

A Novel Monte-Carlo-Sampling-Based Receiver for Large-Scale Uplink Multiuser MIMO Systems

Tanumay Datta, N. Ashok Kumar, A. Chockalingam, *Senior Member, IEEE*, and
B. Sundar Rajan, *Senior Member, IEEE*

Abstract—In this paper, we propose low-complexity algorithms based on Monte Carlo sampling for signal detection and channel estimation 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 a similar number of uplink users. A BS receiver that employs a novel mixed sampling technique (which makes a probabilistic choice between Gibbs sampling and random uniform sampling in each coordinate update) for detection and a Gibbs-sampling-based method for channel estimation is proposed. The algorithm proposed for detection alleviates the stalling problem encountered at high signal-to-noise ratios (SNRs) in conventional Gibbs-sampling-based detection and achieves near-optimal performance in large systems with M -ary quadrature amplitude modulation (M -QAM). A novel ingredient in the detection algorithm that is responsible for achieving near-optimal performance at low complexity is the joint use of a mixed Gibbs sampling (MGS) strategy coupled with a multiple restart (MR) strategy with an efficient restart criterion. Near-optimal detection performance is demonstrated for a large number of BS antennas and users (e.g., 64 and 128 BS antennas and users). The proposed Gibbs-sampling-based channel estimation algorithm refines an initial estimate of the channel obtained during the pilot phase through iterations with the proposed MGS-based detection during the data phase. In time-division duplex systems where channel reciprocity holds, these channel estimates can be used for multiuser MIMO precoding on the downlink. The proposed receiver is shown to achieve good performance and scale well for large dimensions.

Index Terms—Channel estimation, detection, Gibbs sampling, large-scale multiuser multiple-input–multiple-output (MIMO) system, multiple restarts (MRs), randomized sampling, stalling problem.

I. INTRODUCTION

THE CAPACITY of multiple-input–multiple-output (MIMO) wireless channels is known to increase linearly with the minimum number of transmit and receive antennas in rich scattering environments [1]–[5]. Large-scale MIMO

systems with tens to hundreds of antennas have attracted much interest recently [6]–[17]. The motivation to consider these large-scale MIMO systems is the potential to practically realize the theoretically predicted benefits of MIMO, in terms of very high spectral efficiency/sum rates, increased reliability, and power efficiency, through the exploitation of large spatial dimensions. Using a large number of antennas is being recognized as a good approach to fulfilling the increased throughput requirements in future wireless systems. In particular, 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] and on the downlink [9], [14], [15] in these large systems are of interest. Key issues in large multiuser MIMO systems on the downlink include low-complexity precoding strategies and a pilot contamination problem encountered in using nonorthogonal pilot sequences for channel estimation in multicell scenarios [14]. In large multiuser MIMO systems on the uplink, users with one or more antennas simultaneously transmit to the BS having a large number of antennas, and their signals are separated at the BS using their spatial signatures toward the BS. Sophisticated signal processing is required at the BS receiver to extract the signal of each user from the aggregate received signal [4]. Using a large number of BS antennas has been shown to improve the power efficiency of uplink transmissions in multiuser MIMO systems using linear receivers at the BS [16]. Linear receivers, including matched-filter (MF) and MMSE receivers, are shown to be attractive for a very large number of BS antennas [13]. Our focus in this paper is to achieve near-optimal receiver performance at the BS in large multiuser MIMO systems on the uplink at low complexity. The receiver functions that we consider include signal detection and channel estimation. The approach that we adopt for both detection and channel estimation is a Monte Carlo sampling approach.

The uplink multiuser MIMO architecture can be viewed as a point-to-point MIMO system with colocated 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

Manuscript received May 24, 2012; revised March 10, 2013; accepted April 14, 2013. Date of publication April 29, 2013; date of current version September 11, 2013. This work was presented in part at the 2011 IEEE International Conference on Communications, Kyoto, Japan, May 2011 and at the 2012 Information Theory and Applications Workshop, San Diego, CA, USA, February 2012. The review of this paper was coordinated by Dr. C. Yuen.

T. Datta, A. Chockalingam, and B. S. Rajan are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India (e-mail: tan.swapnil@gmail.com; achockal@ece.iisc.ernet.in; bsrajan@ece.iisc.ernet.in).

N. A. Kumar is with Broadcom India Research Private Ltd., Bangalore 560103, India (e-mail: ashokkumarnimmala2@gmail.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TVT.2013.2260572

neighborhood search, including tabu search [8], [10], [18]–[21], probabilistic data association [22], and message passing on graphical models with factor graphs and Markov random fields [23]–[25].

Another interesting class of low-complexity algorithms reported in the context of code-division multiple access and MIMO detection is based on Markov chain Monte Carlo (MCMC) simulation techniques [26]–[33]. MCMC techniques are computational techniques that make use of sampling from probability distributions [34]. MCMC methods have their roots in the Metropolis algorithm, which is 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 complexity. Gibbs sampling is a popular MCMC method. An issue with conventional Gibbs-sampling-based detection, however, is the *stalling problem*, which degrades performance at high SNRs [27]. The 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 a Monte-Carlo-sampling-based detection algorithm that alleviates the stalling problem encountered in conventional Gibbs sampling and achieves near-optimal performance in large systems [37], [38]. A key idea that is instrumental in alleviating the stalling problem is a mixed Gibbs sampling (MGS) strategy that makes a probabilistic choice between conventional Gibbs sampling and random uniform sampling in each coordinate update. An efficient stopping criterion aids complexity reduction. The proposed MGS strategy is shown to achieve near-optimal performance in large multiuser MIMO systems with 16–128 BS antennas and the same or less number of uplink users for 4-quadrature amplitude modulation (4-QAM) [37]. However, we find that this MGS strategy alone is not adequate to achieve near-optimal performance at low complexity for higher order QAM (e.g., 16-QAM and 64-QAM). We show that near-optimal performance is also achieved in higher order QAM if a *multiple restart (MR) strategy* is performed in conjunction with the proposed MGS [38]. We refer to this as the “MGS with MRs (MGS-MR)” strategy. Here, again, an efficient restart criterion aids complexity reduction. The *joint use* of both MGS and MR strategies is found to be crucial in achieving near-optimal performance for higher order QAM in large systems. To our knowledge, this mixed-sampling-based algorithm has not been reported before for detection.

Channel estimation at the BS is an important issue in large multiuser MIMO systems on the uplink. While channel estimation at the BS is needed for uplink signal detection, in time-division duplex systems where channel reciprocity holds, the estimated channel can be also used for precoding purposes on the downlink, avoiding the need for feeding back channel estimates from the users. Our second contribution in this paper is that we propose a Gibbs-sampling-based uplink channel estimation algorithm at the BS receiver. The algorithm employs Gibbs sampling to refine an initial estimate of the channel

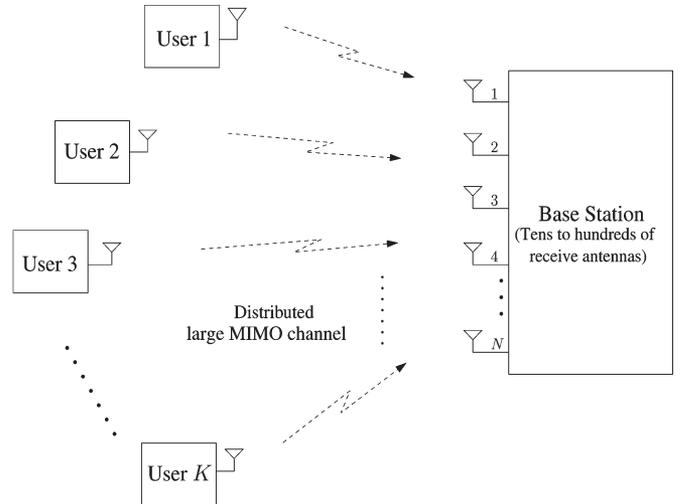


Fig. 1. Large-scale multiuser MIMO system on the uplink.

obtained during the pilot phase, through iterations with the proposed MGS-MR-based detection during the data phase. The algorithm is shown to yield good MSE and bit error rate (BER) performances in large multiuser MIMO systems (e.g., 128 BS antennas and users). BER performance that is close to that with perfect channel knowledge is shown to be achieved.

The remainder of this paper is organized as follows. The uplink multiuser MIMO system model is presented in Section II. The proposed MGS algorithm with and without MRs, and its performance/complexity results are presented in Section III. Section IV presents the proposed Gibbs-sampling-based channel estimation algorithm and its performance. 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 uplink users with one transmit antenna each, i.e., $K \leq N$ (see Fig. 1). N and K are in the range of tens to hundreds. All users transmit symbols from a modulation alphabet \mathbb{B} . Although we consider single-antenna users here, the proposed schemes apply to a general setting where user k can have n_{t_k} transmit antennas and transmit n_{t_k} spatial streams of data subject to $\sum_k n_{t_k} = K$. It is assumed that synchronization and sampling procedures have been carried out, and that the sampled baseband signals are available at the BS receiver. Let $x_k \in \mathbb{B}$ denote the transmitted symbol from user k . Let $\mathbf{x}_c = [x_1, x_2, \dots, x_K]^T$ denote the vector that is comprised of the symbols simultaneously transmitted by all users in one channel use. Let $\mathbf{H}_c \in \mathbb{C}^{N \times K}$, which is given by $\mathbf{H}_c = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_K]$, denote the channel gain matrix, where $\mathbf{h}_k = [h_{1k}, h_{2k}, \dots, h_{Nk}]^T$ is the channel gain vector from user k to the BS, and h_{jk} denotes the channel gain from the k th user to the j th receive antenna at the BS. Assuming rich scattering and adequate spatial separation between the 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 the received power from different users, and $\sigma_k^2 = 1$ corresponds to the

perfect power control scenario. The received signal vector at the BS in a channel use, which is denoted by $\mathbf{y}_c \in \mathbb{C}^N$, can be written as

$$\mathbf{y}_c = \mathbf{H}_c \mathbf{x}_c + \mathbf{n}_c \quad (1)$$

where \mathbf{n}_c is the noise vector whose entries are modeled as independent and identically distributed $\mathcal{CN}(0, \sigma^2)$. We will work with the real-valued system model corresponding to (1), which is given by

$$\mathbf{y}_r = \mathbf{H}_r \mathbf{x}_r + \mathbf{n}_r \quad (2)$$

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

$$\begin{aligned} \mathbf{H}_r &= \begin{bmatrix} \Re(\mathbf{H}_c) & -\Im(\mathbf{H}_c) \\ \Im(\mathbf{H}_c) & \Re(\mathbf{H}_c) \end{bmatrix} \\ \mathbf{y}_r &= \begin{bmatrix} \Re(\mathbf{y}_c) \\ \Im(\mathbf{y}_c) \end{bmatrix} \\ \mathbf{x}_r &= \begin{bmatrix} \Re(\mathbf{x}_c) \\ \Im(\mathbf{x}_c) \end{bmatrix} \\ \mathbf{n}_r &= \begin{bmatrix} \Re(\mathbf{n}_c) \\ \Im(\mathbf{n}_c) \end{bmatrix}. \end{aligned} \quad (3)$$

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

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

For a QAM alphabet \mathbb{B} , the elements of \mathbf{x} will take values from the underlying pulse-amplitude modulation (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

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

where $f(\mathbf{x}) \triangleq \mathbf{x}^T \mathbf{H}^T \mathbf{H} \mathbf{x} - 2\mathbf{y}^T \mathbf{H} \mathbf{x}$ is the ML cost. While the ML detector in (5) is exponentially complex in K (which is prohibitive for large K), the algorithms that we propose in the following have a per-symbol complexity that is quadratic in K , and they achieve near-ML performance as well.

III. PROPOSED MIXED GIBBS SAMPLING ALGORITHM FOR DETECTION

The ML detection problem in (5) can be solved by using MCMC simulations [34]. First, consider a conventional 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

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

Here, we assume perfect knowledge of channel gain matrix \mathbf{H} at the BS receiver. We will relax the perfect channel knowledge assumption by proposing a Gibbs-sampling-based channel estimation algorithm in Section IV.

A. Conventional Gibbs Sampling Algorithm for Detection

In conventional Gibbs-sampling-based detection, the algorithm starts with an initial symbol vector, which is denoted by $\mathbf{x}^{(t=0)}$. The initial vector can be a random vector or an output vector from known detectors, such as MF, zero forcing (ZF), and MMSE detectors. Let t denote the iteration index and i denote the coordinate index, $i = 1, 2, \dots, 2K$. Each iteration consists of $2K$ coordinate updates. In each iteration, $2K$ updates are carried out by sampling from distributions as follows:

$$\begin{aligned} x_1^{(t+1)} &\sim p\left(x_1 \mid x_2^{(t)}, x_3^{(t)}, \dots, x_{2K}^{(t)}, \mathbf{y}, \mathbf{H}\right) \\ x_2^{(t+1)} &\sim p\left(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \dots, x_{2K}^{(t)}, \mathbf{y}, \mathbf{H}\right) \\ x_3^{(t+1)} &\sim p\left(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \dots, x_{2K}^{(t)}, \mathbf{y}, \mathbf{H}\right) \\ &\vdots \\ x_{2K}^{(t+1)} &\sim p\left(x_{2K} \mid x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{2K-1}^{(t+1)}, \mathbf{y}, \mathbf{H}\right). \end{aligned} \quad (7)$$

The updated symbol vector at the end of each iteration is fed back to the next iteration for further coordinate updates. The algorithm is run for a certain number of iterations. The detected symbol vector is chosen to be that symbol vector that has the least ML cost in all the iterations.

A problem with the given conventional Gibbs-sampling-based detection is the stalling problem, which results in BER floors at high SNRs [27]. This is shown in Fig. 2(a) for $K = N = 16$, 4-QAM, a random initial vector, and 256 iterations, where the BER of the conventional Gibbs sampler becomes degraded for SNRs of more than 8 dB. The reason for this flooring is that the algorithm becomes trapped in some poor local solutions for a long time (i.e., for many iterations). This can be observed in Fig. 2(b), which shows an evolution of the ML cost of the state vector in the n th iteration as a function of n for an SNR of 12 dB. Note that the ML cost of the state vector does not change much from iterations 4 to 256, and note that this trapped ML cost is quite poor compared with the ML cost of the sphere-decoder solution. This leads to inferior performance compared with the sphere-decoder performance. Although the chain is guaranteed to converge to the target distribution (6) asymptotically as $n \rightarrow \infty$, stalling occurs and degrades the performance with a finite number of iterations.

B. Motivation for the Proposed MGS

One might think that the most natural target distribution for sampling is the posterior distribution itself, i.e., the distribution of \mathbf{x} , given \mathbf{y} and \mathbf{H} in (6). Gibbs sampling with this posterior distribution indeed guarantees taking us to this target distribution *in the limit* $n \rightarrow \infty$ [34]. However, this is not the appropriate distribution to sample from *if the goal is to minimize the expected number of iterations for finding the correct solution*, as has been demonstrated in [39, pp. 5]. This result was shown in the context of guessing passwords using MCMC. As in the result in [39, pp. 6], the correct target distribution with which one must sample to minimize

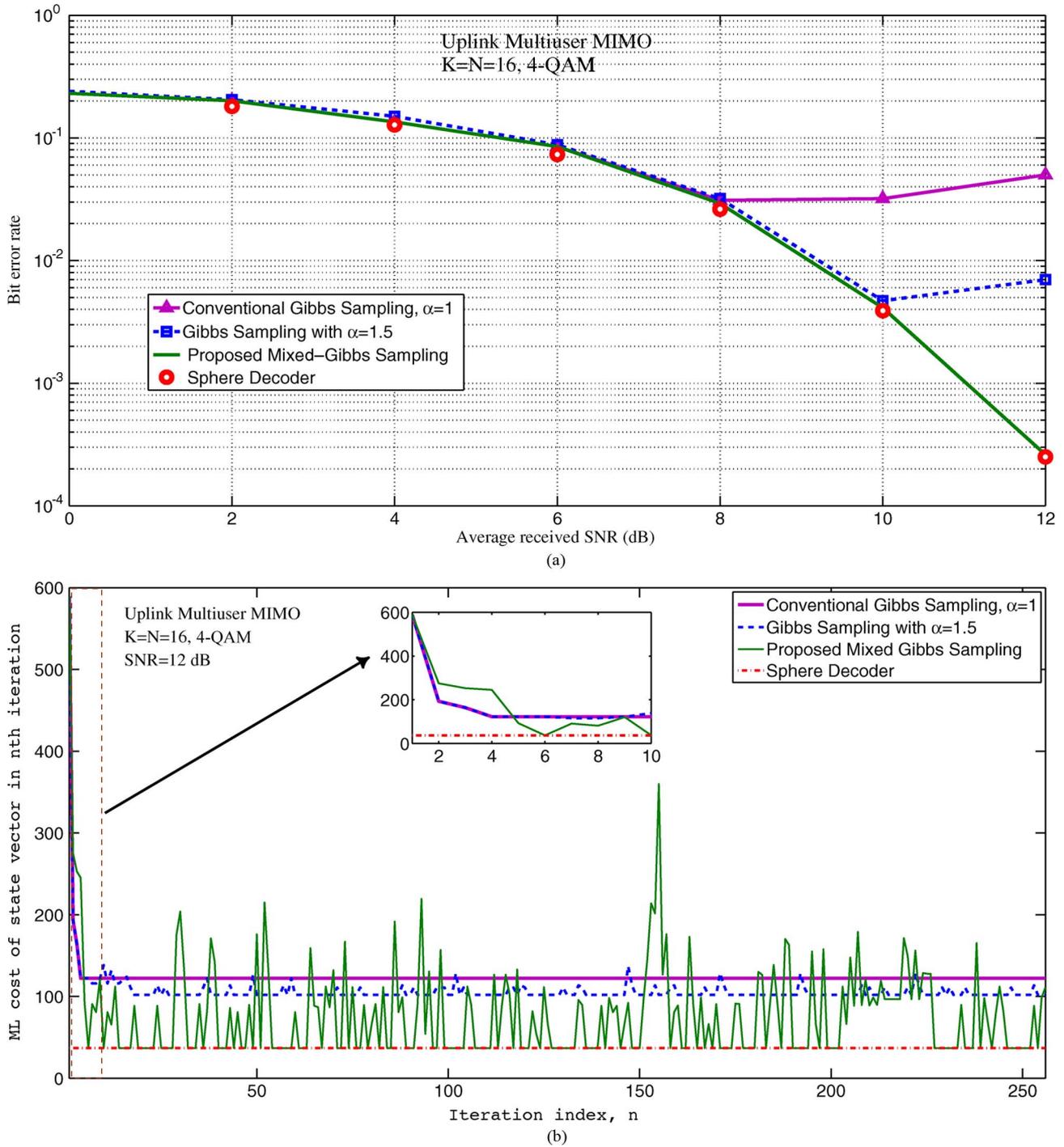


Fig. 2. BER performance and evolution of the ML cost of the state vector in a conventional Gibbs sampler, a Gibbs sampler with $\alpha = 1.5$, the proposed mixed Gibbs sampler, and the sphere decoder for $K = N = 16$ and 4-QAM.

the expected number of iterations is a *tilted* version of the posterior; specifically, it must be proportional to the square root of the posterior, i.e., $(p(x_i|y, \mathbf{H}, \mathbf{x}_{-i}))^{1/2}$ [39]. If there are only a finite number of iterations and we need to maximize the probability of arriving at the correct solution within these iterations, a heuristic is to sample in such a way that minimizes the higher moments of the number of iterations for finding the correct solution (see [40] and [41]). This can be achieved by choosing temperature parameter $\alpha \geq 1$ and by sampling according to $(p(x_i|y, \mathbf{H}, \mathbf{x}_{-i}))^{1/\alpha^2}$. The target distribution for

sampling proposed in [33] for MIMO detection used parameter α , where the target distribution is taken as

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

α represents a tunable positive parameter, which controls the mixing time of the Markov chain; the larger the value of α , the lesser will be the mixing time [33]. Conventional Gibbs sampling results as a special case when $\alpha = 1$. A larger α speeds

up the mixing and serves the purpose of reducing the higher moments of the number of iterations for finding the correct solution. However, the stalling problem persists even with large α . This is illustrated for Gibbs sampling with $\alpha = 1.5$ for an SNR of 12 dB in Fig. 2(b); the corresponding evolution of the ML cost of the state vector shows that the ML cost does not go below a certain value (which is well above the ML cost of the sphere-decoder solution) from iterations 20 to 256. These poor local solutions, in turn, result in degraded BER performance for SNRs of more than 10 dB, as shown in Fig. 2(a) for Gibbs sampling with $\alpha = 1.5$.

Motivated by the given observations, we propose a simple yet effective mixed sampling strategy to avoid local traps, thereby alleviating the stalling problem significantly. To break away from traps that lead to stalling, one needs to use a noisy version of the MCMC procedure. The noisiest is the one with infinite temperature (i.e., $\alpha = \infty$), which randomly and *uniformly* samples from all the possibilities. Our proposal is to use an intelligent *mixture* of: 1) Gibbs sampling with the posterior in (6) (i.e., $\alpha = 1$); and 2) random uniform sampling (i.e., $\alpha = \infty$). We will see (in Figs. 2(a) and 5) that sampling with this mixture distribution is able to achieve near-ML performance for 4-QAM. Another exploratory feature that enables us to get out of local traps when using higher order QAM is to have parallel explorations (i.e., MRs).

C. Proposed MGS

The key idea behind the proposed MGS approach is that, in each coordinate update, instead of updating the $x_i^{(t)}$ s as in the update rule in (7) with probability of 1 as done in conventional Gibbs sampling, we update them as in (7) with probability $1 - q$, and use a different update rule with probability q . The different update rule is as follows. Generate $|\mathbb{A}|$ probability values from the uniform distribution as

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

such that $\sum_{j=1}^{|\mathbb{A}|} p(x_i^{(t)} = j) = 1$; then, sample $x_i^{(t)}$ from this generated probability mass function (pmf). In other words, the proposed mixture distribution for sampling is given by

$$p(x_1, \dots, x_{2K} | \mathbf{y}, \mathbf{H}) \propto (1 - q)\psi(\alpha_1) + q\psi(\alpha_2) \quad (9)$$

where $\psi(\alpha) = \exp(-\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 / \alpha^2 \sigma^2)$, and q is the mixing ratio. The choice of different values for (α_1, α_2) is possible. Note that, with $\alpha_1 = 1$ and $\alpha_2 = \infty$, the first and second distributions in (9) become the true distribution and uniform distribution, respectively. That is, the sampling distribution is a weighted combination of the true distribution and the uniform distribution. We consider the $(\alpha_1 = 1, \alpha_2 = \infty)$ combination throughout this paper as this choice is simple and gives good performance. Note that $q = 0$ in (9) corresponds to the conventional Gibbs sampler, and $q = 1$ corresponds to pure random walk. In the Appendix, we present an analysis of the effect of the mixing ratio q and its optimal choice. Our analysis approach is to define an absorbing Markov chain and to use the property of absorbing Markov chains regarding the expected number

of iterations needed to reach the global minima for the first time. The analysis and results in the Appendix show that the optimum value of q , which minimizes the expected number of iterations needed to reach the global minima for the first time, is the inverse of the number of dimensions in the system. For our system with a complex modulation alphabet, the number of real dimensions is $2K$; therefore, the optimum mixing ratio is $q = 1/2K$.

Fig. 2(b) shows an evolution of the ML cost of the state vector in the n th iteration as a function of iteration index n in the proposed MGS with $q = 1/2K$. It can be observed that, because of the random uniform component in the sampling distribution, the variation of ML cost between successive iterations is quite significant. Two key observations can be made in Fig. 2(b) for the proposed MGS: 1) Unlike in conventional Gibbs sampling and Gibbs sampling with $\alpha = 1.5$, the state vector does not get trapped in local solutions for long; and 2) the quality of the ML cost at several instances in the evolution is very good to the extent that the ML cost of the sphere-decoder solution is almost reached. This enables the sampling from the proposed mixed distribution to almost achieve the BER performance of the sphere decoder, as shown in Fig. 2(a).

Fig. 3 shows the effect of mixing ratio q on the BER performance of the mixed Gibbs sampler for $K = N = 8, 16, 32, 64$, and 4-QAM at an SNR of 10 dB. In Fig. 3, along the lines of the optimal q result in the Appendix, the optimum value of q that minimizes the BER is observed to be $1/2K$. The optimum value of q will be small for large values of K . For $K = N = 64$ in Fig. 2(b), the optimum q is small, i.e., $q_{\text{opt}} = 1/128 = 0.0078$. The BER difference between the cases of $q = 0.0078$ and $q = 0$ for $K = N = 64$ is significant.

1) *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. However, a fixed value of the 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. Stalling is said to have occurred if the ML cost remains unchanged in two consecutive iterations. Once stalling is identified, the algorithm generates positive integer Θ_s (referred to as the *stalling limit*), and the iterations are allowed to continue in stalling mode (i.e., without changing the ML cost) 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 the stalling limit as aforementioned, the algorithm terminates on completing a maximum number of iterations, i.e., 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

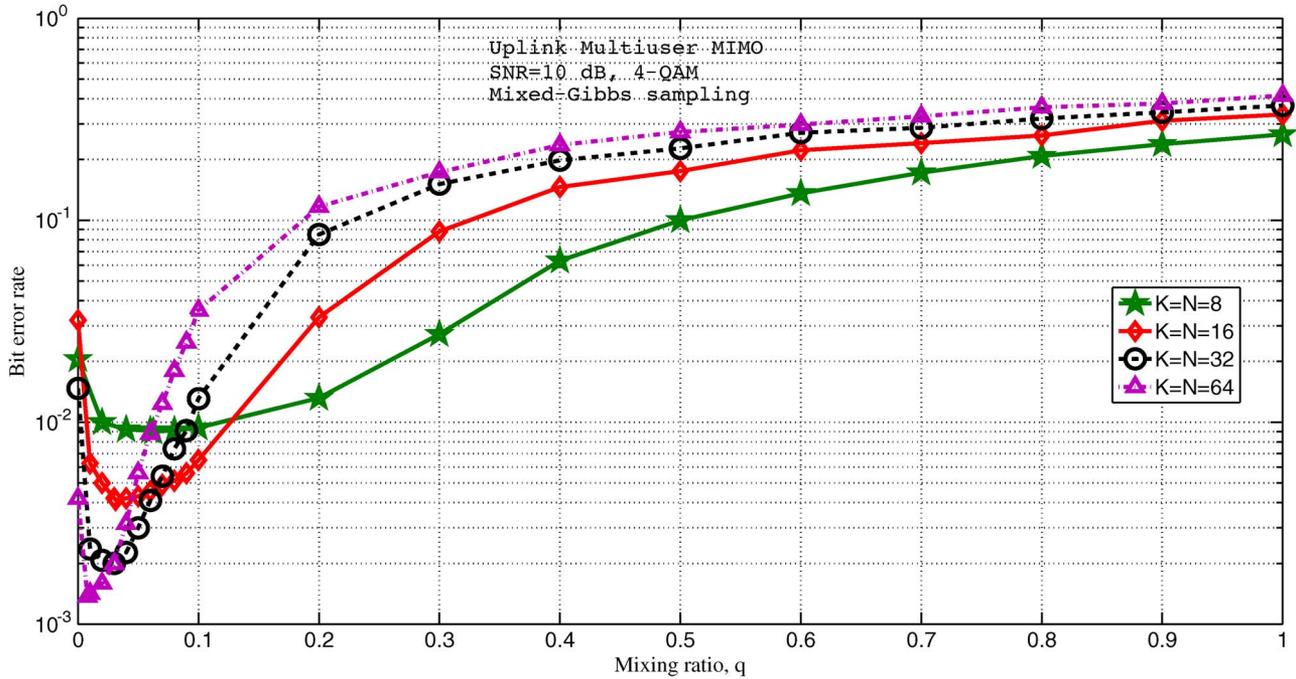


Fig. 3. BER performance of MGS as a function of mixing ratio q for $K = N = 8, 16, 32, 64$; 4-QAM; and $\text{SNR} = 10$ dB.

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 scaled chi-square 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 $\|\mathbf{n}\|^2$, which is scaled by the standard deviation, i.e., the quality metric of vector $\hat{\mathbf{x}}$ is defined as

$$\phi(\hat{\mathbf{x}}) = \frac{\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|^2 - N\sigma^2}{\sqrt{N}\sigma^2}. \quad (10)$$

We refer to the metric in (10) 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 the ML solution. Therefore, from the earlier 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 form¹

$$\Theta_s(\phi(\hat{\mathbf{x}})) = c_1 \exp(\phi(\hat{\mathbf{x}})). \quad (11)$$

The constant c_1 is chosen depending upon the QAM size; a larger value of c_1 is chosen for a larger QAM size. As the QAM size increases, the search space is also increased. Therefore, by choosing a c_1 proportional to the QAM size, we allow more iterations before stopping and thereby allow search over a larger region in the increased search space. In addition, we allow a minimum number of iterations c_{\min} following a stalling event.

¹Intuitively, the number of iterations to wait in stalling mode should be inversely proportional to the reliability of the vector that the process is stalled in, which exponentially decreases with the ML cost of that vector. Hence, we consider an exponential function in (11).

Based on the given discussion, we adopt the following rule to compute the stalling count:

$$\Theta_s(\hat{\mathbf{x}}) = \lceil \max(c_{\min}, c_1 \exp(\phi(\hat{\mathbf{x}}))) \rceil. \quad (12)$$

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 and near-optimal performance. A complete listing of the proposed algorithm incorporating the MGS and stopping criterion ideas is given in Algorithm 1.

Algorithm 1 Proposed MGS algorithm

- 1: **input:** $\mathbf{y}, \mathbf{H}, \mathbf{x}^{(0)}$;
 $\mathbf{x}^{(0)}$: initial vector $\in \mathbb{A}^{2K}$; MAX-ITER: max. no. of iterations;
- 2: $t = 0$; $\mathbf{z} = \mathbf{x}^{(0)}$; $\mathbf{q} = [q_1, q_2, \dots, q_{2K}]$;
- 3: $\beta = f(\mathbf{x}^{(0)})$; $f(\cdot)$: ML cost function; $\Theta_s(\cdot)$: *stalling limit* function;
- 4: **while** $t < \text{MAX-ITER}$ **do**
- 5: **for** $i = 1$ to $2K$ **do**
- 6: generate $\kappa \sim U[0, 1]$
- 7: **if** ($\kappa > q_i$)
- 8: $x_i^{(t+1)} \sim p(x_i | x_1^{(t+1)}, \dots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \dots, x_{2K}^{(t)})$
- 9: **else**
- 10: generate pmf $p(x_i^{(t+1)} = j) \sim U[0, 1]$, $\forall j \in \mathbb{A}$
- 11: sample $x_i^{(t)}$ from this pmf
- 12: **end if**
- 13: **end for**
- 14: $\gamma = f(\mathbf{x}^{(t+1)})$;
- 15: **if** ($\gamma \leq \beta$) **then**
- 16: $\mathbf{z} = \mathbf{x}^{(t+1)}$; $\beta = \gamma$;
- 17: **end if**

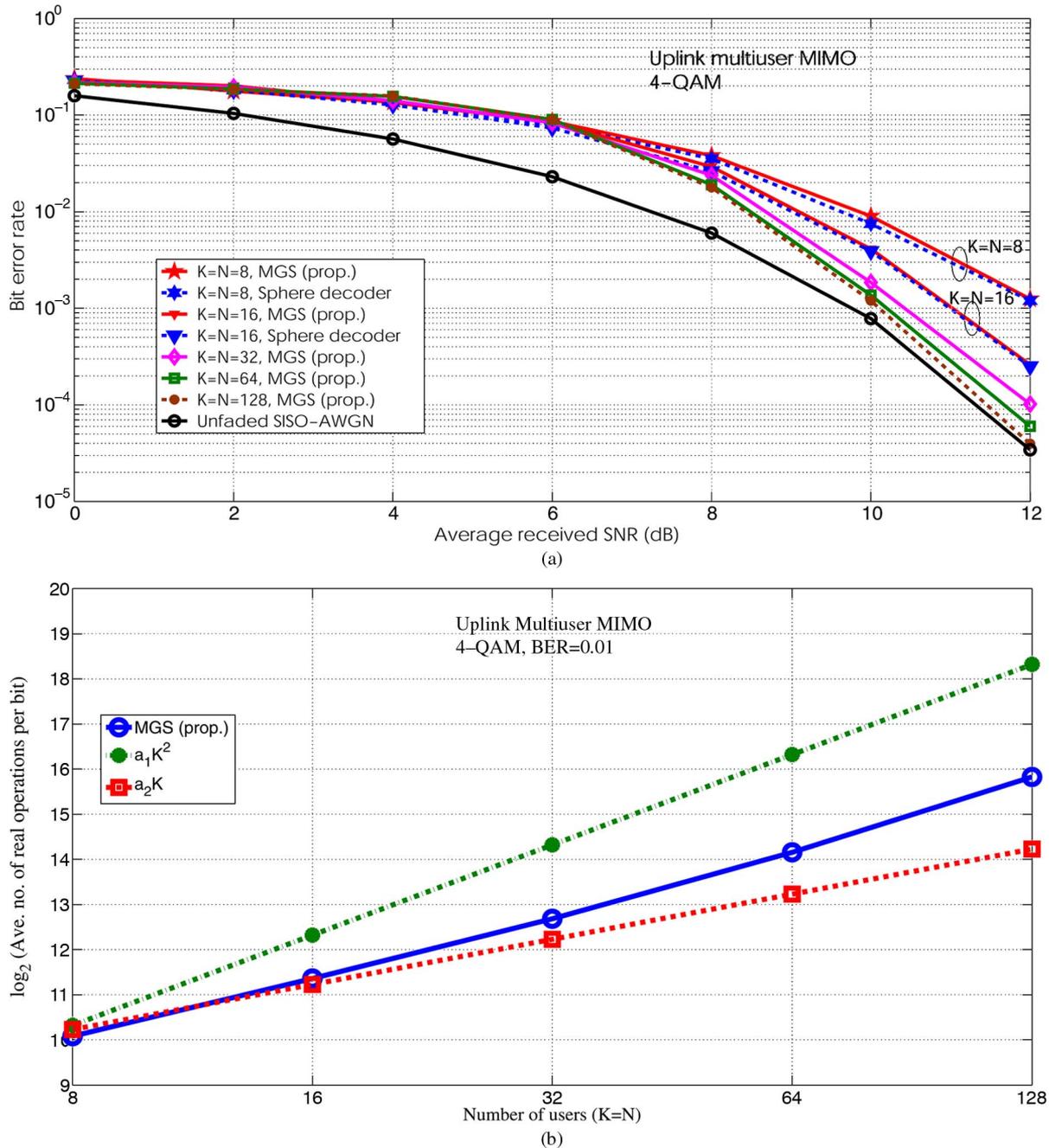


Fig. 4. BER performance and complexity of the proposed MGS algorithm for $K = N = 8, 16, 32, 64,$ and 128 and 4-QAM.

```

18:  $t = t + 1$ ;
19:  $\beta_v^{(t)} = \beta$ ;
20: if  $\beta_v^{(t)} == \beta_v^{(t-1)}$  then
21:   calculate  $\Theta_s(\mathbf{z})$ ;
22:   if  $\Theta_s < t$  then
23:     if  $\beta_v^{(t)} == \beta_v^{(t-\Theta_s)}$  then
24:       goto step 29
25:     end if
26:   end if
27: end if
28: end while
29: output:  $\mathbf{z}$ .     $\mathbf{z}$  : output solution vector
    
```

2) *Performance and Complexity of the MGS Algorithm:* The simulated BER performance and complexity of the proposed MGS-based detection in uplink multiuser MIMO systems with 4-QAM are shown in Fig. 4(a) and (b), respectively. The following parameters are used in the simulations: $c_{\min} = 10$, $c_1 = 20$, $\text{MAX-ITER} = 16K$, $q = 1/2K$, $\sigma_k^2 = 0 \text{ dB } \forall k$, and a random initial vector. Perfect channel knowledge at the BS is assumed. In Fig. 4(a), we see that the proposed MGS detector almost achieves the same performance of a sphere decoder for $K = N = 8$ and 16 . Further, while the sphere decoder is prohibitively complex for more than 32 real dimensions, the proposed MGS algorithm scales very well in complexity. This is shown in Fig. 4(b), where it can be seen that the average per-bit complexity of the MGS detector only quadratically

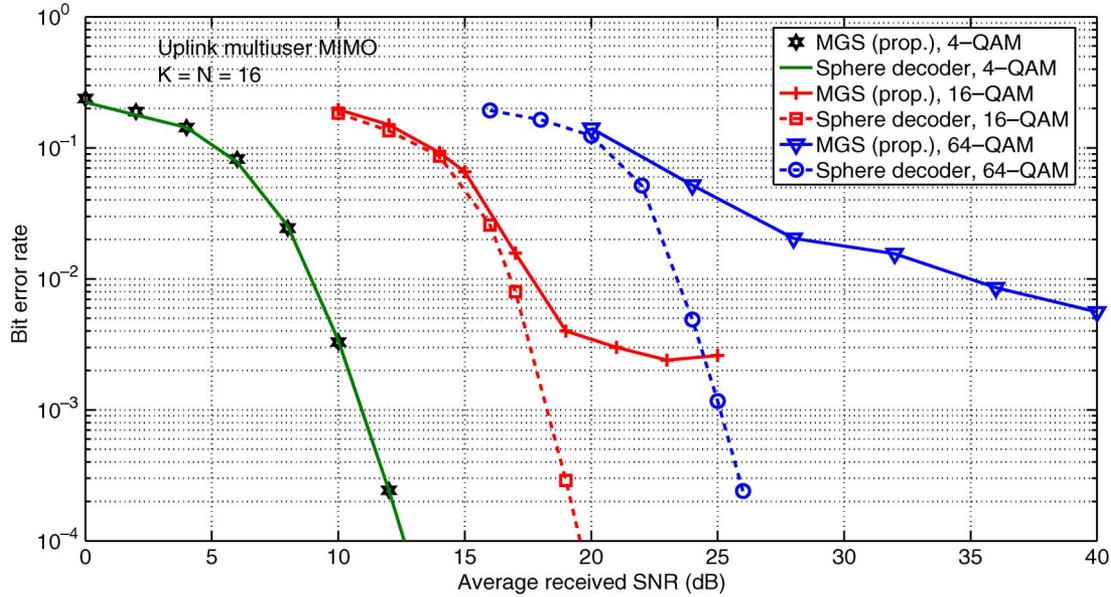


Fig. 5. Comparison between the performance of the MGS algorithm and that of the sphere decoder in uplink multiuser MIMO systems with $K = N = 16$ and 4-/16-/64-QAM.

grows in K (i.e., $O(K^2)$). Because of this low complexity, the MGS algorithm easily scales for $K = N = 32, 64$, and 128 , whose simulated BER performances are also shown in Fig. 4(a). Since sphere-decoder simulation is prohibitive for such large dimensions, we have plotted unfaded single-input–single-output (SISO) additive white Gaussian noise (AWGN) performance as a lower bound on the ML performance for comparison. It can be seen that the MGS detector achieves a performance that is very close to SISO AWGN performance for large $K = N$, e.g., close to within 0.5 dB at a BER of 10^{-3} for $K = N = 128$. This shows the ability of the proposed MGS detector to achieve near-optimal performance in large-scale multiuser MIMO systems.

D. MGS-MR Algorithm for Higher Order QAM

Although the MGS algorithm is very attractive in terms of both performance and complexity for 4-QAM, its performance for higher order QAM is far from optimal. This is shown in Fig. 5, where MGS is seen to achieve sphere-decoder performance for 4-QAM, whereas for 16-QAM and 64-QAM, it performs poorly compared with a sphere decoder. This observation motivates the need for ways to improve MGS performance in higher order QAM. Interestingly, we found that using MRs² coupled with MGS significantly improves the performance and achieves near-ML performance in large systems with higher order QAM.

1) *Effect of MRs*: In Fig. 6(a) and (b), we compare the effect of random MRs in MGS and conventional Gibbs sampling 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 the n th iteration as a function of n . We show the results of this experiment for multiuser MIMO with $K = N = 16$ at an SNR of 11 dB for 4-QAM and of 18 dB for 16-QAM (these SNRs give a BER of about 10^{-3} 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 MGS achieves a much better least ML cost compared with conventional Gibbs sampling. This is because conventional Gibbs sampling becomes locked up in some state (with a very low state transition probability) for a long time without any change in the ML cost in subsequent iterations, whereas the mixed Gibbs sampling strategy is able to exit from such states quickly and give improved ML costs in subsequent iterations. This shows that MGS is preferred over conventional Gibbs sampling. Interestingly, by comparing the least ML costs of 4-QAM and 16-QAM (in Fig. 6(a) and (b), respectively), we see that all the three random initializations could converge to an almost true ML vector cost for 4-QAM within 100 iterations, whereas only initial vector 3 converges to the near true ML cost for 16-QAM and initial vectors 1 and 2 do not. Since any random initialization works well with 4-QAM, MGS is able to achieve near-ML performance without MRs for 4-QAM. However, it is seen that 16-QAM performance is more sensitive to the initialization, which explains the poor performance of MGS without restarts in higher order QAM. The 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 MRs to improve the performance of MGS in higher order QAM.

2) *MGS-MR*: In MGS-MR, we run the basic MGS algorithm multiple times, each time with a different random initial vector, and choose that vector with the least ML cost at the end as the solution vector. Fig. 7 shows the improvement in the BER performance of MGS as the number of restarts R is increased in multiuser MIMO systems with $K = N = 16$ and 16-QAM at SNR = 18 dB. In each restart, 300 iterations are used. It can

²It is noted that MRs, which are also referred to as running multiple parallel Gibbs samplers, have been tried with conventional and other variants of MCMC in [27], [29], and [30]. However, the stalling problem is not fully removed, and near-ML performance is not achieved. It turns out that restarts when coupled with MGS is very effective in achieving near-ML performance.

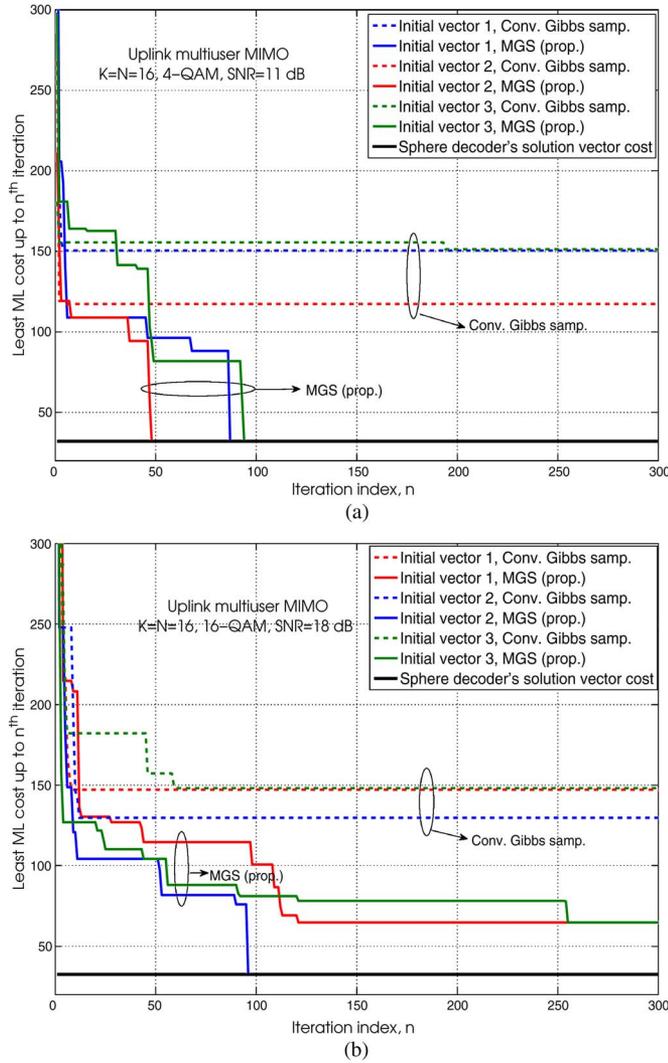


Fig. 6. Least ML cost up to the n th iteration versus n in conventional Gibbs sampling and MGS for different initial vectors in multiuser MIMO systems with $K = N = 16$. (a) 4-QAM and SNR = 11 dB. (b) 16-QAM and SNR = 18 dB.

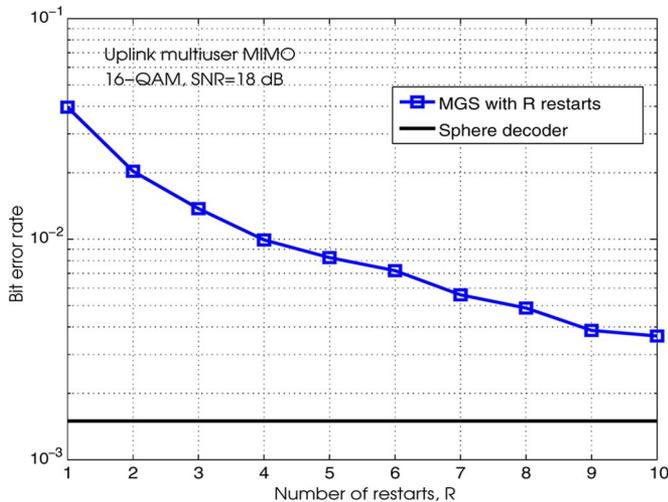


Fig. 7. BER performance of the MGS algorithm as a function of the number of restarts in multiuser MIMO systems with $K = N = 16$ and 16-QAM at SNR = 18 dB.

be observed that, although BER improves with increasing R , a large gap still remains between sphere-decoder performance and MGS performance even with $R = 10$. A larger R could get the MGS performance to be closer to the sphere-decoder performance but at the cost of increased complexity. Whereas a small R results in poor performance, a large R results in high complexity. Therefore, 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 in the following.

3) *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 using the following:

- the standardized ML costs [given by (10)] of solution vectors;
- the number of repetitions of the solution vectors.

The 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 a 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. 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. 8, we plot the histograms of the standardized ML cost of correct and incorrect solution vectors at the output of MGS with restarts in multiuser MIMO systems with $K = N = 8$ and 4-/16-QAM. We judge the correctness of the obtained solution vector from MGS output by running sphere-decoder simulation for the same realizations. It can be observed in Fig. 8 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. However, as the standardized ML cost increases in the positive domain, the reliability of that vector decreases; hence, it would require more repetitions for it to be trusted as the final solution vector. It can be also observed in Fig. 8 that the incorrect density in the case of 16-QAM is much more than that of 4-QAM for the same SNR. Therefore, 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 (integer threshold P) is chosen as per the following expression:

$$P = \lfloor \max(0, c_2 \phi(\tilde{\mathbf{x}})) \rfloor + 1 \tag{13}$$

where $\tilde{\mathbf{x}}$ is the solution vector with minimum ML cost so far, and c_2 is a constant chosen depending on the QAM size; a larger value of c_2 is chosen for larger QAM size.³ Now, denoting R_{\max} to be the maximum number for restarts, the proposed *MGS-MR* algorithm (we refer to this as the *MGS-MR* algorithm) can be stated as follows.

- *Step 1:* Choose an initial vector.

³We can observe in Fig. 8 that the probability of an output vector with positive normalized ML cost being incorrect increases with QAM size. Hence, we require more repetitions of the best vector to consider it to be reliable enough. Therefore, c_2 is chosen proportional to QAM size.

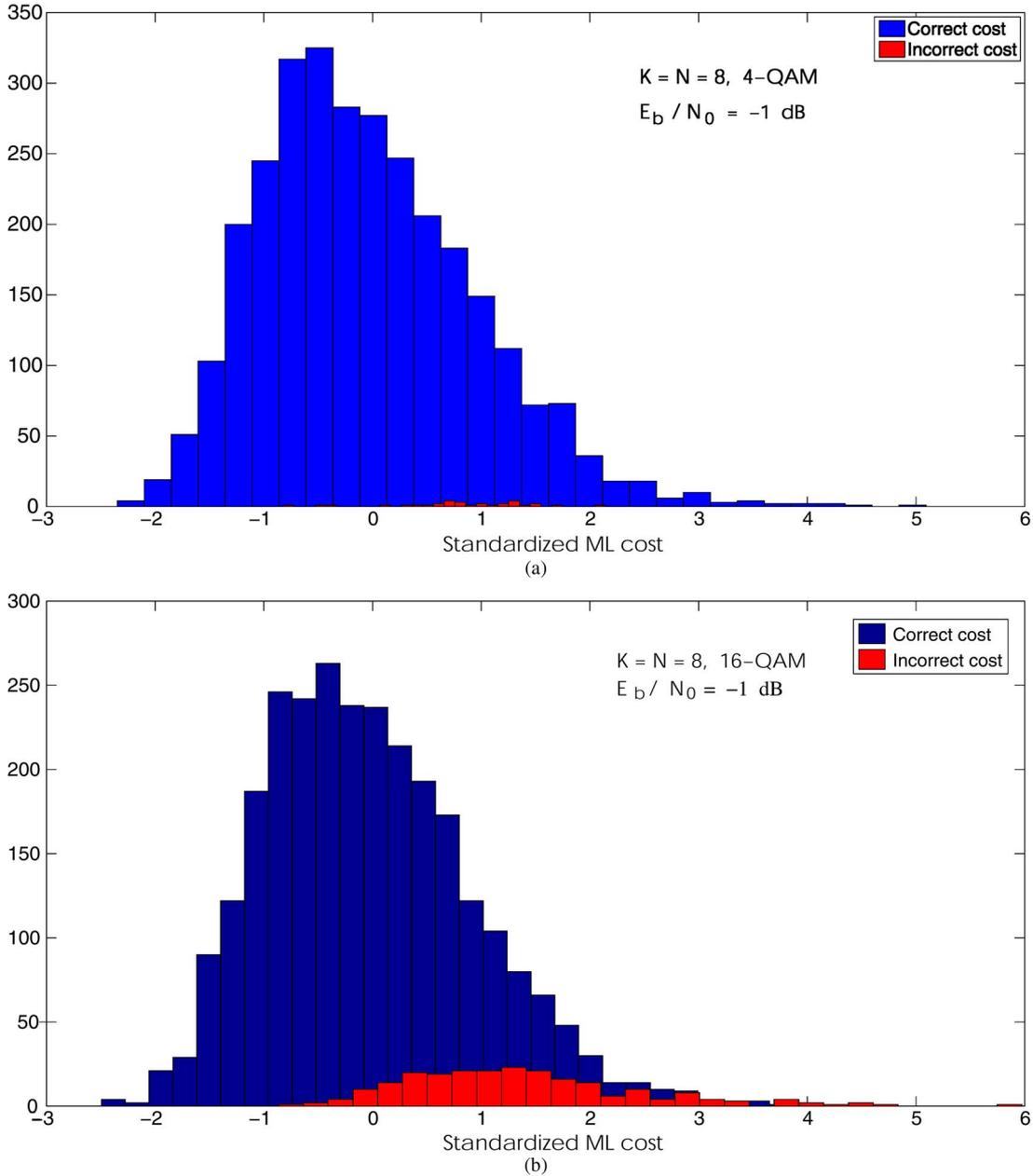


Fig. 8. Histograms of standardized ML costs of correct and incorrect outputs from MGS with restarts in multiuser MIMO with $K = N = 8$ and 4-/16-QAM. (a) 4-QAM. (b) 16-QAM.

- **Step 2:** Run the basic MGS algorithm in Section III-B.
- **Step 3:** Check if the 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 the minimum ML cost obtained so far, find the required number of repetitions needed using (13). 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.

4) *Soft-Decision Value Generation:* As shown, the proposed detection algorithm generates hard-decision outputs. In coded systems, soft-decision values of the bits are preferred as inputs to the channel decoder. Soft-decision values can be generated

from the hard-decision output vector from the MGS-MR algorithm as follows. Let the output of the MGS-MR detector be denoted by $\bar{\mathbf{x}}$. Knowing \mathbf{y} , \mathbf{H} , and $\bar{\mathbf{x}}$, the receiver needs to obtain soft-decision values of every transmitted bit. Let the set \mathbb{A} be partitioned into \mathbb{A}_i^+ and \mathbb{A}_i^- for each i , $i = 1, 2, \dots, \log_2 \sqrt{M}$, where \mathbb{A}_i^+ is the set of all the symbols in \mathbb{A} in which the i th bit is +1, and \mathbb{A}_i^- is the set of all the symbols in \mathbb{A} in which the i th bit is -1. Let the soft value of the i th bit of the k th user be denoted $L_{k,i}$. Let $\bar{\mathbf{x}}_{-k}$ denote the vector containing all elements in $\bar{\mathbf{x}}$ other than \bar{x}_k . Now, $L_{k,i}$ can be obtained as

$$L_{k,i} = \log \left(\frac{\sum_{a \in \mathbb{A}_i^+} p(\bar{x}_k = a | \bar{\mathbf{x}}_{-k}, \mathbf{y}, \mathbf{H})}{\sum_{a \in \mathbb{A}_i^-} p(\bar{x}_k = a | \bar{\mathbf{x}}_{-k}, \mathbf{y}, \mathbf{H})} \right) \quad (14)$$

where $p(\bar{x}_k = a | \bar{\mathbf{x}}_{-k}, \mathbf{y}, \mathbf{H})$ is calculated from (6).

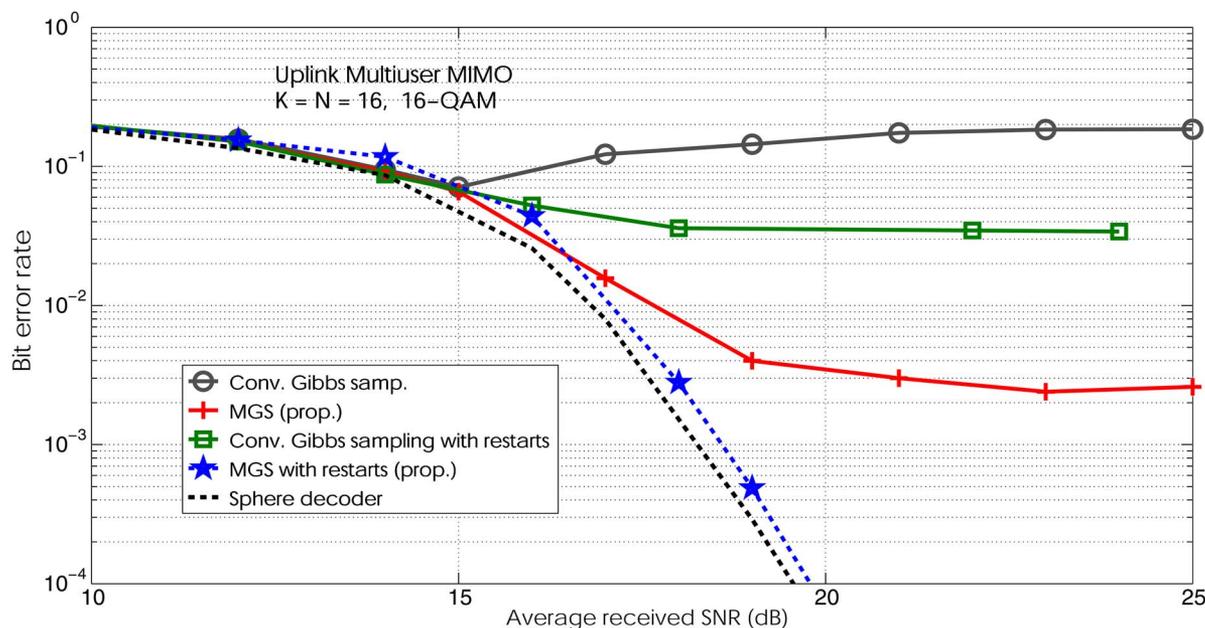


Fig. 9. BER performance comparison between conventional Gibbs sampling (without and with restarts), the proposed MGS (without and with restarts), and the sphere decoder in uplink multiuser MIMO systems with $K = N = 16$ and 16-QAM.

5) *Performance and Complexity of the MGS-MR Algorithm:*

The BER performance and complexity of the MGS-MR algorithm are evaluated through simulations. The following parameters are used in the simulations of MGS and MGS-MR: $c_{\min} = 10$, $c_1 = 10 \log_2 M$ (i.e., $c_1 = 20, 40, 60$ for 4-/16-/64-QAM, respectively), $\text{MAX-ITER} = 8K\sqrt{M}$, $R_{\max} = 50$, $c_2 = 0.5 \log_2 M$, and $q = 1/2K$. In Fig. 9, we compare the BER performance of conventional Gibbs sampling, MGS, MGS-MR, and the sphere decoder in multiuser MIMO with $K = N = 16$ and 16-QAM. In the first start, the MMSE solution vector is used as the initial vector. In the subsequent restarts, random initial vectors are used. For 64-QAM, the mixed sampling is applied only to the one-symbol away neighbors of the previous iteration index; this helps to reduce complexity in 64-QAM. In Fig. 9, it is shown that the performance of the conventional Gibbs sampler, either without or with restarts, is quite poor. That is, using restarts in conventional Gibbs sampling is not of much help. This shows the persistence of the stalling problem. The performance of MGS (without restarts) is better than conventional Gibbs sampling with and without restarts, but its performance still is far from the sphere-decoder performance. This shows that MGS alone (without restarts) is inadequate to alleviate the stalling problem in higher order QAM. However, the MGS when used along with restarts (i.e., MGS-MR) gives strikingly improved performance. In fact, the proposed MGS-MR algorithm almost achieves the sphere-decoder performance (close to within 0.4 dB at a BER of 10^{-3}). This points to the important observations that application of any one of the two features, namely, mixture sampling and restarts, to the conventional 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. Fig. 10(a) shows that the MGS-MR algorithm is able to achieve almost sphere-decoder

performance for 4-/16-/64-QAM in multiuser MIMO systems with $K = N = 16$. Similar performance plots for 4-/16-/64-QAM for $K = N = 32$ are shown in Fig. 10(b), where the performance of the MGS-MR algorithm is seen to be quite close to unfaded SISO AWGN performance, which is a lower bound on true ML performance.

In Table I, we present a comparison of the BER performance and the complexity of the proposed MGS-MR algorithm with those of another detection algorithm based on local search techniques, namely, the random-restart reactive tabu search (R3TS) algorithm [21], which has been reported to have good performance and complexity for large systems. Comparisons are made for systems with $K = N = 16, 32$, and, 4-/16-/64-QAM. Table I shows the complexity measured in an average number of real operations at a BER of 10^{-2} and the SNR required to achieve the BER of 10^{-2} for MGS-MR and R3TS algorithms. It can be seen that the MGS-MR algorithm achieves better performance at lower complexity for $K = N = 16$ with 16-QAM and 64-QAM. In 4-QAM and in $K = N = 32$, MGS-MR achieves the same or slightly better performance than R3TS at some increased complexity.

6) *Performance as a Function of the Loading Factor:* In Figs. 11 and 12, we present the BER and complexity plots as a function of the loading factor $\tau = K/N$, the ratio between the number of uplink users K , and the number of BS antennas N . In these figures, BER and complexity plots for the proposed MGS-MR detector and linear detectors, such as MF, ZF, and MMSE detectors, are presented and compared. Fig. 11 and Fig. 12 shows this comparison for 4-QAM and 16-QAM, respectively. The number of BS antennas N is fixed at 128, and the number of uplink users K is varied from small values up to 128. In Figs. 11 and 12, it is observed that the proposed MGS-MR detector performs better than MF, ZF, and MMSE detectors, moderately better under low loading factors,

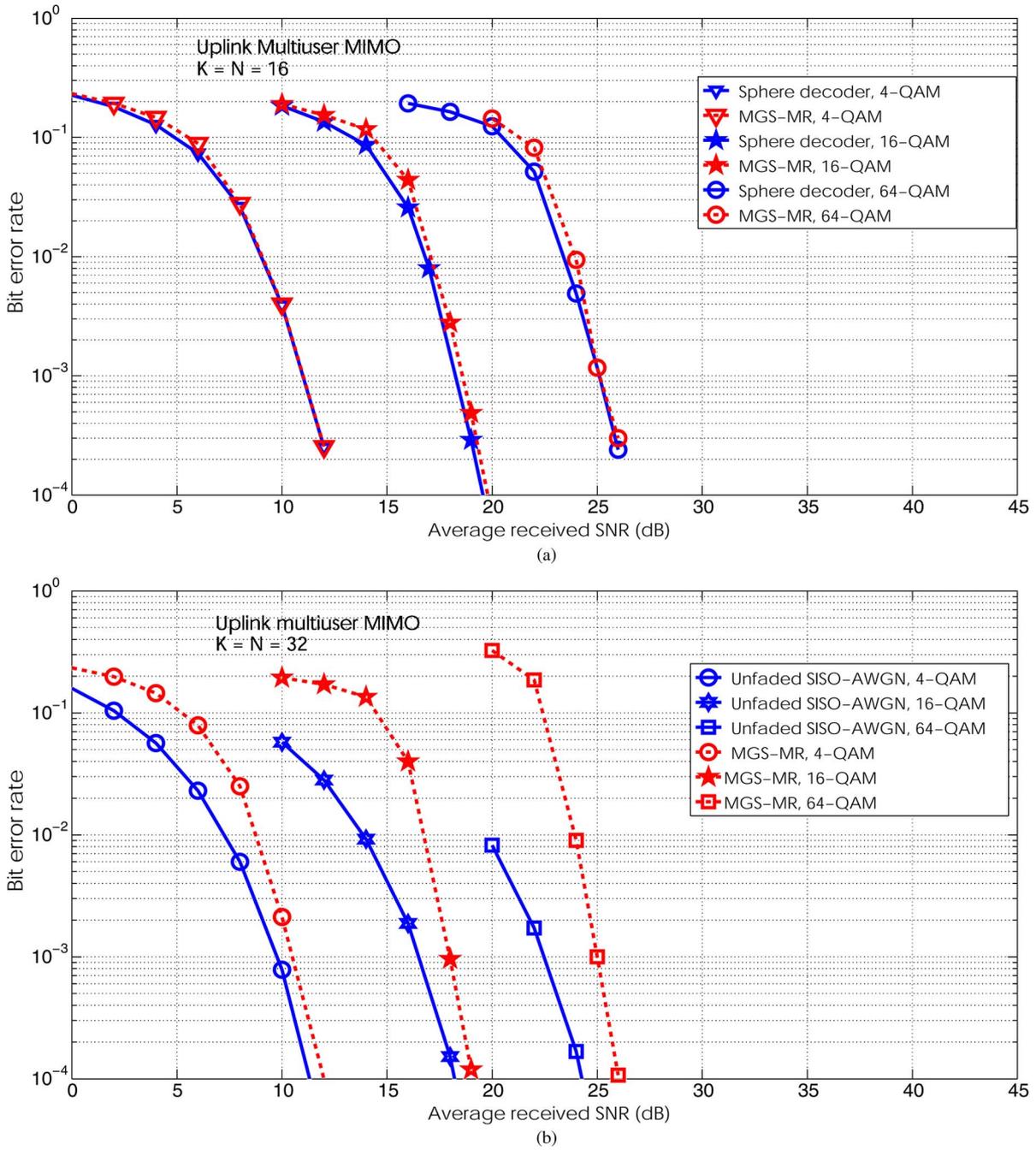


Fig. 10. BER performance of the MGS-MR algorithm in uplink multiuser MIMO systems with $K = N = 16$ and 32, and higher order QAM (4-/16-/64-QAM). (a) $K = N = 16$. (b) $K = N = 32$.

TABLE I
PERFORMANCE AND COMPLEXITY COMPARISON OF THE PROPOSED MGS-MR DETECTOR WITH THE TABU SEARCH-BASED DETECTOR IN [21] FOR $K = N = 16$ AND 32 AND 4-/16-/64-QAM

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	MGS-MR (prop.)	0.1424	9 dB	.848	8.8 dB
	R3TS [21]	0.1877	9 dB	0.6823	8.8 dB
16-QAM	MGS-MR (prop.)	1.7189	17 dB	15.158	16.7 dB
	R3TS [21]	3.968	17 dB	7.40464	17 dB
64-QAM	MGS-MR (prop.)	11.181	24 dB	166.284	24 dB
	R3TS [21]	25.429504	24.2 dB	77.08784	24.1 dB

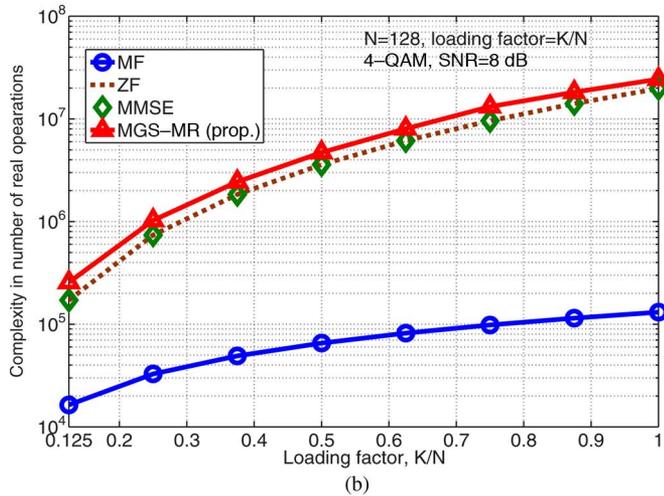
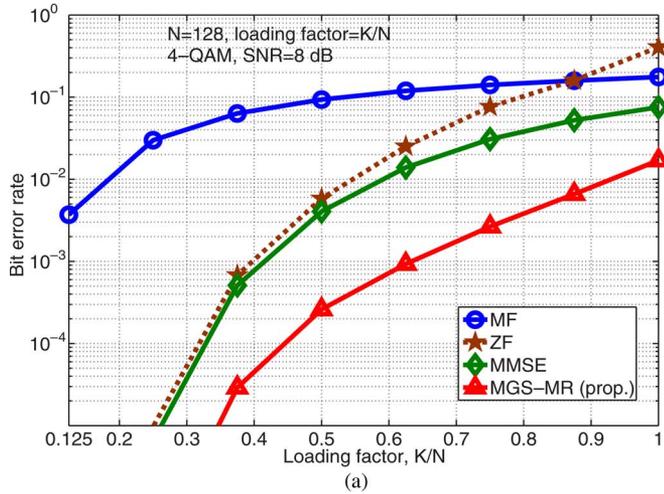


Fig. 11. BER performance and complexity of the proposed MGS-MR detector in comparison with those of linear (MF, ZF, and MMSE) detectors as a function of loading factor $\tau = K/N$. $N = 128$ and 4-QAM. (a) BER. (b) Complexity.

and significantly better (about one to two orders of improved BER) under medium to high loading factors. It is also seen that the complexity increase in MGS-MR detection compared with ZF/MMSE detection is nominal (not orders higher).

7) *Hard-Decision Versus Soft-Decision Performance*: Fig. 13 shows the coded BER performance of the system with rate-1/2 and rate-1/3 turbo codes, 4-QAM, and MGS-MR detection, for $K = 64$ and $N = 128$ (i.e., $\tau = K/N = 0.5$). Performance achieved using hard-decision output from the MGS-MR algorithm and the soft-decision output obtained by the method in Section III-D4 are plotted. In Fig. 13, it can be observed that, by using the soft-decision values generated by the method in Section III-D4, we achieve an improvement of about 1.5 dB in coded BER performance compared with using hard-decision values.

IV. PROPOSED GIBBS-SAMPLING-BASED CHANNEL ESTIMATION

Earlier, we assumed perfect channel knowledge at the BS receiver. Here, we relax the perfect channel knowledge assumption and propose an MCMC channel estimation algorithm.

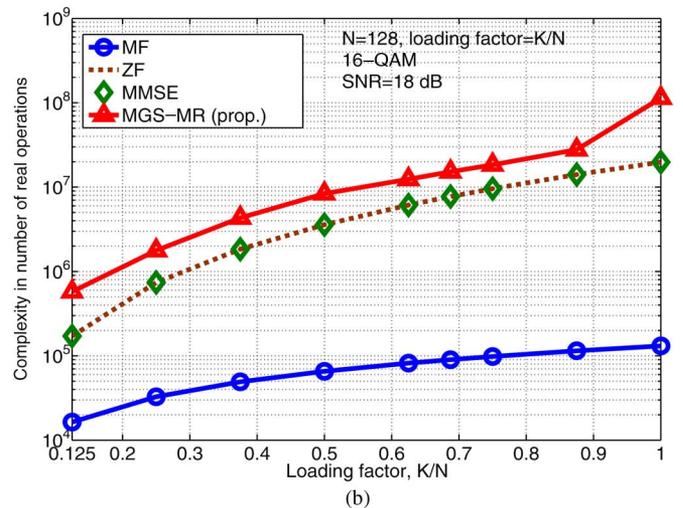
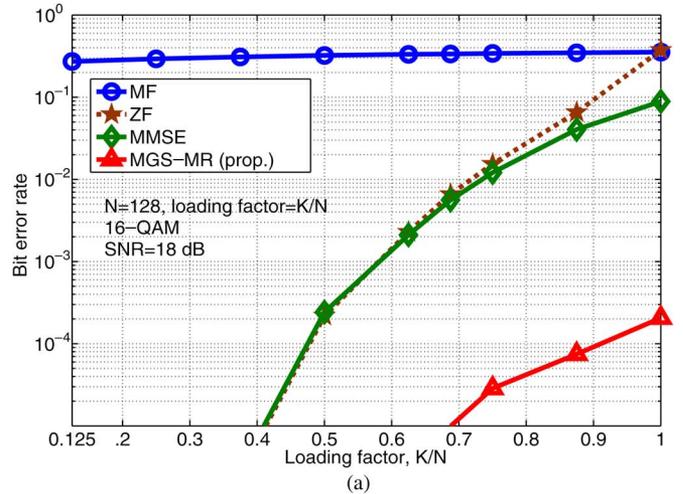


Fig. 12. BER performance and complexity of the proposed MGS-MR detector in comparison with those of linear (MF, ZF, MMSE) detectors as a function of loading factor $\tau = K/N$. $N = 128$, 16-QAM. (a) BER. (b) Complexity.

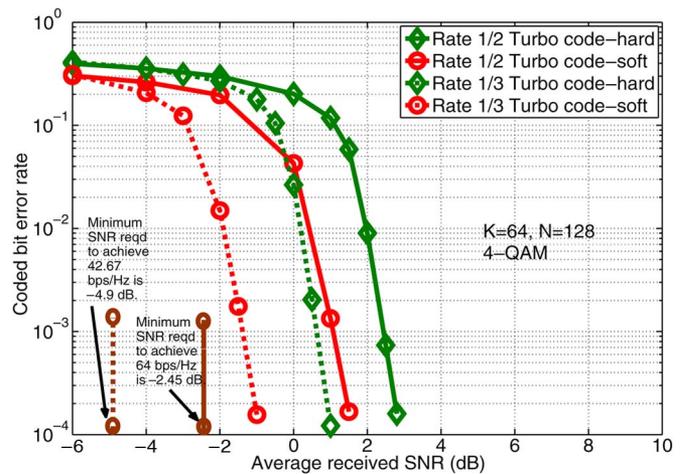


Fig. 13. Coded BER performance with hard-decision and soft-decision. $K = 64$, $N = 128$, $\tau = K/N = 0.5$, 4-QAM, MGS-MR detection, rate-1/2 and rate-1/3 turbo codes.

A. System Model

Consider the uplink multiuser MIMO system model in (1). As in Section II, perfect synchronization among users'

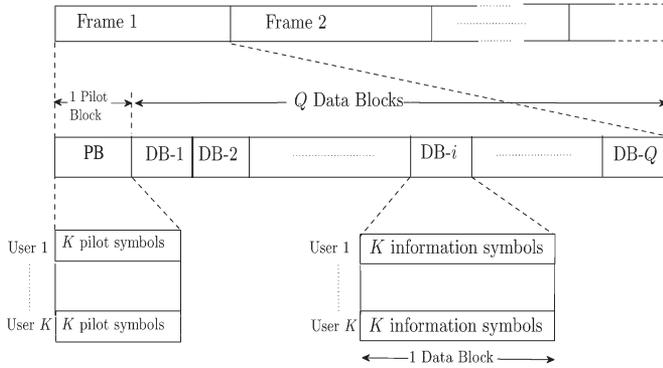


Fig. 14. Frame structure for uplink multiuser MIMO system.

transmissions is assumed. However, the assumption of perfect knowledge of the channel matrix at the BS is relaxed here. The channel matrix is estimated based on a pilot-based channel estimation scheme. Transmission is carried out in frames, where each frame consists of several blocks, as shown in Fig. 14. A slow-fading channel (typical with no/low mobility users) is assumed, where the channel is assumed to be constant over one frame duration. Each frame consists of a pilot block (PB) for the purpose of initial channel estimation, followed by Q data blocks (DBs). The PB consists of K channel uses in which a K -length pilot symbol vector comprising of pilot symbols transmitted from K users (one pilot symbol per user) is received by N receive antennas at the BS. Each DB consists of K channel uses, where K number of K -length information symbol vectors (one data symbol from each user) are transmitted. Taking both pilot and data channel uses into account, the total number of channel uses per frame is $(Q + 1)K$. DBs are detected using the MGS-MR algorithm that uses an initial channel estimate. The detected DBs are iteratively used to refine the channel estimates during the data phase using the proposed Gibbs-sampling-based channel estimation algorithm.

B. Initial Channel Estimate During the Pilot Phase

Let $\mathbf{x}_P^k = [x_P^k(0), x_P^k(1), \dots, x_P^k(K-1)]$ denote the pilot symbol vector transmitted from user k in K channel uses in a frame. Let $\mathbf{X}_P = [(\mathbf{x}_P^1)^T, (\mathbf{x}_P^2)^T, \dots, (\mathbf{x}_P^K)^T]^T$ denote the $K \times K$ pilot matrix formed by the pilot symbol vectors transmitted by the users in the pilot phase. The received signal matrix at the BS, i.e., \mathbf{Y}_P , of size $N \times K$ is given by

$$\mathbf{Y}_P = \mathbf{H}_c \mathbf{X}_P + \mathbf{N}_P \quad (15)$$

where \mathbf{N}_P is the $N \times K$ noise matrix at the BS.

We use the pilot sequence given by

$$\mathbf{x}_P^k = [\mathbf{0}_{(k-1) \times 1} \quad p \quad \mathbf{0}_{(K-k) \times 1}]. \quad (16)$$

We choose $p = \sqrt{KE_s}$, where E_s is the average symbol energy. Using the scaled identity nature of \mathbf{x}_P , an initial channel estimate $\hat{\mathbf{H}}_c$ is obtained as

$$\hat{\mathbf{H}}_c = \mathbf{Y}_P / p. \quad (17)$$

C. Data Detection Using Initial Channel Estimate

Let $\mathbf{x}_i^k = [x_i^k(0), x_i^k(1), \dots, x_i^k(K-1)]$ denote the data symbol vector transmitted from user k in K channel uses during the i th DB in a frame. Let $\mathbf{X}_i = [(\mathbf{x}_i^1)^T, (\mathbf{x}_i^2)^T, \dots, (\mathbf{x}_i^K)^T]^T$ denote the $K \times K$ data matrix formed by the data symbol vectors transmitted by the users in the i th DB during the data phase, i.e., $i = 1, 2, \dots, Q$. The received signal matrix at the BS in the i th DB, i.e., \mathbf{Y}_i of size $N \times K$, is given by

$$\mathbf{Y}_i = \mathbf{H}_c \mathbf{X}_i + \mathbf{N}_i \quad (18)$$

where \mathbf{N}_i is the $N \times K$ noise matrix at the BS during the i th DB. We perform the detection on a vector-by-vector basis using the independence of data symbols transmitted by the users. Let $\mathbf{y}_i^{(t)}$ denote the t th column of \mathbf{Y}_i , $t = 0, 2, \dots, K-1$. Denoting the t th column of \mathbf{X}_i as $\mathbf{x}_i^{(t)} = [x_i^1(t), x_i^2(t), \dots, x_i^K(t)]^T$, we can rewrite the system equation (15) as

$$\mathbf{y}_i^{(t)} = \mathbf{H}_c \mathbf{x}_i^{(t)} + \mathbf{n}_i^{(t)} \quad (19)$$

where $\mathbf{n}_i^{(t)}$ is the t th column of \mathbf{N}_i . The initial channel estimate $\hat{\mathbf{H}}_c$ obtained from (17) is used to detect the transmitted data vectors using the MGS-MR algorithm shown in Section III.

From (15) and (17), we observe that $\hat{\mathbf{H}}_c = \mathbf{H}_c + \mathbf{N}_P / p$. This knowledge about the imperfection of channel estimates is used to calculate the statistics of error-free ML cost required in the MGS-MR algorithm. In Section III, we have observed that, in the case of perfect channel knowledge, the error-free ML cost is nothing but $\|\mathbf{n}^2\|$. In the case of imperfect channel knowledge at the receiver, at channel use t , we have

$$\|\mathbf{y}_i^{(t)} - \hat{\mathbf{H}}_c \mathbf{x}_i^{(t)}\|^2 = \|\mathbf{n}_i^{(t)} - \mathbf{N}_P \mathbf{x}_i^{(t)} / p\|^2.$$

Each entry of the vector $\mathbf{n}_i^{(t)} - \mathbf{N}_P \mathbf{x}_i^{(t)} / p$ has mean zero and variance $2\sigma^2$. Using this knowledge at the receiver, we detect the transmitted data using the MGS-MR algorithm and obtain $\hat{\mathbf{x}}_i^{(t)}$. Let the detected data matrix in DB i be denoted as $\hat{\mathbf{X}}_i = [\hat{\mathbf{x}}_i^{(0)}, \hat{\mathbf{x}}_i^{(1)}, \dots, \hat{\mathbf{x}}_i^{(K-1)}]$.

D. Channel Estimation Using the Gibbs Sampling Algorithm in the Data Phase

Let $\mathbf{Y}_{\text{tot}} = [\mathbf{Y}_P \mathbf{Y}_1 \mathbf{Y}_2 \dots \mathbf{Y}_Q]$, $\mathbf{X}_{\text{tot}} = [\mathbf{X}_P \mathbf{X}_1 \mathbf{X}_2 \dots \mathbf{X}_Q]$, and $\mathbf{N}_{\text{tot}} = [\mathbf{N}_P \mathbf{N}_1 \mathbf{N}_2 \dots \mathbf{N}_Q]$ denote the matrices corresponding to one full frame. We can express \mathbf{Y}_{tot} as

$$\mathbf{Y}_{\text{tot}} = \mathbf{H}_c \mathbf{X}_{\text{tot}} + \mathbf{N}_{\text{tot}}. \quad (20)$$

This system model corresponding to the full frame is converted into a real-valued system model, as done in Section II. That is, (20) can be written as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N} \quad (21)$$

where

$$\begin{aligned} \mathbf{Y} &= \begin{bmatrix} \Re(\mathbf{Y}_{\text{tot}}) & -\Im(\mathbf{Y}_{\text{tot}}) \\ \Im(\mathbf{Y}_{\text{tot}}) & \Re(\mathbf{Y}_{\text{tot}}) \end{bmatrix} \\ \mathbf{H} &= \begin{bmatrix} \Re(\mathbf{H}_c) & -\Im(\mathbf{H}_c) \\ \Im(\mathbf{H}_c) & \Re(\mathbf{H}_c) \end{bmatrix} \\ \mathbf{X} &= \begin{bmatrix} \Re(\mathbf{X}_{\text{tot}}) & -\Im(\mathbf{X}_{\text{tot}}) \\ \Im(\mathbf{X}_{\text{tot}}) & \Re(\mathbf{X}_{\text{tot}}) \end{bmatrix} \\ \mathbf{N} &= \begin{bmatrix} \Re(\mathbf{N}_{\text{tot}}) & -\Im(\mathbf{N}_{\text{tot}}) \\ \Im(\mathbf{N}_{\text{tot}}) & \Re(\mathbf{N}_{\text{tot}}) \end{bmatrix}. \end{aligned}$$

$$\propto \exp(-|g_i|^2) \exp\left(-\frac{\|\mathbf{r} - \sum_{q=1, q \neq i}^{4KN} g_q \widehat{\mathbf{s}}_q - g_i \widehat{\mathbf{s}}_i\|^2}{\sigma^2}\right) \quad (27)$$

$$= \exp\left(-|g_i|^2 - \frac{\|\tilde{\mathbf{r}}^{(i)} - g_i \widehat{\mathbf{s}}_i\|^2}{\sigma^2}\right) \quad (28)$$

$$= \exp\left(-\frac{\|\tilde{\mathbf{r}}^{(i)} - g_i \bar{\mathbf{s}}_i\|^2}{\sigma^2}\right) \quad (29)$$

Equation (21) can be written as

$$\mathbf{Y}^T = \mathbf{X}^T \mathbf{H}^T + \mathbf{N}^T. \quad (22)$$

Vectorizing the matrices \mathbf{Y}^T , \mathbf{H}^T , and \mathbf{N}^T , we define

$$\mathbf{r} \triangleq \text{vec}(\mathbf{Y}^T) \quad \mathbf{g} \triangleq \text{vec}(\mathbf{H}^T) \quad \mathbf{z} \triangleq \text{vec}(\mathbf{N}^T).$$

With the given definitions, (22) can be written in vector form as

$$\mathbf{r} = \underbrace{\mathbf{I}_{2N} \otimes \mathbf{X}^T}_{\triangleq \mathbf{S}} \mathbf{g} + \mathbf{z}. \quad (23)$$

Now, our goal is to estimate \mathbf{g} knowing \mathbf{r} , the estimate of \mathbf{S} , and the statistics of \mathbf{z} using the Gibbs sampling approach. The estimate of \mathbf{S} is obtained as

$$\widehat{\mathbf{S}} = \mathbf{I}_{2N} \otimes \widehat{\mathbf{X}}^T$$

where

$$\widehat{\mathbf{X}} = \begin{bmatrix} \Re(\widehat{\mathbf{X}}_{\text{tot}}) & -\Im(\widehat{\mathbf{X}}_{\text{tot}}) \\ \Im(\widehat{\mathbf{X}}_{\text{tot}}) & \Re(\widehat{\mathbf{X}}_{\text{tot}}) \end{bmatrix}$$

and $\widehat{\mathbf{X}}_{\text{tot}} = [\mathbf{X}_P \ \widehat{\mathbf{X}}_1 \ \widehat{\mathbf{X}}_2 \ \cdots \ \widehat{\mathbf{X}}_Q]$. The initial vector for the algorithm is obtained as

$$\widehat{\mathbf{g}}^{(0)} = \text{vec}(\widehat{\mathbf{H}}^T) \quad (24)$$

where

$$\widehat{\mathbf{H}} = \begin{bmatrix} \Re(\widehat{\mathbf{H}}_c) & -\Im(\widehat{\mathbf{H}}_c) \\ \Im(\widehat{\mathbf{H}}_c) & \Re(\widehat{\mathbf{H}}_c) \end{bmatrix}. \quad (25)$$

1) *Gibbs-Sampling-Based Estimation:* Vector \mathbf{g} is of length $4KN \times 1$. To estimate \mathbf{g} , the algorithm starts with an initial estimate, takes samples from the conditional distribution of each coordinate in \mathbf{g} , and updates the estimate. This is carried out for a certain number of iterations. At the end of the iterations, a weighted average of the previous and current estimates is given as the output.

Let the i th coordinate in \mathbf{g} be denoted by g_i , and let \mathbf{g}_{-i} denote all elements in \mathbf{g} other than the i th element. Let $\widehat{\mathbf{s}}_q$ denote the q th column of $\widehat{\mathbf{S}}$. The conditional probability distribution for the i th coordinate is given by

$$\begin{aligned} p(g_i | \mathbf{r}, \widehat{\mathbf{S}}, \mathbf{g}_{-i}) \\ \propto p(g_i) \cdot p(\mathbf{r} | g_i, \widehat{\mathbf{S}}, \mathbf{g}_{-i}) \end{aligned} \quad (26)$$

where $\tilde{\mathbf{r}}^{(i)} = \mathbf{r} - \sum_{q=1, q \neq i}^{4KN} g_q \widehat{\mathbf{s}}_q$, $\bar{\mathbf{r}}^{(i)} = [\tilde{\mathbf{r}}^{(i)}, 0]^T$, $\bar{\mathbf{s}}_i = [\widehat{\mathbf{s}}_i, \sigma]^T$, and σ^2 is assumed to be known at the receiver. The quantity $\|\tilde{\mathbf{r}}^{(i)} - g_i \bar{\mathbf{s}}_i\|^2$ in (29) is minimized for $g_i = (\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i / \|\bar{\mathbf{s}}_i\|^2$. Hence, we can write

$$\begin{aligned} \|\tilde{\mathbf{r}}^{(i)} - g_i \bar{\mathbf{s}}_i\|^2 &= \left\| \bar{\mathbf{r}}^{(i)} - \left(\frac{(\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i}{\|\bar{\mathbf{s}}_i\|^2} + g_i - \frac{(\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i}{\|\bar{\mathbf{s}}_i\|^2} \right) \bar{\mathbf{s}}_i \right\|^2 \\ &= \left\| \bar{\mathbf{r}}^{(i)} - \frac{(\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i}{\|\bar{\mathbf{s}}_i\|^2} \bar{\mathbf{s}}_i \right\|^2 \\ &\quad + \left(g_i - \frac{(\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i}{\|\bar{\mathbf{s}}_i\|^2} \right)^2 \|\bar{\mathbf{s}}_i\|^2. \end{aligned} \quad (30)$$

Hence

$$p(g_i | \mathbf{r}, \widehat{\mathbf{S}}, \mathbf{g}_{-i}) \propto \exp\left(-\frac{\left(g_i - \frac{(\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i}{\|\bar{\mathbf{s}}_i\|^2}\right)^2}{\frac{\sigma^2}{\|\bar{\mathbf{s}}_i\|^2}}\right) \quad (31)$$

which is Gaussian with mean $\mu_{g_i} = (\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i / \|\bar{\mathbf{s}}_i\|^2$ and variance $\sigma_{g_i}^2 = \sigma^2 / 2 \|\bar{\mathbf{s}}_i\|^2$. Let MAX denote the number of iterations. In each iteration, for each coordinate, the probability distribution specified by its mean and variance has to be calculated to draw samples. Let the mean and variance in the r th iteration and the i th coordinate be denoted as $\mu_{g_i}^{(r)}$ and $\sigma_{g_i}^{2(r)}$, respectively, where $r = 1, 2, \dots, \text{MAX}$, and $i = 1, 2, \dots, 4KN$. We use $\widehat{\mathbf{g}}^{(0)}$ in (24), which is the estimate from the pilot phase, as the initial estimate. In the r th iteration, we obtain $\widehat{\mathbf{g}}^{(r)}$ from $\widehat{\mathbf{g}}^{(r-1)}$ as follows:

- Take $\widehat{\mathbf{g}}^{(r)} = \widehat{\mathbf{g}}^{(r-1)}$.
- Update the i th coordinate of $\widehat{\mathbf{g}}^{(r)}$ by sampling from $\mathcal{N}(\mu_{g_i}^{(r)}, \sigma_{g_i}^{2(r)})$ for all i . Let $\widehat{g}_i^{(r)}$ denote the updated i th coordinate of $\mathbf{g}^{(r)}$.
- Compute weights $\alpha_i^{(r)} = \exp(-(\widehat{g}_i^{(r)} - \mu_{g_i}^{(r)})^2 / 2\sigma_{g_i}^{2(r)})$ for all i . This gives more weight to samples closer to the mean.

After MAX iterations, we compute the final estimate of the i th coordinate, which is denoted by g_i^* , to be the following weighted sum of the estimates from previous and current iterations:

$$g_i^* = \frac{\sum_{r=1}^{\text{MAX}} \alpha_i^{(r)} \widehat{g}_i^{(r)}}{\sum_{r=1}^{\text{MAX}} \alpha_i^{(r)}}. \quad (32)$$

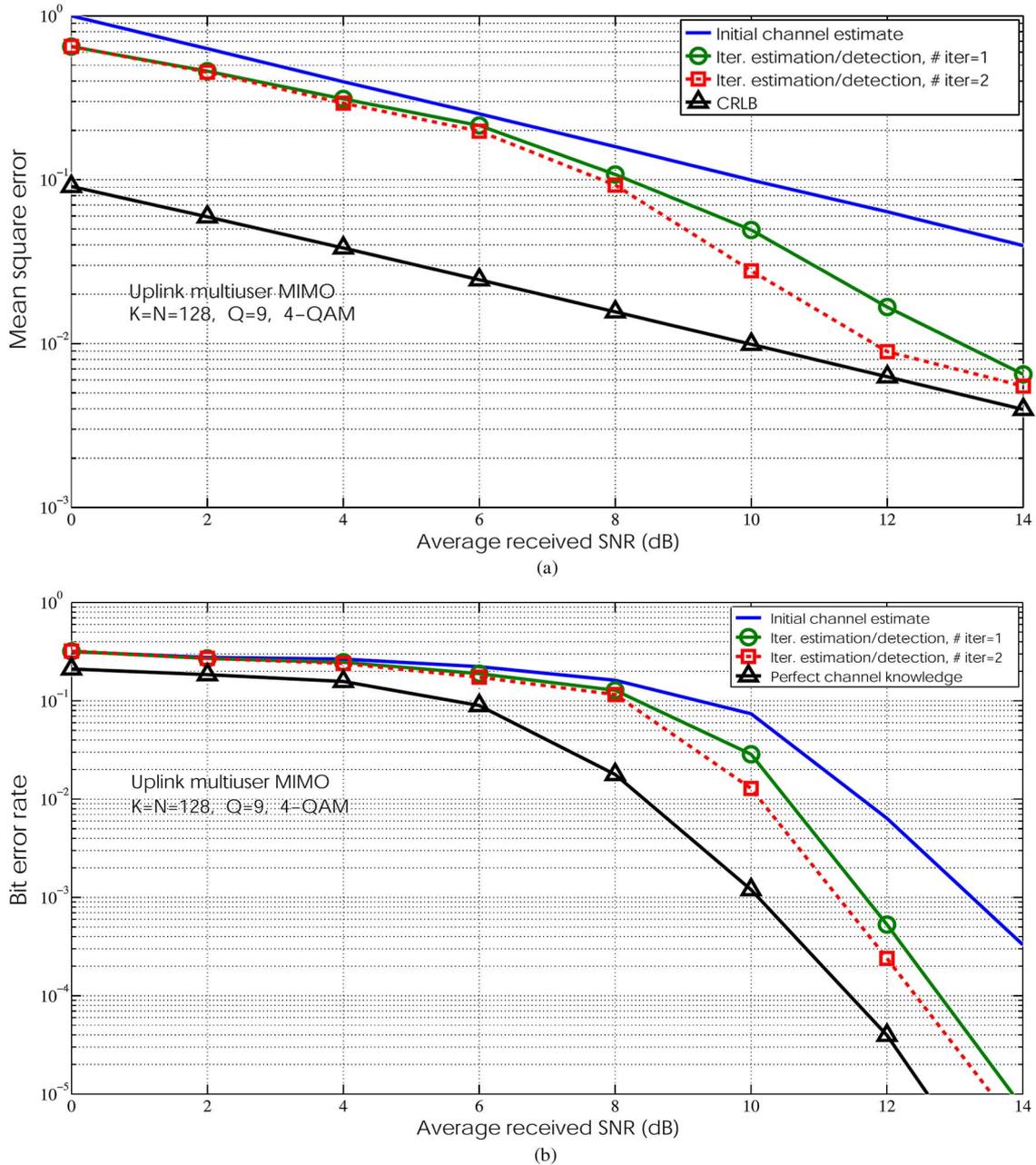


Fig. 15. MSE and BER performance of iterative channel estimation/detection using Gibbs-sampling-based channel estimation and MGS-MR based detection in the uplink multiuser MIMO system with $K = N = 128$, $Q = 9$, and 4-QAM. (a) MSE. (b) BER.

Finally, the updated $2N \times 2K$ channel estimate $\hat{\mathbf{H}}$ is obtained by restructuring $\mathbf{g}^* = [g_1^*, g_2^*, \dots, g_{4KN}^*]^T$ as

$$\hat{\mathbf{H}}(p, q) = g_n^*, \quad p = 1, 2, \dots, 2N; \quad q = 1, 2, \dots, 2K \quad (33)$$

where $n = 2N(p - 1) + q$, and $\hat{\mathbf{H}}(p, q)$ denotes the element in the p th row and q th column of $\hat{\mathbf{H}}$. A full listing of the proposed Gibbs-sampling-based algorithm for channel estimation is given in Algorithm 2.

The matrix $\hat{\mathbf{H}}$ obtained is thus used for data detection using the MGS-MR algorithm. This ends one iteration between channel estimation and detection. The detected data matrix is fed back for channel estimation in the next iteration, whose

output is then used to detect the data matrix again. This iterative channel estimation and detection procedure is carried out for a certain number of iterations.

Algorithm 2 Proposed channel estimation algorithm based on Gibbs sampling

- 1: **input:** $\mathbf{r}, \hat{\mathbf{S}}, \sigma^2, \hat{\mathbf{g}}^{(0)}$: initial vector $\in \mathbb{R}^{4KN}$; MAX: max. no. of iterations;
- 2: $r = 1; \mathbf{g}^*(0) = \hat{\mathbf{g}}^{(0)}; \alpha_i^{(0)} = 0, \forall i = 1, 2, \dots, 4KN$;
- 3: **while** $r < \text{MAX}$ **do**
- 4: $\hat{\mathbf{g}}^{(r)} = \hat{\mathbf{g}}^{(r-1)}$;
- 5: $\tilde{\mathbf{r}}^* = \mathbf{r} - \hat{\mathbf{S}}\hat{\mathbf{g}}^{(r)}$;
- 6: **for** $i = 1$ to $4KN$ **do**

- 7: Compute $\tilde{\mathbf{r}}^{(i)} = \tilde{\mathbf{r}}^* + \hat{g}_i^{(r)} \hat{\mathbf{s}}_i$, $\bar{\mathbf{r}}^{(i)} = [\tilde{\mathbf{r}}^{(i)}, 0]^T$, and $\bar{\mathbf{s}}_i = [\hat{\mathbf{s}}_i, \sigma]^T$;
- 8: Compute $\mu_{g_i}^{(r)} = (\bar{\mathbf{r}}^{(i)})^T \bar{\mathbf{s}}_i / \|\bar{\mathbf{s}}_i\|^2$ and $\sigma_{g_i}^2{}^{(r)} = \sigma^2 / 2\|\bar{\mathbf{s}}_i\|^2$;
- 9: Sample $\hat{g}_i^{(r)} \sim \mathcal{N}(\mu_{g_i}^{(r)}, \sigma_{g_i}^2{}^{(r)})$;
- 10: $\tilde{\mathbf{r}}^* = \tilde{\mathbf{r}}^{(i)} - \hat{g}_i^{(r)} \hat{\mathbf{s}}_i$;
- 11: Compute $\alpha_i^{(r)} = \exp(-(\hat{g}_i^{(r)} - \mu_{g_i}^{(r)})^2 / 2\sigma_{g_i}^2{}^{(r)})$;
- 12: $g_i^*(r) = (\alpha_i^{(r)} \hat{g}_i^{(r)} + (\sum_{z=0}^{r-1} \alpha_i^{(z)}) g_i^*(r-1)) / (\sum_{z=0}^r \alpha_i^{(z)})$;
- 13: **end for**
- 14: $r = r + 1$;
- 15: **end while**
- 16: **output:** $\mathbf{g}^* = \mathbf{g}^*(\text{MAX})$. \mathbf{g}^* : output solution vector

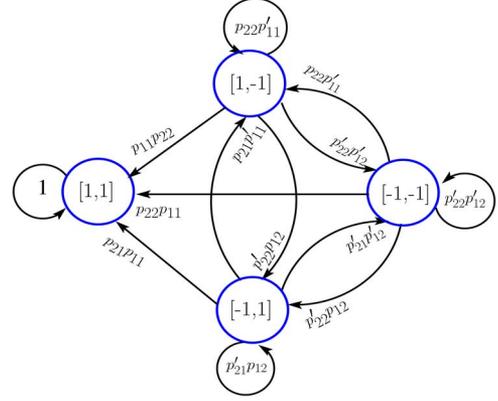


Fig. 16. Markov chain with state transition probabilities for the 2×2 system ($N = K = 2$) with BPSK modulation.

E. Performance Results

In Fig. 15(a), we plot the MSE performance of the iterative channel estimation/detection scheme using the proposed Gibbs-sampling-based channel estimation and MGR-MR-based detection with 4-QAM for $K = N = 128$ and $Q = 9$. In the simulations, the MGS-MR algorithm parameter values used are the same as in Section III-D5. For the channel estimation algorithm, the value of MAX used is 2. The MSEs of the initial channel estimate and the channel estimates after one and two iterations between channel estimation and detection are shown. For comparison, we also plot the Cramer–Rao lower bound (CRLB) for this system. It can be seen that, in the proposed scheme, results show good MSE performance with improved MSE for an increased number of iterations between channel estimation and detection. For the same set of system and algorithm parameters in Fig. 15(a), we plot the BER performance curves in Fig. 15(b). For comparison, we also plot the BER performance with perfect channel knowledge. It can be seen that, with two iterations between channel estimation and detection, the proposed algorithms can achieve 10^{-3} BER within about 1 dB of the performance with perfect channel knowledge.

V. CONCLUSION

We have proposed novel Monte-Carlo-sampling-based detection and channel estimation algorithms that achieved near-optimal performance on the uplink in large-scale multiuser MIMO systems. The proposed MGS detection algorithm with MRs 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: 1) a mixed sampling strategy that gave the algorithm opportunities to quickly exit from stalled solutions and move to better solutions; and 2) random MRs that facilitated the algorithm to seek good solutions in different parts of the solution space. MRs alone (without the random uniform sampling component in the sampling distribution) could not achieve near-ML performance at low complexity. Sampling from the proposed mixture distribution alone (without MRs) could achieve near-ML performance at low complexity in the case of 4-QAM. How-

ever, for higher order QAM (16-/64-QAM) sampling from the mixture distribution alone was not adequate. Joint use of both mixed sampling and MRs was found to be crucial to achieve a near-ML performance for 16-/64-QAM. We also proposed a Gibbs-sampling-based channel estimation algorithm that, in an iterative manner with the MGS-MR detection, achieved performance close to the performance with perfect channel knowledge. We have considered a perfect synchronization and single-cell scenario in this paper. Other system-level issues, including uplink synchronization and multicell operation in large-scale MIMO systems, can be considered as future work.

APPENDIX

ANALYSIS OF THE OPTIMAL CHOICE OF MIXING RATIO q

Here, we analyze the effect of the mixing ratio q in (9) and its optimal choice for the $(\alpha_1 = 1, \alpha_2 = \infty)$ combination using the theory of Markov chains. Note that $q = 0$ corresponds to the conventional Gibbs sampler and that $q = 1$ corresponds to pure random walk. Our analysis approach here is to define an absorbing Markov chain and to use the property of absorbing Markov chains on the expected number of iterations needed to reach the global minima for the first time.

Let us consider the Markov chain where all possible transmit vectors are states and the transition probabilities are calculated from their difference in ML costs. Fig. 16 shows the state-space diagram for a 2×2 system ($N = K = 2$) with binary phase-shift keying (BPSK) modulation alphabet $\mathbb{A} = \{a_1 = 1, a_2 = -1\}$. Let \mathcal{S} denote the set of all possible states, i.e., $\mathcal{S} \triangleq \{[-1, -1]^T, [-1, 1]^T, [1, -1]^T, [1, 1]^T\}$. Without loss of generality, let us assume that $[1, 1]^T$ is the global minima, i.e., the vector with the least ML cost. Let us define $c(\hat{\mathbf{x}}) \triangleq \|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|^2$. We construct the transition probability matrix \mathbf{T} of size $|\mathcal{S}| \times |\mathcal{S}|$, whose (i, j) th element denotes the probability of going from state \mathcal{S}_i to state \mathcal{S}_j . For conventional Gibbs sampling, we have

$$\mathbf{T} = \begin{bmatrix} p'_{22}p'_{12} & p'_{22}p_{12} & p_{22}p'_{11} & p_{22}p_{11} \\ p'_{21}p'_{12} & p'_{21}p_{12} & p_{21}p'_{11} & p_{21}p_{11} \\ p'_{22}p'_{12} & p'_{22}p_{12} & p_{22}p'_{11} & p_{11}p_{22} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (34)$$

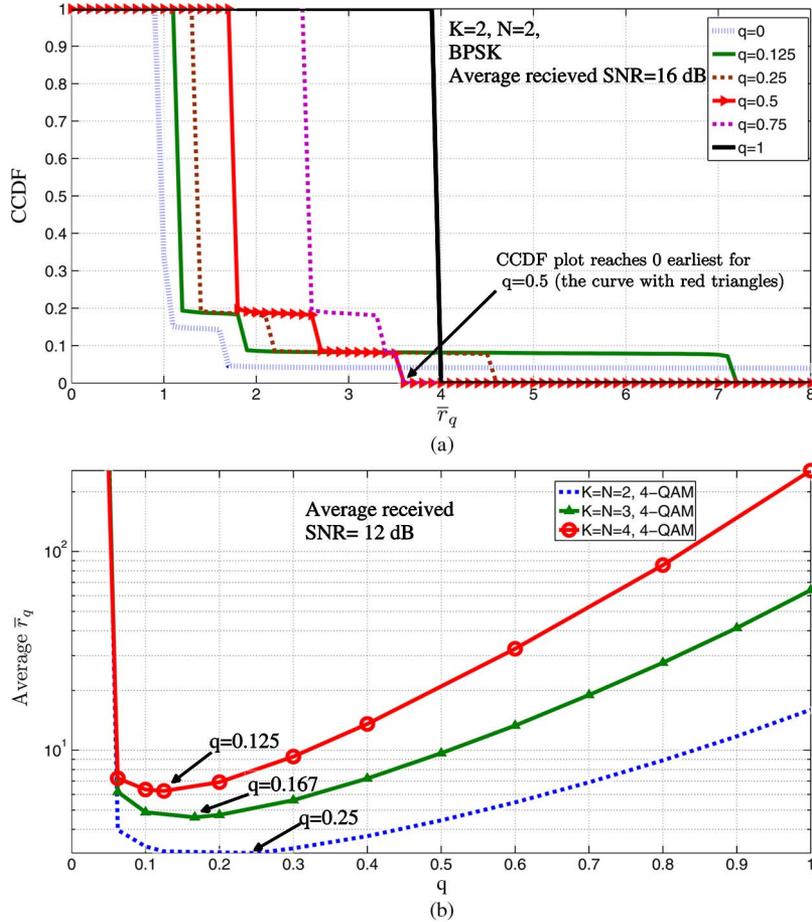


Fig. 17. (a) CCDF of \bar{r}_q for the 2×2 system ($N = K = 2$) with BPSK modulation for different values of q . (b) Average \bar{r}_q as a function of q in 2×2 , 3×3 , and 4×4 systems ($N = K = 2, 3, 4$) with 4-QAM modulation.

where p_{ij} denotes the probability that vector \mathbf{x} with $x_i = a_j$, $j \in \{1, 2\}$, is updated to vector \mathbf{x}' , where the i th coordinate of \mathbf{x}' remains the same as that of \mathbf{x} , and the other coordinate of \mathbf{x}' is 1, and $p'_{i,j} \triangleq 1 - p_{i,j}$. Based on this, p_{11} can be written as

$$p_{11} = \frac{1}{1 + \exp\left(\frac{c([1,1]^T) - c([1,-1]^T)}{\sigma^2}\right)}. \quad (35)$$

Expressions for other $p_{i,j}$ s can be written likewise. The global minima, i.e., state $[1, 1]^T$, is considered an absorbing state as we keep track of the best vector visited so far, and we are interested in finding the number of iterations required to reach global minima for the first time. This makes the Markov chain an absorbing Markov chain. A basic property of an absorbing Markov chain [42, Ch. 11, pp. 418–419] is that the expected number of iterations before being absorbed when starting in transient state S_i is the i th entry of vector

$$\mathbf{r} = \mathbf{N}\mathbf{1} \quad (36)$$

where $\mathbf{1}$ is a vector whose entries are all 1, and \mathbf{N} is the fundamental matrix of the Markov chain, which is given by

$$\mathbf{N} = \sum_{k=0}^{\infty} \mathbf{Q}^k = (\mathbf{I} - \mathbf{Q})^{-1} \quad (37)$$

where \mathbf{Q} is a $(|\mathcal{S}| - 1) \times (|\mathcal{S}| - 1)$ submatrix of \mathbf{T} , such that

$$\mathbf{T} = \begin{bmatrix} \mathbf{Q} & \mathbf{t} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}. \quad (38)$$

Let \bar{r} denote the average of the entries in \mathbf{r} . For a given realization of \mathbf{H} and \mathbf{n} , \bar{r} gives the expected number of steps to reach the global minima averaged over starting points.

Now, depending on the channel matrix and noise vector realization, the ML costs of the states may be ordered in the following three ways (because of symmetry, $[1, -1]^T$ and $[-1, 1]^T$ can exchange places):

- Case 1: $c([1, 1]^T) < c([1, -1]^T) < c([-1, 1]^T) < c([-1, -1]^T)$;
- Case 2: $c([1, 1]^T) < c([1, -1]^T) < c([-1, -1]^T) < c([-1, 1]^T)$;
- Case 3: $c([1, 1]^T) < c([-1, -1]^T) < c([-1, 1]^T) < c([1, -1]^T)$.

As $\sigma \rightarrow 0$, the probability of going from a state to its neighboring state with a lower ML cost tends to 1. Hence, in Cases 1 and 2, the Markov chain starting from any state will reach the global minima in two iterations with high probability. Note that, in Case 3, as $\sigma \rightarrow 0$, $p_{11} \rightarrow 1$, $p_{21} \rightarrow 1$, $p_{12} \rightarrow 0$, and $p_{22} \rightarrow 0$. Therefore, in Case 3, if the Markov chain starts in $[-1, -1]^T$ or $[1, -1]^T$, then the chain is trapped in state $[-1, -1]^T$, i.e.,

$$\mathbf{T}_q = \begin{bmatrix} (q'p'_{22} + \frac{q}{2}) & (q'p'_{12} + \frac{q}{2}) & (q'p'_{22} + \frac{q}{2}) & (q'p'_{12} + \frac{q}{2}) & (q'p'_{22} + \frac{q}{2}) & (q'p'_{11} + \frac{q}{2}) & (q'p'_{22} + \frac{q}{2}) & (q'p'_{11} + \frac{q}{2}) \\ (q'p'_{21} + \frac{q}{2}) & (q'p'_{12} + \frac{q}{2}) & (q'p'_{21} + \frac{q}{2}) & (q'p'_{12} + \frac{q}{2}) & (q'p'_{21} + \frac{q}{2}) & (q'p'_{11} + \frac{q}{2}) & (q'p'_{21} + \frac{q}{2}) & (q'p'_{11} + \frac{q}{2}) \\ (q'p'_{22} + \frac{q}{2}) & (q'p'_{12} + \frac{q}{2}) & (q'p'_{22} + \frac{q}{2}) & (q'p'_{12} + \frac{q}{2}) & (q'p'_{22} + \frac{q}{2}) & (q'p'_{11} + \frac{q}{2}) & (q'p'_{22} + \frac{q}{2}) & (q'p'_{11} + \frac{q}{2}) \\ 0 & & 0 & & 0 & & 0 & 1 \end{bmatrix} \quad (38a)$$

the probability of coming out of this state is very low. This means that, with $q = 0$ (i.e., in conventional Gibbs sampling), the expected number of iterations needed to reach the global minima for the first time will be high if the chain starts from the states $[-1, -1]^T$ or $[1, -1]^T$.

Now, we write the transition probability matrix for the proposed MGS, which is denoted by \mathbf{T}_q , as a function of q , as in (38a), shown at the top of the page, where $q' \triangleq 1 - q$. Note that \mathbf{T}_q specializes to \mathbf{T} in (34) for $q = 0$. \mathbf{T}_q for systems with more number of antennas and other modulation alphabets can be written likewise. For a given realization of \mathbf{H} and \mathbf{n} , the expected number of iterations to reach the global minima for the first time averaged out over the starting points (i.e., \bar{r}_q) can be computed using (38a) and (36).

We computed \bar{r}_q from \mathbf{T}_q for multiple realizations of \mathbf{H} and \mathbf{n} , and obtained the complementary cumulative density function (CCDF) of \bar{r}_q . In Fig. 17(a), we plot the CCDF of \bar{r}_q for $q = 0, 0.125, 0.25, 0.5, 0.75$, and 1 , at an average SNR of 16 dB. The ordinate denotes the probability of not reaching the global minima (i.e., the ML solution). In Fig. 17(a), it can be observed that the CCDF plot reaches zero earliest when $q = 0.5$ compared with other values of q . Note that $q = 0.5$ is the same as the inverse of the number of real dimensions in the 2×2 system ($N = K = 2$) with BPSK, which is 2. Next, in Fig. 17(b), we plot the \bar{r}_q averaged out over \mathbf{H} and \mathbf{n} as a function of q at an average SNR of 12 dB for $2 \times 2, 3 \times 3$, and 4×4 systems ($N = K = 2, 3, 4$) with 4-QAM modulation, where the number of real dimensions is $2K$. The following interesting observations can be made in Fig. 17(b). First, the conventional Gibbs sampler is not optimum, i.e., $q = 0$ does not result in minimum average \bar{r}_q . Second, in pure random walk (i.e., sampling from uniform distribution), the average \bar{r}_q is nothing but 2^{n_d} , where n_d is the number of dimensions in the system; we can see that this result is captured by the analysis by noting in Fig. 17(b) that, for $q = 1$, the average $\bar{r}_q = 16, 64, 256$ for $2 \times 2, 3 \times 3, 4 \times 4$ systems with 4-QAM (i.e., $n_d = 4, 6, 8$), respectively. Finally, the average minimum \bar{r}_q occurs at the inverse of the number of dimensions in the system; this can be seen by noting that the minimum average \bar{r}_q occurs at $q = 1/4 = 0.25, q = 1/6 = 0.167$, and $q = 1/8 = 0.125$ for $2 \times 2, 3 \times 3$, and 4×4 systems with 4-QAM, respectively.

ACKNOWLEDGMENT

The authors would like to thank Prof. R. Sundaresan of the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India, and Prof. V. Borkar of the Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India, for the useful discussions on MCMC techniques.

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, no. 3, pp. 311–335, Mar. 1998.
- [2] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.
- [3] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
- [4] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
- [5] H. Bölcskei, D. Gesbert, and C. B. Papadias, Eds., *Space-Time Wireless Systems: From Array Processing to MIMO Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
- [6] T. L. Marzetta, "How much training is required for multiuser MIMO?" in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Oct./Nov. 2006, pp. 359–363.
- [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. J.*, vol. 9, no. 2, pp. 25–31, Sep. 2007.
- [8] K. V. Vardhan, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity detector for large MIMO systems and multicarrier CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 473–485, Apr. 2008.
- [9] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity precoder for large multiuser MISO systems," in *Proc. IEEE VTC*, Singapore, May 2008, pp. 797–801.
- [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 J. Sel. Topics Signal. Process.*, vol. 3, no. 6, pp. 958–974, Dec. 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," in *Proc. ISCCSP*, Limassol, Cyprus, Mar. 2010, pp. 1–6.
- [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, Nov. 2010.
- [13] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
- [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, Aug. 2011.
- [15] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprasad, "Achieving 'massive MIMO' spectral efficiency with a not-so-large number of antennas," *arXiv:1107.3862v2 [cs.IT]*, pp. 1–38, Sep. 13, 2011.
- [16] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, 2013, to be published.
- [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," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
- [18] B. Cerato and E. Viterbo, "Hardware implementation of a low-complexity detector for large MIMO," in *Proc. IEEE ISCAS*, Taipei, Taiwan, May 2009, pp. 593–596.
- [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. Commun.*, vol. 59, no. 11, pp. 2955–2963, Nov. 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. Lett.*, vol. 14, no. 12, pp. 1107–1109, Dec. 2010.

- [22] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "Low-complexity near-MAP decoding of large non-orthogonal STBCs using PDA," in *Proc. IEEE ISIT*, Seoul, South Korea, Jun./Jul. 2009, pp. 1998–2002.
- [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 J. Sel. Topics Signal Process.*, vol. 5, no. 8, pp. 1497–1511, Dec. 2011.
- [24] J. Goldberger and A. Leshem, "MIMO detection for high-order QAM based on a Gaussian tree approximation," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4973–4982, Aug. 2011.
- [25] C. Knivel, M. Noemm, and P. A. Hoeher, "Low complexity receiver for large-MIMO space time coded systems," in *Proc. IEEE VTC-Fall*, San Francisco, CA, USA, Sep. 2011, pp. 1–5.
- [26] R. Chen, J. S. Liu, and X. Wang, "Convergence analyses and comparisons of Markov chain Monte Carlo algorithms in digital communications," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 255–270, Feb. 2002.
- [27] B. Farhang-Boroujeny, H. Zhu, and Z. Shi, "Markov chain Monte Carlo algorithms for CDMA and MIMO communication systems," *IEEE Trans. Signal Process.*, 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 detection methods," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 669–672, Oct. 2005.
- [29] X. Mao, P. Amini, and B. Farhang-Boroujeny, "Markov chain Monte Carlo MIMO detection methods for high signal-to-noise ratio regimes," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Nov. 2007, pp. 3979–3983.
- [30] S. Akoum, R. Peng, R.-R. Chen, and B. Farhang-Boroujeny, "Markov chain Monte Carlo detection methods for high SNR regimes," in *Proc. IEEE ICC*, Dresden, Germany, Jun. 2009, pp. 1–5.
- [31] R.-R. Chen, R. Peng, and B. Farhang-Boroujeny, "Markov chain Monte Carlo: Applications to MIMO detection and channel equalization," in *Proc. IEEE ITA*, San Diego, CA, USA, Feb. 2009, pp. 44–49.
- [32] R. Peng, R.-R. Chen, and B. Farhang-Boroujeny, "Markov chain Monte Carlo detectors for channels with intersymbol interference," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2206–2217, Apr. 2010.
- [33] M. Hansen, B. Hassibi, A. G. Dimakis, and W. Xu, "Near-optimal detection in MIMO systems using Gibbs sampling," in *Proc. IEEE GLOBECOM*, Honolulu, HI, USA, Dec. 2009, pp. 1–6.
- [34] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
- [35] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. New York, NY, USA: Springer-Verlag, 2004.
- [36] O. Haggstrom, *Finite Markov Chains and Algorithmic Applications*. Cambridge, U.K.: 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," in *Proc. IEEE ICC*, Kyoto, Japan, Jun. 2011, pp. 1–5.
- [38] T. Datta, N. Ashok Kumar, A. Chockalingam, and B. S. Rajan, "A novel MCMC algorithm for near-optimal detection in large-scale uplink multiuser MIMO systems," in *Proc. ITA*, San Diego, CA, USA, Feb. 2012, pp. 69–77.
- [39] M. K. Hanawal and R. Sundaresan, Randomised Attacks on Passwords, DRDO-IISc Programm. on Adv. Res. in Math. Eng., Bangalore, India, Tech. Rep. TR-PME-2010-11. [Online]. Available: http://www.pal.ece.iisc.ernet.in/PAM/docs/techreports/tech_rep10/TR-PME-2010-11.pdf
- [40] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 70–78, Jan. 2011.
- [41] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 99–105, Jan. 1996.
- [42] C. M. Grinstead and J. L. Snell, *Introduction to Probability*, 2nd ed. Providence, RI, USA: Amer. Math. Soc., 1997.



Tanumay Datta received the Bachelor of Engineering degree in electronics and telecommunication engineering from Jadavpur University, Kolkata, India, in 2008. He is currently working toward the Ph.D. degree in electrical communication engineering from the Indian Institute of Science, Bangalore, India.

His current research interests include low-complexity near-optimal detection and estimation algorithms for large multiple-input–multiple-output (MIMO) systems, multiuser MIMO systems, spatial modulation, and wireless relay networks.



N. Ashok Kumar was born in Hyderabad, India. He received the Bachelor of Engineering degree in electronics and communication engineering from the Chaitanya Bharathi Institute of Technology, Hyderabad, in 2009 and the Master of Engineering degree in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 2011.

Since July 2011, he has been working with Broadcom India Research Pvt. Ltd., Bangalore. His research interests include low-complexity detection for large multiple-input–multiple-output systems and wireless network coding.



A. Chockalingam (S'92–M'93–SM'98) was born in Rajapalayam, India. He received the B.E. (Hons.) degree in electronics and communication engineering from the P. S. G. College of Technology, Coimbatore, India, in 1984; the M.Tech. degree (with specialization in satellite communications) from the Indian Institute of Technology, Kharagpur, India, in 1985; and the Ph.D. degree in electrical communication engineering from the Indian Institute of Science (IISc), Bangalore, India, in 1993.

From 1986 to 1993, he worked with the Transmission Research and Development Division, Indian Telephone Industries Ltd., Bangalore. From December 1993 to May 1996, he was a Postdoctoral Fellow and an Assistant Project Scientist with the Department of Electrical and Computer Engineering, University of California, San Diego, CA, USA. From May 1996 to December 1998, he was a Staff Engineer/Manager with the Systems Engineering Group, Qualcomm, Inc., San Diego. Since December 1998, he has been with the Department of Electrical and Communication Engineering, IISc, where he is currently a Professor, working in the area of wireless communications and networking.

Dr. Chockalingam is a Fellow of the Indian National Academy of Engineering, the National Academy of Sciences, India, and the Indian National Science Academy. He served as an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from May 2003 to April 2007. He currently serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He served as a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (*Special Issue on Multiuser Detection for Advanced Communication Systems and Networks*). He also served as a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (*Special Issue on Soft Detection on Wireless Transmission*). He received the Swarnajayanti Fellowship from the Department of Science and Technology, Government of India.



B. Sundar Rajan (S'84–M'91–SM'98) was born in Tamil Nadu, India. He received the B.Sc. degree in mathematics from Madras University, Madras, India, in 1979; the B.Tech degree in electronics from Madras Institute of Technology, Madras, in 1982; and the M.Tech and Ph.D. degrees in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1984 and 1989, respectively.

From 1990 to 1997, he was a Faculty Member with the Department of Electrical Engineering, Indian Institute of Technology, New Delhi, India. Since 1998, he has been a Professor with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India. His primary research interests include space–time coding for multiple-input–multiple-output channels, distributed space–time coding and cooperative communication, coding for multiple-access and relay channels, and network coding.

Dr. Rajan is a member of the American Mathematical Society. He is a Fellow of the Indian National Academy of Engineering, the National Academy of Sciences, India, and the Indian National Science Academy. He is currently an Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS and an Associate Editor of Coding Theory for IEEE TRANSACTIONS ON INFORMATION THEORY. He was an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY (2008–2011) and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2007–2011). He served as Technical Program Co-chair of the IEEE Information Theory Workshop held in Bangalore in 2002. He received the Prof. Rustom Choksi Award from the Indian Institute of Science for excellence in research in engineering for the year 2009, the Institution of Electronics and Telecommunication Engineers Pune Center's S.V.C. Aiya Award for Telecom Education in 2004, and the Best Academic Paper Award at the 2011 IEEE Wireless Communications and Networking Conference.