A Novel MCMC Algorithm for Near-Optimal Detection in Large-Scale Uplink Mulituser MIMO Systems

Tanumay Datta, N. Ashok Kumar, A. Chockalingam, and B. Sundar Rajan
Department of ECE, Indian Institute of Science, Bangalore-560012, India

***Abstract*—In this paper, we propose a low-complexity algorithm based on Markov chain Monte Carlo (MCMC) technique for signal detection on the uplink in large scale multiuser multiple input multiple output (MIMO) systems with tens to hundreds of antennas at the base station (BS) and similar number of uplink users. The algorithm employs a randomized sampling method (which makes a probabilistic choice between Gibbs sampling and random sampling in each iteration) for detection. The proposed algorithm alleviates the stalling problem encountered at high SNRs in conventional MCMC algorithm and achieves near-optimal performance in large systems with M-QAM. A novel ingredient in the algorithm that is responsible for achieving near-optimal performance at low complexities is the joint use of a *randomized MCMC (R-MCMC) strategy* coupled with a *multiple restart strategy* with an efficient restart criterion. Near-optimal detection performance is demonstrated for large number of BS antennas and users (e.g., 64, 128, 256 BS antennas/users).**

Index Terms *– Large-scale multiuser MIMO, Markov chain Monte Carlo technique, Gibbs sampling, detection, stalling problem, randomized sampling, multiple restarts.*

I. Introduction

Capacity of multiple-input multiple-output (MIMO) wireless channels is known to increase linearly with the minimum of the number of transmit and receive antennas [1]- [5]. Large-scale MIMO systems with tens to hundreds of antennas have attracted much interest recently [6]- [17]. The motivation to consider such large-scale MIMO systems is the potential to practically realize the theoretically predicted benefits of MIMO, in terms of very high spectral efficiencies/sum rates, increased reliability and power efficiency, through the exploitation of large spatial dimensions. Use of large number of antennas is getting recognized to be a good approach to fulfill the increased throughput requirements in future wireless systems. Particularly, large multiuser MIMO wireless systems where the base station (BS) has tens to hundreds of antennas and the users have one or more antennas are widely being investigated [9], [12]- [17]. Communications on the uplink [13], [16] as well as on the downlink [9], [14], [15] in such large systems are of interest. Key issues in large multiuser MIMO systems on the downlink include low complexity precoding strategies and pilot contamination problem encountered in using non-orthogonal pilot sequences for channel estimation in multi-cell scenarios [14]. In large multiuser MIMO systems on the uplink, users with one or more antennas transmit simultaneously to the BS having large number of antennas, and their signals are separated at the BS using their spatial signatures towards the BS. Sophisticated signal processing is required at the BS receiver to extract the signal of each user from the aggregate received signal [4]. Use of large number of BS antennas has been shown to improve the power efficiency of uplink transmissions in multiuser MIMO using linear receivers at the BS [16]. Linear receivers including matched filter (MF) and minimum mean square error (MMSE) receivers are shown to be attractive for very large number of BS antennas [13]. Our focus in this paper is on achieving near-optimal detection performance at the BS in large multiuser MIMO systems on the uplink at low complexities. The approach we adopt is the *Markov chain Monte Carlo (MCMC)* approach.

The uplink multiuser MIMO architecture can be viewed as a point-to-point MIMO system with co-located transmit antennas with adequate separation between them (so that there is no or negligible spatial correlation among them), and no cooperation among these transmit antennas [4]. Because of this, receiver algorithms for point-to-point MIMO systems are applicable for receiving uplink multiuser MIMO signals at the BS receiver. Recently, there has been encouraging progress in the development of low-complexity near-optimal MIMO receiver algorithms that can scale well for large dimensions [8], [10], [18]- [25]. These algorithms are based on techniques from local neighborhood search including tabu search [8], [10], [18]- [21], probabilistic data association [22], and message passing on graphical models including factor graphs and Markov random fields [23], [24], [25].

Another interesting class of low-complexity algorithms reported in the context of CDMA and MIMO detection is based on Markov chain Monte Carlo (MCMC) simulation techniques [26]- [33]. MCMC techniques are computational techniques that make use of random numbers [34]. MCMC methods have their roots in the Metropolis algorithm, an attempt by physicists to compute complex integrals by expressing them as expectations for some distribution and then estimating this expectation by drawing samples from that distribution [35], [36]. In MCMC methods, statistical inferences are developed by simulating the underlying processes through Markov chains. By doing so, it becomes possible to reduce exponential detection complexity to linear/polynomial complexities.

An issue with conventional MCMC based detection, how-

ever, is the *stalling problem*, due to which performance degrades at high SNRs [27]. Stalling problem arises because transitions from some states to other states in a Markov chain can occur with very low probability [27].

Our first contribution in this paper is that we propose an MCMC based detection algorithm that alleviates the stalling problem encountered in conventional MCMC and achieves near-optimal performance in large systems. A key idea that is instrumental in alleviating the stalling problem is a *randomized sampling strategy* that makes a probabilistic choice between Gibbs sampling and random sampling in each iteration. An efficient stopping criterion aids complexity reduction. This randomized sampling strategy, referred to as 'randomized MCMC (R-MCMC) strategy', is shown to achieve near-optimal performance in large multiuser MIMO systems with 16 to 256 BS antennas and same number of uplink users for 4-QAM [37]. However, we find that this randomized sampling strategy alone is not adequate to achieve near-optimal performance at low complexities for higher-order QAM (e.g., 16-QAM, 64-QAM). We show that near-optimal performance is achieved in higher-order QAM also if a *multiple restart strategy* is performed in conjunction with R-MCMC. We refer to this strategy as 'R-MCMC with restarts' (R-MCMC-R) strategy. Here again, an efficient restart criterion aids complexity reduction. The *joint use* of both randomized sampling as well as multiple restart strategies is found to be crucial in achieving near-optimal performance for higher-order QAM in large systems. To our knowledge, the closeness to optimal performance achieved by the proposed R-MCMC-R algorithm for tens to hundreds of BS antennas/users with higher-order QAM has not been reported so far using other MCMC based algorithms in the literature.

The rest of the paper is organized as follows. The uplink multiuser MIMO system model is presented in Section II. The proposed R-MCMC algorithm without multiple restarts and its performance/complexity in 4-QAM are presented in Section III. The proposed R-MCMC algorithm that uses multiple restarts (R-MCMC-R) and its performance/complexity in higher-order QAM are presented in Section IV. Conclusions are presented in Section V.

II. System Model

Consider a large-scale multiuser MIMO system on the uplink consisting of a BS with N receive antennas and K homogeneous uplink users with one transmit antenna each, $K \leq N$ (Fig. 1). Extension of the wotk to a system model with non-homogeneous users where different users can have different number of transmit antennas is straightforward. N and K are in the range of tens to hundreds. All users transmit symbols from a modulation alphabet $\mathbb{B}$. It is assumed that synchronization and sampling procedures have been carried out, and that the sampled base band signals are available at the BS receiver. Let $x_k \in \mathbb{B}$ denote the transmitted symbol from user k. Let $\mathbf{x}_c = [x_1, x_2, \cdots, x_K]^T$ denote the vector comprising of the symbols transmitted simultaneously by all users in one channel use. Let $\mathbf{H}_c \in \mathbb{C}^{N \times K}$, given

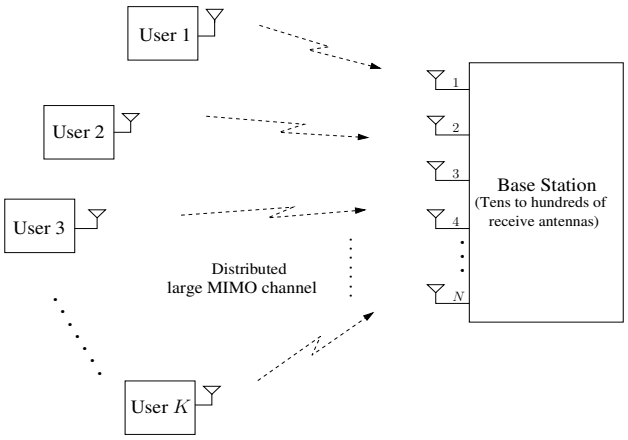


Fig. 1. Large-scale multiuser MIMO system on the uplink.

by $\mathbf{H}_c = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_K]$, denote the channel gain matrix, where $\mathbf{h}_k = [h_{1k}, h_{2k}, \cdots, h_{Nk}]^T$ is the channel gain vector from user k to BS, and h_{jk} denotes the channel gain from kth user to jth receive antenna at the BS. Assuming rich scattering and adequate spatial separation between users and BS antenna elements, $h_{jk}, \forall j$ are assumed to be independent Gaussian with zero mean and σ_k^2 variance such that $\sum_k \sigma_k^2 = K$. σ_k^2 models the imbalance in received powers from different users, and $\sigma_k^2 = 1$ corresponds to the perfect power control scenario. This channel gain model amounts to assuming that the multipath fading between a user and BS is frequency non-selective. Now, the received signal vector at the BS in a channel use, denoted by $\mathbf{y}_c \in \mathbb{C}^N$, can be written as

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

where $\mathbf{n}_c$ is the noise vector whose entries are are modeled as i.i.d. $\mathbb{C}\mathcal{N}(0, \sigma^2)$. We will work with the real-valued system model corresponding to (1), given by

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

where $\mathbf{x}_r \in \mathbb{R}^{2K}$, $\mathbf{H}_r \in \mathbb{R}^{2N \times 2K}$, $\mathbf{y}_r \in \mathbb{R}^{2N}$, $\mathbf{n}_r \in \mathbb{R}^{2N}$ given by

$$\mathbf{H}_r = \begin{bmatrix} \Re(\mathbf{H}_c) & -\Im(\mathbf{H}_c) \\ \Im(\mathbf{H}_c) & \Re(\mathbf{H}_c) \end{bmatrix}, \quad \mathbf{y}_r = \begin{bmatrix} \Re(\mathbf{y}_c) \\ \Im(\mathbf{y}_c) \end{bmatrix},$$

$$\mathbf{x}_r = \begin{bmatrix} \Re(\mathbf{x}_c) \\ \Im(\mathbf{x}_c) \end{bmatrix}, \quad \mathbf{n}_r = \begin{bmatrix} \Re(\mathbf{n}_c) \\ \Im(\mathbf{n}_c) \end{bmatrix}. \tag{3}$$

Dropping the subscript r in (2) for notational simplicity, the real-valued system model is written as

$$\mathbf{y} \quad = \quad \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{4}$$

For a QAM alphabet $\mathbb{B}$, the elements of $\mathbf{x}$ will take values from the underlying PAM alphabet $\mathbb{A}$, i.e., $\mathbf{x} \in \mathbb{A}^{2K}$. The symbols from all the users are jointly detected at the BS. The maximum likelihood (ML) decision rule is given by

$$\mathbf{x}_{ML} \quad = \quad \arg \min_{\mathbf{x} \in \mathbb{A}^{2K}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \quad = \quad \arg \min_{\mathbf{x} \in \mathbb{A}^{2K}} f(\mathbf{x}), \tag{5}$$

where $f(\mathbf{x}) \stackrel{\triangle}{=} \mathbf{x}^T\mathbf{H}^T\mathbf{H}\mathbf{x} - 2\mathbf{y}^T\mathbf{H}\mathbf{x}$ is the ML cost. While the ML detector in (5) is exponentially complex in K (which is prohibitive for large K), the MCMC based algorithm we propose in the next section has a per-symbol complexity that is quadratic in K and it achieves near-ML performance as well.

III. Proposed Randomized-MCMC Algorithm for Detection

The ML detection problem in (5) can be solved by using MCMC simulations [34]. We consider Gibbs sampler, which is an MCMC method used for sampling from distributions of multiple dimensions. In the context of MIMO detection, the joint probability distribution of interest is given by

$$p(x_1, \cdots, x_{2K}|\mathbf{y}, \mathbf{H}) \quad \propto \quad \exp\Big(-\frac{\| \mathbf{y} - \mathbf{H}\mathbf{x} \|^2}{\sigma^2}\Big). \quad (6)$$

We assume perfect knowledge of channel gain matrix $\mathbf{H}$ at the BS receiver.

A. Conventional MCMC algorithm

In conventional Gibbs sampling based detection, referred to as conventional MCMC algorithm, the algorithm starts with an initial symbol vector, denoted by $\mathbf{x}^{(t=0)}$. In each iteration of the algorithm, an updated symbol vector is obtained by sampling from distributions as follows:

$$\begin{aligned}
x_1^{(t+1)} &\sim p(x_1|x_2^{(t)}, x_3^{(t)}, \cdots, x_{2K}^{(t)}),\\
x_2^{(t+1)} &\sim p(x_2|x_1^{(t+1)}, x_3^{(t)}, \cdots, x_{2K}^{(t)}),\\
x_3^{(t+1)} &\sim p(x_3|x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \cdots, x_{2K}^{(t)}),\\
&\vdots\\
x_{2K}^{(t+1)} &\sim p(x_{2K}|x_1^{(t+1)}, x_2^{(t+1)}, \cdots, x_{2K-1}^{(t+1)}). \quad (7)
\end{aligned}$$

The detected symbol vector in a given iteration is chosen to be that symbol vector which has the least ML cost in all the iterations up to that iteration.

Another MCMC algorithm that uses a temperature parameter α and the following joint distribution is presented in [33]:

$$p(x_1, \cdots, x_{2K}|\mathbf{y}, \mathbf{H}) \quad \propto \quad \exp\Big(-\frac{\| \mathbf{y} - \mathbf{H}\mathbf{x} \|^2}{\alpha^2\sigma^2}\Big). \quad (8)$$

The algorithm uses a fixed value of α in all the iterations, with the property that after the Markov chain is mixed, the probability of encountering the optimal solution is only polynomially small (not exponentially small). This algorithm and the conventional MCMC algorithm (which is a special case of $\alpha = 1$) face stalling problem at high SNRs; a problem in which the BER performance gets worse at high SNRs [27].

B. Proposed R-MCMC algorithm

It is noted that the stalling problem occurs due to MCMC iterations getting trapped in poor local solutions, beyond which the ML cost does not improve with increasing iterations for a long time. Motivated by this observation, we propose a simple, yet effective, randomization strategy to avoid such traps. The key idea behind the proposed randomized MCMC (R-MCMC) approach is that, in each iteration, instead of updating $x_i^{(t)}$'s as per the update rule in (7) with probability 1 as done in conventional MCMC, we update them as per (7) with probability $(1 - q_i)$ and use a different update rule with probability $q_i = \frac{1}{2K}$. The different update rule is as follows. Generate $|\mathbb{A}|$ probability values from uniform distribution as

$$p(x_i^{(t)} = j) \sim U[0, 1], \quad \forall j \in \mathbb{A}$$

such that $\sum_{j=1}^{|\mathbb{A}|} p(x_i^{(t)} = j) = 1$, and sample $x_i^{(t)}$ from this generated pmf.

1) Proposed stopping criterion: A suitable termination criterion is needed to stop the algorithm. A simple strategy is to terminate the algorithm after a fixed number of iterations. But a fixed value of number of iterations may not be appropriate for all scenarios. Fixing a large value for the number of iterations can yield good performance, but the complexity increases with the number of iterations. To address this issue, we develop a dynamic stopping criterion that yields good performance without unduly increasing the complexity. The criterion works as follows. A stalling is said to have occurred if the ML cost remains unchanged in two consecutive iterations. Once such a stalling is identified, the algorithm generates a positive integer Θ_s (referred to as the *stalling limit*), and the iterations are allowed to continue in stalling mode (i.e., without ML cost change) up to a maximum of Θ_s iterations from the occurrence of stalling. If a lower ML cost is encountered before Θ_s iterations, the algorithm proceeds with the newly found lower ML cost; else, the algorithm terminates. If termination does not happen through stalling limit as above, the algorithm terminates on completing a maximum number of iterations, MAX-ITER.

The algorithm chooses the value of Θ_s depending on the quality of the stalled ML cost, as follows. A large value for Θ_s is preferred if the quality of the stalled ML cost is poor, because of the available potential for improvement from a poor stalled solution. On the other hand, if the stalled ML cost quality is already good, then a small value of Θ_s is preferred. The quality of a stalled solution is determined in terms of closeness of the stalled ML cost to a value obtained using the statistics (mean and variance) of the ML cost for the case when $\mathbf{x}$ is detected error-free. Note that when $\mathbf{x}$ is detected error-free, the corresponding ML cost is nothing but $\|\mathbf{n}\|^2$, which is Chi-squared distributed with $2N$ degrees of freedom with mean $N\sigma^2$ and variance $N\sigma^4$. We define the quality metric to be the difference between the ML cost of the stalled solution and the mean of $\|\mathbf{n}\|^2$, scaled by the standard deviation, i.e., the quality metric of vector $\hat{\mathbf{x}}$ is defined as

$$\phi(\hat{\mathbf{x}}) = \frac{\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|^2 - N\sigma^2}{\sqrt{N\sigma^2}}. \quad (9)$$

We refer to the metric in (9) as the *standardized ML cost* of solution vector $\hat{\mathbf{x}}$. A small value of $\phi(\hat{\mathbf{x}})$ can be viewed as an indicator of increased closeness of $\hat{\mathbf{x}}$ to ML solution. Therefore, from the previous discussion, it is desired to choose

the stalling limit Θ_s to be an increasing function of $\phi(\hat{\mathbf{x}})$. For this purpose, we choose an exponential function of the form

$$\Theta_s(\phi(\hat{\mathbf{x}})) = c_1 \exp(\phi(\hat{\mathbf{x}})). \tag{10}$$

Also, we allow a minimum number of iterations (c_{min}) following a stalling. Based on the above discussion, we adopt the following rule to compute the stalling count:

$$\Theta_s(\hat{\mathbf{x}}) = \lceil \max\left(c_{min}, c_1 \exp\left(\phi(\hat{\mathbf{x}})\right)\right) \rceil . \tag{11}$$

The constant c_1 is chosen depending upon the QAM size; a larger c_1 is chosen for larger QAM size. As we will see in the performance and complexity results, the proposed randomization in the update rule and the stopping criterion are quite effective in achieving low complexity as well as near-optimal performance.

2) Performance and complexity of the R-MCMC algorithm: The simulated BER performance and complexity of the proposed R-MCMC algorithm in uplink multiuser MIMO systems with 4-QAM are shown in Figs. 2 to 5. The following R-MCMC parameters are used in the simulations: $c_{min} = 10$, $c_1 = 20$, MAX-ITER $= 16K$. Figures 2 to 5(a) are for the case where there is no imbalance in the received powers of all users, i.e., $\sigma_k^2 = 0$ dB $\forall\, k$. Perfect channel knowledge at the BS is assumed. The performance of R-MCMC in multiuser MIMO with $K = N = 16$ is shown in Fig. 2. The performance of the MCMC algorithm using the distribution in (8) with temperature parameter values $\alpha = 1, 1.5, 2, 3$ are also plotted. $16K$ iterations are used in the MCMC algorithm with temperature parameter. Sphere decoder performance is also shown for comparison. It is seen that the performance of MCMC with temperature parameter is very sensitive to the choice of the value of α. For example, for $\alpha = 1, 1.5$, the BER is found to degrade at high SNRs due to stalling problem. For $\alpha = 2$, the performance is better at high SNRs but worse at low SNRs. The proposed R-MCMC performs better than MCMC with temperature parameter (or almost the same) at all SNRs and α values shown. In fact, the performance of R-MCMC is almost the same as the sphere decoder performance. The R-MCMC complexity is, however, significantly lower than the sphere decoding complexity. While sphere decoder gets exponentially complex in K at low SNRs, the R-MCMC complexity (in average number of real operations per bit) is only $O(K^2)$ as can be seen in Fig. 3. Because of this low complexity, the R-MCMC algorithm scales well for large-scale systems with large values of K and N. This is illustrated in Fig. 4 and 5(a) where performance plots for systems up to $K = N = 128$ and 256 are shown. While Fig. 4 shows the BER as a function of SNR, Fig. 5(a) shows the average received SNR required to achieve a target BER of 10^{-3} as a function of $K = N$. Since sphere decoder complexity is prohibitive for hundreds of dimensions, we have plotted unfaded single-input single-output (SISO) AWGN performance as a lower bound on ML performance for comparison. It can be seen that R-MCMC achieves performance which is very close to SISO AWGN performance for large $K = N$, e.g., close to within 0.5 dB at 10^{-3} BER for $K = N = 128$ and 256. This illustrates

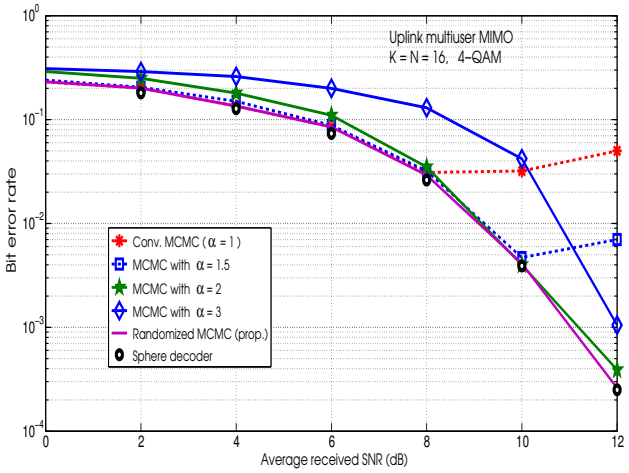


Fig. 2. BER performance of R-MCMC algorithm in comparison with those of sphere decoder and MCMC algorithm with different values of α. $K = N = 16$, 4-QAM, and no power imbalance. Performance of R-MCMC is almost the same as sphere decoder performance.

the achievability of near-optimal performance using R-MCMC for large systems. Figure 5(b) shows the BER performance in multiuser MIMO systems with received power imbalance among different users. The imbalance is simulated by choosing different σ_k^2 for different users, with σ_k^2 being uniformly distributed between -3 dB to 3 dB. Performance in systems with $K = N = 16$ and 128 are plotted with and without power imbalance. It is seen that even with power imbalance R-MCMC achieves almost the same performance as that of sphere decoder for $K = N = 16$.

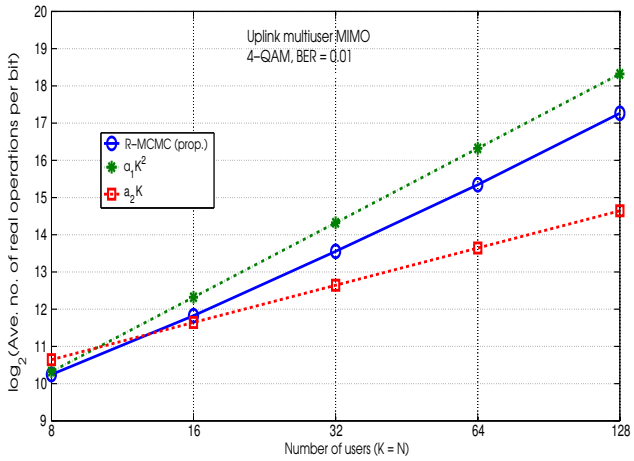


Fig. 3. Complexity of the R-MCMC algorithm in average number of real operations per bit as a function of $K = N$ with 4-QAM and no power imbalance at 10^{-2} BER.

IV. Multi-restart R-MCMC algorithm for higher-order QAM

Although the R-MCMC algorithm is very attractive in terms of both performance as well as complexity for 4-QAM, its performance for higher-order QAM is far from optimal. This is illustrated in Fig. 6, where R-MCMC is seen to achieve sphere decoder performance for 4-QAM, whereas for 16-QAM

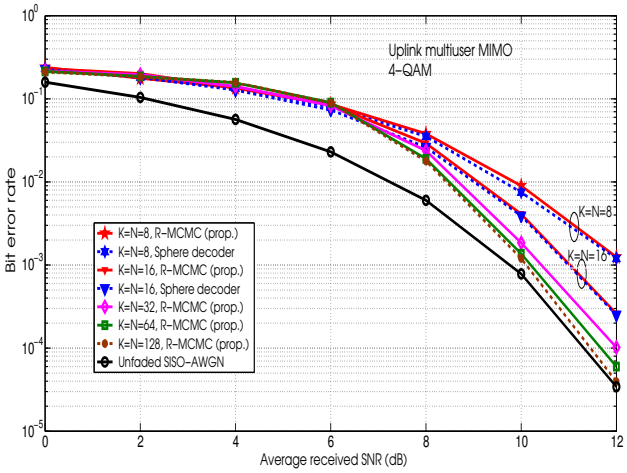


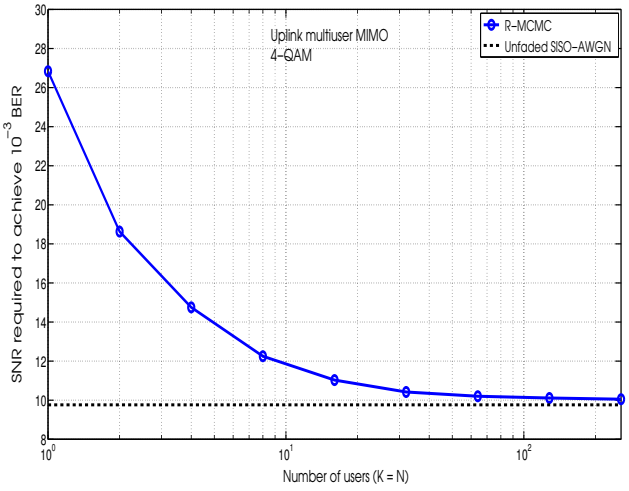
Fig. 4. BER performance of the R-MCMC algorithm in uplink multiuser MIMO with $K = N = 8, 16, 32, 64, 128$, 4-QAM and no power imbalance.

and 64-QAM it performs poorly compared to sphere decoder. This observation motivates the need for ways to improve R-MCMC performance in higher-order QAM. Interestingly, we found that use of multiple restarts[1] coupled with R-MCMC is able to significantly improve performance and achieve near-ML performance in large systems with higher-order QAM.

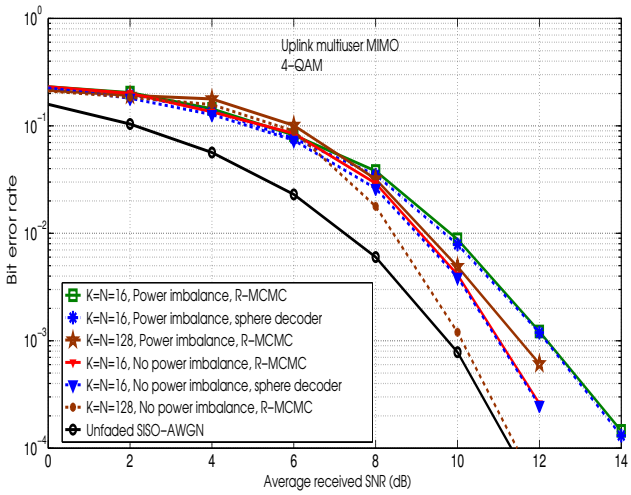
A. Effect of restarts in R-MCMC and conventional MCMC

In Figs. 7(a) and 7(b), we compare the effect of multiple random restarts in R-MCMC and conventional MCMC algorithms for 4-QAM and 16-QAM, respectively. For a given realization of $\mathbf{x}, \mathbf{H}$ and $\mathbf{n}$, we ran both algorithms for three different random initial vectors, and plotted the least ML cost up to nth iteration as a function of n. We show the results of this experiment for multiuser MIMO with $K = N = 16$ at 11 dB SNR for 4-QAM and 18 dB SNR for 16-QAM (these SNRs give about 10^{-3} BER with sphere decoding for 4-QAM and 16-QAM, respectively). The true ML vector cost (obtained through sphere decoder simulation for the same realization) is also plotted. It is seen that R-MCMC achieves much better least ML cost compared to conventional MCMC. This is because conventional MCMC gets locked up in some state (with very low state transition probability) for long time without any change in ML cost in subsequent iterations, whereas the randomized sampling strategy in R-MCMC is able to exit from such states quickly and give improved ML costs in subsequent iterations. This shows that R-MCMC is preferred over conventional MCMC. Even more interestingly, comparing the least ML costs of 4-QAM and 16-QAM (in Figs. 7(a) and (b), respectively), we see that all the three random initializations could converge to almost true ML vector cost for 4-QAM within 100 iterations, whereas

[1]It is noted that multiple restarts, also referred to as running multiple parallel Gibbs samplers, have been tried with conventional and other variants of MCMC in [27], [29], [30]. But the stalling problem is not fully removed and near-ML performance is not achieved. It turns out that restarts when coupled with R-MCMC is very effective in achieving near-ML performance.



(a)



(b)

Fig. 5. (a) Average SNR required to achieve 10^{-3} BER as a function of number of users $(K = N)$ in uplink multiuser MIMO with 4-QAM and no power imbalance. (b) BER performance of the R-MCMC algorithm in uplink multiuser MIMO with $K = N = 16, 128$ and 4-QAM. Power imbalance with σ_k^2's uniformly distributed between -3 dB and 3 dB.

only initial vector 3 converges to near true ML cost for 16-QAM and initial vectors 1 and 2 do not. Since any random initialization works well with 4-QAM, R-MCMC is able to achieve near-ML performance without multiple restarts for 4-QAM. However, it is seen that 16-QAM performance is more sensitive to the initialization, which explains the poor performance of R-MCMC without restarts in higher-order QAM. MMSE vector can be used as an initial vector, but it is not a good initialization for all channel realizations. This points to the possibility of achieving good initializations through multiple restarts to improve the performance of R-MCMC in higher-order QAM.

B. R-MCMC with multiple restarts

In R-MCMC with multiple restarts, we run the basic R-MCMC algorithm multiple times, each time with a different

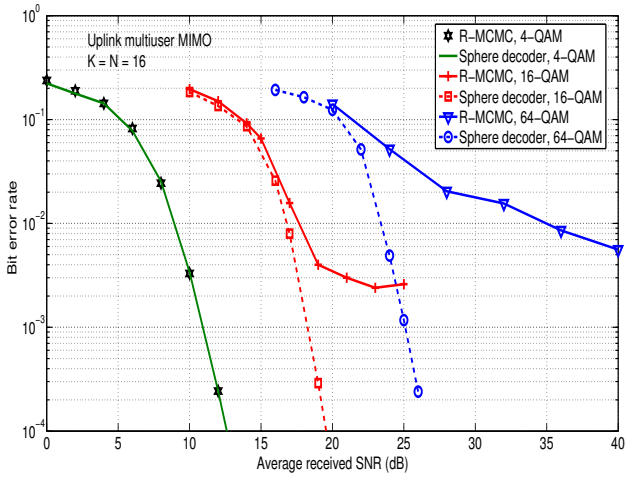


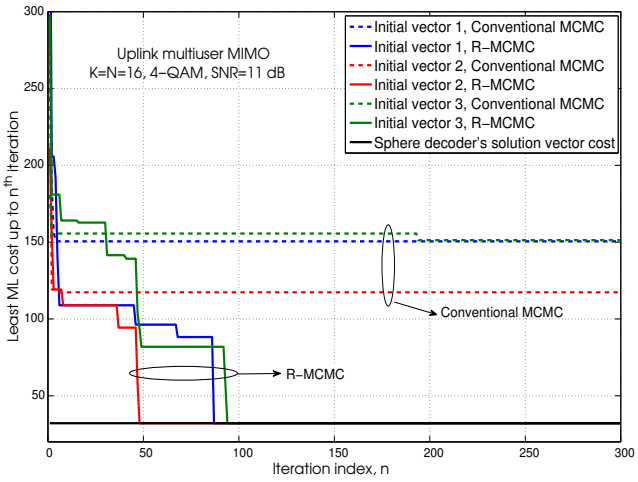
Fig. 6. Comparison between R-MCMC performance and sphere decoder performance in uplink multiuser MIMO with $K = N = 16$ and 4-/16-/64-QAM.

random initial vector, and choose that vector with the least ML cost at the end as the solution vector. Figure 8 shows the improvement in the BER performance of R-MCMC as the number of restarts (R) is increased in multiuser MIMO with $K = N = 16$ and 16-QAM at SNR = 18 dB. 300 iterations are used in each restart. It can be observed that, though BER improves with increasing R, much gap still remains between sphere decoder performance and R-MCMC performance even with $R = 10$. A larger R could get the R-MCMC performance close to sphere decoder performance, but at the cost of increased complexity. While a small R results in poor performance, a large R results in high complexity. So, instead of arbitrarily fixing R, there is a need for a good restart criterion that can significantly enhance the performance without incurring much increase in complexity. We devise one such criterion below.

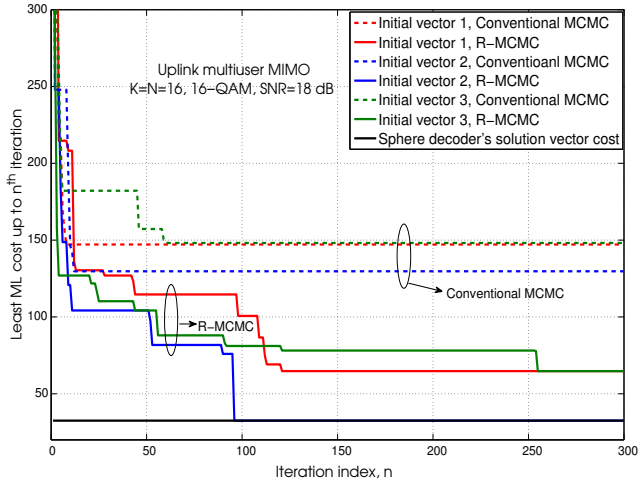
1) Proposed restart criterion: At the end of each restart, we need to decide whether to terminate the algorithm or to go for another restart. To do that, we propose to use

- the standardized ML costs (given by (9)) of solution vectors, and
- the number of repetitions of the solution vectors.

Nearness of the ML costs obtained so far to the error-free ML cost in terms of its statistics can allow the algorithm to get near ML solution. Checking for repetitions can allow restricting the number of restarts, and hence the complexity. We use the minimum standardized ML cost obtained so far and its number of repetitions to decide the credibility of the solution. An integer threshold (P) is defined for the best ML cost obtained so far for the purpose of comparison with the number of repetitions. In Fig 9, we plot histograms of the standardized ML cost of correct and incorrect solution vectors at the output of R-MCMC with restarts in multiuser MIMO with $K = N = 8$ and 4-/16-QAM. We judge the correctness of the obtained solution vector from R-MCMC output by running sphere decoder simulation for the same



(a) 4-QAM, SNR=11 dB



(b) 16-QAM, SNR=18 dB

Fig. 7. Least ML cost up to nth iteration versus n in conventional MCMC and R-MCMC for different initial vectors in multiuser MIMO with $K = N = 16$.

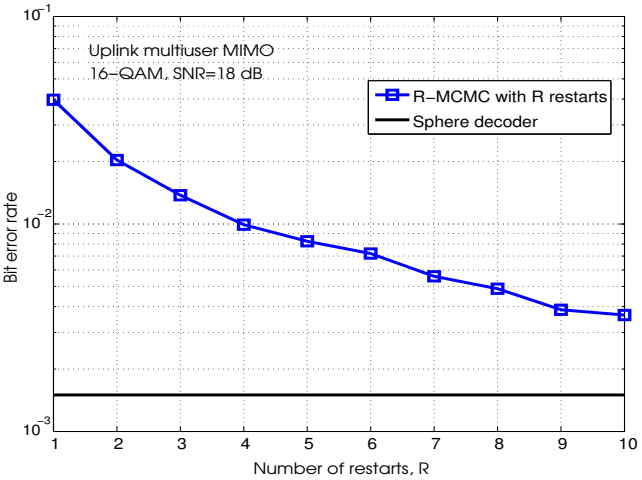


Fig. 8. BER performance of R-MCMC as a function of number of restarts in multiuser MIMO with $K = N = 16$ and 16-QAM at SNR = 18 dB.

realizations. It can be observed in Fig. 9 that the incorrect standardized ML cost density does not stretch into negative values. Hence, if the obtained solution vector has negative standardized ML cost, then it can indeed be correct with high probability. But as the standardized ML cost increases in the positive domain, the reliability of that vector decreases and hence it would require more number of repetitions for it to be trusted as the final solution vector. It can also be observed from Fig. 9 that the incorrect density in case of 16-QAM is much more than that of 4-QAM for the same SNR. So it is desired that, for a standardized ML cost in the positive domain, the number of repetitions needed to declare as the final solution should increase with the QAM size. Accordingly, the number of repetitions needed for termination (P, the integer threshold) is chosen as per the following expression:

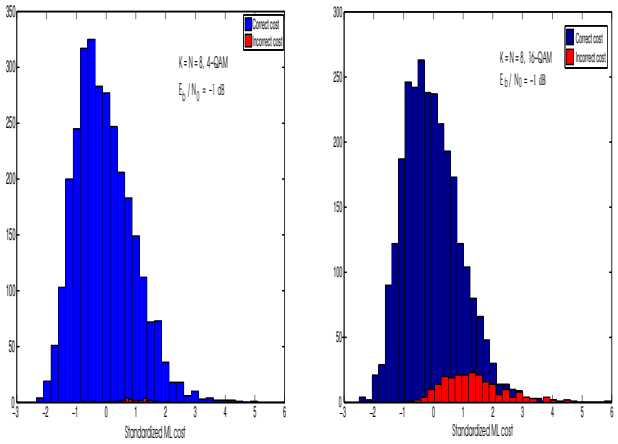
$$P = \lfloor \max\left(0, c_2 \phi(\tilde{\mathbf{x}})\right) \rfloor + 1, \tag{12}$$

where $\tilde{\mathbf{x}}$ is the solution vector with minimum ML cost so far. Now, denoting R_{max} to be the maximum number for restarts, the proposed *R-MCMC with restarts* algorithm (we refer to this as the R-MCMC-R algorithm) can be stated as follows.

- ***Step 1:*** Choose an initial vector.
- ***Step 2:*** Run the basic R-MCMC algorithm in Sec. III-B.
- ***Step 3:*** Check if R_{max} number of restarts are completed. If yes, go to Step 5; else go to Step 4.
- ***Step 4:*** For the solution vector with minimum ML cost obtained so far, find the required number of repetitions needed using (12). Check if the number of repetitions of this solution vector so far is less than the required number of repetitions computed in Step 4. If yes, go to Step 1, else go to Step 5.
- ***Step 5:*** Output the solution vector with the minimum ML cost so far as the final solution.

C. Performance and complexity of the R-MCMC-R Algorithm

The BER performance and complexity of the R-MCMC-R algorithm are evaluated through simulations. The following parameters are used in the simulations of R-MCMC and R-MCMC-R: $c_{min} = 10$, $c_1 = 10 \log_2 M$ (i.e., $c_2 = 20, 40, 60$ for 4-/16-/64-QAM, respectively), MAX-ITER = $8K\sqrt{M}$, $R_{max} = 50$, and $c_2 = 0.5 \log_2 M$. In Fig. 10, we compare the BER performance of conventional MCMC, R-MCMC, R-MCMC-R and sphere decoder in multiuser MIMO with $K = N = 16$ and 16-QAM. In the first restart, MMSE solution vector is used as the initial vector. In the subsequent restarts, random initial vectors are used. For 64-QAM, the randomized sampling is applied only to the one-symbol away neighbors of the previous iteration index; this helps to reduce complexity in 64-QAM. From Fig. 10, it is seen that the performance of conventional MCMC, either without or with restarts, is quite poor. That is, using restarts in conventional MCMC is not of much help. This shows the persistence of the stalling problem. The performance of R-MCMC (without restarts) is better than conventional MCMC with and without restarts, but its performance still is far from sphere decoder performance. This shows that R-MCMC alone (without restarts) is inadequate the



(a) 4-QAM (b) 16-QAM

Fig. 9. Histograms of standardized ML costs of correct and incorrect outputs from R-MCMC with restarts in multiuser MIMO with $K = N = 8$ and 4-/16-QAM.

alleviate the stalling problem in higher-order QAM. However, the randomized sampling in R-MCMC when used along with restarts (i.e., R-MCMC-R) gives strikingly improved performance. In fact, the proposed R-MCMC-R algorithm achieves almost sphere decoder performance (close to within 0.4 dB at 10^{-3} BER). This points to the important observations that application of any one of the two features, namely, randomized sampling and restarts, to the conventional MCMC algorithm is not adequate, and that simultaneous application of both these features is needed to alleviate the stalling problem and achieve near-ML performance in higher-order QAM.

Figure 11(a) shows that the R-MCMC-R algorithm is able to achieve almost sphere decoder performance for 4-/16-/64-QAM in multiuser MIMO with $K = N = 16$. Similar performance plots for 4-/16-/64-QAM for $K = N = 32$ are shown in Fig. 11(b), where the performance of R-MCMC-R algorithm is seen to be quite close to unfaded SISO-AWGN performance, which is a lower bound on true ML performance.

1) Performance/complexity comparison with other detectors: In Table-I, we present a comparison of the BER performance and complexity of the proposed R-MCMC-R algorithm with those of other detectors in the literature. Comparisons are made for systems with $K = N = 16, 32$ and 4-/16-/64-QAM. Detectors considered for comparison include: i) random-restart reactive tabu search (R3TS) algorithm reported in [21], which is a local neighborhood search based algorithm, and ii) fixed-complexity sphere decoder (FSD) reported in [38], which is a sub-optimal variant of sphere decoder whose complexity is fixed regardless of the operating SNR. Table-I shows the complexity measured in average number of real operations at a BER of 10^{-2} and the SNR required to achieve 10^{-2} BER for the above three detection algorithms. It can be seen that both R-MCMC-R and R3TS perform better than FSD. Also, R-MCMC-R achieves the best performance at the lowest complexity compared to R3TS and FSD for $K = N = 16$ with 16-QAM and 64-QAM. In 4-QAM and

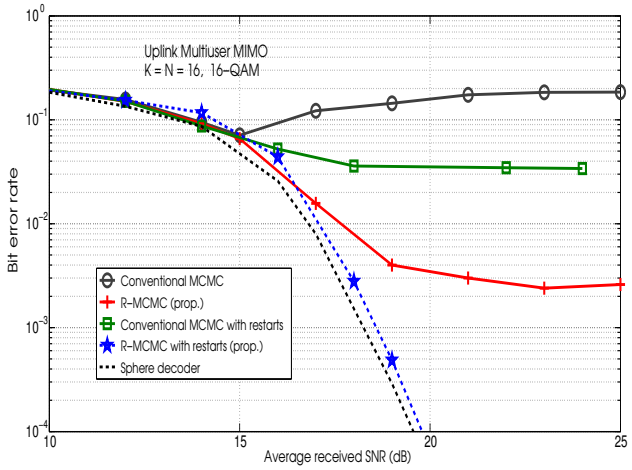


Fig. 10. BER performance between conventional MCMC (without and with restarts), proposed R-MCMC (without and with restarts), and sphere decoder in uplink multiuser MIMO with $K = N = 16$ and 16-QAM.

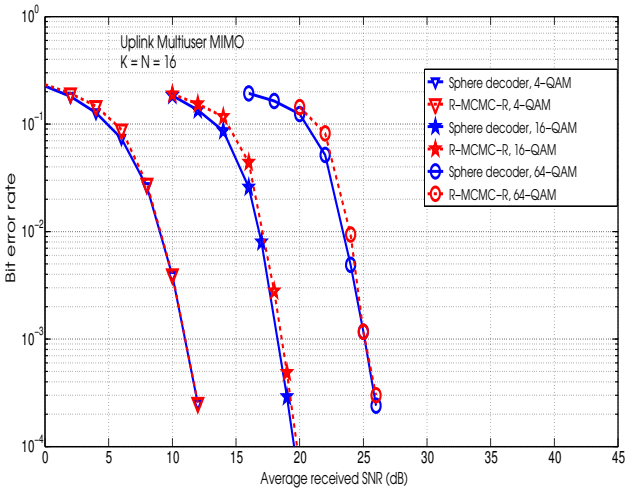
in $K = N = 32$, R-MCMC-R achieves same or slightly better performance than R3TS at some increased complexity compared to R3TS.

V. Conclusion

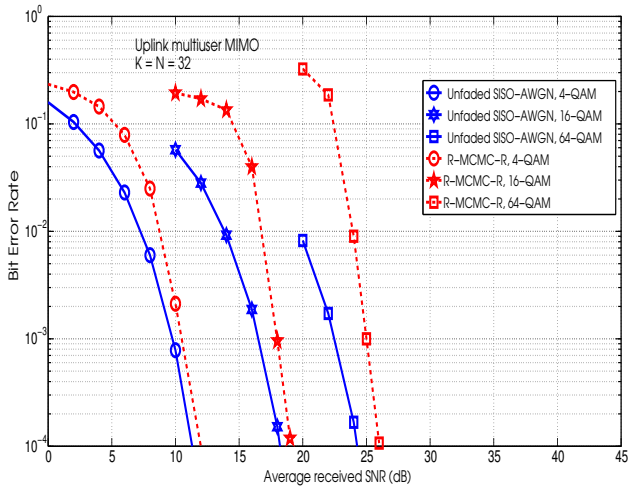
We proposed a novel MCMC based detection algorithm that achieved near-optimal performance on the uplink in large-scale multiuser MIMO systems. The proposed R-MCMC-R algorithm was shown to alleviate the stalling problem and achieve near-ML performance in large systems with tens to hundreds of antennas and higher-order QAM. Key ideas that enabled such attractive performance and complexity include i) a randomized sampling strategy that gave the algorithm opportunities to quickly exit from stalled solutions and move to better solutions, and ii) multiple random restarts that facilitated the algorithm to seek good solutions in different parts of the solution space. Multiple restarts alone (without randomized sampling) could not achieve near-ML performance at low complexity. Randomized sampling alone (without multiple restarts) could achieve near-ML performance at low complexity in the case of 4-QAM. But for higher-order QAM (16-/64-QAM) randomized sampling alone was not adequate. Joint use of both randomized sampling as well as multiple restarts was found to be crucial to achieve near-ML performance for 16-/64-QAM. While simulations were used to establish the attractiveness of the algorithm in performance and complexity, a theoretical analysis that could explain its good performance is important and challenging, which is a topic for future work. We have considered perfect synchronization and single-cell scenario in this paper. Other system level issues including uplink synchronization and multi-cell operation in large-scale MIMO systems can be considered as future work.

References

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, pp. 311-335, March 1998.



(a) $K = N = 16$



(b) $K = N = 32$

Fig. 11. BER performance of R-MCMC-R algorithm in uplink multiuser MIMO with $K = N = 16$ and 32 and higher-order modulation (4-/16-/64-QAM).

[2] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecommun.,* vol. 10, no. 6, pp. 585-595, November 1999.

[3] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*, Cambridge University Press, 2003.

[4] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[5] H. Bölcskei, D. Gesbert, C. B. Papadias, and Alle-Jan van der Veen, editors. *Space-Time Wireless Systems: From Array Processing to MIMO Communications*. Cambridge University Press, 2006.

[6] T. L. Marzetta, "How much training is required for multiuser MIMO?" *Asilomar Conf. on Signals, Systems and Computers*, pp. 359-363, October-November 2006.

[7] H. Taoka and K. Higuchi, "Field experiment on 5-Gbit/s ultra-high-speed packet transmission using MIMO multiplexing in broadband packet radio access," *NTT DoCoMo Tech. Journ.,* vol. 9, no. 2, pp. 25-31, September 2007.

[8] K. V. Vardhan, S. K. Mohammed, A. Chockalingam, B. S. Rajan, "A low-complexity detector for large MIMO systems and multicarrier CDMA systems," *IEEE JSAC Spl. Iss. on Multiuser Detection for Adv. Commun. Sys. & Net.*, vol. 26, no. 3, pp. 473-485, April 2008.

[9] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-

Modulation	Algorithm	Complexity in average number of real operations in $\times 10^6$ and SNR required to achieve 10^{-2} BER			
		$K = N = 16$		$K = N = 32$	
		Complexity	SNR	Complexity	SNR
4-QAM	R-MCMC-R (prop.)	0.1424	9 dB	.848	8.8 dB
	R3TS [21]	0.1877	9 dB	0.6823	8.8 dB
	FSD in [38]	0.1351	10.1 dB	4.9681	10.3 dB
16-QAM	R-MCMC-R (prop.)	1.7189	17 dB	15.158	16.7 dB
	R3TS [21]	3.968	17 dB	7.40464	17 dB
	FSD [38]	4.836432	17.6 dB	4599.5311	17.8 dB
64-QAM	R-MCMC-R (prop.)	11.181	24 dB	166.284	24 dB
	R3TS [21]	25.429504	24.2 dB	77.08784	24.1 dB
	FSD in [38]	305.7204	24.3 dB	$\star$	$\star$

TABLE I
PERFORMANCE AND COMPLEXITY COMPARISON OF PROPOSED R-MCMC-R DETECTOR WITH OTHER DETECTORS IN [21] AND [38] FOR FOR $K = N = 16, 32$ AND 4-/16-/64-QAM. $\star$: NOT SIMULATED DUE TO PROHIBITIVE COMPLEXITY.

complexity precoder for large multiuser MISO systems," *Proc. IEEE VTC'2008*, pp. 797 - 801, Singapore, May 2008.

[10] S. K. Mohammed, A. Zaki, A. Chockalingam, and B. S. Rajan, "High-rate space-time coded large-MIMO systems: Low-complexity detection and channel estimation," *IEEE Jl. Sel. Topics in Sig. Proc. (JSTSP): Spl. Iss. on Managing Complexity in Multiuser MIMO Systems*, vol. 3, no. 6, pp. 958-974, December 2009.

[11] H. Taoka and K. Higuchi, "Experiments on peak spectral efficiency of 50 bps/Hz with 12-by-12 MIMO multiplexing for future broadband packet radio access," *Proc. Intl. Symp. on Commun., Contr., and Sig. Proc. (ISCCSP'2010),* Limassol, Cyprus, March 2010.

[12] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590-3600, November 2010.

[13] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO: How many antennas do we need?," online arXiv:1107.1709v1 [cs.IT] 8 Jul 2011.

[14] J. Jose, A. Ashikhmin, T. Marzetta, and S. Viswanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640-2651, August 2011.

[15] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas," online arXiv:1107.3862v2 [cs.IT] 13 Sep 2011.

[16] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Uplink power efficiency of multiuser MIMO with very large antenna arrays," *Proc. Allerton Conf. on Commun., Contr., and Comput.*, pp. 1272-1279, September 2011.

[17] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *accepted IEEE Signal Processing Magazine.* http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-71581

[18] B. Cerato and E. Viterbo, "Hardware implementation of low-complexity detector for large MIMO," *Proc. IEEE ISCAS'2009*, pp. 593-596, Taipei, May 2009.

[19] P. Li and R. D. Murch, "Multiple output selection-LAS algorithm in large MIMO systems," *IEEE Commun. Lett.,* vol. 14, no. 5, pp. 399-401, May 2010.

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

[21] T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Random-restart reactive tabu search algorithm for detection in large-MIMO systems," *IEEE Commun. Letters*, vol. 14, no.12, pp. 1107-1109, December 2010.

[22] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "Low-complexity near-MAP decoding of large non-orthogonal STBCs using PDA," *Proc. IEEE ISIT'2009*, Seoul, June-July 2009.

[23] P. Som, T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Low-complexity detection in large-dimension MIMO-ISI channels using graphical Models," *IEEE Jl. Sel. Topics in Signal Processing (JSTSP): Special issue on Soft Detection for Wireless Transmission*, vol. 5, no. 8, pp. 1497-1511, December 2011.

[24] J. Goldberger and A. Leshem, "MIMO detection for high-order QAM based on a Gaussian tree approximation," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 4973-4982, August 2011.

[25] C. Knievel, M. Noemm and P. A. Hoeher, "Low complexity receiver for large-MIMO space time coded systems," *Proc. IEEE VTC-Fall'2011*, September 2011.

[26] R. Chen, J. S. Liu, and X. Wang, "Convergence analyses and comparisons of Markov chain Monte Carlo algorithms in digital communications," *IEEE Trans. Sig. Proc.*, vol. 50, no. 2, pp. 255-270, February 2002.

[27] B. Farhang-Boroujeny, H. Zhu, and Z. Shi, "Markov chain Monte Carlo algorithms for CDMA and MIMO communication systems," *IEEE Trans. Sig. Proc.*, vol. 54, no. 5, pp. 1896-1909, May 2006.

[28] H. Zhu, B. Farhang-Boroujeny, and R. R. Chen, "On performance of sphere decoding and Markov chain Monte Carlo methods," *IEEE Sig. Proc. Lett.,* vol. 12, no, 10, pp. 669-672, October 2005.

[29] X. Mao, P. Amini, and B. Farhang-Boroujeny, "Markov chain Monte Carlo MIMO detection methods for high signal-to-noise ratio regimes," *Proc. IEEE GLOBECOM'07*, Washington DC, November 2007.

[30] S. Akoum, R. Peng, R-R. Chen and B. Farhang-Boroujeny, "Markov chain Monte Carlo detection methods for high SNR regimes," *Proc. IEEE GLOBECOM'09*, November-December 2009.

[31] R-R. Chen, R. Peng, and B. Farhang-Boroujeny, "Markov chain Monte Carlo: Applications to MIMO detection and channel equalization," *Proc. IEEE ITA'09*, San Diego, February 2009.

[32] R. Peng, R-R. Chen, and B. Farhang-Boroujeny, "Markov chain Monte Carlo detectors for channels with intersymbol interference," *IEEE Trans. Signal Proc.*, vol. 58, no. 4, pp. 2206-2217, April 2010.

[33] M. Hansen, B. Hassibi, A. G. Dimakis, and W. Xu, "Near-optimal detection in MIMO systems using Gibbs sampling," *Proc. IEEE ICC'2009*, Honolulu, December 2009.

[34] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge Univ. Press, 2003.

[35] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2nd Edition, 2004.

[36] O. Haggsstrom, *Finite Markov Chains and Algorithmic Applications*, Cambridge Univ. Press, 2002.

[37] A. Kumar, S. Chandrasekaran, A. Chockalingam, and B. S. Rajan, "Near-optimal large-MIMO detection using randomized MCMC and randomized search algorithms," *Proc. IEEE ICC'2011*, Kyoto, June 2011.

[38] L. G. Barbero and J. S. Thompson, "Fixing the complexity of the sphere decoder for MIMO detection," *IEEE Trans. Wireless Commun.,* vol. 7, no. 6, pp. 2131-2142, June 2008.