Vithanage, C. M., Soler Garrido, J., Andrieu, C., & Piechocki, R. J. (2008). Tree-based reparameterization with distributional approximations for reduced-complexity MIMO symbol detection. IEEE Transactions on Wireless Communications, 7(11, part 2), 4617 - 4626. 10.1109/T-WC.2008.070640

Tree-Based Reparameterization with Distributional Approximations for Reduced-Complexity MIMO Symbol Detection

C. M. Vithanage, J. Soler-Garrido, C. Andrieu, and R. J. Piechocki

***Abstract*—Detection of spatially multiplexed data transmissions subject to frequency flat fading is considered. Optimal decoders require knowledge of the marginal posterior distributions of the transmitted symbols, but their exact computation is not feasible for practical systems. Hence sub-optimal approaches are generally sought. By recasting this problem into the graphical model framework, we investigate here a recently proposed sub-optimal approach which relies on a tree-based reparameterization principle. For quasi-static fading channels, the resulting decoder complexity has an order which is at most quadratic in the number of transmit antennas. However, in its standard form, the algorithm often fails to converge, severely restricting its practical usability. We here develop a novel methodology to ensure systematic convergence of the algorithm in this communication scenario at the expense of the introduction of a minimal bias on the computation of the symbol marginal posterior probabilities. This bias is quantified theoretically and its innocuity for the problem at hand is ultimately demonstrated through numerical simulations. For a system using 16-QAM modulation with four transmit and receive antennas, the proposed detector achieves a bit-error rate of 10^{-4} requiring only 3dB greater SNR than the optimal method.**

***Index Terms*—Digital communication, decoding, fading channels, MIMO systems, signal detection.**

I. Introduction

IT IS ANTICIPATED that multiple-input multiple-output (MIMO) wireless channels are to be widely adopted in future point-to-point wireless communications systems. Transmission of spatially multiplexed independent data streams from a transmitter with multiple antennas can lead to the extraction of their capacity [1], in the presence of optimal detection at the receiver. Unfortunately, optimal detection has a complexity which is exponential in the number of transmit antennas and hence is practically infeasible for large transmit antenna numbers. Thus sub-optimal detection algorithms such as the V-BLAST (vertical Bell laboratories layered space-time) algorithm of [2] are needed to take advantage of the capacity of MIMO systems. Presently, state of the art in sub-optimal MIMO detectors are the sphere decoders [3], [4], [5], [6]. Although the decoding performance is near optimal, the expected complexity of sphere decoders has been shown to be exponential in the number of transmit antennas for a fixed signal-to-noise ratio (SNR) [7]. Other successful approaches include the successive moment matching to Gaussian distributions approach of [8], the iterative tree based search approach of [9] and the Gibbs sampling based approach of [10].

We define optimal soft detection, or the optimal detection considering a subsequent channel decoder, in a spatial multiplexing system as the computation of the *a posteriori* marginal probability distributions of the symbols transmitted by each antenna. These posterior marginals will be frequently referred to as the APPs in this paper. Given perfect channel state information at the receiver, the posterior joint distribution is readily available up to a proportionality constant, as will be illustrated later on. Thus optimal soft detection reduces to a task of marginalization. Exact marginalization to compute the APPs necessarily involves an enumeration over the domain of the set of symbols transmitted by the antennas. As a consequence, optimal soft detection has a complexity which is exponential in the number of transmit antennas, preventing its use in practical implementations.

In this work, since the problem of calculating the APPs is a one of calculating the node beliefs on a graphical model [11], we exploit recently developed sub-optimal algorithms which avoid the combinatorial complexity described above. In this paper, the probability distributions of concern are represented by undirected graphical models (UGM) [12]. As will be seen later, the joint distribution on which marginalization needs to be performed in this case corresponds to an UGM with cycles. Had the graphical model been cycle free, the required marginal distributions could have been computed with a complexity which is linear in the number of variables using a message passing algorithm such as the sum-product algorithm [13], [14] or the belief propagation algorithm [15]. Even in the presence of cycles, repeated message passing as if there were no cycles, termed loopy belief propagation (LBP), has been shown to produce good approximate marginals in some applications [16], [17]. A reformulation of LBP that has recently been introduced in the literature relies on a tree-based reparameterization (TRP) [18]. This reparameterization has been shown to improve the convergence properties of LBP [18].

Direct application of the TRP principle to the case of MIMO symbol detection leads to an algorithm which has

Manuscript received June 13, 2007; revised November 7, 2007 and January 30, 2008; accepted May 9, 2008. The associate editor coordinating the review of this paper and approving it for publication was L. Yang.

C. M. Vithanage and J. Soler-Garrido are with Toshiba Research Europe Limited, 32, Queen Square, Bristol BS1 4ND, UK (e-mail: cheran@ieee.org, {cheran.vithanage, Josep.Soler}@toshiba-trel.com).

C. Andrieu is with the Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK (e-mail: c.andrieu@bristol.ac.uk).

R. J. Piechocki is with the Department of Electrical & Electronic Engineering, University of Bristol, Bristol BS8 1UB, UK (r.j.piechocki@bristol.ac.uk).

Digital Object Identifier 10.1109/T-WC.2008.070640

the advantage of having an implementation complexity that is at most quadratic in the number of transmit antennas. This is an attractive prospect for practical implementations. Performance of such applications is shown in [19], the work we improve upon here. However, similar to the LBP, the TRP based algorithms also face the problem that for distributions with loopy graphs, the algorithm fails to converge in some instances. In our setup, this sometimes results in an error floor for the bit-error rate (BER) performance.

In this work, we build upon the tree-based reparameterization and develop a principled methodology to ensure the reliable convergence of the detection algorithm for a given system. The methodology builds on the knowledge that convergence properties of TRP based algorithms are intrinsically related to the dependence structure of the joint distribution of interest. We hence introduce a slight modification of the distributions involved, which is shown in practice to dramatically improve the convergence properties of the algorithm while still capturing the important features of the exact distribution. Specifically, the true distribution of interest is approximated by selecting the best, in a sense to be defined later on, distribution out of a family of distributions. It is shown theoretically that the bias introduced by the selection of this alternative distribution is well controlled since the cost function that is minimized in selection automatically enforces that an upper bound on the distance between the actual and possible alternative distributions is also minimized. Simulation results show that excellent decoding performance is achieved as the result of this ensured convergence. For example, in the simulated system using 16-QAM modulation with four transmit and receive antennas, the SNR required to achieve a BER of 10^{-4} was only 3dB greater than that required by the optimal method. This is a much better error rate performance than, for example, the successive interference cancellation based V-BLAST decoding algorithm with the MMSE criterion [20], [21]. Also, this improved performance was achieved while retaining the low complexity offered by the classical TRP based detection method.

The system model and the method of optimal symbol detection will be presented in the next Section. The tree-based reparameterization principle and its application to MIMO symbol detection will be described in Section III. Development and analysis of the proposed method of ensuring convergence of the decoding algorithm is given in Section IV. Section V presents the complexity of the resulting algorithm while Section VI presents numerical simulation results. Finally, Section VII will give the conclusions.

II. System Model and Optimal Symbol Detection

A. Notation

First, let us define some notation which will be used throughout this paper. For a complex scalar s; $s^*, \Re(s), \Im(s)$ will denote its complex conjugate, real part and imaginary part, respectively. Vectors are taken to be column vectors and for a vector $\mathbf{v}$, $(\mathbf{v})_i$ denotes its ith element and $diag(\mathbf{v})$ denotes the diagonal matrix with $\mathbf{v}$ as the diagonal. For a matrix $\mathbf{M}$; $\mathbf{M}^\dagger$, $\mathbf{M}^\ddagger$, $\mathbf{M}^{-1}$ and $Tr(\mathbf{M})$ denotes its transpose, conjugate transpose, inverse and trace, respectively. $\Re(\mathbf{M})$ and $\Im(\mathbf{M})$ denote matrices of dimensions identical to $\mathbf{M}$, composing of the real and imaginary parts of the elements of $\mathbf{M}$, respectively. The (i,j)th element of $\mathbf{M}$ is denoted as $(\mathbf{M})_{i,j}$. $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The Frobenius norm of $\mathbf{M}$, $\sqrt{Tr(\mathbf{M}^\ddagger\mathbf{M})}$ is denoted by $\|\mathbf{M}\|_F$ and when $\mathbf{M}$ is a positive semidefinite matrix, $\|\mathbf{M}\|_2$ denotes its largest eigenvalue. The set of $n \times n$ positive definite matrices will be denoted by $\mathbf{S}^n_{++}$. For an $n \times m$ real matrix $\mathbf{N}$ and a real scalar r, $\min(\mathbf{N}, r)$ denotes the $n \times m$ matrix with the (i,j)th element being minimum of $(\mathbf{N})_{i,j}$ and r. Notation $vec(\mathbf{M})$ is used to denote a vectorization of matrix $\mathbf{M}$. For an $n \times n$ positive definite matrix $\mathbf{M}$, $\mathfrak{u}_T(\mathbf{M})$ refers to a vector consisting of the $\frac{n^2-n}{2}$ strictly upper triangular elements of $\mathbf{M}$ and $\mathcal{F}(\mathbf{M})$ denotes the real vector of length n^2 given by $\mathcal{F}(\mathbf{M}) = \left[vec(\Re(\mathbf{M}))^\dagger \ \mathfrak{u}_T(\Im(\mathbf{M}))^\dagger\right]^\dagger$.

B. System Model

Consider an n_t-transmit antenna, n_r-receive antenna MIMO communication system operating in a channel subject to frequency-flat quasi-static Rayleigh fading, as shown in Fig. 1. For a complex baseband discrete time signal model, for each time instant we have

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}. \tag{1}$$

Here, $\mathbf{y} = (y^1, ..., y^{n_r})^\dagger$ with y^j being the received signal on receiver antenna j and $\mathbf{x} = (x^1, ..., x^{n_t})^\dagger$ with x^i being the modulated symbol transmitted on transmit antenna i. The $n_r \times n_t$ matrix $\mathbf{H}$ has $h^{j,i}$, the channel fading coefficient from transmit antenna i to receive antenna j, as the (j,i)th element. A signal normalization is considered such that each $h^{j,i}$ has a zero-mean circularly-symmetric complex Gaussian distribution with a unit variance. Such probability distributions will be denoted as $\mathcal{CN}(0,1)$. Let the symbols on each antenna be selected from a set $B = \{a_1, ..., a_N\}$ with cardinality $|B| = N$. Finally, $\mathbf{w} = (w^1, ..., w^{n_r})^\dagger$ where w^j is the spatially uncorrelated additive white noise manifesting at the jth receive antenna with probability distribution $\mathcal{CN}(0, N_0)$.

C. Optimal Symbol Detection

We can consider optimal detection suited for a subsequent channel decoding operation as the computation of the marginal distributions $p(x^i|\mathbf{y})$ for each $i \in \{1, \cdots, n_t\}$. Considering no prior information at the decoder about $\mathbf{x}$, the joint posterior distribution of the vector of symbols is available up to a proportionality constant as

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &\propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{H}x\|^2}{N_0}\right) \\ &\propto \exp\left(\frac{2}{N_0}\Re(\mathbf{x}^\ddagger\mathbf{H}^\ddagger\mathbf{y}) - \frac{1}{N_0}\mathbf{x}^\ddagger\mathbf{H}^\ddagger\mathbf{H}\mathbf{x}\right). \end{aligned} \tag{2}$$

From this joint distribution, $p(x^i|\mathbf{y})$ for each i can be computed by a brute force marginalization. Because of the enumeration over the domain of $\mathbf{x}$, which has a size N^{n_t}, the complexity of the optimal algorithm is exponential in the number of transmit antennas, n_t. For large n_t this represents a prohibitive complexity for practical implementation

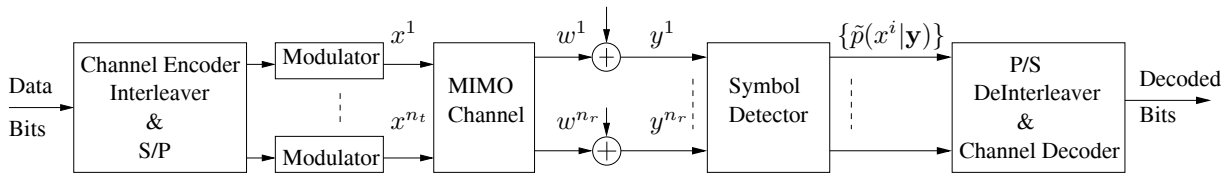

Fig. 1. The System Model.

and reduced-complexity symbol detection methods become necessary.

In the following, we will consider undirected graphical models for the representation of probability distributions [12]. It will be seen in the next Section that probability distributions of the form (2) are represented by UGMs with loops. A reduced-complexity method of obtaining approximate marginals in loopy UGMs, termed the tree-based reparameterization is introduced in [18]. This method is essentially a reformulation of the more popular loopy belief propagation. This TRP principle will be described via UGMs in the next Section followed by a description of its direct application for MIMO symbol detection.

III. Undirected Graphical Models, Tree-Based Reparameterization and the TRP MIMO Detector

Undirected graphical models are graphical representations for probability distributions. They compose of nodes and edges such that there are no multiple edges between any two nodes and there are no edges going from a node to itself [22], [12]. The set of nodes (say V) of the graph denotes component random variables of the joint distribution and the set of edges (say E) indicate some structural relationship, in the joint distribution, between the end nodes of each edge.

We restrict attention to distributions on sets of discrete random variables. Also, let us consider instances where the joint distribution is known at least up to a proportionality constant as the product of a set of functions of subsets of the set of variables. One example of such an UGM is shown in Fig. 2, which is the graph corresponding to (2) for the case of eight transmit antennas. For the particular case of (2), the joint probability distribution factorizes according to

$$p(\mathbf{x}) \propto \prod_{s\in\mathsf{V}} \psi_s(x^s) \prod_{(s,t)\in\mathsf{E}} \psi_{s,t}(x^s, x^t). \tag{3}$$

Where

$$\psi_s(x^s) = \exp\left(\frac{1}{N_0}\left[-|x^s|^2 \left(\mathbf{H}^{\ddagger}\mathbf{H}\right)_{s,s} + 2\Re\left\{(x^s)^* \left(\mathbf{H}^{\ddagger}\mathbf{y}\right)_s\right\}\right]\right), \tag{4}$$

and

$$\psi_{s,t}(x^s, x^t) = \exp\left(-\frac{2}{N_0}\Re\left\{(x^s)^* \left(\mathbf{H}^{\ddagger}\mathbf{H}\right)_{s,t} x^t\right\}\right). \tag{5}$$

Functions such as $\psi_s(x^s)$ and $\psi_{s,t}(x^s, x^t)$ will also be called *potentials* in the following. Now, had the UGM been cycle free, the posterior marginals of variables, $p(x^s|\mathbf{y})$ for $s \in \mathsf{V}$ and pairs of variables $p(x^s, x^t|\mathbf{y})$ for $(s,t) \in \mathsf{E}$ could have been found using a message passing scheme such as the sum-product algorithm, with an $\mathcal{O}\left(|\mathsf{E}|\, N^2\right)$ complexity [13]. The readers are also referred to [12] and [22] for excellent descriptions about UGMs and the sum-product algorithm.

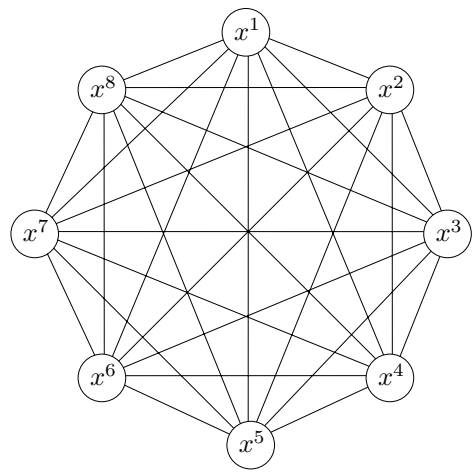

Fig. 2. The undirected graphical model corresponding to $n_t = 8$.

For loopy graphs such as Fig. 2, one sub-optimal reduced-complexity approach is to still apply the message updates of the sum-product algorithm iteratively until coming to a fixed point. This approach, which is termed loopy belief propagation, has shown much empirical success [16], [17]. In the tree-based reparameterization method of [18], which is the focus of this work, again the message updates of the sum-product algorithm are iteratively applied, but in a different manner as described next.

In UGMs such as Fig. 2, one can always identify a collection of edges of the graph which corresponds to a spanning tree. A spanning tree refers to a tree which covers all the nodes of the graph. Some examples are given in the subplots of Fig. 3. We will generically denote a spanning tree as T_k and its constituent edges as E_k, with k being an index on a set of spanning trees. Then, (3) can be decomposed as

$$p(\mathbf{x}) \propto \underbrace{\prod_{s\in\mathsf{V}} \psi_s(x^s) \prod_{(s,t)\in\mathsf{E}_k} \psi_{s,t}(x^s, x^t)}_{\propto\, p^k(\mathbf{x})} \cdot \underbrace{\prod_{(s,t)\in\mathsf{E}\setminus\mathsf{E}_k} \psi_{s,t}(x^s, x^t)}_{\propto\, p^{\setminus k}(\mathbf{x})}. \tag{6}$$

Now, the sum-product algorithm could be executed to marginalize the tree structured "distribution" $p^k(\mathbf{x})$ and find its marginals efficiently. These marginals would be termed *beliefs* in the following. Let the beliefs computed in this

manner be given as $q_s(x^s)$ and $q_{s,t}(x^s, x^t)$ for $s \in \mathsf{V}$ and $(s,t) \in \mathsf{E}_k$. Given these beliefs, it is known [12] that $p^k(\mathbf{x})$ can be *reparameterized* [1] as

$$p^k(\mathbf{x}) = \prod_{s \in \mathsf{V}} q_s(x^s) \prod_{(s,t) \in \mathsf{E}_k} \frac{q_{s,t}(x^s, x^t)}{q_s(x^s)\, q_t(x^t)}. \tag{7}$$

This reparameterization can be substituted back into (6) by setting

$$\psi_s^{new}(x^s) = q_s(x^s) \quad \text{for} \quad s \in \mathsf{V} \tag{8}$$

$$\psi_{s,t}^{new}(x^s, x^t) = \frac{q_{s,t}(x^s, x^t)}{q_s(x^s)\, q_t(x^t)} \quad \text{for} \quad (s,t) \in \mathsf{E}_k, \tag{9}$$

which determines the new potentials for the tree that was considered. In its execution, application of the tree-based reparameterization consists of considering a series of spanning trees. For each spanning tree, the sum-product algorithm is executed to compute the beliefs of the constituent variables and pairs of variables, which in turn keeps reparameterizing parts of (3). After several iterations of the procedure outlined above have been executed, the most recent beliefs $q_s(x^s)$ are output as the approximations to the marginal distributions $p(x^s|\mathbf{y})$.

A. The TRP MIMO detector

The TRP principle described above is used to develop a MIMO symbol detection algorithm in this Section. For the execution of the TRP principle, a sequence of spanning trees $\mathsf{T}_1, ..., \mathsf{T}_k, ..., \mathsf{T}_K$ with corresponding edge sets $\mathsf{E}_1, ..., \mathsf{E}_k, ..., \mathsf{E}_K$ needs to be selected. Instead of arbitrary trees, we select a set of chains $\mathsf{C}_1, ..., \mathsf{C}_k, ..., \mathsf{C}_K$ with each chain spanning all the variables. This standardizes the execution of the sum-product algorithm and will lead to a reduction in implementation complexity. For even n_t, the minimum number of chains to cover all the edges of the graph is $K = \binom{n_t}{2}/(n_t - 1) = n_t/2$. A method to select $n_t/2$ such spanning chains will be presented in Section III-B.

For a given time instant, the decoder begins with the joint distribution (3) with the potentials associated with the singleton and pairwise variables given by (4) and (5). Thereafter, the TRP MIMO detector at the kth iteration selects the chain C_k and the corresponding edge set E_k. Let us assume the chosen chain leads to an ordering of the indices of the variables as $\{k_1 \to ... \to k_{i-1} \to k_i \to k_{i+1} \to ... \to k_{n_t}\}$. Now the sum-product algorithm is used to marginalize this chain structured distribution. In its implementation, the sum-product algorithm carries out a forward recursion through the chain computing a set of forward messages $M^k_{k_i \to k_{i+1}}(x^{k_{i+1}})$ for $i = 1, ..., n_t - 1$, and a backward recursion through the chain computing a set of backward messages $M^k_{k_{i+1} \to k_i}(x^{k_i})$ for $i = n_t - 1, ..., 1$. The formulae for these forward-backward recursions are given below in (10) and (11). Note that from here onwards, for brevity of presentation, we usually choose not to indicate the arguments of functions such as the potentials, beliefs and messages.

$$M^k_{k_i \to k_{i+1}} \propto \sum_{x^{k_i} \in B} \left(\psi^{k-1}_{k_i, k_{i+1}} \psi^{k-1}_{k_i} M^k_{k_{i-1} \to k_i} \right) \tag{10}$$

$$M^k_{k_{i+1} \to k_i} \propto \sum_{x^{k_{i+1}} \in B} \left(\psi^{k-1}_{k_i, k_{i+1}} \psi^{k-1}_{k_{i+1}} M^k_{k_{i+2} \to k_{i+1}} \right) \tag{11}$$

Here, the initializations $M^k_{k_0 \to k_1}(x^{k_1}) = 1$ for $x^{k_1} \in B$ and $M^k_{k_{n_t+1} \to k_{n_t}}(x^{k_{n_t}}) = 1$ for $x^{k_{n_t}} \in B$ are assumed. After a full forward and a backward recursion have been carried out, the beliefs of the singleton and pairwise connected variables of the selected chain are given by:

$$q_{k_i} \propto M_{k_{i-1} \to k_i} \psi^{k-1}_{k_i} M_{k_{i+1} \to k_i}$$

$$q_{k_i, k_{i+1}} \propto M_{k_{i-1} \to k_i} \psi^{k-1}_{k_i} \psi^{k-1}_{k_i, k_{i+1}} \psi^{k-1}_{k_{i+1}} M_{k_{i+2} \to k_{i+1}}.$$

From the reparameterization provided by (7), this enables the computation of the new potentials associated with the vertices and edges of the chain C_k as

$$\psi^k_{k_i} \propto \psi^{k-1}_{k_i} M^k_{k_{i-1} \to k_i} M^k_{k_{i+1} \to k_i} \tag{12}$$

$$\psi^k_{k_i, k_{i+1}} \propto \frac{\psi^{k-1}_{k_i, k_{i+1}}}{M^k_{k_{i+1} \to k_i} M^k_{k_i \to k_{i+1}}}. \tag{13}$$

Note from (5), that initial edge potentials $\psi_{s,t}$ need to be computed only once for a given channel $\mathbf{H}$. Thus, the TRP MIMO detection algorithm is as given below.

1 Select the sequence of spanning chains $\mathsf{C}_1, ..., \mathsf{C}_k, ..., \mathsf{C}_K$ with the corresponding edge sets $\mathsf{E}_1, ..., \mathsf{E}_k, ..., \mathsf{E}_K$.
2 Initialize the edge potentials $\psi^{init}_{s,t} = \psi_{s,t}$ for $(s,t) \in \mathsf{E}$ using (5).
3 For each received signal vector $\mathbf{y}$
   3.1 Set $k = 1$. Initialize $\psi^0_s = \psi_s$ for $s \in \mathsf{V}$ using (4) and set $\psi^0_{s,t} = \psi^{init}_{s,t}$ for $(s,t) \in \mathsf{E}$.
   3.2 Perform message passing along the chain C_k using (10) and (11) and compute the new potentials of the variables associated with the vertices V using (12) and the new pairwise potentials associated with the edges E_k using (13).
   3.3 Increase k by one. If $k > K$ go to Step 3.4, otherwise go to Step 3.2.
   3.4 Output the properly normalized ψ^K_s for $s \in \mathsf{V}$ as the computed approximations to the marginals.

B. A sequence of spanning chains

It is easy to see that there are many possible sets of $n_t/2$ spanning chains which will cover all the edges of graphs such as that in Fig. 2. Here, we assume that n_t is even. Otherwise, the procedure can be easily extended to find a set of $\frac{(n_t+1)}{2}$ spanning chains which cover all the edges of the graph. The identification of one particular set of chains is as follows.

Define the modulo operation $\langle r \rangle_{n_t}$ acting on some $r \in \{-(n_t - 1), \cdots, 0, \cdots, n_t\}$ as $\langle r \rangle_{n_t} = r$ if $r > 0$ and $\langle r \rangle_{n_t} = (r + n_t)$ if $r \leq 0$. Then the sequence of chains $\mathsf{C}_1, ..., \mathsf{C}_k, ..., \mathsf{C}_K$ can be described by the kth chain containing a sequence of variables with the indices ordered as $\{\langle k \rangle_{n_t} \to \langle k+1 \rangle_{n_t} \to \langle k-1 \rangle_{n_t} \to \langle k+2 \rangle_{n_t} \to \cdots \to \langle k + \frac{n_t}{2} \rangle_{n_t}\}$. The resulting selection of four spanning chains for the case of $n_t = 8$ is shown in Fig. 3.

[1] using the terminology of [18]

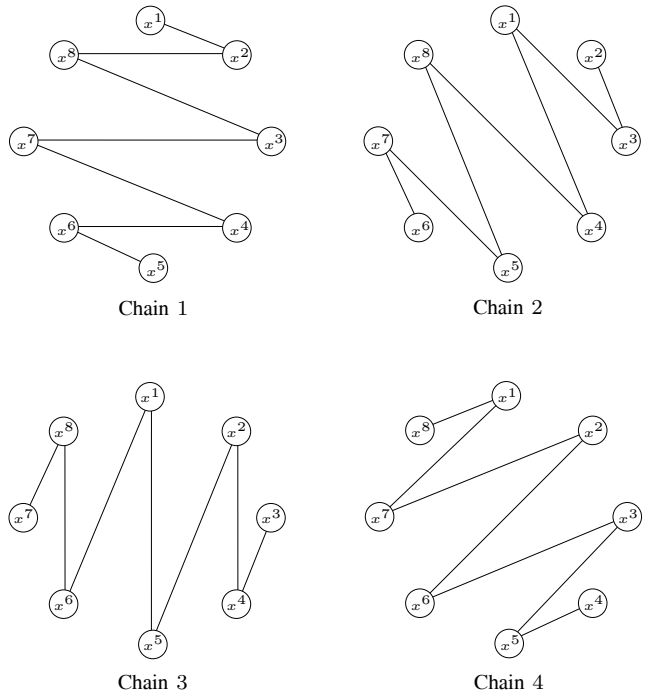

Fig. 3. A set of 4 chains to cover all the edges.

IV. Distributional Approximations for the TRP-Based Detection

The straightforward application of the TRP principle does not always perform well since, as with LBP itself, there exists the issue that the decoding algorithm does not converge for some joint distributions [18]. The result is that for some system setups, such as when $n_t = n_r = 4$, the decoded error rate floors out. This will be illustrated later in Fig. 4. We propose to overcome this problem by selecting a suitable alternative joint distribution on which the TRP based decoder will be executed. Note that use of approximate models have been shown to be specifically beneficial in computation limited situations, in [23].

First, let us process the received signal as follows, which is not going to affect the joint distribution on which the marginalization is to take place:

$$\tilde{\mathbf{y}} = \left(\mathbf{H}^{\ddagger}\mathbf{H}\right)^{-1}\mathbf{H}^{\ddagger}\mathbf{y} = \mathbf{x} + \tilde{\mathbf{n}}\,.$$

Here, it has been assumed that $\mathbf{H}^{\ddagger}\mathbf{H}$ is invertible. Observing that $\tilde{\mathbf{n}} \sim \mathcal{CN}(0, N_0\left(\mathbf{H}^{\ddagger}\mathbf{H}\right)^{-1})$, we can write the joint distribution of (2) as

$$p\left(\mathbf{x}\,|\mathbf{y}\right) \propto \exp\left\{-\left(\tilde{\mathbf{y}}-\mathbf{x}\right)^{\ddagger}\mathbf{C}\left(\tilde{\mathbf{y}}-\mathbf{x}\right)\right\}, \tag{14}$$

where $\mathbf{C} = \frac{\mathbf{H}^{\ddagger}\mathbf{H}}{N_0}$ is almost surely a positive definite matrix. Given this distribution, one can ask the following question: What properties in this distribution are desired such that the convergence of a TRP based detector can be guaranteed? Exact derivation of necessary and sufficient conditions on the potentials for the guaranteed convergence of TRP based detection, is still an open problem. But, there exist sufficient conditions for the guaranteed convergence of LBP based detectors. Examples are [24] and [25]. These works suggest that a limitation on the dynamic ranges of the edge potentials can ensure convergence of the decoding algorithm. For the edge potentials pertaining to (14), this correspond to a limitation on $\max_{i\neq j}\left|(\mathbf{C})_{i,j}\right|$. Since for a positive definite matrix $\mathbf{C}$, $\max_{i,j}\left|(\mathbf{C})_{i,j}\right| \leq \|\mathbf{C}\|_2$, and since its principal minors are positive [26],

$$\max_{i\neq j}\left|(\mathbf{C})_{i,j}\right| < \|\mathbf{C}\|_2\,.$$

Therefore we expect that a suitable constraint on the largest eigenvalue of $\mathbf{C}$ in (14) will ensure the convergence of TRP based marginal computations.

Note that the set of all positive definite matrices parameterizes a family of distributions in that for any $\mathbf{M} \in \mathbf{S}_{++}^{n_t}$,

$$g\left(\mathbf{x};\mathbf{M}\right) \propto \exp\left\{-\left(\tilde{\mathbf{y}}-\mathbf{x}\right)^{\ddagger}\mathbf{M}\left(\tilde{\mathbf{y}}-\mathbf{x}\right)\right\}$$

is a valid distribution in the discrete random vector $\mathbf{x}$. In this work, for a given joint distribution $p\left(\mathbf{x}\,|\mathbf{y}\right) = g\left(\mathbf{x};\mathbf{C}\right)$, we propose to execute the TRP based detection algorithm instead on a distribution $g\left(\mathbf{x};\mathbf{M}_{opt}\right)$. $\mathbf{M}_{opt}$ is the positive definite matrix closest to $g\left(\mathbf{x};\mathbf{C}\right)$ in the Euclidean norm $\|\mathbf{C}-\mathbf{M}_{opt}\|_F$, such that its largest eigenvalue is below a value σ_{th}. Here, σ_{th} is a parameter to be predetermined for guaranteed convergence and the resulting improved performance of TRP based detectors will be demonstrated in Section VI.

A. Selection of the optimal distribution

Given the actual distribution $p\left(\mathbf{x}\,|\mathbf{y}\right) = g\left(\mathbf{x};\mathbf{C}\right)$, we have proposed the execution of the TRP based detection scheme on an alternative distribution $g\left(\mathbf{x},\mathbf{M}_{opt}\right)$ such that the eigenvalues of $\mathbf{M}_{opt}$ are less than a parameter σ_{th} and the cost function $\|\mathbf{C}-\mathbf{M}_{opt}\|_F$ is minimized. The following proposition finds that optimal distribution which also happens to be unique.

Proposition 1: Let $p\left(\mathbf{x}\,|\mathbf{y}\right) = g\left(\mathbf{x};\mathbf{C}\right)$ with $\mathbf{C} \in \mathbf{S}_{++}^{n_t}$. Also let $\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^{\ddagger}$ where $\mathbf{V}$ is unitary and $\mathbf{\Lambda} = diag(\lambda_1, ..., \lambda_{n_t})$. The distributional approximation $g\left(\mathbf{x};\mathbf{M}\right)$ to $p\left(\mathbf{x}\,|\mathbf{y}\right)$ such that $\mathbf{M} \in \mathbf{S}_{++}^{n_t}$, the eigenvalues of $\mathbf{M}$ are less than a value σ_{th} (> 0) and the Frobenius norm $\|\mathbf{C}-\mathbf{M}\|_F$ is minimized, is given by

$$\mathbf{M}_{opt} = \mathbf{V}[\min(\mathbf{\Lambda}, \sigma_{th})]\mathbf{V}^{\ddagger}. \tag{15}$$

Proof: Please see Appendix A. ■

Therefore, in this modified TRP based detection algorithm, the marginalization will take place on the joint distribution $g\left(\mathbf{x};\mathbf{M}_{opt}\right) \propto \exp\{2\Re(\mathbf{x}^{\ddagger}\mathbf{N}\mathbf{y}) - \mathbf{x}^{\ddagger}\mathbf{M}_{opt}\mathbf{x}\}$, where

$$\mathbf{N} = \mathbf{M}_{opt}\left(\mathbf{H}^{\ddagger}\mathbf{H}\right)^{-1}\mathbf{H}^{\ddagger}. \tag{16}$$

The initial potential functions given by this distribution are:

$$\psi_s\left(x^s\right) = \exp(-\left|x^s\right|^2\left(\mathbf{M}_{opt}\right)_{s,s} + 2\Re\left\{\left(x^s\right)^*\left(\mathbf{N}\mathbf{y}\right)_s\right\}), \tag{17}$$

and

$$\psi_{s,t}\left(x^s, x^t\right) = \exp\left(-2\Re\left\{\left(x^s\right)^*\left(\mathbf{M}_{opt}\right)_{s,t}x^t\right\}\right). \tag{18}$$

With the observation that the initial edge potential computations and the computation of the matrices $\mathbf{M}_{opt}$ and $\mathbf{N}$ need to

be done only once for a given channel matrix $\mathbf{H}$, the modified TRP MIMO detection algorithm is as given below.

1 Select the sequence of spanning chains $\mathsf{C}_1, ..., \mathsf{C}_k, ..., \mathsf{C}_K$ with the corresponding edge sets $\mathsf{E}_1, ..., \mathsf{E}_k, ..., \mathsf{E}_K$.
2 Select the parameter σ_{th} and compute $\mathbf{M}_{opt}$ and $\mathbf{N}$ from (15) and (16).
3 Initialize the edge potentials $\psi_{s,t}^{init} = \psi_{s,t}$ for $(s,t) \in \mathsf{E}$ using (18).
4 For each received signal vector $\mathbf{y}$
   4.1 Set $k = 1$. Initialize $\psi_s^0 = \psi_s$ for $s \in \mathsf{V}$ using (17) and set $\psi_{s,t}^0 = \psi_{s,t}^{init}$ for $(s,t) \in \mathsf{E}$.
   4.2 Perform message passing along the chain C_k using (10) and (11) and compute the new potentials of the variables associated with the vertices V using (12) and the new pairwise potentials associated with the edges E_k using (13).
   4.3 Increase k by one. If $k > K$ go to Step 4.4, otherwise go to Step 4.2.
   4.4 Output the properly normalized ψ_s^K for $s \in \mathsf{V}$ as the computed approximations to the marginals.

B. Effect on the distance between distributions by the Frobenius norm reduction

In finding the $\mathbf{M}_{opt}$ with the desired eigenvalue constraints, we are interested in introducing a controlled bias in the marginal computations. This Section shows that the introduced bias is well controlled since the minimization of the cost function $\|\mathbf{C} - \mathbf{M}\|_F$ in finding $g(\mathbf{x}; \mathbf{M}_{opt})$ also ensures that an upper bound on the total variation norm between the distributions $g(\mathbf{x}; \mathbf{C})$ and $g(\mathbf{x}; \mathbf{M})$,

$$TV = \sum_{\mathbf{x} \in B^{n_t}} |g(\mathbf{x}; \mathbf{C}) - g(\mathbf{x}; \mathbf{M})|,$$

is also minimized. Thus the alternative distribution $g(\mathbf{x}; \mathbf{M}_{opt})$ is ensured not to be far away from the actual distribution. Denoting $g(\mathcal{E}; \mathbf{C}) = \sum_{\mathbf{x} \in \mathcal{E}} g(\mathbf{x}; \mathbf{C})$, from [27],

$$\max_{\mathcal{E} \subseteq B^{n_t}} \{g(\mathcal{E}; \mathbf{C}) - g(\mathcal{E}; \mathbf{M})\} = \frac{TV}{2}.$$

Thus a reduction in the total variation norm also leads to a reduction in the variation between, for example, the marginal distributions. Therefore the containment of the TV norm is desirable for the purpose of soft detection considered here.

Consider a parameterization of our family of distributions $g(\mathbf{x}; \mathbf{M})$ given by $\theta_{\mathbf{M}} = \mathcal{F}(\mathbf{M})$, where $\mathcal{F}(\mathbf{M})$ was introduced in Section II-A. The range of $\mathcal{F} : \mathbf{S}_{++}^{n_t} \to \mathbb{R}^{n_t^2}$, denoted by Θ, is given by $\{\mathcal{F}(\mathbf{M}) | \mathbf{M} \in \mathbf{S}_{++}^{n_t}\}$. It can be seen that $\mathcal{F}$ defines a bijection from $\mathbf{S}_{++}^{n_t}$ to Θ, and therefore there is an inverse map $\mathcal{F}^{-1} : \Theta \to \mathbf{S}_{++}^{n_t}$. Thus, each $\theta_{\mathbf{M}}$ refers to a unique positive definite matrix $\mathbf{M}$.

With the parameterization on $\theta_{\mathbf{M}}$, each member of the family of distributions $g(\mathbf{x}; \mathbf{M})$, which also happens to be an exponential family of distributions, can be expressed as

$$g(\mathbf{x}; \mathbf{M}) = g(\mathbf{x}; \theta_{\mathbf{M}}) = \frac{\exp\langle \theta_{\mathbf{M}}, \phi(\mathbf{x}) \rangle}{Z(\theta_{\mathbf{M}})}.$$

Here, the partition function $Z(\theta_{\mathbf{M}})$ is given by $Z(\theta_{\mathbf{M}}) = \sum_{\mathbf{x} \in B^{n_t}} \exp\langle \theta_{\mathbf{M}}, \phi(\mathbf{x}) \rangle$ and the set of sufficient statistics, $\phi(\mathbf{x})$ is given in Appendix C. Now in selecting the distribution $g(\mathbf{x}; \mathbf{M}_{opt})$, the criterion minimized is the Frobenius norm $\|\mathbf{C} - \mathbf{M}\|_F$, which is equivalent to the Euclidean distance $\|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2$. We claim the following proposition.

Proposition 2: Let $g(\mathbf{x}; \mathbf{C})$ and $g(\mathbf{x}; \mathbf{M})$ be two members of the exponential family with sufficient statistics $\phi(\mathbf{x})$, parameterized by the exponential parameters $\theta_{\mathbf{C}}$ and $\theta_{\mathbf{M}}$, respectively. The total variation norm between these two distributions, $TV = \sum_{\mathbf{x}} |g(\mathbf{x}; \mathbf{C}) - g(\mathbf{x}; \mathbf{M})|$ is upper bounded as

$$TV \leq \exp(2\kappa n_t \|\theta_{\mathbf{C}}\|_2) \{\exp(2\kappa n_t \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2) - 1\},$$

where $\kappa = \sup_{\mathbf{x}, i} |(\phi(\mathbf{x}))_i|$.

Proof: Please see Appendix B. ■

Thus an upper bound on the total variation norm between the two distributions monotonically decreases with the reduction in the Euclidean distance between the parameters. Thus the minimization of the Euclidean distance also minimizes this upper bound, which justifies Frobenius norm reduction as the selection criterion in this modified TRP MIMO detector. We note that for *small* $\|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|$, this upper bound decreases linearly with the reduction in $\|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|$ and that the $TV \to 0$ as $\|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\| \to 0$.

V. Orders of Complexities

The optimal method of obtaining the APPs with its enumeration over all the possible configurations of $\mathbf{x}$ can be seen to have a complexity per time instant on the order of $\mathcal{O}(N^{n_t} {n_t}^2)$ assuming $n_t = n_r$. For both TRP based algorithms, the complexity of each algorithm is governed by the message passing operation on each chain decoding, which is having an $\mathcal{O}(KN^2(n_t - 1))$ complexity. From simulation results, the number of TRP iterations needed for convergence in the algorithms is found to be $\mathcal{O}(n_t)$, provided that the decoding is actually going to converge to a fixed point. Usually n_t iterations sufficed in both the classical TRP based detection algorithm as well as in the modified TRP based detection algorithm.

It should be noted that computation of the eigenvalues and eigenvectors, as required for the modified TRP algorithm needs to be performed only once per a given channel matrix. Thus, for quasi-static fading channels, the TRP principle based decoders reduce the system complexity from $\mathcal{O}(N^{n_t} {n_t}^2)$ to $\mathcal{O}(N^2 n_t(n_t - 1))$ with a bit error rate performance as given in the next Section. We note that the message passing involved in each chain decoding is much similar to the application of the BCJR algorithm of [28]. Therefore, for large symbol constellations one can also apply reduced-complexity BCJR variations (e.g. [29]) to accomplish this task at a reduced complexity, thereby making the complexity less than quadratic in N.

For fast fading channels, assuming the channels change independently from one time instant to the next, the $\mathbf{M}_{opt}$ and $\mathbf{N}$ need to be computed at each time instant. Again assuming $n_t = n_r$, this leads to an order of complexity which is cubic in the number of transmit antennas, and is still comparable with the complexity of low complexity decoders such as the successive interference cancellation based V-BLAST

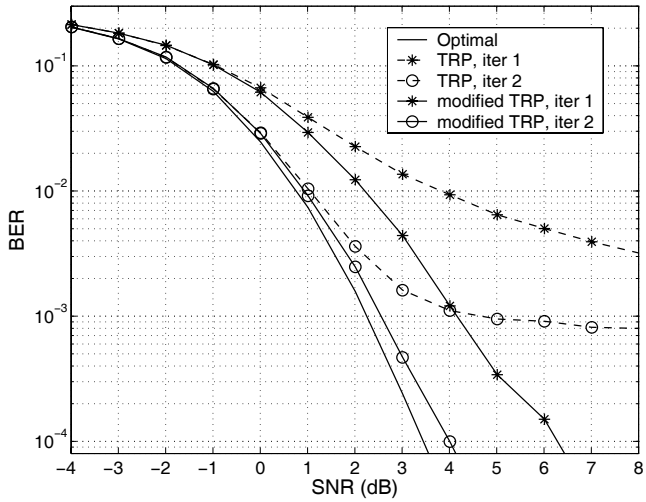

Fig. 4. Bit error rate performance with BPSK transmissions. $n_t = n_r = 4$. Modified TRP decoding is with $\sigma_{th} = 10$.

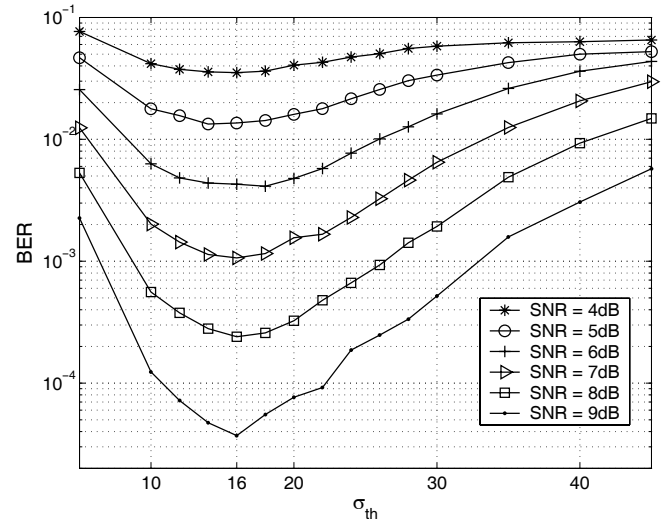

Fig. 5. Bit error rate performance with σ_{th} for QPSK transmissions. $n_t = n_r = 4$. Four modified TRP iterations.

decoding algorithm with the MMSE criterion (fast version of [21]). Obviously, when there are significant correlations between the channels at each time instant, the required eigen decompositions can be done only at suitable time lags when the underlying channels have significantly changed. Also in the case of symbol detection in orthogonal frequency division multiplexed systems, the eigen decompositions in each subcarrier can be appropriately avoided in the presence of significant correlations between the subcarriers.

VI. Simulation Results

In the following simulations, each frame transmission contained 1152 data bits which were encoded by a rate half turbo encoder and passed through a random interleaver. The turbo encoder consisted of two constituent $(5, 7)_8$ convolutional codes. The sequence of chains were selected as described in Section III-B. The turbo decoder performed four iterations of decoding. Each simulation point represents the simulation of at least 10^4 frame transmissions.

Even though the convergence properties of the classical TRP based decoder improve with the number of transmit antennas [19], we observe the simulations in systems with four transmit antennas, which are practically more relevant. Fig. 4 shows the decoding performance for BPSK modulated transmissions with $n_r = 4$. Here, SNR $= E_s/N_0$ and E_s denotes the average energy per transmitted vector symbol $\mathbf{x}$. We can clearly observe the error floor behaviour due to the non convergence of the classical TRP principle based decoding algorithm for some distributions. With the modifications suggested in Section IV, this error floor behaviour is removed and the decoded error rates come close to the optimal performance. Fig. 5 shows the bit error rate performance with σ_{th} for QPSK transmissions operating at different signal to noise energy ratios. With these results $\sigma_{th} = 16$ is selected for decoding QPSK transmissions for $n_t = n_r = 4$ systems. Note that simulations indicate optimal σ_{th} values for ensembles of distributions defined by the number of transmit-receive antennas and the symbol alphabet. But an analytical identification of such optimal values require the exact knowledge of convergence of these

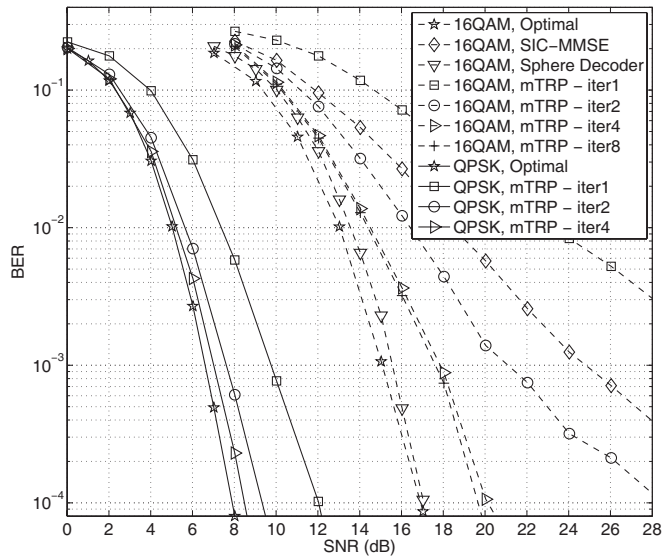

Fig. 6. Bit error rate performance with QPSK and 16-QAM transmissions, where mTRP denotes the modified TRP based decoding algorithm. $n_t = n_r = 4$. SIC-MMSE denotes the successive interference cancellation method using the MMSE criterion. The QPSK system used $\sigma_{th} = 16$ and the 16-QAM system used $\sigma_{th} = 48$.

decoding schemes, which is still an open problem. Thus σ_{th} is selected for each system setup via numerical simulations. Fig. 6 shows the resulting BER performance for QPSK and 16-QAM transmissions. The modified TRP algorithm in the QPSK case was implemented with $\sigma_{th} = 16$, while that in the 16-QAM case was with $\sigma_{th} = 48$. For the 16-QAM systems, we have also plotted the performances of the hard decision making successive interference cancellation method for V-BLAST systems, which uses the minimum mean squared error (MMSE) criterion [20] and the max-log sphere decoder of [6]. The V-BLAST decoder was implemented along with its optimal ordering for the successive interference cancellation.

Fig. 7 shows the results when $n_r = 6$. The modified TRP algorithm in the QPSK case was implemented with $\sigma_{th} = 18$, while that in the 16-QAM case was with $\sigma_{th} = 52$. It

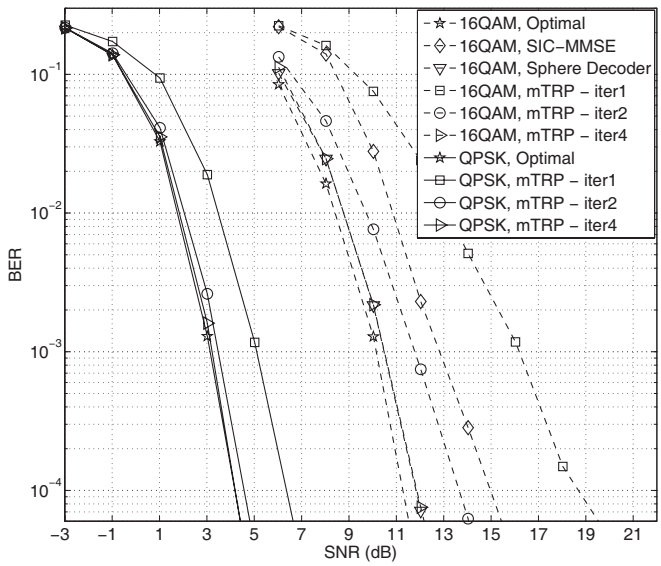

Fig. 7. Bit error rate performance with QPSK and 16-QAM transmissions. $n_t = 4$, $n_r = 6$. The QPSK system used $\sigma_{th} = 18$ and the 16-QAM system used $\sigma_{th} = 52$.

can be observed that the proposed algorithm performs nearly optimally for the QPSK system and within 0.5dB of the optimal performance in the 16-QAM system.

VII. Conclusions

We have investigated the tree-based reparameterization principle for reduced-complexity symbol detection in MIMO spatial multiplexing systems. The resulting algorithm has an $\mathcal{O}(N^2 n_t(n_t-1))$ complexity, which is attractive for practical implementations. Furthermore, a novel methodology has been developed to improve the decoding performance of the classical TRP principle by the selection of a suitable alternative joint distribution. It has been shown theoretically that the bias introduced by the selection of this alternative distribution is well controlled since the cost function that is minimized in selection also minimizes an upper bound on the total variation norm between the actual distribution and possible alternative distributions. Simulation results show the excellent decoding performance due to this ensured convergence while keeping the low complexity offered by the TRP principle.

We also comment here that the proposed method of ensuring the convergence of the TRP principle based algorithms can also be used to improve the convergence properties of other decoding strategies such as the use of loopy belief propagation.

In the developed algorithms, the task of marginalization is broken down into a series of marginalizations on distributions with chain structured undirected graphs. Marginalization on such distributions takes a form similar to the BCJR algorithm. Therefore, for large symbol alphabets, it is further possible to apply reduced-complexity BCJR variations to perform these marginalizations at a lower complexity.

Appendix A
Selection of the Optimal $g(\mathbf{x}, \mathbf{M})$ Subject to the Eigenvalue Constraints on $\mathbf{M}$

Let $\mathbf{M} = \mathbf{U}\mathbf{\Delta}\mathbf{U}^{\ddagger}$, where $\mathbf{U}$ is a unitary matrix and $\mathbf{\Delta} = diag(\delta_1, ..., \delta_{n_t})$. Without loss of generality, we will assume that $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{n_t} > 0$ and $\sigma_{th} \geq \delta_1 \geq \delta_2 \geq \cdots \delta_{n_t} > 0$. Let $\mathbf{R}$ be the unitary matrix such that $\mathbf{R} = \mathbf{V}^{\ddagger}\mathbf{U}$. Consider the objective function

$$\begin{aligned}
\|\mathbf{C}-\mathbf{M}\|_F^2 &= Tr\left[(\mathbf{C}-\mathbf{M})^{\ddagger}(\mathbf{C}-\mathbf{M})\right] \\
&= Tr\left[\mathbf{C}^{\ddagger}\mathbf{C}+\mathbf{M}^{\ddagger}\mathbf{M}-\mathbf{M}^{\ddagger}\mathbf{C}-\mathbf{C}^{\ddagger}\mathbf{M}\right] \\
&= Tr[\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{\ddagger}+\mathbf{U}\mathbf{\Delta}^2\mathbf{U}^{\ddagger} \\
&\quad -\mathbf{V}\mathbf{R}\mathbf{\Delta}\mathbf{R}^{\ddagger}\mathbf{\Lambda}\mathbf{V}^{\ddagger}-\mathbf{V}\mathbf{\Lambda}\mathbf{R}\mathbf{\Delta}\mathbf{R}^{\ddagger}\mathbf{V}^{\ddagger}] \\
&= \sum_{i=1}^{n_t}\delta_i^2+\sum_{i=1}^{n_t}\lambda_i^2-2Tr\left[\mathbf{R}\mathbf{\Delta}\mathbf{R}^{\ddagger}\mathbf{\Lambda}\right]
\end{aligned}$$

Now, we can minimize this squared norm in $\mathbf{R}$ and δ_i for $i = \{1, ..., n_t\}$ to obtain the positive definite matrix closest to $\mathbf{C}$ which satisfies the eigenvalue constraints.

From [30], $Tr\left[\mathbf{R}\mathbf{\Delta}\mathbf{R}^{\ddagger}\mathbf{\Lambda}\right] \leq \sum_{i=1}^{n_t}\delta_i\lambda_i$. Here the equality is achieved for $\mathbf{R} = \mathbf{J}_{n_t}$, where $\mathbf{J}_{n_t}$ is a diagonal matrix with each of the diagonal elements being some complex square root of 1. This means that for any $\mathbf{\Delta}$, $\mathbf{R} = \mathbf{J}_{n_t}$ minimizes the objective function, and due to the commutativity in the multiplication of diagonal matrices, the objective function reduces to

$$\|\mathbf{C}-\mathbf{M}\|_F^2 = \sum_{i=1}^{n_t}(\delta_i-\lambda_i)^2.$$

With the constraints on each δ_i for every i, the solution is simply $\delta_i = \min(\lambda_i, \sigma_{th})$, and noting that $\mathbf{J}_{n_t}\mathbf{J}_{n_t}^{\ddagger} = \mathbf{I}_{n_t}$, the *unique* optimal $\mathbf{M}_{opt}$ satisfying the constraints is

$$\mathbf{M}_{opt} = \mathbf{V}[\min(\mathbf{\Lambda}, \sigma_{th})]\mathbf{V}^{\ddagger}.$$

Appendix B
Upper Bound on the Total Variation Norm

Assuming $\theta^{\gamma} = \gamma\theta_{\mathbf{M}} + (1-\gamma)\theta_{\mathbf{C}}$ and $f_{\theta^{\gamma}}(\mathbf{x}) = \frac{\exp\langle\theta^{\gamma},\phi(\mathbf{x})\rangle}{Z(\theta^{\gamma})}$,

$$\begin{aligned}
TV &= \sum_{\mathbf{x}\in B^{n_t}}\left|\int_0^1\frac{d}{d\gamma}f_{\theta^{\gamma}}(\mathbf{x})\,d\gamma\right| \\
&= \sum_{\mathbf{x}\in B^{n_t}}\left|\int_0^1\xi f_{\theta^{\gamma}}(\mathbf{x})\,d\gamma-\int_0^1\zeta f_{\theta^{\gamma}}(\mathbf{x})\,d\gamma\right|,
\end{aligned}$$

where $\xi = \langle\theta_{\mathbf{M}}-\theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle$ and $\zeta = \frac{1}{Z(\theta^{\gamma})}\frac{d}{d\gamma}Z(\theta^{\gamma})$. Thus

$$\begin{aligned}
TV &\leq \underbrace{\sum_{\mathbf{x}\in B^{n_t}}\left|\int_0^1\langle\theta_{\mathbf{M}}-\theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle f_{\theta^{\gamma}}(x)\,d\gamma\right|}_{V} \\
&\quad + \underbrace{\sum_{\mathbf{x}\in B^{n_t}}\left|\int_0^1 f_{\theta^{\gamma}}(\mathbf{x})\frac{1}{Z(\theta^{\gamma})}\frac{d}{d\gamma}Z(\theta^{\gamma})\,d\gamma\right|}_{W}. \quad (19)
\end{aligned}$$

Now, using the Cauchy-Schwarz inequality,

$$\begin{aligned} V &\leq \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2 \sum_{\mathbf{x}\in B^{n_t}} \int_0^1 \|\phi(\mathbf{x})\|_2 f_{\theta^\gamma}(\mathbf{x})\, d\gamma \\ &= \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2 \int_0^1 E_{f_{\theta^\gamma}}\left\{\|\phi(\mathbf{x})\|_2\right\} d\gamma. \end{aligned}$$

Since $\frac{d}{d\gamma} Z(\theta^\gamma) = \sum_{\mathbf{x}\in B^{n_t}} \exp\langle \theta^\gamma, \phi(\mathbf{x})\rangle \langle \theta_{\mathbf{M}} - \theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle$,

$$\begin{aligned} W &= \sum_{\mathbf{y}} \left| \int_0^1 f_{\theta^\gamma}(\mathbf{y}) \left[\sum_{\mathbf{x}} f_{\theta^\gamma}(\mathbf{x}) \langle \theta_{\mathbf{M}} - \theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle \right] d\gamma \right| \\ &\leq \sum_{\mathbf{y}} \int_0^1 f_{\theta^\gamma}(\mathbf{y}) \left[\sum_{\mathbf{x}} f_{\theta^\gamma}(\mathbf{x}) \left|\langle \theta_{\mathbf{M}} - \theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle\right| \right] d\gamma \\ &\leq \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2 \int_0^1 E_{f_{\theta^\gamma}}\left\{\|\phi(\mathbf{x})\|_2\right\} d\gamma \end{aligned}$$

Therefore from (19),

$$TV \leq 2\|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2 \int_0^1 E_{f_{\theta^\gamma}}\left\{\|\phi(\mathbf{x})\|_2\right\} d\gamma. \tag{20}$$

Let $\kappa = \sup_{\mathbf{x},i} |(\phi(x))_i|$. We will also denote the l_∞ norm by $\|\bullet\|_\infty$ and noting that $\|\phi(\mathbf{x})\|_2 \leq n_t \|\phi(\mathbf{x})\|_\infty \leq \kappa n_t$ [26], note the inequalities: $-\kappa n_t \|\theta\|_2 \leq \langle \theta, \phi(\mathbf{x})\rangle \leq \kappa n_t \|\theta\|_2$. For $\gamma \in [0,1]$,

$$\begin{aligned} Z(\theta^\gamma) &= \sum_{\mathbf{x}} \exp\langle \gamma(\theta_{\mathbf{M}} - \theta_{\mathbf{C}}) + \theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle \\ &= \sum_{\mathbf{x}} \exp\langle \theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle \exp\langle \gamma(\theta_{\mathbf{M}} - \theta_{\mathbf{C}}), \phi(\mathbf{x})\rangle \\ &\geq N^{n_t} \exp(-\kappa n_t \|\theta_{\mathbf{C}}\|_2) \exp(-\kappa n_t \gamma \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2). \end{aligned}$$

$$\begin{aligned} &E_{f_{\theta^\gamma}}\left\{\|\phi(\mathbf{x})\|_2\right\} \\ &= \sum_{\mathbf{x}} \|\phi(\mathbf{x})\|_2 \frac{\exp\langle \gamma(\theta_{\mathbf{M}} - \theta_{\mathbf{C}}), \phi(\mathbf{x})\rangle}{Z(\theta^\gamma)} \exp\langle \theta_{\mathbf{C}}, \phi(\mathbf{x})\rangle \\ &\leq \kappa n_t \frac{\exp(\kappa n_t \gamma \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2)}{Z(\theta^\gamma)} \exp(\kappa n_t \|\theta_{\mathbf{C}}\|_2) N^{n_t}. \end{aligned}$$

This results in the inequality

$$E_{f_{\theta^\gamma}}\left\{\|\phi(\mathbf{x})\|_2\right\} \leq \kappa n_t \exp(2\kappa n_t \gamma \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2) \cdot \exp(2\kappa n_t \|\theta_{\mathbf{C}}\|_2).$$

Thus, from (20), the following upper bound on the total variation norm results:

$$TV \leq \exp(2\kappa n_t \|\theta_{\mathbf{C}}\|_2) \left\{\exp(2\kappa n_t \|\theta_{\mathbf{C}} - \theta_{\mathbf{M}}\|_2) - 1\right\}.$$

APPENDIX C
SUFFICIENT STATISTICS OF $g(\mathbf{x}; \theta_{\mathbf{M}})$

Consider the distribution $g(\mathbf{x}; \mathbf{M}) \propto \exp\left\{\tilde{\mathbf{y}}^\ddagger \mathbf{M}\mathbf{x} + \mathbf{x}^\ddagger \mathbf{M}\tilde{\mathbf{y}} - \mathbf{x}^\ddagger \mathbf{M}\mathbf{x}\right\}$. With the parameterization $\theta_{\mathbf{M}} = \mathcal{F}(\mathbf{M})$, the term $(\tilde{\mathbf{y}}^\ddagger \mathbf{M}\mathbf{x} + \mathbf{x}^\ddagger \mathbf{M}\tilde{\mathbf{y}} - \mathbf{x}^\ddagger \mathbf{M}\mathbf{x})$ can be expressed as a real valued inner product of the form $\langle \theta_{\mathbf{M}}, \phi(\mathbf{x})\rangle$ by defining the vector $\phi(\mathbf{x})$ of length n_t^2 to be given in (21), where $l = \frac{n_t^2 - n_t}{2}$, and we have made use of the two l length vectors

$$\begin{aligned} \varkappa &= \left(\overbrace{1\,1\cdots 1}^{n_t-1}\ \overbrace{2\,2\cdots 2}^{n_t-2} \cdots (n_t-1) \right) \\ \varsigma &= (2\,3\cdots(n_t-1)\ 3\,4\cdots(n_t-2)\cdots(n_t-1)). \end{aligned}$$

ACKNOWLEDGMENT

The authors wish to thank the Directors of the Telecommunications Research Laboratory of Toshiba Research Europe Limited for their support and the anonymous reviewers, whose comments improved the presentation of this work. The first author also thanks Dr. Yugang Jia for the helpful discussions.

REFERENCES

[1] E. Telatar, "Capacity of the multi antenna Gaussian channel," *European Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.

[2] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V–BLAST: an architecture for realizing very high data rates over the rich scattering wireless channel," in *Proc. URSI International Symposium on Signals, Systems and Electronics*, Italy, Oct. 1998, pp. 295–300.

[3] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1639–1642, July 1999.

[4] B. M. Hochwald and S. ten Brink, "Achieving near–capacity on a multiple–antenna channel," *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 389–399, Mar. 2003.

[5] R. Wang and G. B. Giannakis, "Approaching MIMO channel capacity with soft detection based on hard sphere decoding," *IEEE Trans. Commun.*, vol. 54, no. 4, pp. 587–590, Apr. 2006.

[6] M. S. Yee, "Max–log–MAP sphere decoder," in *Proc. IEEE International conference on Acoustics, Speech and Signal Processing*, Philadelphia, Mar. 2005.

[7] J. Jalden and B. Ottersten, "On the complexity of sphere decoding in digital communications," *IEEE Trans. Signal Processing*, vol. 53, no. 4, pp. 1474–1484, Apr. 2005.

[8] Y. Jia, C. Andrieu, R. J. Piechocki, and M. Sandell, "Gaussian approximation based mixture reduction for near optimum detection in MIMO systems," *IEEE Commun. Lett.*, vol. 9, no. 11, pp. 997–999, Nov. 2005.

[9] Y. L. C. de Jong and T. J. Willink, "Iterative tree search detection for MIMO wireless systems," *IEEE Trans. Commun.*, vol. 53, no. 6, pp. 930–935, June 2005.

[10] H. Zhu, B. Farhang-Boroujeny, and R. Chen, "On performance of sphere decoding and Markov chain Monte Carlo detection methods," *IEEE Signal Processing Lett.*, pp. 669–672, Oct. 2005.

[11] O. Shental, A. J. Weiss, N. Shental, and Y. Weiss, "Generalized belief propagation receiver for near–opimal detection of two–dimensional channels with memory," in *Proc. IEEE Information Theory Workshop*, San Antonio, Texas, 2004, pp. 225–229.

[12] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, Probabilistic networks and expert systems, *Statistics for Engineering and Information Science*. New York: Springer–Verlag, 1999.

[13] M. I. Jordan, "Graphical models," *Statistical Science*, vol. 19, no. 1, pp. 140–155, Feb. 2004.

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

[15] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 1988.

[16] B. Frey, R. Koetter, and N. Petrovic, "Very loopy belief propagation for unwrapping phase images," in *Proc. Neural Information Processing Systems Conference*, 2002.

[17] B. J. Frey and D. J. C. MacKay, "A revolution: belief propagation in graphs with cycles," in *Proc. Neural Information Processing Systems Conference*, 1997.

[18] M. J. Wainwright, T. S. Jaakkola, and A. S. Wilsky, "Tree–based reparameterization framework for analysis of sum–product and related algorithms," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1120–1146, May 2003.

$$(\phi(\mathbf{x}))_i = \begin{cases} 2\Re\{(\tilde{\mathbf{y}})^*_{\varkappa_i}(\mathbf{x})_{\varsigma_i} + (\mathbf{x})^*_{\varkappa_i}(\tilde{\mathbf{y}})_{\varsigma_i} - (\mathbf{x})^*_{\varkappa_i}(\mathbf{x})_{\varsigma_i}\}\ ; i = 1, ..., l \\ (-2)\Im\{(\tilde{\mathbf{y}})^*_{\varkappa_{i-l}}(\mathbf{x})_{\varsigma_{i-l}} + (\mathbf{x})^*_{\varkappa_{i-l}}(\tilde{\mathbf{y}})_{\varsigma_{i-l}} - (\mathbf{x})^*_{\varkappa_{i-l}}(\mathbf{x})_{\varsigma_{i-l}}\}\ ; i = l+1, ..., 2l \\ 2\Re((\tilde{\mathbf{y}})^*_{i-2l}(\mathbf{x})_{i-2l}) - |(\mathbf{x})_{i-2l}|^2\ \ ; i = 2l+1, ..., n_t^2 \end{cases} \tag{21}$$

[19] C. M. Vithanage, C. Andrieu, R. J. Piechocki, and J. P. Coon, "Tree–based reparameterization for symbol detection in spatially multiplexed MIMO systems in frequency flat fading," in *Proc. Seventh IEEE Workshop in Signal Processing Advances in Wireless Communications*, Cannes, France, July 2006.

[20] B. Hassibi, "An efficient square–root algorithm for BLAST," in *Proc. IEEE International conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000.

[21] J. Benesty, Y. Huang, and J. Chen, "A fast recursive algorithm for optimum sequential signal detection in a BLAST system," *IEEE Trans. Signal Processing*, vol. 51, no. 7, pp. 1722–1730, July 2003.

[22] S. L. Lauritzen, *Graphical Models*. New York: Oxford University Press, 1996.

[23] M. J. Wainwright, "Estimating the "wrong" graphical model: benefits in the computation–limited setting," *J. Machine Learning Rresearch*, vol. 7, pp. 1829–1859, 2006.

[24] J. M. Mooij and H. J. Kappen, "Sufficient conditions for convergence of the sum–product algorithm," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4422–4437, Dec. 2007.

[25] A. T. Ihler, J. W. Fisher III, and A. S. Willsky, "Loopy belief propagation: Convergence and effects of message errors," *J. Machine Learning Research*, vol. 6, pp. 905–936, May 2005.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 2006.

[28] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. 20, pp. 284–287, Mar. 1974.

[29] C. M. Vithanage, C. Andrieu, and R. J. Piechocki, "Approximate inference in hidden Markov models using iterative active state selection," *IEEE Signal Processing Lett.*, vol. 13, no. 2, Feb. 2006.

[30] J. B. Lasserre, "A trace inequality for matrix product," *IEEE Trans. Automat. Contr.*, vol. 40, no. 8, pp. 1500–1501, Aug. 1995.