MIMO Detection via Gaussian Mixture Expectation Propagation: A Bayesian Machine Learning Approach for High-Order High-Dimensional MIMO Systems

1st Shachar Shayovitz
Toga Networks
Hod Hasharon, Israel
shachar.shayovitz@huawei.com

2nd Doron Ezri
Toga Networks
Hod Hasharon, Israel
doron.ezri@huawei.com

3rd Yoav Levinbook
Toga Networks
Hod Hasharon, Israel
yoav.levinbook@huawei.com

***Abstract*—MIMO systems can simultaneously transmit multiple data streams within the same frequency band, thus exploiting the spatial dimension to enhance performance. MIMO detection poses considerable challenges due to the interference and noise introduced by the concurrent transmission of multiple streams. Efficient Uplink (UL) MIMO detection algorithms are crucial for decoding these signals accurately and ensuring robust communication. In this paper a MIMO detection algorithm is proposed which improves over the Expectation Propagation (EP) algorithm. The proposed algorithm is based on a Gaussian Mixture Model (GMM) approximation for Belief Propagation (BP) and EP messages. The GMM messages better approximate the data prior when EP fails to do so and thus improve detection. This algorithm outperforms state of the art detection algorithms while maintaining low computational complexity.**

***Index Terms*—MIMO Detection, Bayesian Machine Learning, Belief Propagation, Expectation Propagation, Gaussian Mixture Models**

I. Introduction

In recent years, Multiple-Input Multiple-Output (MIMO) technology has emerged as a cornerstone in modern wireless communication systems, offering significant improvements in spectral efficiency, reliability, and capacity [1], [2]. By leveraging multiple antennas at both the transmitter and receiver, MIMO systems can simultaneously transmit multiple data streams within the same frequency band, thus exploiting the spatial dimension to enhance performance.

The detection of signals in MIMO systems, however, poses considerable challenges due to the interference and noise introduced by the concurrent transmission of multiple streams. This task is increasingly challenging as the dimensions of the system (number of transmit and receive antennas) and the constellation order (the number of symbols in the modulation scheme) grow.

Traditional detection methods, such as Maximum Likelihood Detection (MLD), provide optimal performance but are computationally prohibitive for large-scale MIMO systems. The complexity of MLD grows exponentially with the number of antennas and the constellation order. For instance, in a system with N_t transmit antennas and M-QAM modulation, the ML detector needs to evaluate M^{N_t} possible transmitted symbol combinations. This exponential growth makes MLD infeasible for large MIMO systems, especially in real-time applications. As a result, suboptimal yet computationally efficient algorithms, including Linear Detectors (e.g., Zero Forcing, Minimum Mean Squared Error), Successive Interference Cancellation, and various heuristic and probabilistic approaches, have been extensively studied [3]–[11].

Message-passing algorithms have gained prominence in MIMO detection due to their ability to efficiently handle the inherent complexity of these systems. These algorithms, inspired by Belief Propagation (BP) [12] in graphical models, iteratively exchange messages between nodes in a Factor Graph (FG) representing the joint posterior probability distribution of the transmitted symbols given received observations. This iterative approach facilitates the computation of marginal posterior distribution over the transmitted symbols, enabling effective detection and decoding.

Among the various message-passing techniques, the Sum-Product Algorithm (SPA) and its variants, such as the Approximate Message Passing (AMP) and the Generalized Belief Propagation (GBP), have been extensively studied. These algorithms [13]–[15] leverage the structure of the MIMO channel model to simplify the detection process, achieving a balance between computational complexity and detection accuracy.

Expectation Propagation (EP) [16], a Bayesian Machine Learning technique which is a more recent development in the field, has also shown promise in MIMO detection. EP extends traditional message-passing algorithms by approximating the posterior distribution with a series of moment matching steps, which can improve the accuracy of the approximations. This technique has been successfully applied to various problems in signal processing and machine learning, demonstrating its potential in enhancing the performance of MIMO detection systems.

In summary, the combination of exponential complexity, high-dimensional search space, amplified noise and interference effects, and the need for scalable algorithms make MIMO detection in large dimensions with high constellation orders a formidable challenge. This motivates the ongoing research into more efficient and robust detection algorithms that can operate effectively in these challenging scenarios.

In this paper, a MIMO detection algorithm based on a GMM approximation for EP is proposed. It is observed that the true prior for the data symbols is a discrete uniform distribution and thus the Gaussian prior used in Linear Minimum Mean Square Error (LMMSE) and EP is not accurate. We propose to approximate certain EP messages as GMMs and improve the resulting posterior accuracy.

II. System Model

In this section, the mathematical model for the received UL signal is defined. Let n be the number of data symbols from some constellation (for simplicity we assume the same constellation for all symbols), where $\mathcal{A}$ denotes the set of symbols in the constellation and E_s is the mean symbol energy. The transmitted symbol vector $\mathbf{u} \in \mathcal{A}^n$ is an $n \times 1$ i.i.d. vector. The symbols are transmitted over a flat-fading complex MIMO channel defined by $\mathbf{H} \in \mathbb{C}^{m \times n}$, where each coefficient is drawn according to a proper complex zero-mean unit-variance Gaussian distribution and m is the number of receiving antennas.

The channel output $\mathbf{y} \in \mathbb{C}^m$ is given by:

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

where $\mathbf{n}$ is an additive white circular-symmetric complex Gaussian noise vector with independent zero-mean components and σ_w^2 variance.

Given the model in (1), the posterior probability for the transmitted symbol vector $\mathbf{u}$:

$$p\left(\mathbf{u}|\mathbf{y}\right) \propto p\left(\mathbf{y}|\mathbf{u}\right) p\left(\mathbf{u}\right) \propto \mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{u},\ \sigma_w^2\mathbf{I})\Pi_{i=1}^{n} \mathbb{1}_{u_i \in \mathcal{A}} \tag{2}$$

where $\mathbb{1}_{u_i \in \mathcal{A}}$ is the indicator function that takes value one if $u_i \in \mathcal{A}$ and zero otherwise. Note that $p(\mathbf{u}) \propto \Pi_{i=1}^{n} \mathbb{1}_{u_i \in \mathcal{A}}$ is uniform across all points in $\mathcal{A}^n$.

Similarly to the formulation in [13], the MIMO model is transformed from complex valued to real valued by an equivalent double-sized real-valued representation. This representation is obtained by considering the real and imaginary parts separately. The complex notations are the same as the real notations to the keep the notations uncluttered.

The minimum symbol error optimal detection rule is the Maximum A-Posteriori (MAP) detector which requires the knowledge of the posterior distribution $p\left(u_i|\mathbf{y}\right)$. This posterior can be calculated using marginalization on the joint posterior proposed in (2):

$$p\left(u_i|\mathbf{y}\right) \propto \int_{\mathbf{u}^{-i}} \mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I})\Pi_{j=1}^{N} \mathbb{1}_{u_j \in \mathcal{A}}\, d\mathbf{u}^{-i} \tag{3}$$

where $\mathbf{u}^{-i} = [u_1, u_2, ..., u_{i-1}, u_{i+1}, ..., u_m]$ is the vector of all symbols except u_i.

The multi-dimensional integral in (3) is intractable and approximations are needed in order to find a solution. This is exactly where the BP algorithm comes in and provides an iterative message passing algorithm which computes an approximation of this integral for all u_i. Since this work is focused on the MAP detector for the marginal posterior, we will use the Sum Product Algorithm [17] which is a specific variant of the BP algorithm for computing marginal posteriors.

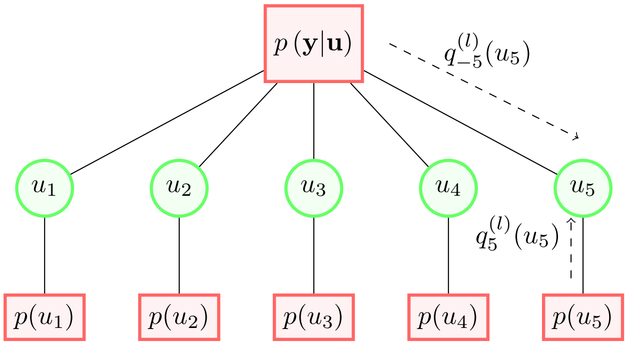

Fig. 1. Factor Graph Representation for 5 data streams MIMO Channel

In order to define the SPA messages, we first introduce the reader to Factor Graphs. A Factor Graph (FG) is a bipartite graph representing the factorization of a function. In probability theory and its applications, FGs are used to represent factorization of a probability distribution function, enabling efficient computations, such as the computation of marginal distributions through the sum–product algorithm. The FG for the posterior in (2) is shown in Fig.1 for $m = 5$. The SPA messages are defined in [17] for a general FG.

III. MIMO Detection via Expectation Propagation

A straightforward implementation of SPA for the FG described Fig.1 is still too complex since the distribution $\mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{u},\ \sigma_w^2\mathbf{I})$ does not factor. In [13], Expectation Propagation (EP) [16] is used in order to perform low complexity MIMO detection for the same FG. EP is an approximation of the BP algorithm [12]. It is an iterative message passing algorithm which estimates the posterior on each data stream u_i while improving the posterior with every iteration. The EP messages are Gaussians which propagate through the FG and iteratively refine the posterior of each data symbol. Effectively the mean and variance are the messages being passed between the nodes in the FG. In this section, for mathematical completeness, we will detail the messages of the EP MIMO detection algorithm as provided in [13].

A. Cavity Update

For all iterations $l \geq 0$, EP uses Gaussian priors for the messages $q_i^{(l)}\left(u_i\right)$ and updates their parameters (mean and covariance) every iteration:

$$q_i^{(l)}\left(u_i\right) = \mathcal{N}(u_i; \gamma_i^{(l)} \Lambda_i^{-1^{(l)}}, \Lambda_i^{-1^{(l)}}) \tag{4}$$

The Gaussian priors propagate upwards in the FG as can be seen for $i = 5$ in Fig.1. All these messages (priors) are multiplied at the function node and the joint posterior is updated:

$$q^{(l)}\left(\mathbf{u}|\mathbf{y}\right) \propto \mathcal{N}(\mathbf{y};\mathbf{H}\mathbf{u},\sigma_w^2\mathbf{I})\Pi_{i=1}^{N} q_i^{(l)}\left(u_i\right) \tag{5}$$

Multiplication of Gaussians produces a Gaussian with the following mean and covariance:

$$\mathbf{\Sigma}^{(\mathbf{l})} = \left(\sigma_w^{-2}\mathbf{H}^T\mathbf{H} + \text{diag}\left(\mathbf{\Lambda}^{(l)}\right)\right)^{-1} \tag{6}$$

$$\boldsymbol{\mu}^{(l)} = \mathbf{\Sigma}^{(l)}\left(\sigma_w^{-2}\mathbf{H}^T\mathbf{y} + \boldsymbol{\gamma}^{(l)}\right) \tag{7}$$

The resulting Gaussian mean (7) is effectively the LMMSE estimator based on the updated Gaussian priors. For $l = 0$, the priors are zero mean Gaussians and therefore we get the conventional MIMO LMMSE detector.

The messages propagating downwards in the FG (denoted cavity in [13]) are computed based on the BP algorithm:

$$q_{-i}^{(l)}\left(u_i|\mathbf{y}\right) \propto \int_{\mathbf{u}^{-i}} \mathcal{N}(\mathbf{y};\mathbf{H}\mathbf{u},\sigma_w^2\mathbf{I})\Pi_{j\neq i} q_j^{(l)}\left(u_j\right)\, d\mathbf{u}^{-i} \tag{8}$$

Since the priors are Gaussians, then (8) is a Gaussian with a simple analytical expression. The resulting Gaussian takes the corresponding elements from the mean vector, $\mu_i^{(l)}$, in (7) and covariance matrix diagonal, $\sigma_i^{2^{(l)}}$, in (6):

$$q_{-i}^{(l)}\left(u_i|\mathbf{y}\right) = \mathcal{N}(u_i; t_i^{(l)}, h_i^{2^{(l)}}) \tag{9}$$

where:

$$h_i^{2(l)} = \frac{\sigma_i^{2(l)}}{1-\sigma_i^{2(l)}\Lambda_i^{(l)}}$$

$$t_i^{(l)} = h_i^{2(l)}\left(\frac{\mu_i^{(l)}}{\sigma_i^{2(l)}} - \gamma_i^{(l)}\right)$$

The i-th cavity takes into account the channel model and the information from all the priors associated with data symbols other than the i-th symbol.

In this paper we consider, for the ease of presentation, the uncoded case and an extension to the coded case is straightforward. The cavity messages are used to compute Log Likelihood Ratios (LLRs) which are sent to an error correcting code such as LDPC (Turbo decoding [18]).

B. Prior Update

The main innovation in [13] is the method by which the prior, $q_i^{(l+1)}\left(u_i\right)$, is updated. In order to keep $q_i^{(l+1)}\left(u_i\right)$ Gaussian, the following approximation is done on the joint posterior of u_i:

$$q_i^{(l+1)}\left(u_i\right) q_{-i}^{(l)}\left(u_i\right) \approx q_{-i}^{(l)}\left(u_i\right)\mathbb{1}_{u_i\in\mathcal{A}} \tag{10}$$

The right hand side of (10) is not a Gaussian and thus a projection to the Gaussian parametric family needs to be taken (moment matching minimizes the Kullback Liebler Divergence (KLD)).

$$\mathcal{N}(u_i;\mu_{p_i}^{(l)},\sigma_{p_i}^{2^{(l)}}) = \text{proj}\left(q_{-i}^{(l)}\left(u_i\right)\mathbb{1}_{u_i\in\mathcal{A}}\right) \tag{11}$$

where $\mu_{p_i}^{(l)}$ and $\sigma_{p_i}^{2^{(l)}}$ are the mean variance generated using moment matching.

The updated prior is then computed:

$$q_i^{(l+1)}\left(u_i\right) = \frac{\text{proj}\left(q_{-i}^{(l)}\left(u_i\right)\mathbb{1}_{u_i\in\mathcal{A}}\right)}{q_{-i}^{(l)}\left(u_i\right)} \tag{12}$$

Therefore:

$$\begin{aligned} q_i^{(l+1)}\left(u_i\right) &= \frac{\mathcal{N}(u_i;\mu_{p_i}^{(l)},\sigma_{p_i}^{2^{(l)}})}{\mathcal{N}(u_i;t_i^{(l)},h_i^{2^{(l)}})} \\ &\propto \mathcal{N}(u_i;\gamma_i^{(l+1)}\Lambda_i^{-1^{(l+1)}},\Lambda_i^{-1^{(l+1)}}) \end{aligned} \tag{13}$$

where:

$$\gamma_i^{(l+1)} = \frac{\mu_{p_i}^{(l)}}{\sigma_{p_i}^{2^{(l)}}} - \frac{t_i^{(l)}}{h_i^{2(l)}}$$

$$\Lambda_i^{(l+1)} = \frac{1}{\sigma_{p_i}^{2^{(l)}}} - \frac{1}{h_i^{2(l)}}$$

The process repeats with the updated prior using (4).

IV. Gaussian Mixture Messages

However, the Gaussian division in (13) can produce a Gaussian with a negative variance when $\sigma_{p_i}^{2^{(l)}} > h_i^{2(l)}$, meaning that the updated prior, $q_i^{(l+1)}\left(u_i\right)$, is not a normalized Gaussian. In [13], this is mitigated by identifying the negative variance prior messages and replacing them with zero mean and E_s variance Gaussians. Effectively, it means that the updated prior is uninformative and does not utilize the cavity information for the subsequent iteration.

In this section, we propose to use a GMM and not a single Gaussian when negative variances arise. We first realize that if only one constellation symbol has a dominant probability in the joint posterior, then $\sigma_{p_i}^{2^{(l)}} \leq h_i^{2(l)}$. Therefore, when $\sigma_{p_i}^{2^{(l)}} > h_i^{2(l)}$, there may be several constellation symbols which are similarly likely. An illustration of this is provided in Fig.2 where we can see the true prior (delta functions), cavity message, joint posterior and the Gaussian approximation for the joint posterior. It can be observed that in this case the variance of the true posterior is larger than the cavity ($\sigma_{p_i}^{2^{(l)}} > h_i^{2(l)}$), which results with a negative variance for the updated prior. As observed, the multiplication of the prior and cavity distributions differs from a single Gaussian distribution if the variance (uncertainty) of the cavity is larger than the distance between two adjacent real constellation points. Our approach is to improve this approximation so the updated prior will provide refined information to the computation of the updated cavity.

We propose the following approximation for the true prior:

$$\mathbb{1}_{u_i\in\mathcal{A}} \approx \sum_{k=1}^{K}\alpha_k\mathcal{N}\left(u_i;a_k,\sigma_0^2\right) \tag{14}$$

where $\sigma_0^2, a_k \in \mathcal{A}$ and α_k are the variance, mean of a Gaussian around a real value constellation point and the

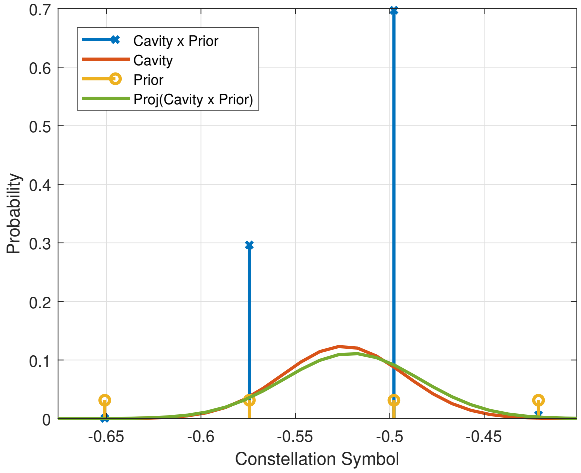

Fig. 2. Gaussian Approximation for Messages

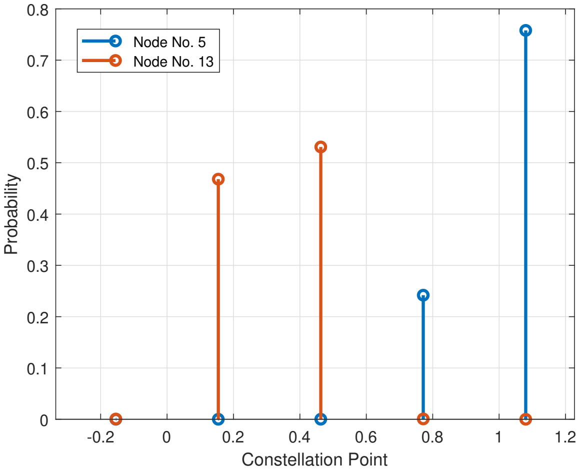

Fig. 3. Joint Posterior for Negative Variance Variable Nodes

mixing coefficient respectively. This approximation is accurate when $\sigma_0^2 \to 0$ and $\alpha_k = 1$ for all k.

Using the mixture approximation (14), we can write the prior message for u_s:

$$q_s^{(l+1)}(u_s) = \sum_{k=1}^{K} \alpha_k q_{s,k}^{(l+1)}(u_s) \tag{15}$$

where:

$$q_{s,k}^{(l+1)}(u_s) = \mathcal{N}(u_s; \gamma_{s,k}^{(l+1)} \Lambda_{s,k}^{-1^{(l+1)}}, \Lambda_{s,k}^{-1^{(l+1)}}) \tag{16}$$

and $a_k = \gamma_{s,k}^{(l+1)} \Lambda_{s,k}^{-1^{(l+1)}}$ and $\sigma_0^2 = \Lambda_{s,k}^{-1^{(l+1)}}$.

A. Selection of Variable Node for GMM Approximation

Ideally we would like to model *all* the variable nodes with negative variance with GMMs but this will incur an unreasonable computational burden. The subsequent cavity computation will include an integral on a multiplication of several mixtures and the complexity becomes exponential. Taking into account complexity considerations, the mixture approximation may be used for a limited number of variable nodes with negative updated prior variance.

In this work, we chose the nodes (corresponding to negative variance) with the lowest entropy of their joint posterior, $q_{-i}^{(l)}(u_i) \mathbb{1}_{u_i \in \mathcal{A}}$. The idea is that these nodes have the lowest uncertainty but their respective EP prior approximation is very uncertain. Therefore, if we model them using the true prior (mixture) then we can infer them more easily than higher entropy posteriors. In order to illustrate this idea, an example from the EP process is provided, where the variable nodes 5 and 13 have a negative variance. In Fig.3 the joint posteriors for both variable nodes are plotted. The EP algorithm [13] will approximate both of them using the same Gaussian. If we can only model one of them using a GMM, we argue that selecting node 5 would be better. This is since the updated prior will provide better information for the subsequent EP process. This approach was compared to random selection empirically and provided superior results.

The nodes u_s which are selected for mixture message approximation can be changed per message passing iteration. We note also that the prior update for the other nodes is unchanged from the EP algorithm.

B. Cavity Update with a Gaussian Mixture

As described in the previous section, the Gaussian priors (and mixture) propagate upwards in the FG as can be seen for $i = 5$ in Fig.1. All these messages are multiplied at the function node and the joint posterior is updated. In the following derivation, we consider a single GMM message in the FG for u_s. The extension to more GMM messages is straightforward. Based on the BP messages, the cavity for all the nodes *except* u_s is computed using the following expression:

$$q_{-i}^{(l+1)}(u_i) = \sum_{k=1}^{K} \alpha_k \int_{\mathbf{u}^{-i}} \mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) \Pi_{j \neq s,i} q_j^{(l+1)}(u_j)\, q_{s,k}^{(l+1)}(u_s)\, d\mathbf{u}^{-i} \tag{17}$$

We can re-write (17) as:

$$q_{-i}^{(l+1)}(u_i) = \sum_{k=1}^{K} \alpha_k q_k(\mathbf{y}) \int_{\mathbf{u}^{-i}} \frac{q_k^{(l+1)}(\mathbf{u}|\mathbf{y})}{q_i^{(l+1)}(u_i)} d\mathbf{u}^{-i} \tag{18}$$

where:

$$q_k^{(l+1)}(\mathbf{u}|\mathbf{y}) = \frac{\mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) \Pi_{j \neq s} q_j^{(l+1)}(u_j)\, q_{s,k}^{(l+1)}(u_s)}{q_k(\mathbf{y})} \tag{19}$$

and

$$q_k(\mathbf{y}) = \int_{\mathbf{u}} \mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) \Pi_{j \neq s} q_j^{(l+1)}(u_j)\, q_{s,k}^{(l+1)}(u_s)\, d\mathbf{u} \tag{20}$$

The cavity (18) is a GMM where the mixing coefficients are $\alpha_k q_k(\mathbf{y})$ and $q_k(\mathbf{y})$ are the likelihoods for the observed data, $\mathbf{y}$, based on each mixture component. We observe that (19) is a normalized Gaussian with the following mean and covariance:

$$\mathbf{\Sigma}_k^{(l+1)} = \left(\sigma_w^{-2}\mathbf{H}^T\mathbf{H} + \text{diag}\left(\mathbf{\Lambda}_k^{(l+1)}\right)\right)^{-1} \tag{21}$$

$$\mu_{\mathbf{k}}^{(\mathbf{l+1})} = \mathbf{\Sigma}_k^{(l+1)}\left(\sigma_w^{-2}\mathbf{H}^T\mathbf{y} + \gamma_k^{(l+1)}\right) \tag{22}$$

where $\mathbf{\Lambda}_{s,k}^{-1^{(l+1)}} = \sigma_0^2$ which means that $\mathbf{\Sigma}_k^{(l+1)}$ is the same for all mixture components, k, and thus can be computed once. The mean $\mu_k^{(l+1)}$ changes for different constellation points (different $\gamma_k^{(l+1)}$ per k). However, its computation does not require matrix inversion and thus is linear with the number mixture components.

We compute $q_k(\mathbf{y})$ using the fact that $\mathbf{y}$ is Gaussian and thus $\mu_{\mathbf{y}} = \mathbf{H}\mu_{\mathbf{u}}$ and $\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{H}\mathbf{\Sigma}_{\mathbf{u}}\mathbf{H}^T + \sigma_w^2\mathbf{I}$. Therefore:

$$q_k(\mathbf{y}) \propto e^{-(\mathbf{y}-\mathbf{H}\mu_{\mathbf{u}})^T\left(\mathbf{H}\mathbf{\Lambda}^{(l+1)^{-1}}\mathbf{H}^T+\sigma_w^2\mathbf{I}\right)^{-1}(\mathbf{y}-\mathbf{H}\mu_{\mathbf{u}})} \tag{23}$$

The reason the Gaussian normalization factor is not taken into account is due to the fact that the covariance, $\mathbf{\Sigma}_{\mathbf{y}}$ is the same for all mixture components. Currently, per EP iteration, the following matrix inversion is performed: $\left(\sigma_w^{-2}\mathbf{H}^T\mathbf{H} + \text{diag}\left(\mathbf{\Lambda}_k^{(l+1)}\right)\right)^{-1}$. However, in (23) a different inversion is needed: $\left(\mathbf{H}\mathbf{\Sigma}_{\mathbf{u}}\mathbf{H}^T + \sigma_w^2\mathbf{I}\right)^{-1}$. In order to reduce complexity and avoid another direct matrix inversion, we look at the probability of the sufficient statistics, $\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}$:

$$q_k(\mathbf{y}) \propto e^{-\mathbf{e}^T\left(\mathbf{\Sigma}_{\mathbf{u}}+\sigma_w^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\right)^{-1}\mathbf{e}} \tag{24}$$

where:

$$\mathbf{e} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y} - \mu_{\mathbf{u}} \tag{25}$$

Using the Matrix Inversion Lemma (MIL) [19]:

$$\begin{aligned}\left(\mathbf{\Sigma}_{\mathbf{u}} + \sigma_w^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\right)^{-1} = \text{diag}\left(\mathbf{\Lambda}_k^{(l+1)}\right) - \\ \text{diag}\left(\mathbf{\Lambda}_k^{(l+1)}\right)\mathbf{\Sigma}_k^{(l+1)}\text{diag}\left(\mathbf{\Lambda}_k^{(l+1)}\right)\end{aligned} \tag{26}$$

which uses the inverted matrix already computed and is a multiplication of diagonal matrices which can be further simplified.

Finally, the updated cavity $q_{-i}^{(l+1)}(u_i)$ is approximated using a Gaussian and thus projection using moment matching of a Gaussian mixture to a single Gaussian is used along with the means and variances from (21) and (22):

$$\begin{aligned}&q_{-i}^{(l+1)}(u_i) = \\ &\text{proj}\left(\sum_{k=1}^{K}\alpha_k q_k(\mathbf{y})\frac{\mathcal{N}(u_i;\mu_{i,k}^{(l+1)},\sigma_{i,k}^{2^{(l+1)}})}{\mathcal{N}(u_i;\gamma_i^{(l+1)}\Lambda_i^{-1^{(l+1)}},\Lambda_i^{-1^{(l+1)}})}\right)\end{aligned} \tag{27}$$

We can re-write,

$$q_{-i}^{(l+1)}(u_i) = \text{proj}\left(\sum_{k=1}^{K}\alpha_k q_k(\mathbf{y})\mathcal{N}\left(y;t_{i,k}^{(l+1)},h_{i,k}^{2^{(l+1)}}\right)\right) \tag{28}$$

where:

$$h_{i,k}^{2^{(l+1)}} = \frac{\sigma_{i,k}^{2^{(l+1)}}}{1-\sigma_{i,k}^{2^{(l+1)}}\Lambda_i^{(l+1)}}$$

$$t_{i,k}^{(l+1)} = h_{i,k}^{2^{(l+1)}}\left(\frac{\mu_{i,k}^{(l+1)}}{\sigma_{i,k}^{2^{(l+1)}}} - \gamma_i^{(l+1)}\right)$$

Using moment matching (minimal Kullback Liebler Divergence (KLD)):

$$q_{-i}^{(l+1)}(u_i) = \mathcal{N}\left(y;t_i^{(l+1)},h_i^{2^{(l+1)}}\right) \tag{29}$$

where:

$$t_i^{(l+1)} = \sum_{k=1}^{K}\alpha_k q_k(\mathbf{y})\,t_{i,k}^{(l+1)}$$

$$h_i^{2^{(l+1)}} = \sum_{k=1}^{K}\alpha_k q_k(\mathbf{y})\left(h_{i,k}^{2^{(l+1)}} + \left(t_{i,k}^{(l+1)} - t_i^{(l+1)}\right)^2\right)$$

Note that $h_{i,k}^{2^{(l+1)}}$ is fixed for all k and that can be used to reduce complexity.

In order to compute the cavity for u_s, the following BP message is used:

$$\begin{aligned}q_{-s}^{(l+1)}(u_s) &= \int_{\mathbf{u}^{-s}}\mathcal{N}(\mathbf{y};\mathbf{Hu},\,\sigma_w^2 I)\Pi_{i\neq s}\mathcal{N}(u_i;\gamma_i^{(l+1)}\Lambda_i^{-1^{(l+1)}},\Lambda_i^{-1^{(l+1)}})\,d\mathbf{u}^{-s} \\ &= \frac{\mathcal{N}(u_s;\gamma_{s,0}^{(l+1)}\Lambda_{s,0}^{-1^{(l+1)}},\Lambda_{s,0}^{-1^{(l+1)}})\int_{\mathbf{u}^{-s}}\mathcal{N}(\mathbf{y};\mathbf{Hu},\sigma_w^2 I)\Pi_{i\neq s}\mathcal{N}(u_i;\gamma_i^{(l+1)}\Lambda_i^{-1^{(l+1)}},\Lambda_i^{-1^{(l+1)}})d\mathbf{u}^{-s}}{\mathcal{N}(u_s;\gamma_{s,0}^{(l+1)}\Lambda_{s,0}^{-1^{(l+1)}},\Lambda_{s,0}^{-1^{(l+1)}})}\end{aligned}$$

where $\mathcal{N}(u_s;\gamma_{s,0}^{(l+1)}\Lambda_{s,0}^{-1^{(l+1)}},\Lambda_{s,0}^{-1^{(l+1)}})$ is essentially a dummy prior which is used in order to simplify the computations and use the variances as the diagonal elements in the covariance of a joint posterior. From a computational complexity standpoint we have already computed the covariance and mean associated with this joint posterior (numerator) in (21) and (22).

Therefore,

$$q_{-s}^{(l+1)}(u_s) = \mathcal{N}(y;t_s^{(l+1)},h_s^{2^{(l+1)}}) \tag{30}$$

where:

$$h_s^{2(l+1)} = \frac{\sigma_{s,0}^{2^{(l+1)}})}{1-\sigma_{s,0}^{2^{(l+1)}}\Lambda_{s,0}^{(l+1)}}$$

$$t_s^{(l+1)} = h_s^{2(l+1)}\left(\frac{\mu_{s,0}^{(l+1)}}{\sigma_{s,0}^{2^{(l+1)}}} - \gamma_{s,0}^{(l+1)}\right)$$

Note that the covariance and mean of the joint posterior for u_s does not require matrix inversion. We will insert the dummy variance $\Lambda_{s,0}^{-1^{(l+1)}}$ into the vector $\mathbf{\Lambda}^{-1^{(l+1)}}$ and in order to compute the new covariance, apply MIL. Since the matrix diag $\left(\mathbf{\Lambda}_k^{(l+1)}\right)$ can be viewed as a sum of rank-1 matrices, the change for u_s is on a single rank-1 matrix.

In order to improve the robustness of the algorithm, in [13] it is suggested to smooth the parameter update of the updated priors. In our algorithm we propose to also pass the cavity messages through a low pass filter in order to improve robustness:

$$h_i^{2(l+1)} = \beta h_i^{2(l+1)} + (1-\beta)h_i^{2(l)}$$

$$t_i^{(l+1)} = \beta t_i^{(l+1)} + (1-\beta)t_i^{(l)}$$

where $\beta \in [0, 1]$. We also use all the other techniques for numerical stability as described in [13].

V. Complexity

In this section, the computational complexity of the proposed algorithm is discussed and compared to other low complexity MIMO detection schemes. The mixture algorithm, denoted GMEP (Gaussian Mixture Expectation Propagation), shares most computational blocks with the EP algorithm [13]. The most computationally demanding block is the matrix inversion which is invoked the same number of times in both algorithms ($\mathcal{O}(n^3)$). The most significant added complexity of GMEP on top of EP is the matrix by vector multiplication in (22): $\mathbf{\Sigma}_k^{(l+1)}\gamma_k^{(l+1)}$, which is of order $\mathcal{O}(n^2 M)$, where $M \leq |\mathcal{A}|$ is the number of mixture components and in fact much smaller than $|\mathcal{A}|$ as shown in simulations. All other matrix multiplications are performed once per variable node and do not depend on the mixture order M. The comparison is detailed in Table I, where L is the number of iterations.

It is clear that GMEP becomes attractive compared to EP only for $M \ll n$. We have also included an order of magnitude for the complexity of GMEP with 2 mixture messages. Due to the multiplication of messages in the cavity computation, there are M^2 multiplication per node in this case.

TABLE I
Computational Complexity

Algorithm	Computational Complexity
LMMSE	$\mathcal{O}(n^3)$
EP [13]	$\mathcal{O}(n^3 L + n\lvert\mathcal{A}\rvert L)$
GMEP (1 Node Mix)	$\mathcal{O}(n^3 L + n\lvert\mathcal{A}\rvert L + n^2 ML)$
GMEP (2 Node Mix)	$\mathcal{O}(n^3 L + n\lvert\mathcal{A}\rvert L + 2n^2 ML + n^2 M^2 L)$

VI. Performance Analysis

In this section, we examine the performance of the GMEP algorithm for MIMO detection in high-order high-dimensional scenarios. We have used a MATLAB based simulation and averaged our results for 10^5 symbols. We consider two scenarios of increasing dimension: $n = m = 8$ and $n = m = 12$. We specifically look at scenarios with $n = m$ since LMMSE does not perform well there. The detector performance is shown in terms of the Symbol Error Eate (SER) as a function of the SNR. In both scenarios, we compare the performance of GMEP, EP, Zero Forcing (ZF) [20] and LMMSE [20].

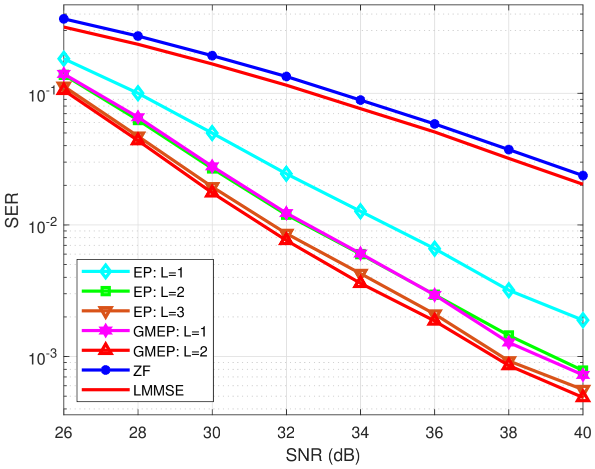

Fig. 4. Comparison of various detectors in a 8x8 system, 64-QAM

Also, we have opted to reduce the number of mixture components using:

$$\alpha_k = \begin{cases} 1, & q_{-s}^{(l)}(u_s = k) > 10^{-3} \\ 0, & \text{otherwise} \end{cases}$$

This choice has empirically provided reasonable results and reduced the mixture order significantly (for 256-QAM, approximately 2 components in a mixture for SER $= 2 \cdot 10^{-3}$). For the GMM message smoothing we have used $\beta = 0.8$ for $L = 2$ and $\beta = 1$ for $L = 1$. Also, we have allowed two messages to be GMMs when needed (if more than 1 prior update is negative).

We first consider $n = m = 8$ for a 64-QAM constellation. In Fig. 4, we can observe that GMEP with $L = 2$ and $L = 1$ outperform all the other algorithms including EP with $L = 2$. For $L = 1$, GMEP has a 2.5dB gain over EP with $L = 1$ and for $L = 2$, GMEP has a gain of 1.5dB over EP with $L = 2$. We also note that GMEP with $L = 2$ is better than EP with $L = 3$ which has higher complexity than GMEP. Both EP and GMEP significantly outperform LMMSE and ZF as expected.

Next, we consider a scenario with $n = m = 12$ with 256-QAM. In Fig. 5, we can see that GMEP with $L = 1$ has comparable performance to EP with $L = 2$ but since $n > M^2$ for high SNR as shown in Fig.6, the complexity of GMEP is lower and thus favorable. The same happens for EP with $L = 3$ and GMEP with $L = 2$, but the gap is larger than for the $n - m - 8$ case. Moreover, for $L = 1$, GMEP is 3dB better than EP and for $L = 2$, GMEP is approximately 2dB better than EP.

We have analyzed the expected mixture order for a variable node in Fig.6. We can see that the expected mixture order decreases as the SNR increases and as the iteration increases, as expected. It is also clear that for the same SNR, the expected

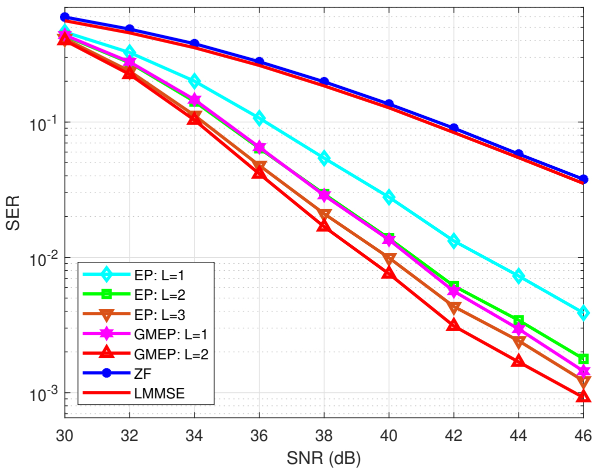

Fig. 5. Comparison of various detectors in a 12x12 system, 256-QAM

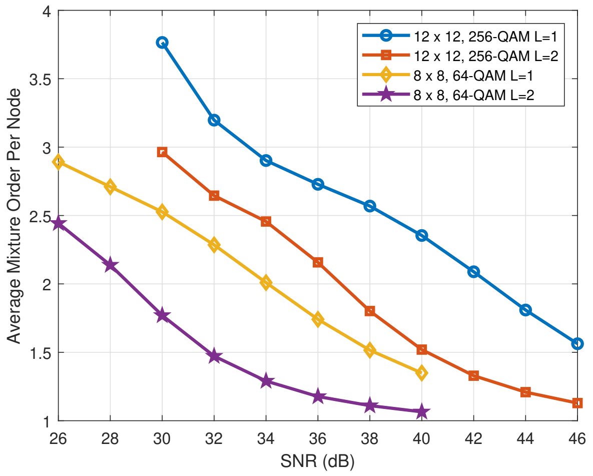

Fig. 6. Average mixture order (M) per variable node

mixture order is larger for a larger constellation size. The complexity of GMEP is dominated by $\mathcal{O}(n^2M^2)$ as opposed to EP's $\mathcal{O}(n^3)$ as detailed in Table I. Therefore, GMEP with L iterations has lower complexity than EP with $L+1$ iterations (for sufficiently $M^2 < n$). In terms of SER performance, it can be seen in Fig.4 and Fig.5, that GMEP $L = 1$ and EP $L = 2$ have similar SER, thus GMEP is favorable.

VII. Discussion

In this paper a MIMO detection algorithm based on a GMM approximation for EP messages was proposed. The manipulation of the GMM messages is performed with a particular focus on balancing the trade-off between computational complexity and detection performance. Through empirical simulations, we demonstrate the efficacy of our proposed scheme in achieving superior detection performance while maintaining practical feasibility for real-time implementation. Our findings highlight the potential of advanced MIMO detection algorithms to significantly enhance the performance of future wireless communication systems, especially for large array order and symbol constellation MIMO systems. In addition, our algorithm provides soft information on the symbols thus enabling seamless integration to a coded communications system.

References

[1] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of mimo channels," *IEEE Journal on selected areas in Communications*, vol. 21, no. 5, pp. 684–702, 2003.

[2] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Transactions on information theory*, vol. 49, no. 5, pp. 1073–1096, 2003.

[3] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for mimo detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1702–1715, 2020.

[4] J. Boutros, N. Gresset, L. Brunel, and M. Fossorier, "Soft-input soft-output lattice sphere decoder for linear channels," in *GLOBECOM'03. IEEE Global Telecommunications Conference (IEEE Cat. No. 03CH37489)*, vol. 3. IEEE, 2003, pp. 1583–1587.

[5] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bolcskei, "Vlsi implementation of mimo detection using the sphere decoding algorithm," *IEEE Journal of solid-state circuits*, vol. 40, no. 7, pp. 1566–1577, 2005.

[6] J. Goldberger and A. Leshem, "Mimo detection for high-order qam based on a gaussian tree approximation," *IEEE transactions on information theory*, vol. 57, no. 8, pp. 4973–4982, 2011.

[7] J. Goldberger, "Improved mimo detection based on successive tree approximations," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 2004–2008.

[8] Z. Guo and P. Nilsson, "Algorithm and implementation of the k-best sphere decoding for mimo detection," *IEEE Journal on selected areas in communications*, vol. 24, no. 3, pp. 491–503, 2006.

[9] T.-H. Liu and Y.-L. Y. Liu, "Modified fast recursive algorithm for efficient mmse-sic detection of the v-blast system," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3713–3717, 2008.

[10] C. Studer, A. Burg, and H. Bolcskei, "Soft-output sphere decoding: Algorithms and vlsi implementation," *IEEE Journal on selected areas in Communications*, vol. 26, no. 2, pp. 290–300, 2008.

[11] P. Švač, F. Meyer, E. Riegler, and F. Hlawatsch, "Soft-heuristic detectors for large mimo systems," *IEEE Transactions on Signal Processing*, vol. 61, no. 18, pp. 4573–4586, 2013.

[12] J. S. Yedidia, W. T. Freeman, Y. Weiss *et al.*, "Understanding belief propagation and its generalizations," *Exploring artificial intelligence in the new millennium*, vol. 8, no. 236–239, pp. 0018–9448, 2003.

[13] J. Cespedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional mimo systems," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2840–2849, 2014.

[14] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large mimo detection via approximate message passing," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1227–1231.

[15] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser mimo-ofdm systems using approximate message passing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 902–915, 2014.

[16] T. P. Minka, "Expectation propagation for approximate bayesian inference," *arXiv preprint arXiv:1301.2294*, 2013.

[17] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.

[18] A. G. Uchoa, C. T. Healy, and R. C. de Lamare, "Iterative detection and decoding algorithms for mimo systems in block-fading channels using ldpc codes," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2735–2741, 2015.

[19] M. A. Woodbury, *Inverting modified matrices*. Department of Statistics, Princeton University, 1950.

[20] D. Tse and P. Viswanath, "Fundamentals of wireless communication12," *Notes*, p. 583, 2004.