Markov Chain Monte Carlo Detection Methods for High SNR Regimes

Salam Akoum, Ronghui Peng, Rong-Rong Chen and Behrouz Farhang-Boroujeny

Department of Electrical and Computer Engineering

University of Utah, Salt Lake City, Utah 84112

e-mail: {akoum, peng, rchen and farhang}@ece.utah.edu

Abstract—**Statistical detectors that are based on Markov chain Monte Carlo (MCMC) simulators have emerged as promising low-complexity solutions to both multiple-input multiple-output (MIMO) and code division multiple access (CDMA) communication systems. While these types of detectors achieve unprecedented near capacity performance, i.e., when operated in low signal-to-noise ratio (SNR) regime, they exhibit a serious problem at medium to high SNR regimes, referred to as the "stalling" problem. In this paper, we investigate the sources of this degradation and propose a new search strategy called *constrained MCMC* to remedy the issue of stalling.**

Index Terms—**MIMO channel, MCMC detection, LLR values, ML detection, Sphere Decoding, Zero forcing.**

I. INTRODUCTION

Code division multiple access (CDMA) and multiple-input multiple-output (MIMO) communication systems are two different, but fundamentally very similar communication techniques. To achieve the full capacity of the channel, the multiuser data, in a CDMA system, as well as the parallel data substreams, in a MIMO system, have to be jointly detected at the receiver [1]. The detector generates soft information in the form of log likelihood ratio (LLR) values and passes them to the channel decoder to make the final decision on the transmitted uncoded bits. The soft information may be cycled back from the channel decoder to the detector in what is known as a turbo loop to improve on the performance of the receiver [2]. The channel code can be a simple convolutional code or a more advanced turbo or a low-density parity check code (LDPC).

The optimum MIMO detector - the maximum likelihood (ML) detector - performs an exhaustive search over all vector symbols, rendering the decoding complexity of the receiver exponential in the number of bits per channel use[1]. To avoid this complexity, researchers have proposed a number of suboptimum detectors such as zero forcing (ZF) [3], minimum mean square error (MMSE) [4], and ordered successive interference cancellation (SIC) - MMSE equalizers [5]. These methods reduce the complexity of detectors at the cost of significant performance loss. Meanwhile, to achieve near-capacity performance, more elegant detectors have been proposed. The list sphere decoding (LSD) [2] and other tree search methods [6] form a class of detectors whose goal is to select only a subset of the bit combinations at each channel use as a candidate list, and use this list for the computation of the soft information out of the detector. The candidate list is obtained through a deterministic approach. Although the size of the list may be significantly smaller than the signal space (the number of all possible bit combinations), it still grows exponentially with the number of bits per channel use [6].

The Markov chain Monte Carlo (MCMC) simulation, [7], is an alternative search technique that may also be used to generate a candidate list [8], [9]. This method is different from the tree search methods in two ways: (i) it is a stochastic search; (ii) the growth of the size of the list and thus the complexity of the MIMO detector is not exponential in the number of bits per channel use. In fact, the complexity of the MCMC detector only grows slightly faster than linear. However, past studies have shown that while the MCMC detector performs very well in low SNR (near capacity) regime, it suffers from an error floor or a degradation in performance as SNR increases. The authors in [10] noted that this problem is alleviated when the underlying Markov chain is properly initialized. This controlled initialization strategy was however unable to completely remove the error floor at high SNR, without significantly increasing the complexity of the detector. In this paper, we approach the problem from another angle, focusing on the sources of the degradation of the MCMC algorithm in terms of the probability of transition in the Markov chain. We propose a number of amendments that lead to significantly improved results with a relatively small number of samples. We note that the complexity of MIMO/CDMA detectors is proportional to the number of samples extracted by the Gibbs samples (GSs) - the mechanism used for implementation of MCMC - and thus reduction of the number of samples is equivalent to reduction of the cost, e.g., power and silicon area in an ASIC design.

This paper is organized as follows. The system model is introduced in Section II. Without loss of generality, we restrict our discussion to a MIMO system; extension to a CDMA system is straightforward. In Section III, we introduce the methods used to compute the LLR values. The shortcomings of the solutions proposed in [8] and [10] are discussed in Section IV. The novel search strategies proposed in this work are presented in Section V. Simulation results are presented in Section VI. The concluding remarks and some further comments are made in Section VII.

---

[1]The number of bits per channel use is equal to the number of users/transmit antennae times the constellation size of the transmitted data symbols.

II. SYSTEM MODEL

We consider a MIMO system with M_t antennae at the transmitter and M_r antennae at the receiver. At the transmitter, the information word is first encoded by a forward error correction (FEC) code. The output of the channel encoder, after passing through an interleaver, is divided into blocks of M bits. These blocks form a vector sequence $\mathbf{b}(n) = [b_1(n) \ b_2(n) \ \cdots \ b_M(n)]^\mathrm{T}$, where n is the time index. Each $\mathbf{b}(n)$ is then mapped to the transmit symbol $\mathbf{d}(n) = [d_1(n) \ d_2(n) \ \cdots \ d_{M_\mathrm{t}}(n)]^\mathrm{T}$. We assume that each element of $\mathbf{d}(n)$ carries $M_c = M/M_\mathrm{t}$ coded bits and thus is chosen from a 2^{M_c}-ary QAM/PSK constellation. We note that each value of n corresponds to one channel use and, hence, during each channel use, M coded bits are being transmitted.

Assuming a flat fading channel, the received signal at each channel use can be modeled as

$$\mathbf{y} = \mathbf{Hd} + \mathbf{n}, \tag{1}$$

where $\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_{M_\mathrm{r}}]^\mathrm{T}$ is an $M_\mathrm{r} \times 1$ vector, $\mathbf{H}$ is the $M_\mathrm{r} \times M_\mathrm{t}$ channel gain matrix assumed perfectly known at the receiver and $\mathbf{n}$ is the channel noise. We assume that $\mathbf{n}$ is a Gaussian vector with zero mean and covariance matrix $E[\mathbf{nn}^\mathrm{H}] = \sigma_n^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and σ_n^2 is the variance of each element of $\mathbf{n}$.

III. THE DETECTION METHODS

The LLR values generated by the MIMO detector at the receiver can be expressed as

$$\lambda_1(b_k) = \ln \frac{P(b_k = +1 | \mathbf{y}, \lambda_2^e(\mathbf{b}))}{P(b_k = -1 | \mathbf{y}, \lambda_2^e(\mathbf{b}))} \tag{2}$$

where b_k is the kth element of $\mathbf{b}$, and $\lambda_2^e(\mathbf{b})$ is the vector of extrinsic information (if available) from the channel decoder, on the elements of $\mathbf{b}$. The extrinsic information out of the detector, $\lambda_1^e(b_k) = \lambda_1(b_k) - \lambda_2^e(b_k)$, is then formed and passed to the channel decoder. Following the max-log approximation, the LLR value expression can be written as, [2]

$$\lambda_1^e(b_k) = \max_{\mathbf{b} \in \beta_k^+} \left\{ -\frac{||\mathbf{y} - \mathbf{Hd}||^2}{2\sigma^2} + \frac{1}{2} \mathbf{b}_{-k}^T \lambda_2^e(\mathbf{b}_{-k}) \right\}$$
$$- \max_{\mathbf{b} \in \beta_k^-} \left\{ -\frac{||\mathbf{y} - \mathbf{Hd}||^2}{2\sigma^2} + \frac{1}{2} \mathbf{b}_{-k}^T \lambda_2^e(\mathbf{b}_{-k}) \right\}. \tag{3}$$

β_k^+ and β_k^- are, respectively, the set of symbol vectors having $b_k = +1$ and $b_k = -1$. In the absence of extrinsic information from the channel decoder, i.e., when $\lambda_2^e(\mathbf{b}_{-k}) = 0$, (3) reduces to

$$\lambda_1(b_k) = \frac{1}{2\sigma^2} \left(\min_{\mathbf{b} \in \beta_k^-} ||\mathbf{y} - \mathbf{Hd}||^2 - \min_{\mathbf{b} \in \beta_k^+} ||\mathbf{y} - \mathbf{Hd}||^2 \right). \tag{4}$$

An exact implementation of (3) or (4) has a complexity that grows exponentially with M. This obviously will become prohibitively complex as the number of bits per channel use increases. For instance, if we consider a 4 transmit, 4 receive

antennae MIMO system, employing 64-QAM modulation, there will be a total of $4 \times 6 = 24$ bits per channel use and consequently $16,777,216$ possible signals to query from. Fortunately, there are ways around this complexity issue. The key idea is to develop a search algorithm that limits the choices of $\mathbf{b}$ (or, equivalently, $\mathbf{d}$) to those corresponding to the small values of the norm $||\mathbf{y} - \mathbf{Hd}||^2$. For the LSD method, this is done by searching for the points that satisfy the inequality $||\mathbf{y} - \mathbf{Hd}|| < r$, i.e., by confining the search area to the points within a hyper-sphere with a radius r, [2]. In the MCMC algorithm, on the other hand, a stochastic search procedure is followed to limit the search space of $\mathbf{b}$. The Gibbs sampler algorithm, [8], randomly crawls through the transmitted symbol space searching for the signals that minimize the norm $||\mathbf{y} - \mathbf{Hd}||^2$, and consequently forming subsets to replace β_k^+ and β_k^- in (4).

IV. THE DILEMMA OF HIGH SNR

The authors in [8], [10], [11] noted the problem of MCMC in high SNR regimes and proposed several methods to work around it. They argued that, at high SNR, the Markov chain associated with the MCMC method becomes nearly reducible, i.e., the transition probabilities out of certain states or a cluster of states become fairly negligible. Hence, the GS algorithm, if initialized randomly, may iterate between states that do not include the signals that result in small distances $||\mathbf{y} - \mathbf{Hd}||^2$.

To improve the performance of the detector, the intuitive solution was first to run several GSs in parallel, each initialized randomly from a different state, [8]. The idea is to provide several initial points for the GS so as to increase the chance that the algorithm iterates close to the optimum solution. This method, although greatly helps in reducing the error floor, does not guarantee that the MCMC detector will converge to the right solution. Other solutions were therefore suggested to solve the problem. One strategy, also proposed in [8], assumes a higher noise variance than the actual one when running the GSs but uses the actual noise variance when calculating the LLR values. This method proves useful for medium values of SNR but loses its effectiveness as SNR increases.

The authors in [10] suggested another search strategy based on clever initialization of the GSs. They proved empirically that the problem is alleviated if one of the parallel GSs is initialized to a state that corresponds to a small norm $||\mathbf{e}||^2 = ||\mathbf{y} - \mathbf{Hd}||^2$. One simple, yet effective, initialization that achieves this goal is the zero-forcing (ZF) solution. Other techniques, such as the MMSE, may also be used for initialization, [10]. Moreover, while the solution from ZF or MMSE or any other such method for initialization of one of the GSs improves the performance of the MIMO detector, a number of other GSs, randomly initialized, are also needed in parallel, for boosting the performance of the MCMC.

The first objective of this paper is to develop insights into the operation of the MCMC simulators that will guide us to the root of the problem of the MCMC detector, and justify why the above solutions do or do not achieve good performance.
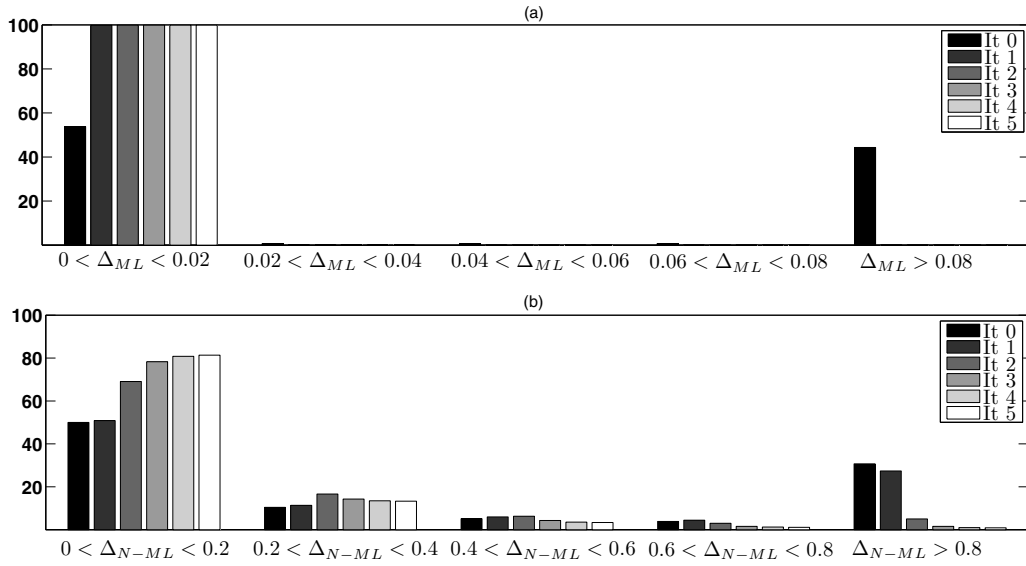
Fig. 1. Bar chart obtained using 5 GSs (1 initialized with ZF, 4 initialized randomly) at 30 dB for a 2 Tx, 2 Rx, 64 QAM system. The figure shows the distances for the different iterations of the GS. (a)Δ_{ML}. (b)$\Delta_{\mathrm{N-ML}}$

To this end, we make the key observation that

$$\min\left(\min_{\mathbf{d}_{k+}}||\mathbf{y}-\mathbf{H}\mathbf{d}_{k+}||^2, \min_{\mathbf{d}_{k-}}||\mathbf{y}-\mathbf{H}\mathbf{d}_{k-}||^2\right)$$
$$= \min_{\mathbf{d}}||\mathbf{y}-\mathbf{H}\mathbf{d}||^2. \qquad (5)$$

In other words, the smaller of the two norms on the right-hand side of (4) is equal to the minimum value of $||\mathbf{y}-\mathbf{H}\mathbf{d}||^2$ over all possible choices of $\mathbf{d}$ (or, equivalently, $\mathbf{b}$). We recall that the vector $\hat{\mathbf{d}}$ that minimizes $||\mathbf{y}-\mathbf{H}\mathbf{d}||^2$ is called the maximum likelihood (ML) estimate of $\mathbf{d}$. Therefore, for convenience of reference, we call the minimum of the two norms in (4) the *ML norm* and the corresponding maximum, the *non-ML (N-ML) norm*.

Furthermore, we note that the GS, by design, searches for the states that correspond to the smaller values of $||\mathbf{y}-\mathbf{H}\mathbf{d}||^2$. Hence, there is a good probability that at least one of the parallel GSs finds the ML norm correctly. However, it is quite likely that all of the GSs miss a fair percentage of the N-ML norms. Moreover, the use of the ZF solution for initialization of one of the GSs is most beneficial in increasing the chance of finding the ML norm correctly. The use of randomized GSs is intended to try other parts of the Markov chain, in an effort to find good estimates of a larger percentage of the N-ML norms.

To quantify these observations, we present empirical results that show the ability of the MCMC detector in [10] to compute satisfactory ML and N-ML norms. We consider a two transmit, two receive MIMO system employing 64-point gray coded QAM alphabet. We find the actual ML and N-ML norms, D_{ML} and $D_{\mathrm{N-ML}}$, respectively, and we compare them with their respective estimates obtained through the MCMC detector. We

evaluate the differences

$$\Delta_{\mathrm{ML}} = \hat{D}_{\mathrm{ML}} - D_{\mathrm{ML}}, \qquad (6)$$

and

$$\Delta_{\mathrm{N-ML}} = \hat{D}_{\mathrm{N-ML}} - D_{\mathrm{N-ML}}, \qquad (7)$$

for different channel realizations $\mathbf{H}$. Following the methods in [10], five GSs are run in parallel. One GS is initialized using the ZF solution and the four others are left uncontrolled. Fig. 1 presents bar charts that show how Δ_{ML} and $\Delta_{\mathrm{N-ML}}$ evolve as the number of iterations of five parallel GSs increases. As seen, the ML distance is calculated correctly for a very large percentage of the channel realizations, suggesting that the MCMC problem is almost completely solved when proper initializations such as the ZF solution is used. However, Fig. 1(b) tells a different story. The MCMC detector, as used in [10], is unable to find good estimates for the N-ML norms. This translates into an inability to compute reliable soft information and therefore, a loss in performance. We further note that increasing the GS complexity by running it for a large number of iterations does not help. Hence, a more solid method has to be used to boost the MCMC capabilities.

V. SEARCH STRATEGIES FOR ESTIMATING THE N-ML NORMS

We devise a method to control the two factors of the soft information output of the MCMC algorithm without increasing the complexity of the MIMO detector. We benefit from initializing one of the Gibbs samplers by the ZF solution to compute the ML norm and invest the rest of the GSs in searching for the N-ML norms.

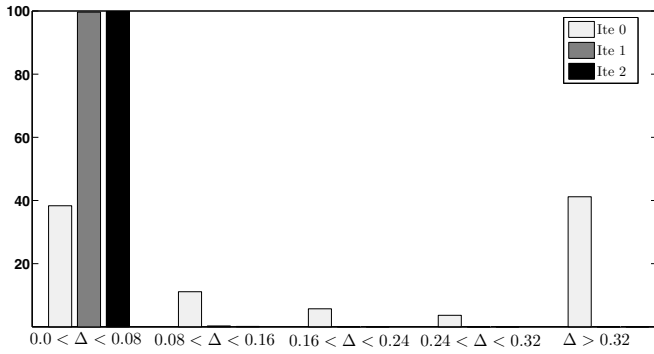Starting with the ZF solution, we run the GS for one or two iterations in order to get a good estimate of the ML

Fig. 2. Bar chart of the distances $\Delta = \Delta_{\mathrm{N-ML}}$ obtained using the C-MCMC algorithm at 30 dB for a 2 Tx, 2 Rx, 64 QAM system.

distance. We call this estimate the ML state. We then use this ML state to obtain better N-ML norms. To this end, we run a number of GSs in parallel, one for each bit of **b**. For the k-th GS, we flip the k-th bit of the ML state and keep it fixed for all the steps in the GS iteration. The k-th GS is used to refine the N-ML norm estimation for the k-th bit. Furthermore, following the near reducibility of the Markov chain, it is necessary to ensure that the distance $||\mathbf{y} - \mathbf{Hd}||^2$ is minimally disturbed away from the ML norm. This implies that, in addition to flipping the k-th bit, other bits in **b** may need to be adjusted according to Gray coding, such that the resulting vector stays in the neighborhood of the ML state. Such neighboring states can be tabulated and a table look-up method can be used for an efficient implementation of the MCMC detector. This manoeuver, as witnessed by simulations results, guarantees that we stay in the vicinity of the optimum state that we are seeking. We call this new MCMC algorithm the constrained MCMC (C-MCMC).

To quantify the performance of the constrained MCMC algorithm, we present a bar plot similar to those in Fig. 1, and for the same MIMO setup. We run one GS initialized with ZF solution, followed by 12 GSs in parallel, each for 3 iterations. We evaluate the difference $\Delta_{\mathrm{N-ML}}$ for one of the bits in the vector **b**. Fig. 2 clearly shows that the N-ML norm is obtained almost perfectly after only the second iteration of the GSs, confirming the superior ability of the C-MCMC algorithm to compute good N-ML distances and consequently, good LLR values.

VI. SIMULATION RESULTS

In order to assess the error rate performance of the constrained MCMC detector, we present simulation results of the proposed search strategies and compare them with the MCMC detector of [10] and the optimum ML detector. Figures 3 and 4 plot the bit error rate (BER) versus E_b/N_0 for a number of practical MIMO setups for two transmit, two receive antenna systems and four transmit, four receive MIMO systems, respectively. The plots show two different quadrature amplitude modulation (QAM) symbol mappings for each setup.

We first consider initialization with the ZF solution for both the MCMC implementation of [10] and the C-MCMC algorithm. To ensure fairness in the comparison, we run the same number of Gibbs sampler iterations for both algorithms. For example, if we have 16 bits per channel use, the constrained algorithm requires 16 GS $\times 2 + 1$ GS $\times 1 = 33$ GS iterations. The MCMC algorithm of [10] is run using 11 GSs, each of depth 3. It is clear from the figures that the constrained algorithm is at least one order of magnitude better than the non-controlled MCMC algorithm of [10]. However, the error floor characteristic of the latter algorithm seems to plague the C-MCMC algorithm when initialized with ZF as well, especially at very high SNR levels.

It is well known that the suboptimal detectors, such as the ZF solution, used to initialize the GS algorithm are very much affected by the condition number of the channel matrix, [12]. This subsequently degrades the performance of the MCMC detector. This phenomenon is aggravated in high SNR regimes because the propagation of error from the ML norm evaluation to the computation of the N-ML distances prevents the C-MCMC detector from eliminating the error floor. The problem is more pronounced for lower order constellation such as 16-QAM, as opposed to 64-QAM or 256-QAM. This can be explained by the fact that for larger constellations, the points in the state space are closer to each other and thus the MCMC algorithm is more likely to iterate between several states, hence yielding higher chance of reaching the optimal or a near optimal state. This problem can be alleviated through a deterministic computation of the ML distance. Such an approach is robust to channel ill-conditioning and leads to near optimal ML solutions [12]. The outcome of the ML initialized detector is shown in the BER plots labeled SD-C-MCMC. Here, we use the hard sphere decoding method [13] to find the near-optimum ML distance and use the C-MCMC algorithm to get the LLR values. Observe that the SD initializer used has much less complexity than the LSD algorithm, because in the latter case, the radius of the sphere has to be constantly increased, especially in the presence of bad channels, rendering the LSD's complexity exponential in the number of bits per channel use. On the other hand, we use the hard sphere decoder (HSD) to compute the ML sample only, and we confine the radius to a small value.

Moreover, initialization with the HSD is only necessary for a small percentage of the channel realizations. Our experiments show that "bad channels" occur for roughly 5% of the total number of channel realizations, and consequently, we suggest using the HSD and the ZF methods interchangeably to further reduce the complexity of the algorithm. The performance of the latter search strategy is shown in the BER plots labeled ZF-SD-C-MCMC (Figure 4). For the ZF-SD-C-MCMC, we use the HSD initializer for the cases where the condition number of the channel is larger than 20. The ZF solution is used in all other cases. As can be depicted in the plots, the performance of the ZF-SD-C-MCMC detector is comparable to that of the SD-C-MCMC algorithm, for a fraction of its complexity.

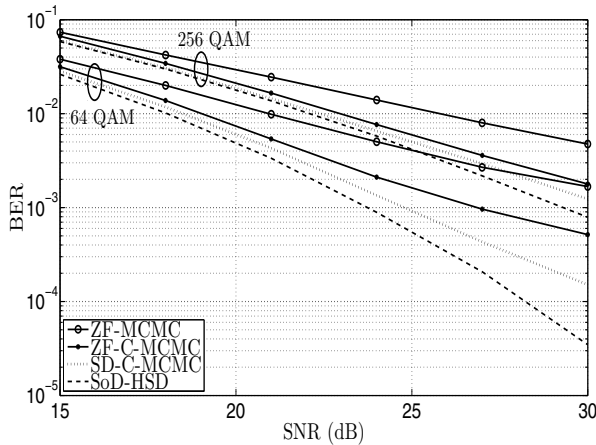Finally, the optimum ML solution (max-log ML detector) is

Fig. 3. Bit error rate plots that show the performance of the C-MCMC detector, initialized with ZF and SD, and compared to the MCMC algorithm and the optimum detector. The Figure shows the results for a two transmit, two receive MIMO system, employing 64 QAM and 256 QAM constellations.
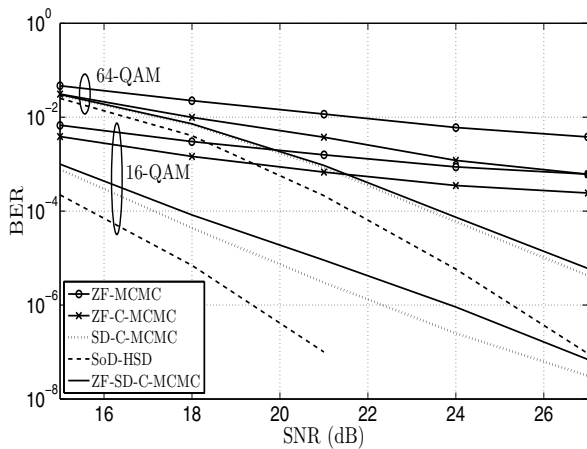

Fig. 4. Bit error rate plots that show the performance of the C-MCMC detector, initialized with ZF and SD, and compared to the MCMC algorithm, the ZF-SD-C-MCMC algorithm and the optimum detector. The Figure shows the results for a four transmit, four receive MIMO system, employing 16 QAM and 64 QAM constellations.

presented as a performance benchmark. It is obtained through a deterministic search strategy that follows the soft detector based on hard sphere decoder (SoD-HSD) method proposed in [14].

VII. CONCLUSION

The near reducibility of the Markov chain at high SNR causes the GS algorithm to stall in local minimum states and consequently leads to a severe degradation in the performance of the algorithm. Controlled initialization methods such as the ZF solution improve the ability of the algorithm to reach minimum ML states, but fails to control both terms of the LLR value equation. The constrained MCMC algorithm is a novel implementation of the MCMC detector that ensures proper

control of the ML and N-ML distances through running Gibbs samplers that target both distances.

The C-MCMC algorithm, when initialized with suboptimal solutions in the presence of ill-conditioned channels, falls short of harnessing the performance of the optimum detector. Judicially switching between initialization with the hard sphere decoding solution and the ZF solution, according to the condition number of the channel, guarantees a good tradeoff between optimum performance and minimal implementation complexity. Consequently, the C-MCMC detector should be established as the detector of choice when the system size prohibits the use of the optimum detector.

REFERENCES

[1] S. Verdú, "Minimum probability of error for asynchronous gaussian multiple-access channels," *IEEE Transactions on Information Theory*, vol. 32, pp. 85–96, January 1986.

[2] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE transactions on communications*, vol. 51, no. 3, pp. 389–399, March 2003.

[3] G. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Technical Journal*, vol. 1 (2), pp. 41 – 59, August 1996.

[4] H. V. Poor and S. Verdu, "Probability of error in mmse multiuse detection," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 255–270, February 1997.

[5] P. W. Wolniansky, G. Foshini, G. Golden, and R. Valenzuela, "V-blast: An architecture for realizing very high data rates over the rich-scattering wireless channel," *1998 URSI International Symposium on Signals, Systems, and Electronics*, pp. 295–300, 1998.

[6] J. Luo, K. Pattipati, P. Willett, and G. Levchuk, "Fast optimal and suboptimal any-time algorithms for cdma multiuser detection based on brach and bound," *IEEE Transactions on Communications*, vol. 52, no. 4, pp. 632–642, April 2004.

[7] C. Robert and G. Casella, *Monte Carlo statistical Methods*. Springer-Verlag, New York, 1999.

[8] B. Farhang-Boroujeny, H. Zhu, and Z. Shi, "Markov chain monte carlo algorithms for cdma and mimo systems," *IEEE transactions on Signal Processing*, vol. 54, no. 5, pp. 1896 – 1909, May 2006.

[9] H. Zhu, B. Farhang-Boroujeny, and R. Chen, "On performance of sphere decoding and markov chain monte carlo detection methods," *IEEE Signal Processing Letters*, pp. 669–672, October 2005.

[10] X. Mao, P. Amini, and B. Farhang-Boroujeny, "Markov chain monte carlo mimo detection methods for high signal-to-noise ratio regimes," *IEEE GlobComm*, pp. 3979 – 3983, November 26-30 2007.

[11] S. Laraway and B. Farhang-Boroujeny, "Implementation of a markov chain monte carlo based multiuser/mimo detector," *Submitted to IEEE Trans. Signal Proc.*, 2006.

[12] H. Artes, D. Seethaler, and F. Hlawatsch, "Efficient detection algorithms for mimo channels: A geometrical approach to approximate ml detection," *IEEE Transactions on Signal Processing*, vol. 51, no. 11, November 2003.

[13] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Mathematics of Computation*, vol. 44, no. 237, pp. 429–440, 1985.

[14] R. Wang and G. Giannakis, "Approaching mimo channel capacity with soft detection based on hard sphere decoding," *IEEE Transactions on Communications*, vol. 54, no. 4, pp. 587–590, April 2006.