This document is published in:

IEEE Transactions on Communications, 62 (2014) 8-August, pp. 2840-2849

DOI: 10.1109/TCOMM.2014.2332349

Expectation Propagation Detection for High-Order High-Dimensional MIMO Systems

Javier Céspedes, Pablo M. Olmos, *Member, IEEE*, Matilde Sánchez-Fernández, *Senior Member, IEEE*, and Fernando Perez-Cruz, *Senior Member, IEEE*

***Abstract*—Modern communications systems use multiple-input multiple-output (MIMO) and high-order QAM constellations for maximizing spectral efficiency. However, as the number of antennas and the order of the constellation grow, the design of efficient and low-complexity MIMO receivers possesses big technical challenges. For example, symbol detection can no longer rely on maximum likelihood detection or sphere-decoding methods, as their complexity increases exponentially with the number of transmitters/receivers. In this paper, we propose a low-complexity high-accuracy MIMO symbol detector based on the Expectation Propagation (EP) algorithm. EP allows approximating iteratively at polynomial-time the posterior distribution of the transmitted symbols. We also show that our EP MIMO detector outperforms classic and state-of-the-art solutions reducing the symbol error rate at a reduced computational complexity.**

***Index Terms*—High-dimensional MIMO communication systems, high-order QAM, low complexity, expectation propagation.**

I. Introduction

MULTIPLE-INPUT MULTIPLE-OUTPUT (MIMO) systems are getting to a mature stage with a significant deployment in several wireless communication systems [1]. MIMO systems increase capacity (throughput) through the multiplexing gain, improve reliability (reduced symbol error rate and outage) and augment transmission range thanks to the diversity (or array) gain [2], [3]. These gains scale with the dimension of the MIMO system, roughly with the number of transmit/receive elements. However, some limitations prevent the widespread deployment of high-dimensional MIMO systems. Specifically, spatial restrictions due to deploying a large number of radiating elements close by and the complexity (energy consumption) of the signal processing at both ends. Despite these issues, novel studies [4], [5] suggest benefits from incorporating a very large number of transmitting/receiving elements and they point out some feasible solutions for its practical implementation.

Symbol estimation and detection is a particularly sensitive process in high-order high-dimensional systems. Even in an additive white Gaussian noise MIMO scenario, assuming perfect channel state information (CSI), a memoryless channel and uniformly distributed transmitted symbols, the maximum likelihood (ML) detector needs to explore all possible transmitted vectors, i.e.,

$$\hat{\mathbf{u}}_{\mathrm{ML}} = \arg\max_{\mathbf{u}\in\mathcal{A}^n} p(\mathbf{u}|\mathbf{y}) = \arg\min_{\mathbf{u}\in\mathcal{A}^n} \|\mathbf{H}\mathbf{u}-\mathbf{y}\|^2 \qquad (1)$$

where $\mathbf{u}$ is the transmitted symbol vector taken from an n-dimensional alphabet $\mathcal{A}^n$ of order $|\mathcal{A}|$, $\mathbf{H}$ is the $m \times n$ MIMO complex channel matrix and $\mathbf{y}$ is channel observation vector. The ML detector in (1) is NP-hard [6] and constitutes a bottleneck for high-order high-dimensional MIMO systems. Sphere decoding (SD) methods try to replicate the performance of ML by solving the minimization in a sub-space of $\mathcal{A}^n$ [7]–[12]. However, the dimension of this sub-space must grow rapidly with n, the modulation order and the inverse of the signal-to-noise ratio (SNR) to maintain the good performance, making prohibitive its computational complexity in very large MIMO systems. In particular, the genetic soft-heuristic algorithm (GSA) detector recently proposed in [11], despite outperforming some of the best sphere decoding methods such as SUMIS [12], is only effective for low-order constellations (BPSK and QPSK).

Linear detectors (LD), such as the minimum-mean-squared error (MMSE) [13], have been widely adopted, because of their polynomial-time complexity (an $n \times n$ matrix inversion is the leading computational complexity term). MMSE detection performance can be significantly improved in large MIMO systems following a divide-and-conquer approach, namely successive interference cancellation (MMSE-SIC) [14], [15], at a higher computational complexity but still $\mathcal{O}(n^3)$. The performance of LD detectors and LD-SIC detectors can be further improved with lattice reduction techniques (LR) [16].

Random step methods such as Tabu Search (TS) [17] typically require to compute the MMSE solution to then perform an iterative descend method by evaluating $\|\mathbf{H}\mathbf{u}-\mathbf{y}\|^2$ in a certain neighborhood. While TS has been shown to achieve near-ML performance for large n and low order constellations (QPSK) with a complexity $\mathcal{O}(n^3 L)$, where L is the number of iterations, it shows poor performance for high-order constellations even for unbounded L. Layered Tabu Search (LTS) [18] improves the

This work has been partly funded by the Spanish Min-istry of Science and Innovation with the projects GRE3NSYST (TEC2011- 29006-C03-03) and ALCIT (TEC2012-38800-C03-01) and by the program CONSOLIDER-INGENIO 2010 under the project COMONSENS (CSD 2008-00010). The associate editor coordinating the review of this paper and approving it for publication was A. Ghrayeb.

J. Céspedes, P. M. Olmos, and M. Sánchez-Fernández are with the Signal Theory & Communications Department, Universidad Carlos III de Madrid, Madrid, Spain (e-mail: jcespedes@tsc.uc3m.es; olmos@tsc.uc3m.es; mati@tsc.uc3m.es).

F. Perez-Cruz is with the Signal Theory & Communications Department, Universidad Carlos III de Madrid, and also with Bell Labs, Alcatel-Lucent, New Providence, NJ 07974 USA (e-mail: fernando@tsc.uc3m.es; fernando.perez-cruz@alcatel-lucent.com).

TS performance for higher-order constellations by performing detection in a layered manner, where the TS algorithm is applied at each one of the n-th layers, but its complexity scales as $\mathcal{O}(n^4 L)$. Besides, to keep good performance, the number of iterations L has to grow rapidly with the constellation order, e.g. it is set to $L = 20$ for QPSK and $L = 200$ for 64-QAM in [18].

The GTA algorithm in [19] constructs a Gaussian tree approximation of the posterior distribution and relies on Belief Propagation (BP) for approximating the posterior distribution of the transmitted symbol, i.e. $p(\mathbf{u}|\mathbf{y})$. The GTA enhanced by successive interference cancellation (GTA-SIC) [20], at similar complexity to MMSE-SIC, improves GTA, MMSE-SIC and the improved MMSE-SIC using lattice reduction techniques [16] and it can be considered one of the state-of-the-art solution for efficient detection in large MIMO systems.

In this paper, we propose the Expectation Propagation (EP) algorithm [21]–[23] as a low-complexity and high-accuracy solution for symbol detection in high constellation order, high-dimensional MIMO systems. EP generalizes BP in two ways. First, EP can naturally and efficiently work with continuous distributions by moment matching (BP needs to propagate the full distribution) and it powerfully deals with more complex and versatile approximating functions, e.g., tree or forests. For instance, in the context of LDPC channel decoding, we have proposed EP to construct a Markov-tree discrete approximation to the posterior distribution of the coded bits, obtaining accurate estimates to the marginal probability for each coded bit and improving the BP solution for finite-length LDPC codes [24]–[27].

To the authors' best knowledge this is the first time EP is applied to MIMO detection. Using EP, we construct a Gaussian approximation to the posterior distribution of the transmitted symbol vector, i.e. $q_{\mathrm{EP}}(\mathbf{u}) \approx p(\mathbf{u}|\mathbf{y})$. EP follows an iterative procedure to construct $q_{\mathrm{EP}}(\mathbf{u})$ that aims to match the first two moments for each MIMO dimension, whose direct computation from $p(\mathbf{u}|\mathbf{y})$ becomes computationally prohibitive for large n. The EP convergence, as in BP, is not guaranteed for loopy graphs, but it has never been an issue in our numerous simulations. Even for a very large MIMO order, like $n = 250$ transmitting antennas, an excellent approximation is achieved with less than 10 iterations and the complexity per iteration is dominated by a $n \times n$ matrix inversion. Simulation results show that the number of required iterations does not scale as we increase the dimension or the constellation order, and thus the total algorithm complexity remains $\mathcal{O}(n^3)$. Performance evaluation of the proposed EP algorithm shows remarkable improvement compared to GTA-SIC and other approaches in the literature with comparable complexity. In addition, EP is a soft-output algorithm that additionally provides a posterior probability estimate for each received symbol, which can be naturally fed to modern channel decoders [28], while the approaches proposed in the previous paragraph cannot provide such an estimate (or would not be accurate).

In this paper, we focus on high-order constellation, high-dimensional MIMO systems. The proposed detector does not impose any specific relation between the number of antennas or any channel statistics. However, for performance evaluation we focus on a single-user MIMO scenario with $n = m$. If we keep constant one of the channel matrix dimensions and we let the other go to infinity, it is known that the simplest linear processing techniques are optimal [4], [29] under certain channel conditions. For instance, if the channel matrix can be considered orthogonal, in the uplink the receiver can simply apply a matched filter to decouple the information of each of the transmitting antennas, and drastically reduce the complexity of the detector. However, the convergence to this behavior has been shown to be slow [30] requiring that the number of transmitting antennas is significantly smaller than the number of receiving antennas ($n \ll m$). Besides, in many realistic channel environments an infinitely large number of antennas in one of the communication sides does not lead to fully orthogonal channels [30]. For all of the above, complex detectors would be needed and thus EP can be regarded as an alternative to MMSE or any other proposed detector in the literature.

The paper proceeds as follows. In Section II we show the system model and present previous approaches in the literature. In Sections III and IV we detail the EP algorithm, and specifically tailor it to detecting the symbols in a MIMO system, and analyze its computational complexity. In Section V exhaustive experimental results are presented. We conclude the paper in Section VI.

II. Low Complexity MIMO Symbol Detection

Let n be the number of transmitters and assume all of them transmit symbols from the same M-QAM constellation[1], where $\mathcal{A}$ denotes the set of symbols of the constellation and E_s the mean symbol energy. The transmitted symbol vector is a $n \times 1$ i.i.d. vector $\mathbf{u} = [u_1, u_2, \ldots, u_n]^\top = \mathbf{a} + j\mathbf{b}$, where each component $u_i = a_i + jb_i \in \mathcal{A}$.

The symbols are transmitted over a flat-fading complex MIMO channel defined by the $m \times n$ matrix $\mathbf{H}$, 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 $\mathbf{y} = [y_1, y_2, \ldots, y_m] \in \mathbb{C}^m$ is given by

$$\mathbf{y} = \mathbf{H}\mathbf{u} + \mathbf{w} \tag{2}$$

where $\mathbf{w}$ is an additive white circular-symmetric complex Gaussian noise vector $\mathbf{w} = [w_1, w_2, \ldots, w_m]^\top$ with independent zero-mean components and σ_w^2-variance.

Given the model above, the posterior probability of the transmitted symbol vector $\mathbf{u}$ has the following expression:

$$p(\mathbf{u}|\mathbf{y}) = \frac{p\left(\mathbf{y}|\mathbf{u}\right)p(\mathbf{u})}{p(\mathbf{y})} \propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) \prod_{i=1}^{n} \mathbb{I}_{u_i \in \mathcal{A}} \tag{3}$$

where $\mathbb{I}_{u_i \in \mathcal{A}}$ is the indicator function that takes value one if $u_i \in \mathcal{A}$ and zero otherwise. Note that $p(\mathbf{u}) \propto \prod_{i=1}^{n} \mathbb{I}_{u_i \in \mathcal{A}}$ is uniform across all points in $\mathcal{A}^n$, although non-uniform signaling could be handled by any of the proposed or

[1] None of the reviewed or proposed algorithms need the input constellations to coincide across inputs, but we assume so to simplify the notation.

reviewed algorithms in this paper. The signal-to-noise ratio is defined as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{nE_s}{\sigma_w^2}\right). \tag{4}$$

Inference in graphical model is typically presented using real-valued random variables, instead of complex-valued variables used in signal processing for communications, and we believe the EP algorithm is better understood that way. Consequently, we first reformulate the complex-valued MIMO system into a real-valued one, before presenting the EP detector. The system model in (2) can be translated into an equivalent double-sized real-valued representation that is obtained by considering the real $\mathcal{R}(\cdot)$ and imaginary parts $\mathcal{I}(\cdot)$ separately. We define $\tilde{\mathbf{u}} = [\mathbf{a}^\top \quad \mathbf{b}^\top]^\top$, $\tilde{\mathbf{y}} = [\mathcal{R}(\mathbf{y})^\top \quad \mathcal{I}(\mathbf{y})^\top]^\top$, $\tilde{\mathbf{w}} = [\mathcal{R}(\mathbf{w})^\top \quad \mathcal{I}(\mathbf{w})^\top]^\top$ and

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathcal{R}(\mathbf{H}) & -\mathcal{I}(\mathbf{H}) \\ \mathcal{I}(\mathbf{H}) & \mathcal{R}(\mathbf{H}) \end{bmatrix}. \tag{5}$$

The channel model can now be written as follows:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\tilde{\mathbf{u}} + \tilde{\mathbf{w}}, \tag{6}$$

$$p(\tilde{\mathbf{u}}|\tilde{\mathbf{y}}) \propto \mathcal{N}\left(\tilde{\mathbf{y}} : \tilde{\mathbf{H}}\tilde{\mathbf{u}}, \sigma_{\tilde{w}}^2\mathbf{I}\right)\prod_{i=1}^{2n}\mathbb{I}_{\tilde{u}_i\in\tilde{\mathcal{A}}}, \tag{7}$$

where $\sigma_{\tilde{w}}^2 = \sigma_w^2/2$ is the variance of the real and imaginary components of the noise, $\tilde{\mathcal{A}}$ is the alphabet for the real and imaginary components of the symmetric M-QAM signal with energy $\tilde{E}_s = E_s/2$. In the rest of this paper we adopt the real-valued channel model formulation in (6) and (7) and we drop the model indicator $\widetilde{(\cdot)}$ to keep the notation uncluttered.

In the rest of the paper, the operator $\mathrm{diag}(\cdot)$ when applied to a vector, e.g. $\mathrm{diag}(\mathbf{x})$, returns a diagonal matrix with diagonal given by $\mathbf{x}$ and for a given square matrix $\mathbf{X}$, e.g. $\mathrm{diag}(\mathbf{X})$, denotes its diagonal vector (just as Matlab would do it).

A. MMSE Detector and Successive Interference Cancellation

The MMSE detector [13] first proceeds by computing

$$\boldsymbol{\mu}_{\mathrm{MMSE}} = \left(\mathbf{H}^\top\mathbf{H} + \frac{\sigma_w^2}{E_s}\mathbf{I}\right)^{-1}\mathbf{H}^\top\mathbf{y} \tag{8}$$

and it then performs a component-wise hard decision by projecting each component of $\boldsymbol{\mu}_{\mathrm{MMSE}}$ into the corresponding QAM constellation:

$$\hat{u}_{i,\mathrm{MMSE}} = \arg\min_{u_i\in\mathcal{A}} |u_i - \mu_{i,\mathrm{MMSE}}|^2 \tag{9}$$

The complexity is dominated by the matrix inversion in (8), given by $\mathcal{O}(n^3)$ [13]. To intuitively relate the MMSE solution with the GTA and EP detectors, it is interesting to present the MMSE solution $\boldsymbol{\mu}_{\mathrm{MMSE}}$ in (8) as the mode of an approximation to the posterior probability $p(\mathbf{u}|\mathbf{y})$ in (3)[31], that we denote by $q_{\mathrm{MMSE}}(\mathbf{u})$. The posterior approximate is directly obtained by replacing the discrete uniform prior $p(\mathbf{u})$ in (3) by a zero-mean and E_s-variance independent Gaussian distribution:

$$q_{\mathrm{MMSE}}(\mathbf{u}) \propto \mathcal{N}\left(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I}\right)\prod_i \mathcal{N}(u_i : 0, E_s). \tag{10}$$

Since $q_{\mathrm{MMSE}}(\mathbf{u})$ is now Gaussian distributed the mode and the mean coincide and a simple calculation shows that

$$\mathbb{E}_{q_{\mathrm{MMSE}}}[\mathbf{u}] = \boldsymbol{\mu}_{\mathrm{MMSE}}. \tag{11}$$

The MMSE detector provides poor performance, because the multidimensional Gaussian approximation in (10) is not a sensible model for large MIMO systems with high-order constellations. The MMSE performance is significantly improved by successive interference cancellation, yielding the so-called MMSE-SIC [14], [15]. Iteratively, we only decide over the component with the smallest diagonal element in the covariance matrix in (8) and remove its effect in the channel output. After each iteration, we update the received vector

$$\mathbf{y}^{(\ell+1)} = \mathbf{y}^{(\ell)} - \mathbf{h}_i\hat{u}_{i,\mathrm{MMSE}}^{(\ell)} \tag{12}$$

where $\mathbf{h}_i$ denotes the i-th column of $\mathbf{H}$ and its effect is removed from the channel matrix given the current decision, i.e. $\hat{u}_{i,\mathrm{MMSE}}^{(\ell)}$, and we drop $\mathbf{h}_i$ from $\mathbf{H}$. In a nutshell, MMSE-SIC improves the MMSE detector, because we use a one-dimensional Gaussian approximation per iteration and we decide only over the component that we have more certainty. Despite MMSE-SIC requires to perform n times a MMSE matrix inversion similar to that of in (8), the algorithm's complexity can be lower down to $\mathcal{O}(n^3)$[15] by efficiently computing the matrix inversion at each iteration using the matrix inversion lemma (a rank-one update given the inverted matrix from the previous iteration).

B. GTA and GTA-SIC

The Gaussian tree approximation was first proposed in [19] as a feasible method to improve the MMSE-SIC solution for MIMO detection. GTA is based on the following idea: given the posterior (3), we first ignore the discrete nature of the prior $p(\mathbf{u})$ and replace it by a non-informative prior:

$$\begin{aligned} p_{\mathrm{n-i}}(\mathbf{u}|\mathbf{y}) &\propto \mathcal{N}\left(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I}\right) \\ &= \mathcal{N}\left(\mathbf{u} : \boldsymbol{z}, \sigma_w^2(\mathbf{H}^\top\mathbf{H})^{-1}\right), \end{aligned} \tag{13}$$

where $\boldsymbol{z} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{y}$. Consider the family of all possible Gaussian distributions with probability density functions that factorize according to a certain tree graph, i.e. any Gaussian distribution with pdf $g(\mathbf{u})$ such that

$$g(\mathbf{u}) = \prod_i g\left(u_i|u_{p(i)}\right), \tag{14}$$

where $p(i)$ is the set of parents of u_i and the associated factor graph is cycle-free. Now, GTA finds the distribution in such family that minimizes the Kullback-Leibler divergence $D_{\mathrm{KL}}(p_{\mathrm{n-i}}(\mathbf{u}|\mathbf{y})\|g(\mathbf{u}))$. Provided that $p_{\mathrm{n-i}}(\mathbf{u}|\mathbf{y})$ is also Gaussian, the solution for $g_{\mathrm{GTA}}(\mathbf{u})$ is known in closed-form

[19] and it can be computed at cost $\mathcal{O}(n^2)$. Finally, we go back to the original posterior $p(\mathbf{u}|\mathbf{y})$ in (3) and replace the Gaussian term by the Gaussian tree distribution $g_{\mathrm{GTA}}(\mathbf{u})$:

$$q_{\mathrm{GTA}}(\mathbf{u}|\mathbf{y}) \propto \prod_i g_{\mathrm{GTA}}\left(u_i|u_{p(i)}\right) \prod_i \mathbb{I}_{u_i \in \mathcal{A}}. \qquad (15)$$

Since $q_{\mathrm{GTA}}(\mathbf{u})$ is a tree factor graph, we can use Belief Propagation to compute the symbol marginals that are then used for decision. BP over the factor graph $q_{\mathrm{GTA}}(\mathbf{u})$ has a complexity $\mathcal{O}(n^2|\mathcal{A}|^2)$. While the overall complexity is dominated by the matrix inversion $(\mathbf{H}^\top\mathbf{H})^{-1}$, the overhead incurred to compute the tree approximation $g_{\mathrm{GTA}}(\mathbf{u})$ and running BP is not negligible for typical-sized MIMO systems. While the GTA performance is similar to MMSE-SIC for low and medium signal-to-noise ratio (SNR), GTA outperforms MMSE-SIC for high SNR and it has a significant lower computational complexity [19].

Recently, Goldberger has shown in [20] that successive interference cancellation substantially improves the GTA performance, in line with MMSE-SIC improvements. The procedure described before is repeated n times, since per iteration we only decide over the symbol that has the least uncertainty and its effect is canceled from the system as in (12). Evaluating the n matrix inversions during GTA-SIC, using the techniques proposed in [15] to efficiently implement MMSE-SIC, requires $\mathcal{O}(n^3)$ iterations and performing n times the Gaussian tree approximation and running BP have a cost of $\mathcal{O}(\sum_{k=1}^{n} k^2) \approx \mathcal{O}(n^3)$ operations for sufficiently large n. Results reported for GTA-SIC in [20] shows that it is able to outperform the best linear detectors for MIMO detection proposed in the literature in the past years, such as MMSE and MMSE-SIC with lattice reduction using the Lenstra-Lenstra-Lovász (LLL) algorithm [32], [33].

III. Expectation Propagation

Expectation Propagation [21]–[23], [34] is a technique in Bayesian machine learning for approximating posterior beliefs with exponential family distributions[2]. Suppose we are given some statistical model with latent variables $\mathbf{x} \in \Omega^d$ that factors in the following way

$$p(\mathbf{x}) \propto f(\mathbf{x}) \prod_{i=1}^{I} t_i(\mathbf{x}), \qquad (16)$$

where $f(\mathbf{x})$ belongs to an exponential family $\mathcal{F}$ with sufficient statistics $\Phi(\mathbf{x}) = \{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_S(\mathbf{x})\}$ and $t_i(\mathbf{x})$ $i = 1, \ldots, I$ are nonnegative factors. For instance, if $\mathcal{F}$ is the multivariate Gaussian family, $\Phi(\mathbf{x}) = \{x_i, x_i x_j\}_{i,j=1}^d$. Assume now that performing inference over the distribution $p(\mathbf{x})$ in (16) is analytically intractable or prohibitively complex. In this scenario, EP provides a general-purpose framework to construct a tractable approximation to $p(\mathbf{x})$ by a distribution $q(\mathbf{x})$ from $\mathcal{F}$. The resemblance between $q(\mathbf{x})$ and $p(\mathbf{x})$ is achieved by designing $q(\mathbf{x})$ such that

$$\mathbb{E}_{q(\mathbf{x})}\left[\phi_j(\mathbf{x})\right] = \mathbb{E}_{p(\mathbf{x})}\left[\phi_j(\mathbf{x})\right] \quad j = 1, \ldots, S, \qquad (17)$$

where $\mathbb{E}_{q(\mathbf{x})}[\cdot]$ denotes expectation with respect to the distribution $q(\mathbf{x})$. Equation (17) is known as the *moment matching* condition. When both $q(\mathbf{x})$ and $p(\mathbf{x})$ are defined over the same support space and measure, the moment matching condition in (17) is equivalent to finding $q(\mathbf{x})$ in $\mathcal{F}$ that minimizes the Kullback-Leibler divergence with $p(\mathbf{x})$, i.e.

$$q(\mathbf{x}) = \arg\min_{q'(\mathbf{x}) \in \mathcal{F}} D_{\mathrm{KL}}\left(p(\mathbf{x}) \| q'(\mathbf{x})\right). \qquad (18)$$

One naïve approach to find $q(\mathbf{x})$ would be to first compute the moments $\mathbb{E}_{p(\mathbf{x})}[\phi_j(\mathbf{x})]$ for $j = 1, \ldots, S$ and second to construct $q(\mathbf{x})$ according to them. By assumption, this is not a viable option since we cannot do inference over $p(\mathbf{x})$. To overcome this problem, Minka proposed a sequential EP algorithm to iteratively approach the solution in (17) at polynomial time complexity [21], [36]. The main idea behind the sequential EP algorithm is the fact that, while performing inference over $p(\mathbf{x})$ in (16) is intractable, we typically are able to perform inference over a distribution of the form

$$\hat{p}_i(\mathbf{x}) \propto f(\mathbf{x}) t_i(\mathbf{x}), \qquad (19)$$

in which there is only present one of the factors $t_i(\mathbf{x})$ $i = 1, \ldots, I$ in (16) that do not belong to the exponential family $\mathcal{F}$. The sequential EP algorithm is as follows. First, assume the following factorization for $q(\mathbf{x}) \in \mathcal{F}$

$$q(\mathbf{x}) = f(\mathbf{x}) \prod_{i=1}^{I} \tilde{t}_i(\mathbf{x}), \qquad (20)$$

where $\tilde{t}_i(\mathbf{x}) \in \mathcal{F}$ for $i = 1, \ldots, I$. Note that we have simply replaced each one of the $t_i(\mathbf{x})$ factors in (16) by a member $\tilde{t}_i(\mathbf{x})$ of $\mathcal{F}$. Given an initial proposal $q^{(0)}(\mathbf{x})$ and being $q^{(\ell)}(\mathbf{x})$ the approximation to $q(\mathbf{x})$ in (18) at iteration ℓ, $q^{(\ell+1)}(\mathbf{x})$ is obtained by updating each one of the $\tilde{t}_i(\mathbf{x})$ factors independently. For $i = 1, \ldots, I$,

1) Compute the *cavity* distribution

$$q^{(\ell)\setminus i}(\mathbf{x}) \doteq \frac{q^{(\ell)}(\mathbf{x})}{\tilde{t}_i(\mathbf{x})} \in \mathcal{F}. \qquad (21)$$

2) Compute the distribution $\hat{p}_i(\mathbf{x}) \propto t_i(\mathbf{x}) q^{(\ell)\setminus i}(\mathbf{x})$, and find

$$\mathbb{E}_{\hat{p}_i(\mathbf{x})}\left[\phi_j(\mathbf{x})\right] \qquad (22)$$

for $j = 1, \ldots, S$.

3) The refined factor $\tilde{t}_i^{\mathrm{new}}(\mathbf{x})$ is obtained so that

$$\mathbb{E}_{\tilde{t}_i^{\mathrm{new}}(\mathbf{x}) q^{(\ell)\setminus i}(\mathbf{x})}\left[\phi_j(\mathbf{x})\right] \qquad (23)$$

coincides with (22) for $j = 1, \ldots, S$.

The sequential EP algorithm is run until a convergence criterion is met or a maximum number of iterations is reached. As shown in [34], this algorithm can be interpreted as a coordinate

[2] A comprehensive introduction to exponential families and their properties can be found in [35].

gradient descent over the parameter space of the $q(\mathbf{x})$ distribution to find a saddle point of a certain energy function. As such, the convergence to a saddle point is not guaranteed [37]. Nonetheless, sequential EP has been shown to achieve accurate results, typically close to the moment matching solution, in a wide range of applications [21], [23].

As shown in [22] and [23], if a factor $t_i(\cdot)$ in (16) only depends on a subset $\mathbf{x}_i \in \Omega^{d_i}$ of the $\mathbf{x}$ components, $d_i < d$, then the approximate factor $\tilde{t}_i(\mathbf{x}_i)$ in (21) is defined over the same domain and its update at each iteration can be alternatively performed over the marginal distribution $q(\mathbf{x}_i)$. An example of this alternative procedure is the EP approximation to the MIMO symbol posterior distribution $p(\mathbf{u}|\mathbf{y})$ in (3) that we present in detail in the next section.

IV. The Expectation Propagation MIMO Detector

The MMSE approximation to the true posterior distribution in (10) replaces the prior over the transmitted symbols by a zero-mean independent component-wise Gaussian whose variance equals the QAM symbol mean energy. Intuitively it might make sense to chose the parameters of the Gaussian prior this way, because it matches the first two moments of the input distribution. However it is certainly not the best choice, as we are interested in matching the posterior distribution to optimally detect the transmitted symbols. In this paper we propose to approximate the symbol posterior distribution $p(\mathbf{u}|\mathbf{y})$ by a Gaussian approximation $q_{\text{EP}}(\mathbf{u}) = \mathcal{N}(\mathbf{u} : \boldsymbol{\mu}_{\text{EP}}, \boldsymbol{\Sigma}_{\text{EP}})$ that is optimized using the EP framework. Thus, the optimal EP solution will be

$$\boldsymbol{\mu}_{\text{EP}} = \mathbb{E}_{p(\mathbf{u}|\mathbf{y})}[\mathbf{u}], \tag{24}$$

$$\Sigma_{\text{EP}} = \text{CoVar}_{p(\mathbf{u}|\mathbf{y})}[\mathbf{u}]. \tag{25}$$

While the direct computation of the $p(\mathbf{u}|\mathbf{y})$ moments requires $|\mathcal{A}|^n$ operations, the sequential EP update rules [22], [23] allows to iteratively approximate (24) and (25) at polynomial complexity with n. Once the iterative method has stopped, the EP detector (EPD) computes the hard output $\hat{\mathbf{u}}_{\text{EP}}$ by independently deciding on each component:

$$\hat{u}_{i,\text{EP}} = \arg\min_{u_i \in \mathcal{A}} \left| u_i - \mu_{i,\text{EP}} \right|^2 \tag{26}$$

for $i = 1, \dots, 2n$.

A. Parallel EP Iterative Method

In the following we present the formulation of the EP update rules according to [22], [23]. Given the factorization of the posterior in (3), we replace each one of the non-Gaussian factors by an unnormalized Gaussian:

$$q(\mathbf{u}) \propto \mathcal{N}\left(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}\right) \prod_{i=1}^{2n} \mathrm{e}^{\gamma_i u_i - \frac{1}{2}\Lambda_i u_i^2}, \tag{27}$$

where γ_i and $\Lambda_i > 0$ are real constants. For any vector $\boldsymbol{\gamma} \in \mathbb{R}^{2n}$ and $\boldsymbol{\Lambda} \in \mathbb{R}_+^{2n}$, $q(\mathbf{u})$ is a Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{H} + \text{diag}(\boldsymbol{\Lambda})\right)^{-1}, \tag{28}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{y} + \boldsymbol{\gamma}\right), \tag{29}$$

The EP iterative method approximates the solution in (24) and (25) at polynomial complexity by recursively updating the pairs (γ_i, Λ_i), $i = 1, \dots, 2n$. For each input dimension, we use a single non-Gaussian factor from the posterior (3) at each iteration. We initialize $\gamma_i = 0$ and $\Lambda_i = E_s^{-1}$ for all i (this would give the MMSE solution). At each EP iteration all pairs $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ for $i = 1, \dots, 2n$ are updated in parallel, where ℓ denotes the EP iteration. Given the i-th marginal of the distribution $q^{(\ell)}(\mathbf{u})$, namely $q_i^{(\ell)}(u_i) = \mathcal{N}(u_i : \mu_i^{(\ell)}, \sigma_i^{2(\ell)})$, the pair $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ is computed as follows:

1) Compute the cavity marginal

$$q^{(\ell)\backslash i}(u_i) = \frac{q^{(\ell)}(u_i)}{\exp\left(\gamma_i^{(\ell)} u_i - \frac{1}{2}\Lambda_i^{(\ell)} u_i^2\right)} \sim \mathcal{N}\left(u_i : t_i^{(\ell)}, h_i^{2(\ell)}\right), \tag{30}$$

where

$$h_i^{2(\ell)} = \frac{\sigma_i^{2(\ell)}}{\left(1 - \sigma_i^{2(\ell)}\Lambda_i^{(\ell)}\right)}, \tag{31}$$

$$t_i^{(\ell)} = h_i^{2(\ell)}\left(\frac{\mu_i^{(\ell)}}{\sigma_i^{2(\ell)}} - \gamma_i^{(\ell)}\right). \tag{32}$$

2) Compute the mean $\mu_{p_i}^{(\ell)}$ and variance $\sigma_{p_i}^{2(\ell)}$ of the distribution

$$\hat{p}^{(\ell)}(u_i) \propto q^{(\ell)\backslash i}(u_i)\mathbb{I}_{u_i \in \mathcal{A}_i}. \tag{33}$$

3) Finally, the pair $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ is updated so that the following unnormalized Gaussian distribution

$$q^{(\ell)\backslash i}(u_i)\exp\left(\gamma_i^{(\ell+1)} u_i - \frac{1}{2}\Lambda_i^{(\ell+1)} u_i^2\right), \tag{34}$$

has mean and variance equal to $\mu_{p_i}^{(\ell)}$ and $\sigma_{p_i}^{2(\ell)}$. A simple calculation shows that the solution is given by

$$\Lambda_i^{(\ell+1)} = \frac{1}{\sigma_{p_i}^{2(\ell)}} - \frac{1}{h_i^{2(\ell)}}, \tag{35}$$

$$\gamma_i^{(\ell+1)} = \frac{\mu_{p_i}^{(\ell)}}{\sigma_{p_i}^{2(\ell)}} - \frac{t_i^{(\ell)}}{h_i^{2(\ell)}}. \tag{36}$$

The parameter update in (35) may return a negative $\Lambda_i^{(\ell+1)}$, which should be positive, because it is a precision (inverse variance) term. This result just means that there is no pair $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ that places the variance of the Gaussian in (34) at $\sigma_{p_i}^{2(\ell)}$. In this case, we simply keep the previous values for these parameters, i.e. $\gamma_i^{(\ell+1)} = \gamma_i^\ell$ and $\Lambda_i^{(\ell+1)} = \Lambda_i^\ell$, and update all the other pairs, $(\gamma_j^{(\ell+1)}, \Lambda_j^{(\ell+1)})$ for $j \neq i$.

Note that the update rules described above only need the marginal for each component. Given $\boldsymbol{\gamma}^{(\ell)}$ and $\boldsymbol{\Lambda}^{(\ell)}$ and once

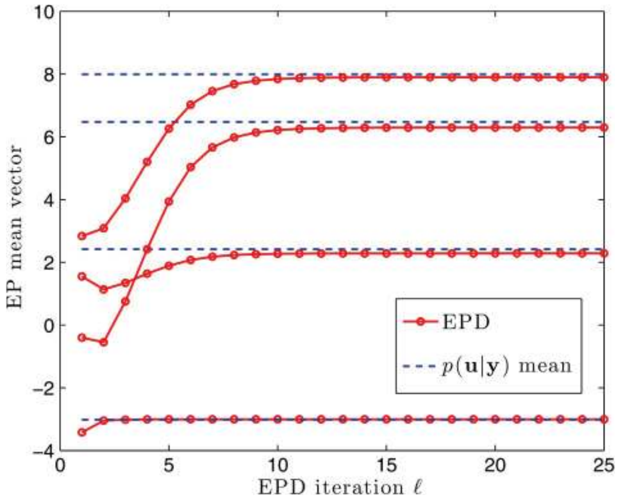


Fig. 1. Evolution of each component of the EP mean $\boldsymbol{\mu}^{(\ell)}$ in (29) as EP iterates for a $n = m = 2$ scenario with a 256-QAM constellation and SNR = 15 dB. In blue dashed lines, we indicate the mean of the true posterior $p(\mathbf{u}|\mathbf{y})$.

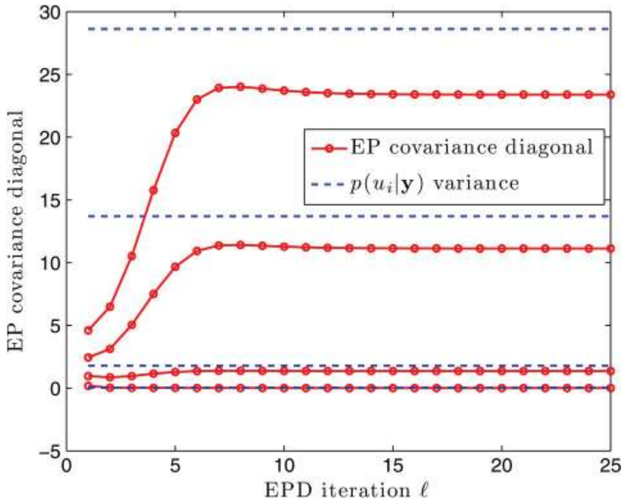


Fig. 2. Evolution of each component of the EP covariance $\boldsymbol{\Sigma}^{(\ell)}$ in (28) as EP iterates for a $n = m = 2$ scenario with a 256-QAM constellation and SNR = 15 dB. In bue dashed lines, we indicate the variance of the marginal symbol posterior $p(u_i|\mathbf{y})$.

we have computed $\boldsymbol{\Sigma}^{(\ell)}$ and $\boldsymbol{\mu}^{(\ell)}$ using (28) and (29), then all $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ pairs for $i = 1, \ldots, 2n$ can be updated in parallel. Finally, to improve the robustness of the algorithm, in [22] and [38] it is suggested to smooth the parameter update (i.e., a low-pass filter) in (35) and (36) by a convex combination with the former value, namely

$$\gamma_i^{(\ell+1)} = \beta \left(\frac{\mu_{p_i}^{(\ell)}}{\sigma_{p_i}^{2(\ell)}} - \frac{t_i^{(\ell)}}{h_i^{2(\ell)}} \right) + (1-\beta)\gamma_i^{(\ell)}, \qquad (37)$$

$$\Lambda_i^{(\ell+1)} = \beta \left(\frac{1}{\sigma_{p_i}^{2(\ell)}} - \frac{1}{h_i^{2(\ell)}} \right) + (1-\beta)\Lambda_i^{(\ell)}, \qquad (38)$$

for some $\beta \in [0,1]$ and we have set in our experiments $\beta = 0.2$. We halt the algorithm when the mean and covariance component-wise variation is less than 10^{-4} between two consecutive iterations or a maximum number of iterations has been reached. Also, to avoid numerical instabilities, we have set $\sigma_{p_i}^{2(\ell)} = \max(\epsilon, \mathrm{Var}_{\hat{p}_i}[u_i])$, where $\epsilon = 5 \times 10^{-7}$ in our experiments is a small constant that sets the minimum variance allowed per component.

B. Matching the Posterior Moments

To illustrate the EP ability to match the moments of the true posterior $p(\mathbf{u}|\mathbf{y})$, we consider a low dimensional scenario where we are able to compute mean and covariance matrix of $p(\mathbf{u}|\mathbf{y})$. In Fig. 1, in solid red lines we show an example of the evolution of the components of the EP mean vector $\boldsymbol{\mu}^{(\ell)}$ in (29) as EP iterates for a given channel observation $\mathbf{y}$ in a $n = m = 2$ scenario with a 256-QAM constellation and SNR = 15 dB. Note that the MMSE estimate would be the EP solution at iteration 1. In blue dashed lines, we indicate the mean of the posterior $p(\mathbf{u}|\mathbf{y})$ (real and imaginary parts). As shown, in 10 iterations, the EPD already provides an accurate estimate of the mean of the posterior distribution $p(\mathbf{u}|\mathbf{y})$ in (3). For the same scenario, Fig. 2 shows an example of the evolution of the diagonal components of the EP covariance matrix $\boldsymbol{\Sigma}^{(\ell)}$ in (28) as EP iterates. In blue dashed lines, we indicate the real and imaginary values of the variance of the marginal symbol posterior $p(u_i|\mathbf{y})$.

Despite the fact that the EP iterative method does not guarantee convergence to the exact moment matching solution, the distribution $q(\mathbf{u})$ constructed is able to present moments very close to the posterior true mean and variance. As shown in Fig. 1 and Fig. 2, the EP algorithm, besides accurately matching the posterior mean, provides a reliable measure of the uncertainty per symbol, identifying which symbols can be decided with high grade of confidence and for which ones the risk of error in hard decision is large. Neither GTA-SIC nor MMSE-SIC provide such kind of soft-output information.

C. EPD Complexity

The complexity of EP per iteration is dominated by the computation of the covariance matrix in (28) and the mean vector in (29). The complexity of this step is identical to the MMSE and GTA posterior covariance matrix computation and mean vector in, respectively, (11) and (13). Once the EP marginals $q^{(\ell)}(u_i)$ for $i = 1, \ldots, 2n$ have been computed, the parallel update of all pairs $(\gamma_i^{\ell}, \Lambda_i^{\ell}) \leftarrow (\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ for $i = 1, \ldots, 2n$ has a small computational complexity, linear in $n|\mathcal{A}|$. Thus, if EP is run L iterations, the final complexity is $\mathcal{O}(n^3 L + n|\mathcal{A}|L)$. The comparison of this complexity with the complexity of GTA-SIC and MMSE-SIC depends on the channel time varying characteristics:

- In a static block fading channel where the channel matrix $\mathbf{H}$ is constant during T consecutive symbol times, the MMSE-SIC matrix inversion only has to be computed once and thus the complexity of detecting the T blocks of n symbols is given by $\mathcal{O}(n^3 + Tn^2)$[4]. The computation of the tree approximations in (15) for GTA-SIC has to be done for each channel observation $\mathbf{y}$ and thus, the complexity to detect the T blocks of n symbols is $\mathcal{O}(Tn^3)$. Similarly, all the EPD processing depends on the channel observation vector and, hence, its complexity is $\mathcal{O}(n^3LT + n|\mathcal{A}|LT)$. Then, the EP detector is approximately L times more complex than GTA-SIC and LT times more complex than MMSE-SIC and MMSE.

- In a quasi-static block fading channel model where the fading coefficients do not change within one time symbol, but vary every symbol time [39], namely $T = 1$ in the former scenario, the EP complexity to detect each block of n symbols is approximately L times more complex than GTA-SIC, MMSE-SIC, and MMSE.

As we show in the next Section, regardless the dimension of the MIMO scenario, the EP detector is able to provide a remarkable performance improvement with respect to GTA-SIC. Regarding the actual number of EP iterations required, we also show that EPD only needs a few iterations and no improvement can be observed between $L = 10$ and $L = 100$. Furthermore, in high-dimension scenarios, a noticeable performance gain with respect to GTA-SIC is already reported with only $L = 2$ iterations. Namely, twice the complexity of GTA-SIC.

V. Experimental Results

In this section, we illustrate the performance of the EPD algorithm for MIMO detection in high-order high-dimensional scenarios. We have averaged our results for 5000 realizations of the channel matrix. We consider five scenarios of increasing dimension: $n = m = 12$, $n = m = 32$, $n = m = 64$, $n = m = 100$ and $n = m = 250$. The detector performance is shown in terms of the symbol error rate (SER) as a function of the SNR defined in (4).

In all scenarios we compare EPD with MMSE, GTA and GTA-SIC[3]. Scenario $n = m = 12$ also allows an implementation of ML based on the Schnorr-Euchner variant [40] of sphere decoding. In our analysis, we do not include comparison with respect MMSE-SIC since its tends to overlap with GTA in most of the SNR range [19]. Besides, we do not include performance results for MMSE and MMSE-SIC detectors with LR techniques. In [20], it is shown that GTA-SIC is able to improve MMSE-SIC with lattice reduction using the LLL algorithm, which achieves the best results to date among LR based MIMO detection methods [32], [33], [41].

We first consider a scenario with $n = m = 12$ antennas and 16-QAM modulation. In Fig. 3, we compare the performance of EPD with $L = 100$ iterations (EPD 100), $L = 10$ iterations (EPD 10) and with only $L = 2$ iterations (EPD 2) with GTA-SIC, GTA and MMSE. Also, ML performance is provided to show how far we are from the optimal solution. First, note that running EPD for 100 iterations does not result in an appreciable gain in performance with respect to the case $L = 10$. We can observe that EPD 100 only outperforms EPD in 0.1 dB for SER $= 10^{-3}$ and its complexity is ten times higher. Compared to the ML solution we are far about 3 dB for SER $= 10^{-3}$. Compared to other sub-optimal methods, EPD with 10 iterations is able to improve the GTA-SIC performance in 1 dB for SER $= 10^{-3}$.

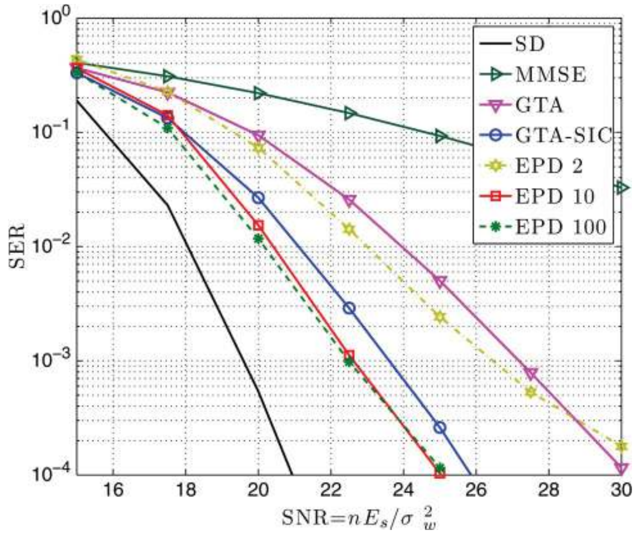


Fig. 3. SER performance of EPD with $L = 100$, $L = 10$ and $L = 2$ iterations, GTA-SIC, GTA MMSE and ML for the case $n = m = 12$ and a 16-QAM constellation.

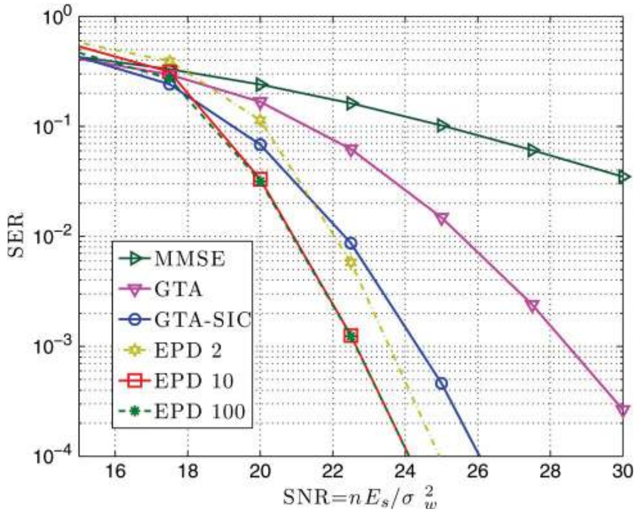
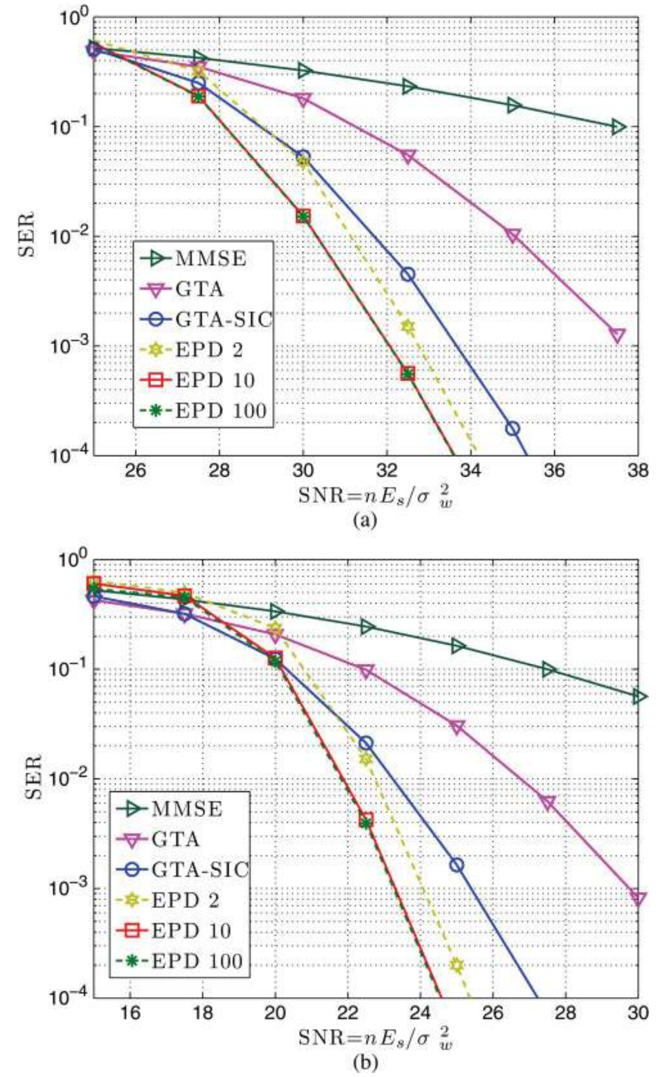


Fig. 4. SER performance of EPD with $L = 100$, $L = 10$ and $L = 2$ iterations, GTA-SIC, GTA and MMSE for the case $n = m = 32$ and a 16-QAM constellation.

A similar study is done in Fig. 4 for a scenario with $n = m = 32$ antennas and 16-QAM modulation, excluding the ML solution, which we are now unable to compute. First, note that again running EPD for 100 iterations does not result in an appreciable gain in performance with respect to the case $L = 10$. Besides, with 10 iterations, EPD is able to improve the GTA-SIC performance in 1.8 dB for SER $= 10^{-3}$. Compared with the $n = m = 12$ scenario, the gap between EPD 10 and GTA-SIC is significantly augmented. Indeed, for SNRs above 20.5 dB and with only two iterations, the EP detector outperforms GTA-SIC. For SER $= 10^{-3}$ and $L = 2$, it exhibits a gain of 0.8 dB. Therefore, for SNR above 20.5 dB, we can modulate the number of iterations in EP between $L = 2$ and $L = 10$ according to our complexity constraints without degrading the performance above the GTA-SIC curve.

Now we examine the case $n = m = 64$. In Fig. 5(a), we include performance results for the case of 16-QAM constellation. As in the 32 × 32 scenario, EPD 10 achieves the same

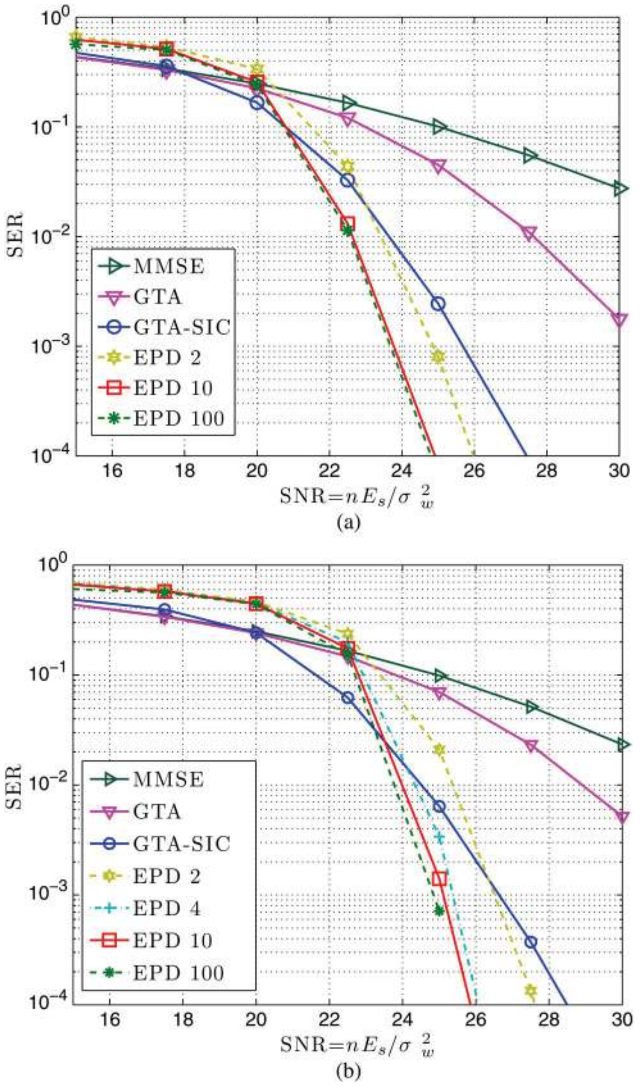
[3] The C code for GTA and GTA-SIC can be accessed in the author's web site, see [20] for details.



(a)

(b)

Fig. 5. SER performance of EPD with $L = 100$, $L = 10$ and $L = 2$ iterations, GTA-SIC, GTA and MMSE for the case $n = m = 64$ and (a) 16-QAM constellation and (b) 64-QAM constellation.



(a)

(b)

Fig. 6. SER performance for the case (a) $n = m = 100$ and (b) $n = m = 250$ and 16-QAM constellation.

performance that EPD 100 while clearly improving GTA-SIC. For SER $= 10^{-3}$, gains in SNR of 2.1 dB for $L = 10$ and of 1.3 dB for $L = 2$ are reported. In Fig. 5(b), we increase the constellation to a 64-QAM with similar performance results. In this case, the measured gain with respect to GTA-SIC is of 1.6 dB when $L = 10$ and 0.9 dB when $L = 2$ at a SER $= 10^{-3}$.

Similar conclusions can be drawn from Fig. 6, where performance results are shown for the case $n = m = 100$ (a) and $n = m = 250$ (b) with 16-QAM constellation. The gain between EPD with 10 iterations and GTA-SIC is of 2 dB. For the $n = m = 250$ case, despite GTA-SIC outperforms EPD 2 in most of the SNR range, with 4 iterations the EP detector already exhibits a gain of 1.7 dB with respect to GTA-SIC. Besides, for this scenario, we can appreciate a small gap, close to 0.1 dB, between EPD with $L = 100$ and $L = 10$ iterations. We have observed this gap vanishes by running EP up to 15 iterations.

The EP detector is not only to be the best detector in all scenarios, but, for fixed constellation, it is able to increase the gain with respect to GTA-SIC as the number of antennas grows, achieving gains up to 2 dB for hundreds of antennas. The best results are achieved within 10 iterations, namely ten times the GTA-SIC complexity, but performance gains at moderate SNRs can be achieved with only 2–4 iterations. Note also that, compared to other alternatives in the literature such as random step methods [17], [18], the EP detector exhibits an excellent performance even though the the number of iterations does not scale with the number of antennas or the constellation size. Consequently, EPD emerges as a powerful and efficient method to implement the receiver detector in high-order high-dimensional MIMO scenarios.

VI. Conclusion

The design of efficient large MIMO digital receivers is a challenging open problem. In this paper, we focus on symbol estimation and detection in MIMO systems when the number of antennas is very large and we work with high-order constellations. Classical methods such as Zero Forcing and MMSE present poor performance. Modern detection methods based on Successive Interference Cancellation, such as MMSE-SIC, or lattice reduction techniques using the LLL algorithm achieve a significant improvement, yet they are still far from the optimal maximum likelihood performance. Recently, the GTA-SIC algorithm has been proposed to outperform MMSE-SIC-LLL and it provides remarkable results in large MIMO low-cost detection. In this paper, we put forward a symbol detector based on the Expectation Propagation algorithm to solve this problem. EP is a powerful approximate inference technique to construct tractable approximations to a given probability distribution. We have shown that in a few iterations EP converges to a Gaussian distribution whose mean and covariance matrix are close to the corresponding moments in the true posterior distribution of the transmitted symbol vector, thus constituting an excellent tool to perform symbol decision. The EP method is robust and fast: regardless the number of antennas or the constellation order, EP with only ten iterations is able to outperform GTA-SIC, achieving SNR gains that grow with n for fixed constellations. Even with only two iterations, EP outperforms GTA-SIC at moderate SNR.

References

[1] Q. Li *et al.*, "MIMO techniques in WiMAX and LTE: A feature overview," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 86–92, May 2010.

[2] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO Channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun. 2003.

[3] L. Zheng, P. Viswanath, and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1095, May 2003.

[4] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[5] H. Huh, G. Caire, H. Papadopoulos, and S. Ramprashad, "Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, Sep. 2012.

[6] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, no. 170, pp. 463–471, Apr. 1985.

[7] A. Burg *et al.*, "VLSI implementation of MIMO detection using the sphere decoding algorithm," *IEEE J. Solid-State Circuits*, vol. 40, no. 7, pp. 1566–1577, Jul. 2005.

[8] J. Boutros, N. Gresset, L. Brunel, and M. Fossorier, "Soft-input soft-output lattice sphere decoder for linear channels," in *Proc. IEEE GLOBECOM*, 2003, vol. 3, pp. 1583–1587.

[9] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-best sphere decoding for MIMO detection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 491–503, Mar. 2006.

[10] C. Studer, A. Burg, and H. Bolcskei, "Soft-output sphere decoding: Algorithms and VLSI implementation," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 290–300, Feb. 2008.

[11] P. Svac, F. Meyer, E. Riegler, and F. Hlawatsch, "Soft-heuristic detectors for large MIMO systems," *IEEE Trans. Signal Process.*, vol. 61, no. 18, pp. 4573–4586, Sep. 2013.

[12] M. Cirkic and E. G. Larsson, "SUMIS: A near-optimal soft-ouput MIMO detector at low and fixed complexity," [Online]. Available: http://arxiv.org/abs/1207.3316

[13] G. Caire, 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] G. Golden, C. J. Foschini, R. Valenzuela, and P. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electron. Lett.*, vol. 35, no. 1, pp. 14–16, Jan. 1999.

[15] T. Liu and Y.-L. Liu, "Modified fast recursive algorithm for efficient MMSE-SIC detection of the V-BLAST system," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3713–3717, Oct. 2008.

[16] Q. Zhou and X. Ma, "Element-based lattice reduction algorithms for large MIMO detection," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 274–286, Feb. 2013.

[17] H. Zhao, H. Long, and W. Wang, "Tabu search detection for MIMO systems," in *Proc. IEEE 18th Int. Symp. PIMRC*, 2007, pp. 1–5.

[18] N. Srinidhi, T. Datta, A. Chockalingam, and B. Rajan, "Layered tabu search algorithm for large-MIMO detection and a lower bound on ML performance," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 2955–2963, Nov. 2011.

[19] 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.

[20] J. Goldberger, "Improved MIMO detection based on successive tree approximations," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, 2013, pp. 2004–2008.

[21] T. Minka, "Expectation propagation for approximate bayesian inference," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 362–369.

[22] M. W. Seeger, "Expectation Propagation For Exponential Families," Univ. Calif., Berkeley, CA, USA, Tech. Rep., 2005.

[23] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Mach. Learn. Res.*, vol. 9, pp. 759–813, Apr. 2008.

[24] P. Olmos, J. Murillo-Fuentes, and F. Perez-Cruz, "Tree-structure expectation propagation for LDPC decoding over the BEC," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3354–3377, Jun. 2013.

[25] L. Salamanca, P. Olmos, J. Murillo-Fuentes, and F. Perez-Cruz, "Tree-structured expectation propagation for LDPC decoding over BMS channels," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4086–4095, Oct. 2013.

[26] P. Olmos, J. J. Murillo-Fuentes, and F. Pérez-Cruz, "Tree-structured expectation propagation for decoding finite-length LDPC codes," *IEEE Commun. Lett.*, vol. 15, no. 2, pp. 235–237, Feb. 2011.

[27] P. M. Olmos, L. Salamanca, J. J. Murillo-Fuentes, and F. Pérez-Cruz, "On the design of LDPC-convolutional ensembles using the TEP decoder," *IEEE Commun. Lett.*, vol. 16, no. 5, pp. 726–729, May 2012.

[28] T. J. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, Mar. 2008.

[29] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[30] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[31] X. Wang and V. Poor, *Wireless Communications: Advanced Techniques for Signal Reception*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2003.

[32] X. Ma and W. Zhang, "Performance analysis for MIMO systems with lattice-reduction aided linear equalization," *IEEE Trans. Commun.*, vol. 56, no. 2, pp. 309–318, Feb. 2008.

[33] Y. H. Gan and W.-H. Mow, "Complex lattice reduction algorithms for low-complexity MIMO detection," in *Proc. IEEE GLOBECOM*, 2005, vol. 5, pp. 2953–2957.

[34] M. Opper and O. Winther, "Expectation consistent approximate inference," *J. Mach. Learn. Res.*, vol. 6, pp. 2177–2204, Dec. 2005.

[35] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1/2, Nov. 2008.

[36] T. P. Minka, "A family of algorithms for approximate Bayesian Inference," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, USA, 2001.

[37] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[38] T. Minka, "The EP Energy Function And Minimization Schemes," MIT MediaLab., Cambridge, MA, USA, Tech. Rep., 2001.

[39] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *J. Commun. Netw.*, vol. 15, no. 4, pp. 338–351, 2013.

[40] C. P. Schnorr and M. Euchner, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," *Math. Programm.*, vol. 66, no. 1–3, pp. 181–199, Aug. 1994.

[41] D. Wübben, D. Seethaler, J. Jaldéen, and G. Matz, "Lattice reduction: A survey with applications in wireless communications," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 70–91, May 2011.

Javier Céspedes was born in Ciudad Real, Spain, in 1988. He received the B.Sc. and M.Sc. degrees in 2011 and 2013, respectively, from University Carlos III de Madrid where he is currently a Ph.D. student. His current research interests are focused in the application of machine learning methods to wireless MIMO systems.

Pablo M. Olmos (M'11) was born in Granada, Spain, in 1984. He received the M.Sc./B.Sc.degrees in telecommunication engineering in 2008 and the Ph.D. degree in telecommunication engineering in 2011 from the University of Sevilla. He is currently an Assistant Professor at University Carlos III de Madrid. He has visited Princeton University, École Polytechnique Fédérale de Lausanne (EPFL), and Notre Dame University as an Invited Researcher. His research interests lie in approximate inference methods for Bayesian learning and its applications to coding and information theory and digital communications. A detailed CV and list of publications can be accessed at http://www.tsc.uc3m.es/olmos.

Matilde Sánchez-Fernández (SM'14) received the M.Sc. degree in telecommunications engineering and the Ph.D. degree from Universidad Politécnica de Madrid, Madrid, Spain, in 1996 and 2001, respectively. In 2000, she joined the Universidad Carlos III de Madrid, Madrid, where she has been an Associate Professor since 2009 teaching several undergraduate and graduate courses (M.Sc. and Ph.D.) related to communication theory and digital communications. Previously, she was a Telecommunication Engineer with Telefónica. She performed several research stays at the Information and Telecommunication Technology Center, The University of Kansas, Lawrence, KS, USA (1998), Bell Laboratories, Crawford Hill, NJ, USA (2003–2006), Centre Tecnològic de Telecomunicacions de Catalunya, Barcelona, Spain (2007), and Princeton University, Princeton, NJ, USA (2011). Her current research interests are multiple-input-multiple-output techniques, wireless communications, and simulation and modeling of communication systems, and in these fields. She has (co)authored more than 50 contributions to international journals and conferences.

Fernando Perez-Cruz (SM'06) was born in Sevilla, Spain, in 1973. He received the Ph.D. degree in electrical engineering in 2000 from the Technical University of Madrid and the M.Sc./B.Sc. degrees in electrical engineering from the University of Sevilla in 1996. He is an Associate Professor with the Department of Signal Theory and Communication, University Carlos III in Madrid. He has been a Visiting Professor at Princeton University under the sponsorship of a Marie Curie Fellowship. He has held positions at the Gatsby Unit (London), Max Planck Institute for Biological Cybernetics (Tuebingen), BioWulf Technologies (New York), and the Technical University of Madrid and Alcala University (Madrid). His current research interest lies in machine learning and information theory and its application to signal processing and communications. He has authored over 90 contributions to international journals and conferences. He has also co-authored a book on digital communications. A detailed CV and list of publications can be accessed at http://www.tsc.uc3m.es/&sim;fernando.