follows from the Gaussian pdf multiplication rule with \( \hat{z}_{ki} \) and \( \Sigma_{ki} \) given in (68), respectively. It follows that \( \mu_{k1} = 0 \) and \( C_{k1} = \sum_{i=1}^{|S_k|} \pi_{k1}^{(i)} (s_k^{(i)})^H \otimes I_N \otimes I_N \). A consequence of (69) and (68), \( \mu_{k0} = y \) and \( C_{k0} = \sigma^2 I_{NT} + \sum_{j\neq k} \sum_{i=1}^{|S_k|} \pi_{j}^{(i)} (s_j^{(i)})^H \otimes I_N \). Let

\[
R_k := \sum_{i=1}^{|S_k|} \pi_{k1}^{(i)} s_k^{(i)} s_k^{(i)H}, \quad \text{and} \quad Q_k := \sum_{l=1, l\neq k}^K R_l + \sigma^2 I_T,
\]

then \( C_{k1} = R_k \otimes I_N \) and \( C_{k0} = Q_k \otimes I_N \). It follows that the posterior update (53) of the EP scheme can be written as

\[
\hat{p}(s_k = s_k^{(i_k)}| y) = \frac{\mathcal{N}(0; y, (s_k^{(i_k)})^H + Q_k \otimes I_N)}{\sum_{i=1}^{|S_k|} \mathcal{N}(0; y, (s_k^{(i)})^H + Q_k \otimes I_N)} \quad \text{for} \quad i_k \in [|S_k|].
\]

Note that this update can be computed efficiently using

\[
\mathcal{N}(0; y, (s_k^{(i_k)})^H + Q_k \otimes I_N) \propto \frac{1}{(1 + s_k^{(i_k)H}Q_k^{-1}s_k^{(i_k)})^N} \exp\left(\frac{\|YHQ_k^{-1}s_k^{(i_k)}\|^2}{1 + s_k^{(i_k)H}Q_k^{-1}s_k^{(i_k)}}\right).
\]

This simplified scheme can be alternatively constructed as follows. The vector form of the received signal (34) is rewritten as

\[
y = (s_k \otimes I_N)h_k + \sum_{l=1, l\neq k}^K (s_l \otimes I_N)h_l + w.
\]

The second term \( t_k := \sum_{l=1, l\neq k}^K (s_l \otimes I_N)h_l \) is the interference from other users while decoding the signal of user \( k \). Since the signals \( s_l \) are independent of the channels \( h_l \) and the channels \( h_l \) have zero mean, we have that \( \mathbb{E}[t_k] = 0 \). The covariance matrix of \( t_k \) is \( \mathbb{E}[t_k t_k^H] = \sum_{l \neq k} \mathbb{E}[s_l s_l^H] \otimes I_N = \sum_{l \neq k} R_l \otimes I_N \). If we treat the interference term \( t_k \) as a Gaussian vector with the same mean
Algorithm 2: MMSE-SIC for probabilistic non-coherent symbol detection

set the maximal number of iterations \( t_{\text{max}} \);
initialize the posteriors \( \hat{p}(s_k|Y) \) and compute \( R_k = \mathbb{E}_{\hat{p}(s_k|Y)}[s_k s_k^H] \) for \( k \in [K] \);
\( t \leftarrow 0 \);
repeat
  \( t \leftarrow t + 1 \);
  for \( k \leftarrow 1 \) to \( K \) do
    compute \( Q_k = \sum_{l=1, l \neq k}^K R_l + \sigma^2 I_T \);
    update \( \hat{p}(s_k|Y) \) according to (74);
    update \( R_k = \mathbb{E}_{\hat{p}(s_k|Y)}[s_k s_k^H] \);
  end
until convergence or \( t = t_{\text{max}} \);
return \( \hat{p}(s_k|Y) \) for \( k \in \{1, \ldots, K\} \)

and covariance matrix \( [1] \) then \( t_k + w \sim \mathcal{N}(0, Q_k \otimes I_N) \). The single-user likelihood under this approximation is

\[
\hat{p}(y|s_k) = \mathcal{N}(y; 0, (s_k s_k^H + Q_k) \otimes I_N).
\] (77)

With this and Lemma \( [1] \) the update of the approximate posterior \( \hat{p}(s_k|Y) \propto \hat{p}(y|s_k) \) then coincides with (74). The matrix \( R_k \) is then recalculated with the updated value of \( \hat{p}(s_k|Y) \). The matrices \( Q_l \) are updated accordingly, and then used to update \( \hat{p}(s_l|y), l \neq k \).

In short, the derived simplification of the EP scheme above iteratively MMSE-estimates the signal \( z_k \) 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 and summarize it in Algorithm 2. In particular, as for the general EP scheme, we can start with the non-informative initialization \( \hat{p}(s_k = s|Y) = \frac{1}{|S_k|} \mathbb{1}\{s \in S_k\} \).

\( [1] \) Another choice is to treat each \( s_l, l \neq k \), as a Gaussian. With this choice, however, the interference term \( t_k \) is a product of Gaussians which makes the approximate single-user likelihood difficult to evaluate.
VI. IMPLEMENTATION ASPECTS

A. Complexity

In the following complexity analysis, for notational convenience, we assume that $T$ and $N$ are of the same order as $K$, and all individual codebooks have cardinality of the order $O(2^B)$.

For the exact marginalization (6), the complexity of computing $p(Y|S)$ in (3) is $O(K^3)$, coming from the inverse of $\sigma^2 I_K + S' S$. Since this must be computed for all possible $S \in \prod_{k=1}^K S_k$, the total complexity is $O(K^3 2^K)$.

For the EP scheme (Algorithm 1), the dominant operation is the inverse of the $NT \times NT$ matrix $(s_k^{(ik)} s_k^{(ik)\text{H}}) \otimes I_N + C_{k0}$ (with complexity $O(K^6)$) for each $k \in [K]$ and each $i_k \in |S_k|$. This is done in (53), (50), and (52) to update $\hat{n}_{k1}^k$, $\Sigma_{ki}$, and $\hat{z}_{ki}$, respectively. The complexity of computing $\hat{z}_k$, $\Sigma_k$, $\mu_{k1}$, $C_{k1}$, $\{\mu_{j0}\}_{j=1,j\neq k}^K$, and $\{C_{j0}\}_{j=1,j\neq k}^K$ are all of lower order. Therefore, the complexity per iteration of the EP algorithm is given by $O(K^7 2^B)$.

For the MMSE-SIC scheme (Algorithm 2), the complexity per iteration is dominated by the computation of $\hat{p}(s_k|Y)$ in (74). However, as opposed to the EP scheme, it is possible to avoid matrix inversion for each $i_k$: using (75), only the inverse of $Q_k$ is computed, which requires $O(K^3)$ operations. Given $Q_k^{-1}$, the complexity of computing the right-hand side of (75) is then $O(K^2)$ for each $i_k \in |S_k|$. Therefore, the complexity of computing $\hat{p}(s_k|Y)$ is $O(K^3 + K^2 2^B)$ for each user $k \in [K]$. Computing $Q_k$ and $R_k$ both have lower complexity order. Finally, the complexity per iteration of the MMSE-SIC algorithm is given by $O(K^4 + K^3 2^B)$.

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

| Detector | Complexity order |
|-----------|------------------|
| Optimal (with exact marginalization) | $O(K^3 2^K B)$ |
| MMSE-SIC | $O(K^4 t_{\text{max}} + K^3 2^B t_{\text{max}})$ |
| EP | $O(K^7 2^B t_{\text{max}})$ |

$t_{\text{max}}$ denotes the number of iterations.

Although the complexity of the EP scheme is greater than that of MMSE-SIC and POCIS, this is well compensated by the performance gain shown in Section VII. Observe that the key feature that reduces the complexity of MMSE-SIC w.r.t. EP is that the matrix $C_{k0}$ is approximated as a
Kronecker product $\mathcal{C}_{k0} \otimes \mathbf{I}_N$ with $\mathcal{C}_{k0} = \mathbf{Q}_k$, and then $\mathcal{N}(0; \mathbf{\mu}_{k0}, (\mathbf{s}_k^{(i_k)} \mathbf{s}_k^{(i_k)\dagger}) \otimes \mathbf{I}_N + \mathcal{C}_{k0})$ can be expressed as in (75). As explained in Section V, the Kronecker product structure of $\mathcal{C}_{k0}$ stems from a simplification in the derivation of the message $(\mathbf{\mu}_{k1}, \mathcal{C}_{k1})$. Alternatively, we can directly fit $\mathcal{C}_{k0}$ to the form of a Kronecker product without imposing such simplification. This can be done, for example, by solving the least squares

$$\min_{\mathcal{C}_{k0} \in \mathbb{C}^T \times T} \| \mathcal{C}_{k0} - \bar{\mathcal{C}}_{k0} \otimes \mathbf{I}_N \|_F^2$$

as formulated in [24, Sec. 4]. Let $\mathcal{C}_{k0}(i, j)$ 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 $\mathcal{C}_{k0}$. Let $\bar{c}_{ij}$ be the element in row $i$ and column $j$ of $\mathcal{C}_{k0}$. It follows that

$$\| \mathcal{C}_{k0} - \bar{\mathcal{C}}_{k0} \otimes \mathbf{I}_N \|_F^2 = \sum_{i=1}^{T} \sum_{j=1}^{T} \| \mathcal{C}_{k0}(i, j) - \bar{c}_{ij} \mathbf{I}_N \|_F^2$$

$$\bar{c}_{ij} \mathbf{I}_N = \sum_{i=1}^{T} \sum_{j=1}^{T} \| \mathcal{C}_{k0}(i, j) \|_F^2 - \bar{c}_{ij} \text{tr} \{ \mathcal{C}_{k0}(i, j)^\dagger \} - \bar{c}_{ij} \text{tr} \{ \mathcal{C}_{k0}(i, j) \} + N|\bar{c}_{ij}|^2. \quad (79)$$

Observe that $\| \mathcal{C}_{k0} - \bar{\mathcal{C}}_{k0} \otimes \mathbf{I}_N \|_F^2$ is the sum of convex quadratic functions of $\bar{c}_{ij}$. By setting the partials $\frac{\partial \| \mathcal{C}_{k0} - \bar{\mathcal{C}}_{k0} \otimes \mathbf{I}_N \|_F^2}{\partial \bar{c}_{ij}}$ to zeros, the optimal $\bar{\mathcal{C}}_{k0}$ is given by $\bar{c}_{ij} = \frac{1}{N} \text{tr} \{ \mathcal{C}_{k0}(i, j) \}$. With this approximation $\mathcal{C}_{k0} \approx \bar{\mathcal{C}}_{k0} \otimes \mathbf{I}_N$, the complexity of EP becomes equivalent to that of MMSE-SIC.

**B. Stabilization**

We discuss some possible numerical problems in the EP algorithm and our solutions. First, in (59), since the $NT \times NT$ matrix $\mathbf{\Sigma}_k$ 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 $\pi_k^{(i)}$ are sufficiently larger than zero. The singularity of $\mathbf{\Sigma}_k$ can also arise from the codebook structure. For example, the codebooks proposed in [16] are precoded versions of a constellation in $G(\mathbb{C}^{T-K+1}, 1)$ and the maximal rank of $\mathbf{\Sigma}_k$ is $N(T - K + 1) \leq NT$. To avoid the inverse of $\mathbf{\Sigma}_k$, we express $\mathcal{C}_{k1}$ in (61) and $\mathbf{\mu}_{k1}$ in (62) respectively as

$$\mathcal{C}_{k1} = -\mathcal{C}_{k0}(\mathbf{\Sigma}_k - \mathcal{C}_{k0})^{-1}\mathbf{\Sigma}_k, \quad (80)$$

$$\mathbf{\mu}_{k1} = \mathcal{C}_{k0}(\mathbf{\Sigma}_k - \mathcal{C}_{k0})^{-1}\left(\mathbf{\Sigma}_k - \sum_{i=1}^{\left|\mathcal{S}_k\right|} \pi_k^{(i)} \mathbf{\Sigma}_{k1}\right)\mathcal{C}_{k0}^{-1}\mathbf{\mu}_{k0}. \quad (81)$$

Another problem is that $\mathcal{C}_{k1}$ is not guaranteed to be positive definite even if both $\mathcal{C}_{k0}$ and $\mathbf{\Sigma}_k$ are. When $\mathcal{C}_{k1}$ is not positive definite, from (68), $\mathcal{C}_{k0}$ can have negative eigenvalues, which,
through (53), 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 diverges. Note that this “negative variance” problem is common in EP (see, e.g., [18, Sec.3.2.1], [25, Sec.5.3]). There is no generally accepted solution and one normally resorts to various heuristic manipulations which are adapted to each problem. In our problem, to control the eigenvalues of $C_{k1}$, we modify (80) by first eigendecomposing $-C_{k0}(\Sigma_k - C_{k0})^{-1}\Sigma_k = V\Lambda V^{-1}$, then computing $C_{k1}$ as $C_{k1} = V|\Lambda|V^{-1}$, where $|\Lambda|$ is the element-wise absolute value of $\Lambda$. This manipulation by replacing the variance parameters by their absolute values was also used in [26].

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 $S_k$. For example, if $\hat{\pi}_{k1}^{(i_k)}$ is close to a Kronecker-delta distribution with a single mode at $i_0$, then (52) and (50) implies that $\Sigma_k$ is approximately $\Sigma_{k0}$, and then from (61), $C_{k1} \approx (s_k^{(i_0)} s_k^{(i_0)\dagger}) \otimes I_N$. In this case, almost absolute certainty is placed on the symbol $s_k^{(i_0)}$, and the algorithm will not be able to change significantly the messages in the subsequent iterations. This implies convergence, but can be problematic when the mode of $\hat{\pi}_{k1}^{(i_k)}$ 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 $\mathcal{N}(\mathbf{z}_k; \mu_{k1}, C_{k1})$ and $\mathcal{N}(\mathbf{z}_k; \mu_{k0}, C_{k0})$. That is, with a damping factor $\eta \in [0; 1]$, at iteration $t$ and for each user $k$, we update

$$C_{k1}(t) = \eta V(t)|\Lambda(t)|V^{-1}(t) + (1 - \eta)C_{k1}(t-1),$$

$$\mu_{k1}(t) = \eta C_{k0}(t-1)(\Sigma_k(t) - C_{k0}(t-1))^{-1}\left(\Sigma_k(t) - \sum_{i=1}^{|S_k|} \pi_{k1}^{(i)}(t)\Sigma_{k1}(t)\right)C_{k0}^{-1}(t-1)\mu_{k0}(t-1) + (1 - \eta)\mu_{k1}(t-1),$$

$$C_{l0}(t) = \eta\left(\sigma^2 I_{NT} + \sum_{j=1,j\neq l}^K C_{j1}(t)\right) + (1 - \eta)C_{l0}(t-1), \quad \forall l \neq k,$$

$$\mu_{l0}(t) = \eta\left(\mathbf{y} - \sum_{j=1,j\neq l}^K \mu_{j1}(t)\right) + (1 - \eta)\mu_{l0}(t-1), \quad \forall l \neq k.$$

In short, we stabilize the EP message updates by replacing (82), (83), (84), and (85) for (61), (62), (68), and (69), respectively. For MMSE-SIC, we damp the update of $Q_k$ and $R_k$ in a similar manner as $Q_k(t) = \eta\left(\sum_{l=1,l\neq k}^K R_l(t-1) + \sigma^2 I_T\right) + (1 - \eta)Q_k(t-1)$ and $R_k(t) = \eta\sum_{i_k=1}^{|S_k|} \pi_{k1}^{(i_k)}(t)s_k^{(i_k)}(s_k^{(i_k)})^\dagger + (1 - \eta)R_k(t-1).$
VII. Performance Evaluation

In this section, we evaluate the performance of our proposed schemes, namely EP and MMSE-SIC, for a given set of individual codebooks. We assume that $B_1 = \ldots B_K =: B$.

A. Test codebooks, a state-of-the-art detector, and benchmarks

We consider the codebook design proposed in [16]. This design imposes a geometric separation between the codebooks of different users through a set of precoders $U_k$ uniquely defined for each user $k \in [K]$. Specifically, starting with a Grassmannian codebook $\mathcal{D} = \{d^{(1)}, d^{(2)}, \ldots, d^{(2^B)}\}$ in $G(\mathbb{C}^{T-K+1}, 1)$, the individual codebook $S_k$ of user $k$ is generated as $s_k^{(i)} = \frac{u_k d^{(i)}}{\|u_k d^{(i)}\|}, i \in [2^B]$.

We consider the precoders $U_k$ defined in [16, Eq.11] and two candidates for $\mathcal{D}$:

1) The numerically optimized codebook generated by solving the max-min distance criteria

$$\max_{d^{(i)} \in G(\mathbb{C}^{T-K+1}, 1), i = 1, \ldots, 2^B} \min_{1 \leq i < j \leq 2^B} d(d^{(i)}, d^{(j)}), \quad (86)$$

where $d(d^{(i)}, d^{(j)}) = \sqrt{1 - |d^{(i)H} d^{(j)}|^2}$ is the chordal distance between two Grassmannian points represented by $d^{(i)}$ and $d^{(j)}$. A codebook with maximal minimum pairwise distance leads to low symbol error rate in the absence of the interference. In our simulation, we approximate the optimization above by $\min_{\mathcal{D}} \log \sum_{1 \leq i < j \leq 2^B} \exp \left(\frac{|d^{(i)H} d^{(j)}|}{\epsilon}\right)$ with a small $\epsilon$ for smoothness, then solve it using gradient descent on the Grassmann manifold using the Manopt toolbox [27]. We label this codebook using the iterative scheme in [28] which propagates the binary labels along the edges of the neighboring graph.

2) The cube-split codebook proposed in [29], [12]. This structured codebook has good distance properties and allows for low-complexity single-user decoding and a simple yet effective binary labeling scheme.

We take the binary labels of the symbols in $\mathcal{D}$ for the corresponding symbols in $S_k$.

Exploiting the precoder structure, [16] introduce a detector [16, 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 a user $k$, POCIS first estimates the row space of the interference $\sum_{l=1,l\neq k}^K s_l h_l^T$ based on the precoders and projects the received signal onto the orthogonal complement of this row space. It then performs a single-user detection to obtain point estimates.

$^2$The codebook $\mathcal{D}$ can be different for different users. Here, we consider a common codebook.
of the transmitted signals. 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. Since the estimation of the row space and the column space of the interference both has complexity $O(K^3)$, and the single-user likelihood computation has complexity $O(K^22^B)$, the complexity per iteration of the POCIS detector is $O(K^4 + K^32^B)$, equivalent to the MMSE-SIC 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 and MMSE-SIC.

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 $s_k$ of user $k$, the knowledge of the signals $s_l$ (but not the channel $h_l$) of all the interfering users $l \neq k$. With this genie-aided information, optimal ML decoding (4) can be performed by keeping $s_l$ fixed for all $l \neq k$ and searching for the best $s_k$ in $S_k$, 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-SIC, and POCIS.

In the following, we set the number of iterations for EP and MMSE-SIC as 20 and for POCIS as 3 since it quickly converges. From numerical evaluation, we have verified that EP operates better under a low damping factor while MMSE-SIC operates better under a high one. Therefore, we set the damping factor as $\eta = 0.3$ for EP and $\eta = 0.8$ for MMSE-SIC.

### B. Achievable Rate

We first plot the mismatched sum-rate lower bound $R_{GMI}$ of the system calculated as in (10) for $T = 6$, $K = 3$, $N \in \{4, 8\}$, and different constellation sizes $2^B$ in Fig. 2. In Fig. 2a, we consider $B = 4$ bits/symbol and use the numerically optimized codebook for $D$. The rates achieved with EP and MMSE-SIC detectors are very close to the achievable rate of the system (with optimal detector). In Fig. 2b, we consider a larger constellation size with $B = 8$ bits/symbol, and use the cube-split codebook for $D$. The rates achieved with EP and MMSE-SIC detectors are not far from the rate achieved with the genie-aided detector. The rate achieved with the POCIS detector is much lower than the others in the lower SNR regime and converges slower to the limit $\frac{BK}{T}$ bits/channel use, even though it was designed specifically for the considered codebook structure.
Fig. 2. The mismatched rate of the system with EP, MMSE-SIC, and POCIS detectors in comparison with the optimal detector or a genie-aided detector for coherence time $T = 6$, $K = 3$ users, and $N \in \{4, 8\}$ receive antennas.

C. Symbol error rates of hard detection

Next, we use the outputs of EP, MMSE-SIC and POCIS for a maximum-a-posteriori (MAP) hard detection. We evaluate the performance in terms of symbol error rate (SER).

In Fig. (3), we consider the same setting as in Fig. 2-a, for which the optimal ML detector is feasible. We observe that the SER of the EP and MMSE-SIC detectors are not much higher than that of the optimal detector, especially in the lower SNR regime. EP performs better than MMSE-SIC when the SNR increases. Both EP and MMSE-SIC outperform POCIS.

In Fig. 4, we consider the same setting as in Fig. 2-b, for which we use the genie-aided detector as a benchmark. The performance of EP is very close to this genie-aided detector and better than MMSE-SIC at $\text{SNR} \geq 10$ dB. Both EP and MMSE-SIC are better than POCIS. We also show the SER of two other multiple-access schemes with the same transmitted data rate. First, we consider a non-coherent time division multiple access (TDMA) scheme where each user transmits from a cube-split constellation of size $2^{BK}$ in $G\left(C^T, 1\right)$ in a round-robin manner. Second, we consider a coherent pilot-based scheme in which the users send orthogonal pilots in the first 3 channel uses and QAM data symbols in the remaining 3 channel uses, the receiver performs MMSE channel estimation based on the pilot symbols and MMSE equalization on the data symbols using the channel estimate. These latter two schemes are outperformed by the non-coherent multiple-access scheme with EP, MMSE-SIC, and POCIS detectors.
Fig. 3. The symbol error rate of the system with EP, MMSE-SIC, and POCIS detectors in comparison with the optimal detector for $T = 6$, $K = 3$, $N \in \{4, 8\}$ and a transmission rate of $B = 4$ bits/user/coherence block.

Fig. 4. The symbol error rate of the system with 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 a transmission rate of $B = 8$ bits/user/coherence block.

D. Bit error rates with a channel code

In this subsection, we use the cube-split codebook for $D$. We integrate a standard symmetric parallel concatenated rate-1/3 turbo code [21]. The turbo encoder accepts packets of 1008 bits; the turbo decoder computes the bit-wise LLR from the soft outputs of the detection scheme and performs 10 decoding iterations for each coded bit packet.

In Fig. 5, 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 (6). The BER of MMSE-SIC vanishes slower with the SNR than the other schemes, and becomes better than POCIS as $K$ and $N$ increase. 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 $3 \text{ dB}$ and $4 \text{ dB}$, respectively. We also observe that the genie-aided detector only provides a rough benchmark w.r.t. the optimal detector.

Finally, in Fig. [6] we consider $T = 6$, $K = 3$, $N = 4$, and compare the BER (with the same turbo code) for different constellation sizes. 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 due to the residual negative effect (after damping) of the phenomenon that all the mass of the pmf $\pi_{k1}^{(0)}$ 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. [5] the performance of MMSE-SIC is significantly improved as the number of receive antennas increases from $N = 3$ to $N = 4$.

VIII. Conclusion

We proposed an expectation propagation based scheme and a MMSE-SIC scheme for soft-output multi-user detection in non-coherent SIMO multiple access channel. The latter scheme
Fig. 6. The bit error rate with turbo codes of EP, MMSE-SIC, POCIS, and the optimal/genie-aided detector for $T = 6$, $K = 3$, and $N = 4$.

can be interpreted as a simplification of the former. Both schemes are shown to achieve good performance, especially the expectation propagation scheme, 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. The advantage of the expectation propagation scheme is more significant when the number of user and/or the constellation size increase.

APPENDIX

A. Properties of the Gaussian pdf

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

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

2) Gaussian pdf multiplication rule:

$$
\mathcal{N}(\mathbf{x}; \mu_1, \Sigma_1)\mathcal{N}(\mathbf{x}; \mu_2, \Sigma_2) = \mathcal{N}(\mathbf{x}; \mu_{\text{new}}, \Sigma_{\text{new}})\mathcal{N}(\mathbf{0}; \mu_1 - \mu_2, \Sigma_1 + \Sigma_2),
$$

where $\mu_{\text{new}} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$ and $\Sigma_{\text{new}} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$.

Proof. The first part follows readily from the definition of $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$. For the second part, we express the Gaussian pdf as

$$
\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{\exp(-\mathbf{x}^\dagger \Sigma^{-1} \mathbf{x} + \mathbf{x}^\dagger \Sigma^{-1} \mu + \mu^\dagger \Sigma^{-1} \mathbf{x} - \mu^\dagger \Sigma^{-1} \mu)}{\pi^n \det(\Sigma)}.
$$

(88)
Thus
\[ N(x; \mu_1, \Sigma_1)N(x; \mu_2, \Sigma_2) \]
\[ = \exp \left( \begin{array}{c}
-x^T(\Sigma_1^{-1} + \Sigma_2^{-1})x + x^T(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) + (\mu_1^T\Sigma_1^{-1} + \mu_2^T\Sigma_2^{-1})x - \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2
\end{array} \right) \]
\[ \pi^{2n}\det(\Sigma_1)^{1/2}\det(\Sigma_2)^{1/2} \]
\[ = \exp \left( \begin{array}{c}
-\left(x - \mu_{\text{new}}\right)^T(\Sigma_{\text{new}}^{-1})\left(x - \mu_{\text{new}}\right) - \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2
\end{array} \right) \]
\[ \pi^{2n}\det(\Sigma_1)^{1/2}\det(\Sigma_2)^{1/2} \]
\[ = C(\mu_1, \mu_2, \Sigma_1, \Sigma_2) \exp \left( -(x - \mu_{\text{new}})^T\Sigma_{\text{new}}^{-1}(x - \mu_{\text{new}}) \right) \pi^{n}\det(\Sigma_{\text{new}}) \]
\[ = C(\mu_1, \mu_2, \Sigma_1, \Sigma_2)N(x; \mu_{\text{new}}, \Sigma_{\text{new}}), \]
where \( \mu_{\text{new}} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2), \Sigma_{\text{new}} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}; \) in the second equality, we used the fact that \( \Sigma_{\text{new}}^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1} \) and \( \Sigma_{\text{new}}^{-1}\mu_{\text{new}} = \Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2; \) and the scaling factor
\[ C(\mu_1, \mu_2, \Sigma_1, \Sigma_2) = \frac{\det(\Sigma_{\text{new}})}{\pi^n\det(\Sigma_1)^{1/2}\det(\Sigma_2)^{1/2}} \exp \left( \mu_{\text{new}}^T\Sigma_{\text{new}}^{-1}\mu_{\text{new}} - \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2 \right). \]

Using the identities \( (A^{-1} + B^{-1})^{-1} = A(A + B)^{-1}B \) and \( (I + AB)^{-1} = I - A(I + BA)^{-1}B, \) after some manipulations, we deduce that \( C(\mu_1, \mu_2, \Sigma_1, \Sigma_2) = N(0; \mu_1 - \mu_2, \Sigma_1 + \Sigma_2). \) This completes the proof of the Gaussian pdf multiplication rule. \( \square \)

**B. Proof of Proposition 1**

Using the natural logarithm for the KL divergence, we can derive that
\[ D(q_\alpha(x) || p(x)) = \int q_\alpha(x) \ln \frac{q_\alpha(x)}{p(x)} \, dx \]
\[ = \int q_\alpha(x) \ln \frac{q_\alpha(x)}{\prod_\beta p_\beta(x_\beta)} \, dx \]
\[ = \sum_\beta \int q_\alpha(x) \ln \frac{1}{p_\beta(x_\beta)} \, dx + c_0 \]
\[ = \sum_{\beta \in \mathcal{G}_\alpha} \int q_\alpha(x) \ln \frac{1}{p_\beta(x_\beta)} \, dx + \sum_{\beta \not\in \mathcal{G}_\alpha} \int q_\alpha(x) \ln \frac{1}{p_\beta(x_\beta)} \, dx + c_0 \]
\[ = \sum_{\beta \in \mathcal{G}_\alpha} \int q_\alpha(x) \ln \frac{1}{p_\beta(x_\beta)} \, dx + \sum_{\beta \not\in \mathcal{G}_\alpha} \int \hat{p}_\beta(x_\beta) \ln \frac{1}{p_\beta(x_\beta)} \, dx + c_0 \]
\[ = -\sum_{\beta \in \mathcal{G}_\alpha} \int q_\alpha(x) \left[ \gamma_\beta^T\phi(x_\beta) - A_\beta(\gamma_\beta) \right] \, dx + \sum_{\beta \not\in \mathcal{G}_\alpha} D(\hat{p}_\beta || p_\beta) + c_0 \]
\[ = \sum_{\beta \in \mathcal{G}_\alpha} \left( A_\beta(\gamma_\beta) - \gamma_\beta^T E_{q_\alpha} [\phi(x_\beta)] \right) + \sum_{\beta \not\in \mathcal{G}_\alpha} D(\hat{p}_\beta || p_\beta) + c_0. \]
where (98) follows from
\[
q_\alpha(x) = \frac{\psi_\alpha(x)}{m_\alpha(x)} \left[ \prod_{\beta \in \Omega_\alpha} \hat{p}_\beta(x) \right] \left[ \prod_{\beta \notin \Omega_\alpha} \hat{p}_\beta(x) \right],
\]
and (99) follows from (16). From (100), we can see that the optimization (21) of \( p \) decouples over \( p_\beta \), and the optimal distribution can be expressed as \( \hat{p}^{\text{new}}_{\alpha,\beta}(x) = \prod_{\beta} \hat{p}^{\text{new}}_{\alpha,\beta}(x) \). For \( \beta \notin \Omega_\alpha \), the minimum of \( D(\hat{p}_\beta || \hat{p}_\beta) \) is simply 0 and achieved with \( \hat{p}^{\text{new}}_{\alpha,\beta}(x) = \hat{p}_\beta(x) \). For \( \beta \in \Omega_\alpha \), since the log-partition function \( A_\beta(\gamma) \) is convex in \( \gamma \) (see, e.g., [30] Lemma 1), the minimum of \( A_\beta(\gamma) - \gamma^T E_{\hat{p}_\beta} [\phi(x)] \) is achieved at the value of \( \gamma \) where its gradient is zero. Using the well-known property of the log-partition function, \( \nabla_{\gamma} A_\beta(\gamma) = E_{\hat{p}_\beta} [\phi(x)] \), we get that the zero-gradient equation is equivalent to the moment matching criterion \( E_{\hat{p}^{\text{new}}_{\alpha,\beta}} [\phi(x)] = E_{q_\alpha} [\phi(x)] \). This completes the proof.

REFERENCES

[1] I. 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] E. G. Larsson, “Massive MIMO for 5G: Overview and the road ahead,” in 51st Annual Conference on Information Sciences and Systems (CISS), Mar. 2017.
[4] 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.
[5] S. Verdú, “Computational complexity of optimum multiuser detection,” Algorithmica, vol. 4, no. 1, pp. 303–312, Jun 1989.
[6] B. L. Hughes, “Differential space-time modulation,” IEEE Trans. Inf. Theory, vol. 46, no. 7, pp. 2567–2578, Nov 2000.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.

[15] J. Cáspedes, P. M. Olmos, M. Sánchez-Fernández, 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.

[16] 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.

[17] 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.

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

[19] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, “Approximate inference techniques with expectation constraints,” J. Stat. Mech: Theory Exp., p. 11015, 2005.

[20] P. J. Bickel and K. A. Doksum, Mathematical statistics : Basic ideas and selected topics. Holden-Day San Francisco, 1977.

[21] 3rd Generation Partnership Project (3GPP), Multiplexing and channel coding, Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Std. 3GPP TS 36.212 V8.0.0 (2007-09).

[22] A. Ganti, A. Lapidoth, and I. E. Telatar, “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.

[23] 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.

[24] C. F. Van Loan and N. Pitsianis, “Approximation with Kronecker products,” in Linear algebra for large scale and real-time applications. Springer, 1993, pp. 293–314.

[25] A. Vehtari, A. Gelman, T. Sivula, P. Jylänki, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. Robert, “Expectation propagation as a way of life: A framework for bayesian inference on partitioned data,” arXiv preprint arXiv:1412.4869, 2014.

[26] P. Sun, C. Zhang, Z. Wang, C. N. Manchón, 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.

[27] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, “Manopt, a Matlab toolbox for optimization on manifolds,” Journal of Machine Learning Research, vol. 15, pp. 1455–1459, 2014. [Online]. Available: http://www.manopt.org

[28] S. Baluja and M. Covell, “Neighborhood preserving codes for assigning point labels: Applications to stochastic search,” in International Conference on Computational Science, 2013, pp. 956–965.

[29] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, “Design and analysis of a practical codebook for non-coherent communications,” in 51st Asilomar Conference on Signals, Systems, and Computers, CA, USA, Oct. 2017, pp. 1237–1241.

[30] 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.