Massive MIMO Detection using MMSE-SIC and Expectation Propagation

Amen Memmi, McGill ID: 260755070, e-mail: amen.memmi@mail.mcgill.ca

***Abstract*—Massive multiple-input multiple-output (MIMO) systems constitute a promising technique for higher spectral efficiency of next-generation cellular telecommunication networks. However the increase in antennas number comes at the price of increased complexity of the MIMO detector, which grows rapidly with the number of antennas and becomes prohibitive for large scale systems. Hence the need for low complexity detection algorithms that ensure near optimal performances. In this work we elaborate the theory, implement and evaluate performances of two important MIMO estimation and detection algorithms. The first one is the well known MMSE algorithm supported with successive interference cancellation. The second is founded on another paradigm of graphical modeling: expectation propagation via message passing.**

I. Introduction

A. Framework

The recent rapid growth of wireless data traffic pushes for higher spectral efficiency in MIMO systems and standards. This is leading to the need to increases both the number of antennas and the alphabet size[1]. For example, the wireless LAN standard IEEE 802.11ac [2] enables up to 8 spatial streams through a MIMO channel and supports up to 256-ary QAM. Also, the 5G, next-generation cellular network, will probably be using carrier frequencies above 6 GHz [3], which allows an increase in the number of antennas with half-wavelength spacing in a small form factor. Thus massive MIMO systems are expected to deploy a very large number (from hundreds to thousands) of antennas (say N) at the base station (BS) to serve few tens of users (say M) simultaneously. However, this increase in dimensionality comes at the price of increased complexity of the MIMO detector, which grows rapidly with the number of antennas. For instance, Maximum likelihood detection (MLD), known as the optimal detection technique, has a computational complexity which increase exponentially with the number of antennas and thus quickly becomes prohibitive for large numbers.

This pushed the research to investigate and suggest few detection algorithms that come with lower complexity while enabling near optimal performances.

B. Literature survey

Detecting symbols becomes a particularly sensitive process when dealing with high-order high-dimensional systems. The MLD detector needs to explore all possible transmitted vectors and has an exponential complexity $|\mathcal{A}|^N$, where $|\mathcal{A}|$ denotes the size of the modulation constellation $\mathcal{A}$. It constitutes thus a bottleneck for high-order high-dimensional MIMO systems. Sphere decoding (SD) methods try to replicate the performance of ML by searching in a sub-space of $|\mathcal{A}|^N$, rather than $|\mathcal{A}|^N$, (Since the full tree search has the same complexity as ML detection,) One advantage of sphere decoding is that it can, by choosing an appropriate value of radius or list size, provide a trade-off between performance and complexity [4]. However, the dimension of this subspace 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. Linear detectors (LD), such as the minimum-mean-squared error (MMSE) can achieve near-optimal bit error rate (BER) performance [5]. They have been widely adopted because of their polynomial-time complexity: an N x N matrix inversion is the leading computational complexity term ($O(N^3)$). MMSE detection performance can be significantly improved in large MIMO systems following a divide-and-conquer approach, namely successive interference cancellation (MMSE-SIC) [6], at a higher computational complexity but still $O(N^3)$.

When powerful error-correcting codes approaching to the channel capacity such as low-density parity check (LDPC) codes or turbo codes are utilized in transmission, their channel decoders require soft-outputs of the detector, i.e., marginal probabilities. Belief propagation (BP) is a practical and powerful way to calculate marginal probabilities; this efficient calculation is performed via message-passing on a factor graph composed of variable and observation nodes. In BP algorithm, messages (beliefs) are iteratively sent from one node to neighboring nodes. BP-based detection has been studied in the literature where factor graphs are defined by channel matrices between transmitting and receiving antennas. When applying the BP algorithm [7] or the sum-product algorithm [8] to such graphs, the complexity is as high as the ML or MAP detector. This is mainly due to the metric computation and the marginalization operation required for the message update at the observation nodes. To reduce the computational complexity, approximations of the message passed and/or the graph (pruning) have been investigated in the literature. A Low-Complexity Gaussian message passing iterative detector for massive MU-MIMO systems was suggested in [9]. Authors of [10] suggested expectation propagation detection for high-order high-dimensional MIMO systems. In [11], the complexity of approximate message passing algorithm has been reduced through expectation propagation, first order approximation and central limit theorem. Complexity of the order of $O(MN)$ has been achieved for near optimal BER performance after few iterations.

II. System Model and Problem Formulation

We consider an $N \times M$ massive MIMO system where N and M are the numbers of transmitters and receivers, respectfully. Let $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N]$ be the $N \times 1$ vector of independent and identically distributed (i.i.d.) transmitted complex symbols drawn from a constellation set $\mathcal{A}$. The vector received after demodulation and sampling at the receiver can be written as follows:

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{H}}\widetilde{\mathbf{x}} + \widetilde{\mathbf{n}}, \tag{1}$$

where $\widetilde{\mathbf{n}}$ is the additive white Gaussian noise (AWGN) with elements assumed to be i.i.d complex Gaussian random variables with mean zero and variance σ^2, i.e. $\sim \mathcal{CN}(0, \sigma_n^2)$. The entries of the channel gain matrix $\widetilde{\mathbf{H}}$ are i.i.d. drawn according to a proper complex zero-mean unit-variance Gaussian distribution $\mathcal{CN}(0,1)$. Instead of complex-valued variables used in signal processing for communications, real-valued random variables are typically used to present inference in graphical model. For this purpose, we first reformulate the complex-valued MIMO system into a real-valued one, before presenting the Gaussian approximated message passing (GAMP) detector. Without loss of generality, the system model in 1 can be translated into an equivalent double-sized real-valued representation as follows:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{2}$$

where:

$$\mathbf{y} = \begin{bmatrix} \Re(\widetilde{\mathbf{y}}) \\ \Im(\widetilde{\mathbf{y}}) \end{bmatrix}_{2N\times 1}, \mathbf{x} = \begin{bmatrix} \Re(\widetilde{\mathbf{x}}) \\ \Im(\widetilde{\mathbf{x}}) \end{bmatrix}_{2N\times 1}, \mathbf{n} = \begin{bmatrix} \Re(\widetilde{\mathbf{n}}) \\ \Im(\widetilde{\mathbf{n}}) \end{bmatrix}_{2N\times 1}$$

and

$$\mathbf{H} = \begin{bmatrix} \Re(\widetilde{\mathbf{H}}) & \Im(\widetilde{\mathbf{H}})^T \\ -\Im(\widetilde{\mathbf{H}})^T & \Re(\widetilde{\mathbf{H}})^T \end{bmatrix}_{2N\times 2K}. \tag{3}$$

$\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts of $(\cdot)$, respectively. The task of multi-user detection at the BS is to estimate the transmitted signal vector x from the received signal vector y. Noting that the channel matrix $\mathbf{H}$ can be usually obtained by time-domain and/or frequency-domain training pilots [12], we assume that the BS knows the Channel State Information (CSI).

III. MIMO detection

A. MMSE Detector and Successive Interference Cancellation

It is well known that MMSE detection is optimal under MSE measure when the sources are Gaussian distributed [13].The MMSE detector [15] first proceeds by computing

$$\mu_{MMSE} = \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma_n^2}{E_s}\mathbf{I}\right)^{-1}\mathbf{H}^T\mathbf{y}, \tag{4}$$

and it then performs a component-wise hard decision by projecting each component of μ_{MMSE} into the corresponding constellation:

$$\hat{x}_{k,\text{MMSE}} = \arg\min_{x_k \in \mathcal{A}} \left|x_k - \mu_{k,\text{MMSE}}\right|^2 \tag{5}$$

The computational complexity here is dominated by the matrix inversion in (4), given by $O(n^3)$ [14]. For large MIMO systems (and especially with high-order constellations), the MMSE detector provides poor performance. This could be significantly improved by successive interference cancellation, yielding the so-called MMSE-SIC [15], [16]. Iteratively, we only decide over the component with the smallest diagonal element in the covariance matrix in (4) and remove its effect in the channel output. After each iteration, we update the received vector:

$$\mathbf{y}^{(\ell+1)} = \mathbf{y}^{(\ell)} - \mathbf{h}_k\hat{x}_{k,\text{MMSE}}^{(\ell)} \tag{6}$$

where $\mathbf{h}_k$ represents the k^{th} column of $\mathbf{H}$ and its effect is removed from the channel matrix given the current decision, i.e. $\hat{x}_{k,\text{MMSE}}^{(\ell)}$, and we drop $\mathbf{h}_k$ 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.

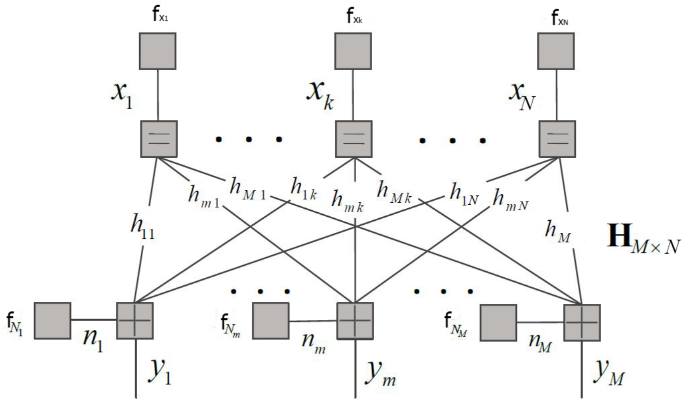

Fig. 1: Factor graph representatiion of Gaussian approximated message passing iterative detection for MIMO systems.

B. Graphical model based detection: Gaussian approximate message passing

1) Factor graph and algorithm: . In this work, we consider Gaussian approximate message passing based on a pairwise factor graph for the MIMO system. Fig. 1 gives the factor graph of MIMO system. The channel parameter from transmitter k to receiver m is h_{mk}, the distribution constraints of each source and noise are denoted by f_{x_k} and f_{N_m}, respectively. The process is very similar to the Belief Propagation (BP) decoding process of LDPC code presented in book [18]. The differences are the nature of the different passed messages on each edge (the mean and variance of a Gaussian distribution) and the different message update rules. The basic idea that allows the use of such factor graph for detection is the assumption we can model the posterior distributions as a product of factors using Gaussian random variable proprieties.

2) Expectation propagation: Expectation Propagation is a technique widely used in machine learning applications based on Bayesian networks. It mainly serves for approximating posterior beliefs with exponential family distributions. To devolop how EP works, let's consider:

$$p(\mathbf{x}) \propto f(\mathbf{x}) \prod_{i=1}^{I} t_i(\mathbf{x}), \tag{7}$$

where $f(\mathbf{x})$ is drawn from an exponential family $\mathcal{F}$ defined by sufficient statistics $[\Phi_1(x), \Phi_2(x), \ldots, \Phi_N(x)]$ and t_i are nonnegative factors. Depending on t_i, performing inference over the distribution $p(\mathbf{x})$ could be analytically intractable or prohibitively complex in general case. Here, EP provides a solution to construct a tractable approximation to $p(\mathbf{x})$ by a distribution $q(\mathbf{x})$ from the same $\mathcal{F}$. This basically done through *moment matching* of the two distributions:

$$\mathbb{E}_{q(\mathbf{x})}\left[\phi_j(\mathbf{x})\right] = \mathbb{E}_{p(\mathbf{x})}\left[\phi_j(\mathbf{x})\right], \quad j = 1, \ldots, S. \tag{8}$$

where $\mathbb{E}_{q(\mathbf{x})}$ denotes expectation with respect to the distribution $q(\mathbf{x})$. Given both $p(\mathbf{x})$ and $q(\mathbf{x})$ are defined over the same support space and measure, moment matching is equivalent to finding the minimizer of the Kullback-Leibler divergence between the two distributions, i.e.

$$q(\mathbf{x}) = \arg\min_{q'(\mathbf{x})\in\mathcal{F}} D_{\mathrm{KL}}\left(p(\mathbf{x}) \| q'(\mathbf{x})\right) 0 \tag{9}$$

3) Expectation propagation MIMO Detector: The aim is to compute the posteriori distribution to be able to get the extrinsic log likelihood ratios (LLRs) of the received symbols based on which the detection would be done. The proposition here is to approximate this symbol posterior distribution $p(\mathbf{x}|\mathbf{y})$ by a Gaussian approximation $q_{\mathrm{EP}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} : \mu_{\mathrm{EP}}, \Sigma_{\mathrm{EP}})$ that is optimized using the EP framework. Thus optimally, EP solution would be

$$\mu_{\mathrm{EP}} = \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[\mathbf{x}], \tag{10}$$

$$\Sigma_{\mathrm{EP}} = \mathrm{CoVar}_{p(\mathbf{x}|\mathbf{y})}[\mathbf{x}]. \tag{11}$$

Complexity of direct computation of the moments is $|\mathcal{A}|^N$. Using EP to iteratively approximate them is done at a polynomial complexity with N. After iterations' end, the EP detector computes the hard output $\hat{\mathbf{x}}_{\mathrm{EP}}$ by independently deciding on each component, for $i = 1, 2, \ldots, 2N$, $\hat{x}_{i,\mathrm{EP}} = \arg\min_{x_i\in\mathcal{A}} \left|x_i - \mu_{i,\mathrm{EP}}\right|^2$.
Given the system model, the posterior probability of the transmitted symbol vector Formula has the following expression:the factorization of the posterior in (3), we replace each one of the non-Gaussian factors by an unnormalized Gaussian:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p\left(\mathbf{y}|\mathbf{u}\right)p(\mathbf{x})}{p(\mathbf{y})} \propto \mathcal{N}(\mathbf{y} : \mathbf{Hu}, \sigma_n^2\mathbf{I}) \prod_{i=1}^{n} \mathbb{I}_{x_i\in\mathcal{A}} \tag{12}$$

where $\mathbb{I}_{x_i\in\mathcal{A}}$ is the indicator function that takes value one if $x_i \in \mathcal{A}$ and zero otherwise. We replace and approximate these ”bad" indicator functions by unnormalized Gaussian ones:

$$q(\mathbf{x}) \propto \mathcal{N}\left(\mathbf{y} : \mathbf{Hx}, \sigma_n^2\mathbf{I}\right) \prod_{i=1}^{2n} \mathrm{e}^{\gamma_i x_i - \frac{1}{2}\tau_i x_i^2} \tag{13}$$

where γ and τ are real constants. Mean vector μ and covariance matrix Σ relative to $q(\mathbf{x})$ are then

$$\Sigma = \left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{H} + \mathrm{diag}(\tau)\right)^{-1}, \tag{14}$$

$$\mu = \Sigma\left(\sigma_n^{-2}\mathbf{H}^\top\mathbf{y} + \gamma\right) \tag{15}$$

10 and 11 are approximated through EP by recursively updating $(\gamma_i, \Sigma_i)_i$. We initialize $\gamma_i = 0$ and $\tau_i = E - s^{-1})$ for all i, which is exactly the MMSE solution. At each EP iteration ℓ, all pairs $(\gamma_i^{(\ell+1)}, \Sigma_i^{(\ell+1)})$ for $i = 1, 2, \ldots, 2N$ are updated in parallel. Using the i^{th} marginal of the distribution $q(\mathbf{x})$, namely $q_i^{(\ell)}(x_i) = \mathcal{N}(x_i : \mu_{\mathrm{i}}^{(\ell)}, \sigma_{\mathrm{i}}^{2(\ell)})$, the update rules are the following:

- $q^{(\ell)\backslash i}(x_i) = \frac{q^{(\ell)}(x_i)}{\exp\left(\gamma_i^{(\ell)} u_i - \frac{1}{2}\tau_i^{(\ell)} x_i^2\right)} \sim \mathcal{N}\left(x_i : t_i^{(\ell)}, h_i^{2(\ell)}\right)$, where $h_i^{(2\ell)} = \frac{\sigma_n^{-2}}{(1-\sigma_n^2 \tau_i^{\ell})}$ and $t_i^{(\ell)} = h_i^{2(\ell)} \left(\frac{\mu_i^{(\ell)}}{\sigma_i^{2(\ell)}} - \gamma_i^{(\ell)} \right)$.
- $\hat{p}^{(\ell)}(x_i) \propto q^{(\ell)\backslash i}(x_i) \mathbb{I}_{x_i \in \mathcal{A}_i}$.
- Update the pair $(\gamma_i^{(\ell+1)}, \Sigma_i^{(\ell+1)})$ so that the following unnormalized Gaussian distribution $\tau_i^{(\ell+1)} = \frac{1}{\sigma_{p_i}^{2(\ell)}} - \frac{1}{h_i^{2(\ell)}} \gamma_i^{(\ell+1)} = \frac{\mu_{p_i}^{(\ell)}}{\sigma_{p_i}^{2(\ell)}} - \frac{t_i^{(\ell)}}{h_i^{2(\ell)}}$.

C. *Results & discussion*

1) Complexity analysis: Despite MMSE-SIC requires to perform N times a MMSE matrix inversion similar to that of in (4), which would rise the algorithm's complexity to $O(n^4)$, some works like [17] proved that it can be lowered down to $O(n^2 m)$, and thus the same as regular MMSE. This is done through efficiently computing the matrix inversion at each iteration by making only a rank-one update given the inverted matrix from the previous iteration. On the other hand, EP requires computationally, $8NM$ multiplications for each iteration. Therefore, the complexity is lower and evaluated at $O(NMn_{iter})$, where n_{iter} is the number of iterations. GAMP using EP is able to significantly reduce computational complexity, and we will see in the following subsection that it preserve performances and even improve them.

2) Precision performance analysis: In this subsection, we illustrate the performance of the MMSE-SIC and EP algorithms for MIMO detection. We have averaged our results for 10000 realizations of the Rayleigh fading MIMO channel matrix. We consider diffrent scenarios of increasing dimensions.The detectors performance is shown in terms of the bit error rate (BER) as a function of the SNR. First, we consider small scale configuration 2 × 16 to evaluate performance of MMSE detector. We also implemented another less complex GAMP (calling it just AMP) that is able to match the performance of MMSE detector in 6 iterations, as it can be seen in figure 2. The advantage of AMP is its simplicity since it can solve the above problem and only requires three lines of code and costs only $O(MN{*}n_{iter})$.

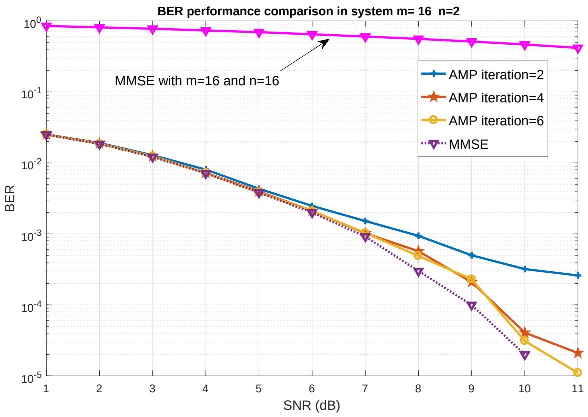

Fig. 2: BER performance versus SNR with multiple iterations in a 2 × 16 MIMO system with QPSK and MMSE detection .

The same figure 2 proves also that MMSE becomes inefficient for higher number of transmitting antennas: BER is higher than 0.5 even for high SNR.

Second, we consider a large-scale MIMO system with $M = N = 64$ users and antennas at the receiver. We make comparisons between the proposed EP message passing algoriths and the MMSE-SIC algorithm. The BER performance of the considered algorithms is shown in Fig. 3. For QPSK transmission, EP achieves the same performance as that of the MMSE-SIC algorithm with 6 iterations, although the latter outperforms them with 3 iterations. Figure 4 gives as an idea about performances for 16-QAM. MMSE-SIC algorithm outperforms the proposed EP algorithm with 6 iterations, but with more iterations, e.g., 12 iterations, EP reaches the MBF, matched filter bound (obtained by exactly removing all the interferences is used as a lower bound on the BER) and outperform the MMSE-SIC algorithm about 0.7 dB at BER= 10^{-4}. The performance improvement may be due to the different way of forming the interference estimate. Figure 4 mainly presents the BER performance of the EP algorithm and the MMSE-SIC algorithm with the number of iterations. It can be see that 5 iterations and 9 iterations are enough for the EP algorithm to achieve the MFBs of QPSK and 16QAM at , respectively. While the EP algorithm can uniformly improve the performance with iterations and achieve the MFB in the low region, the MMSE-SIC algorithm cannot achieve the MFB in the low region with iterations.

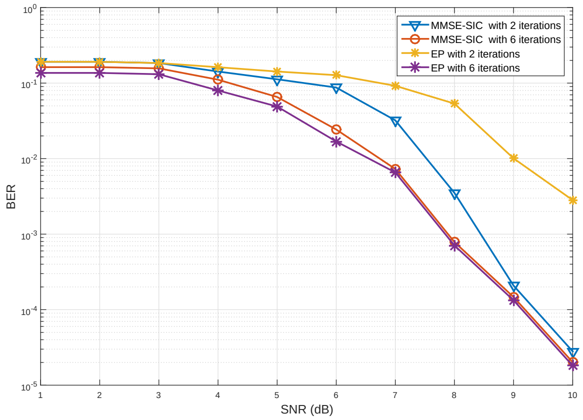

Fig. 3: BER performance versus SNR with multiple iterations in a 64 × 64 MIMO system with QPSK and 16QAM. (a) EP; (b) MMSE-SIC .

D. Conclusion

E. Conclusion

For the detection of large-scale MIMO systems, we have evaluated MMSE, AMP, MMSE-SIC and EP algorithm. While MMSE showed inefficiency when MIMO dimension increases, both MMSE-SIC and managed to perform well in terms of bit error rate. The two had comaparable BER performances with EP performing better with more iterations and having the edge of a smaller complexity of $O(MN * n_{iter})$ compared to $O(N^3)$ for MMSE-SIC. Further approximations could be considered to extend EP algorithm like first order approximation of the messages or using the central limit theorem for higher number of antennas. Those ideas could provide the possibility to operate with less computational complexity offering desirable tradeoff between performance and complexity.

References

[1] Shaoshi Yang, Lajos Hanzo, “Fifty Years of MIMO Detection: The Road to Large-Scale MIMOs", Communications Surveys & Tutorials IEEE, vol. 17, pp. 1941-1988, 2015.

[2] “IEEE.standard.02.11ac-2103", 2013, [online] Available: https://standards.ieee.org/findstds/standard/802.11ac-2013.html.

[3] Tentative 3GPP Timeline for 5G, Mar. 2015, [online] Available: http://www.3gpp.org/news-events/3gpp-news/1674-timeline_5g.

[4] C. Studer, A. Burg, and H. Bolcskei, “Soft-output sphere decoding: Algorithms and VLSI implementation," IEEE Jour. Select. Areas in Comm., vol. 26, no. 2, pp. 290-300, Feb. 2008.

[5] 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.

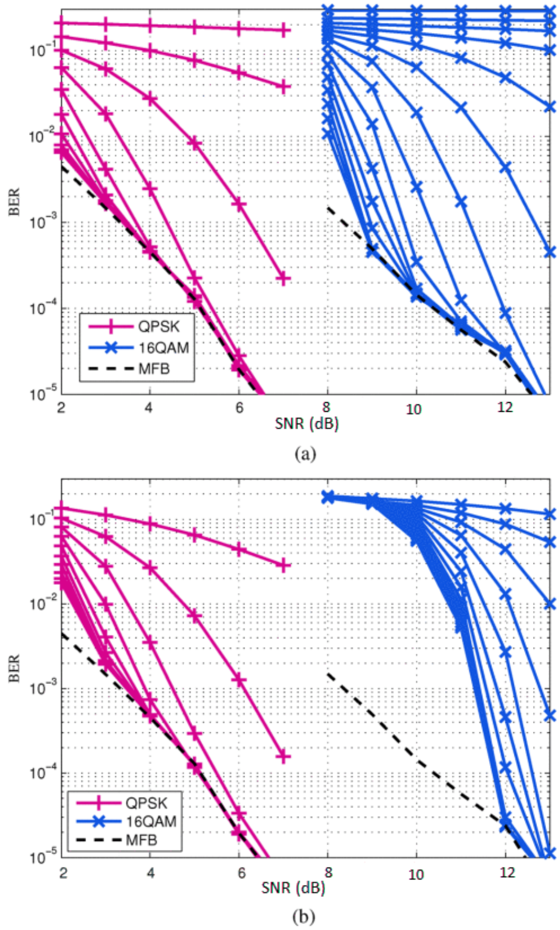

Fig. 4: BER performance versus SNR with multiple iterations in a 64 × 64 MIMO system with QPSK and 16QAM. (a) EP; (b) MMSE-SIC .

[6] 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.

[7] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann, 1987.

[8] F. Kschischang, B. Frey, and H. A. Loeliger, “Factor graphs and the sum-product algorithm," IEEE Trans. Info. Th., vol. 47, no. 2, pp. 498-519, Feb. 2001.

[9] L. Liu, C. Yuen, Y. L. Guan, Y. Li, Y. Su, “A low-complexity Gaussian message passing iterative detector for massive MU-MIMO systems", Proc. IEEE Int. Conf. Inf. Commun. Signal Process. (ICICS), pp. 1-5, Dec. 2015.

[10] J. Céspedes, P. M. Olmos, M. Sanchez-Fernandez, F. Perez-Cruz, “Expectation propagation detection for high-order high-dimensional MIMO systems", IEEE Trans. Commun., vol. 62, no. 8, pp. 2840-2849, Aug. 2014.

[11] S. Wu, L. Kuang, Z. Ni, J. Lu, D. D. Huang, Q. Guo, “Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing", IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 902-915, Oct. 2014.

[12] L. Dai, Z. Wang, and Z. Yang, “Spectrally efficient time-

frequency training OFDM for mobile large-scale MIMO systems,âĂİ IEEE J. Sel. Areas Commun., vol. 31, no. 2, pp. 251-263, Feb. 2013.

[13] S. Verdu, *Multiuser Detection*. Cambridge, UK: Cambridge University Press, 1998.

[14] 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.

[15] 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.

[16] 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.

[17] 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, 2008

[18] R. E. William and S. Lin, *Channel Codes Classical and Modern*, Cambridge University Press, 2009.