Approaching MIMO Channel Capacity With Soft Detection Based on Hard Sphere Decoding

Renqiu Wang, *Student Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Hard sphere decoding (HSD) has well-appreciated merits for near-optimal demodulation of multiuser, block single-antenna or multi-antenna transmissions over multi-input multi-output (MIMO) channels. At increased complexity, a soft version of sphere decoding (SD), so-termed list SD (LSD), has been recently applied to coded layered space–time (LST) systems enabling them to approach the capacity of MIMO channels. By introducing a novel bit-level multi-stream coded LST transmitter along with a soft-to-hard conversion at the decoder, we show how to achieve the near-capacity performance of LSD, and even outperform it as the size of the block to be decoded (M) increases. Specifically, for binary real LST codes, we develop *exact* max-log-based SD schemes with $M + 1$ HSD steps, and an approximate alternative with only one HSD step to trade off performance for average complexity. These schemes apply directly to the real and imaginary parts of quaternary phase-shift keying signaling, and also to quadrature amplitude modulation signaling after incorporating an appropriate interference estimation and cancellation module. We corroborate our near-optimal soft detection (SoD) algorithms based on HSD (SoD-HSD) with simulations.**

***Index Terms*—Multi-input multi-output (MIMO) detection, soft decoding, sphere decoding, turbo decoding.**

I. Introduction

IN WIRELESS communications, quite often we wish to estimate an $M \times 1$ information-bearing symbol vector $\mathbf{s}$ from an $N \times 1$ data vector $\mathbf{y}$ obeying the multi-input multi-output (MIMO) matrix–vector model

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{s}$ has entries belonging to a finite alphabet $\mathcal{S}$ of size $|\mathcal{S}|$, $\mathbf{H}$ is a known $N \times M$ real or complex matrix, and $\mathbf{n}$ is an $N \times 1$ Gaussian noise vector. Suppose that each entry s_m of the symbol vector $\mathbf{s}$ in (1) is drawn from an $M_b \times 1$ binary vector $\mathbf{x}^{(m)}$ with ± 1 entries by constellation mapping, and let the $M_b M \times 1$ vector $\mathbf{x} := [\mathbf{x}^{(1)^T}, \mathbf{x}^{(2)^T}, \ldots, \mathbf{x}^{(M)^T}]^T$ be the binary representation of $\mathbf{s}$, where ^{T} denotes transposition. Under the assumptions that the entries $\{x_k\}_{k=1}^{MM_b}$ of $\mathbf{x}$ are independent and the noise is white Gaussian, invoking Bayes' theorem and the max-log approximation [1], we can approximate the extrinsic information of x_k as [2]

$$\lambda_E(x_k|\mathbf{y}) = \left(\frac{1}{2}\right) \min_{\mathbf{x}\in\mathbb{X}_{k,-1}} \left\{\frac{1}{\sigma^2}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 - \mathbf{x}^T\lambda_A\right\} - \left(\frac{1}{2}\right) \min_{\mathbf{x}\in\mathbb{X}_{k,+1}} \left\{\frac{1}{\sigma^2}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 - \mathbf{x}^T\boldsymbol{\lambda}_A\right\} - \lambda_A(x_k) \tag{2}$$

where σ^2 is the noise variance, and $\boldsymbol{\lambda}_A := [\lambda_A(x_1), \ldots, \lambda_A(x_{MM_b})]^T$, with $\lambda_A(x_k) := \ln[P(x_k = +1)/P(x_k = -1)]$ denotes the *a priori* information of x_k; $\mathbb{X}_{k,+1} := \{\mathbf{x}|x_k = +1\}$ and $\mathbb{X}_{k,-1} := \{\mathbf{x}|x_k = -1\}$.

In [2], a soft, so-called list sphere decoding (LSD) algorithm was derived to avoid the exhaustive search by restricting the search inside a preset sphere. A list of $\mathbf{s}$ candidates is obtained during the search, based on which the extrinsic information is calculated. LSD has been shown to achieve capacity-approaching performance in a coded layered space–time (LST) system [2]. Alternative works aiming at reduced-complexity LSD include, but are not limited to [3], where approximations of the *a posteriori* probability calculation for high-dimensional MIMO systems are obtained by merging a *channel* list and a *prior* list, which may outperform LSD at lower complexity.

In this letter, we derive a soft-to-hard transformation for binary constellations to convert the max-log-based maximum *a posteriori* (MAP) (max-MAP) decoding problem with real block codes to a set of hard sphere decoding (HSD) problems (Section II-A). In addition to providing an *exact* max-MAP decoder, we also derive an approximate alternative to further reduce complexity down to the order of a single HSD (Section II-B). In Section III, we apply our soft detection (SoD) based on HSD (SoD-HSD) schemes to a *bit-level multistream* coded LST system for quadrature amplitude modulation (QAM) signaling. Section IV provides comparisons and demonstrates by simulations that MIMO channel capacity can indeed be approached.

II. SoD Based on HSD

A. Soft-to-Hard Conversion—Exact Scheme

Capitalizing on the fact that the columns of $\mathbf{H}$ are linearly independent with high probability, we can almost always find a vector $\tilde{\mathbf{y}}$ satisfying

$$2\mathbf{H}^T\tilde{\mathbf{y}} = \sigma^2\boldsymbol{\lambda}_A. \tag{3}$$

Paper approved by N. C. Beaulieu, the Editor for Wireless Communication Theory of the IEEE Communications Society. Manuscript received August 5, 2003; revised December 6, 2004 and August 10, 2005. This work was supported through collaborative participation in the Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD 19-01-2-0011. This paper was presented in part at the Wireless Communications and Networking ConferenceAtlanta, GA, March 2004.

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: renqiu@ece.umn.edu; georgios@ece.umn.edu).

Digital Object Identifier 10.1109/TCOMM.2006.873064

Using (3) and focusing on the binary case (where $\mathbf{s} = \mathbf{x}$), we can convert the calculation of the extrinsic information of x_k to the following two integer least-square (ILS) problems:

$$\lambda_E(x_k|\mathbf{y}) = -\frac{1}{2\sigma^2} \min_{\mathbf{x}\in\mathbb{X}_{k,+1}} \|\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{H}\mathbf{x}\|^2 + \frac{1}{2\sigma^2} \min_{\mathbf{x}\in\mathbb{X}_{k,-1}} \|\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{H}\mathbf{x}\|^2 - \lambda_A(x_k). \quad (4)$$

Define $\hat{\mathbf{x}}_{\text{map}} := \arg\min_{\mathbf{x}\in\mathbb{X}} \|\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{H}\mathbf{x}\|^2$ and $\hat{\mathbf{x}}_k := \arg\min_{x_k=-\hat{x}_{k,\text{map}};\mathbf{x}_{[k]}\in\mathbb{X}_{[k]}} \|\mathbf{y} + \tilde{\mathbf{y}} + \hat{x}_{k,\text{map}}\mathbf{h}_k - \mathbf{H}_{[k]}\mathbf{x}_{[k]}\|^2$ where $\hat{x}_{k,\text{map}}$ is the kth entry of $\hat{\mathbf{x}}_{\text{map}}$, $\mathbf{h}_k$ is the kth column of $\mathbf{H}$, and $\mathbf{H}_{[k]}$ is the submatrix obtained from $\mathbf{H}$ after omitting $\mathbf{h}_k$. To obtain the extrinsic information for the entire vector $\mathbf{x}$, we need one HSD step with block size M to find $\hat{\mathbf{x}}_{\text{map}}$, and M HSD steps with block size $(M-1)$ to find $\{\hat{\mathbf{x}}_k\}_{k=1}^{M}$. Notice that the extrinsic information values we obtained are exact under the max-log approximation.

B. Approximate Decoding Scheme

With our *exact* max-MAP decoder as a starting point, we can further reduce the number of HSD steps to only one after invoking the following approximation. Let $\check{\mathcal{X}}_k^{[2]}$ denote the set of vectors that have one or two entries different from $\hat{\mathbf{x}}_{\text{map}}$ including x_k; i.e., $\check{\mathcal{X}}_k^{[2]} := \{\mathbf{x}|x_k = -\hat{x}_{k,\text{map}}, (1/2)\sum_{i=1,i\neq k}^{M} |x_i - \hat{x}_{i,\text{map}}| = 0 \text{ or } 1\}$. The set $\check{\mathcal{X}}_k^{[2]}$ has M elements. Since at high signal-to-noise ratio (SNR), $\hat{\mathbf{x}}_k \in \mathcal{X}_k^{[2]}$ with high probability, we can approximate $\lambda_E(x_k|\mathbf{y})$ as

$$\lambda_E(x_k|\mathbf{y}) \approx -\frac{\hat{x}_{k,\text{map}}}{2\sigma^2} \|\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{H}\hat{\mathbf{x}}_{\text{map}}\|^2 + \frac{\hat{x}_{k,\text{map}}}{2\sigma^2} \min_{\mathbf{x}\in\check{\mathcal{X}}_k^{[2]}} \|\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{H}\mathbf{x}\|^2 - \lambda_A(x_k). \quad (5)$$

To obtain the extrinsic information for the entire vector $\mathbf{x}$ here, we need only one HSD step with block size M to find $\hat{\mathbf{x}}_{\text{map}}$, and M searching steps to obtain the "best vectors" minimizing $\|\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{H}\mathbf{x}\|^2$, respectively, in the sets $\{\check{\mathcal{X}}_k^{[2]}\}_{k=1}^{M}$. These "best vectors" are the approximations of $\{\hat{\mathbf{x}}_k\}_{k=1}^{M}$. Each approximation of $\hat{\mathbf{x}}_k$ has complexity linear in M, and the total complexity for $\{\hat{\mathbf{x}}_k\}_{k=1}^{M}$ is $O(M^2)$, which is negligible compared with the HSD complexity.

To further improve accuracy, we can approximate each $\hat{\mathbf{x}}_k$ with the "best vector" minimizing $\|\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{H}\mathbf{x}\|^2$ in the set $\check{\mathcal{X}}_k^{[3]} := \{\mathbf{x}|x_k = -\hat{x}_{k,\text{map}}, (1/2)\sum_{i=1,i\neq k}^{M} |x_i - \hat{x}_{i,\text{map}}| = 0, 1, \text{ or } 2\}$. The set $\{\check{\mathcal{X}}_k^{[3]}\}_{k=1}^{M}$ has $M + (M-1)(M-2)$ elements. Therefore, the complexity in approximating $\{\hat{\mathbf{x}}_k\}_{k=1}^{M}$ is $O(M^3)$, which is close to the order of a single HSD. Similar approximations by flipping more entries of $\hat{\mathbf{x}}_{\text{map}}$ is not recommended, because complexity increases accordingly.

Remark: Notice that although our SoD-HSD has adopted HSD to find $\hat{\mathbf{x}}_{\text{map}}$ and $\hat{\mathbf{x}}_k$, it can be easily extended to other near-optimal or suboptimal hard MIMO detection algorithms, such as the semidefinite programming (SoD-SDP) [4], or the m-algorithm [3], [5]. These extensions are not detailed here due to space limitations.

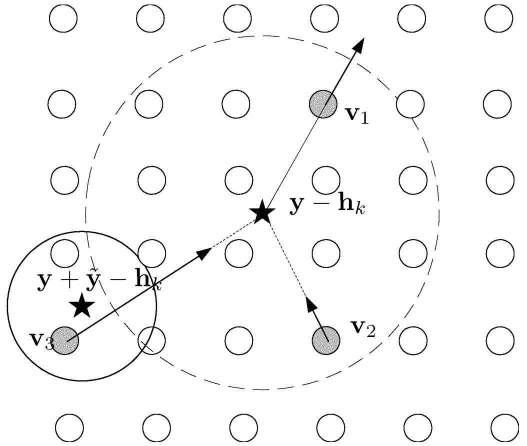
$\mathbf{v}_1$
$\mathbf{y} - \mathbf{h}_k$
$\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{h}_k$
$\mathbf{v}_3$
$\mathbf{v}_2$

Fig. 1. Illustration of LSD and SoD-HSD schemes.

C. Comparison With LSD in the Binary Case

Fig. 1 illustrates the difference that the HSD step effects between LSD [2] and our SoD-HSD scheme, in finding the "best vector" $\hat{\mathbf{x}}_{k,+1}$ for $x_k = +1$. LSD searches in the sphere $(\mathbf{s} - \hat{\mathbf{s}})^T\mathbf{H}^T\mathbf{H}(\mathbf{s} - \hat{\mathbf{s}}) \leq r^2$ to obtain a list of candidates, among which one is used to obtain $\hat{\mathbf{x}}_{k,+1}$. In Fig. 1, small circles denote all possible candidates $\mathbf{v}$ in the ensemble $\mathbb{X}_{k,+1}$, the large dash circle denotes the sphere, and the star denotes its center $\mathbf{y} - \mathbf{h}_k$. The linear correction term $\sigma^2\mathbf{x}^T\boldsymbol{\lambda}_A$ actually perturbs the location of $\mathbf{v}$. LSD only checks points inside the sphere, and thus concludes that the "best point" is $\mathbf{v}_2$. However, in this example, the real optimal solution is $\mathbf{v}_3$, which unfortunately is excluded from the sphere. There is yet another challenging case for LSD; namely, that sometimes all candidates in the list have only value $+1$ (or -1) for a certain bit. In this case, LSD fails to produce even an approximation. A preset value (e.g., ±8) is used instead, in this case.

Our SoD-HSD scheme, on the other hand, avoids these problems by taking a different approach. Instead of fixing the center and perturbing the point every time, it shifts the center to a new position $\mathbf{y} + \tilde{\mathbf{y}} - \mathbf{h}_k$. For each point $\mathbf{v}$, we only need to measure the distance of $\mathbf{v}$ from the center. So, instead of choosing a large sphere, as in LSD (the large dashed circle), the sphere can be chosen as small as we possibly can (the smaller solid bold circle). This way, $\hat{\mathbf{x}}_{k,+1}$ can be easily found using HSD. After the HSD algorithm was first described in [6] and refined in [7], several schemes have been developed to improve it [2], [8]–[11]. One advantage of our scheme is that all existing improvements for HSD can be directly adopted by SoD-HSD. However, this is not the case for LSD.

Although one HSD step has lower complexity than the list search in LSD, our overall SoD-HSD scheme entails additional complexity. Indeed, there are total $M + 1$ HSD steps in soft decoding an $M \times 1$ binary vector $\mathbf{x}$, each of which requires either one QR decomposition of $\mathbf{H}$ incurring complexity $C_{\text{QR}}(M) = O(M^3)$, or one QR decomposition of $\mathbf{H}_{[k]} (k = 1, \ldots, M)$ with complexity $C_{\text{QR}}(M-1) = O((M-1)^3)$. Our SoD-HSD also entails extra calculation of $\tilde{\mathbf{y}}$ which incurs complexity $C_{\tilde{\mathbf{y}}}(M) = O(M^3)$. In summary, the average complexity of SoD-HSD is approximately $C_{\text{SoD-HSD}}(M) = C_{\text{QR}}(M) + MC_{\text{QR}}(M-1)$

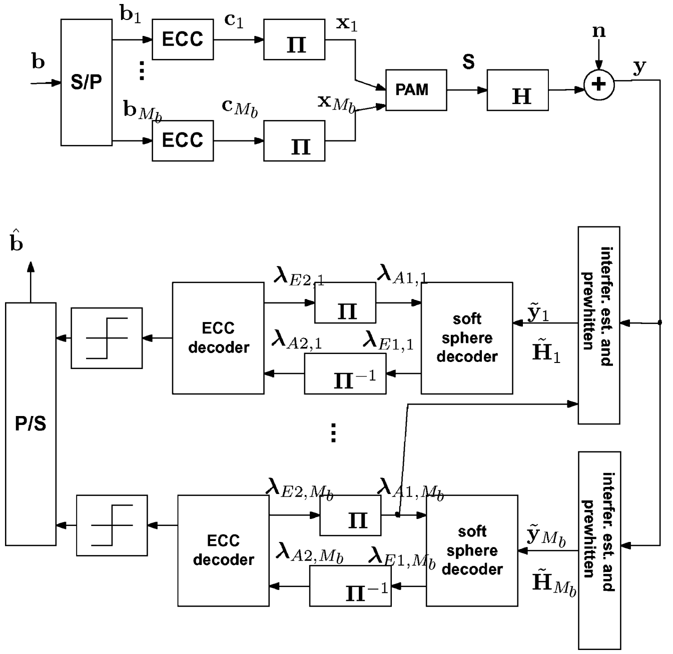
b
S/P
$\mathbf{b}_1$
$\mathbf{b}_{M_b}$
ECC
ECC
$\mathbf{c}_1$
$\mathbf{c}_{M_b}$
Π
Π
$\mathbf{x}_1$
$\mathbf{x}_{M_b}$
PAM
S
H
n
y
$\hat{\mathbf{b}}$
P/S
ECC decoder
ECC decoder
$\boldsymbol{\lambda}_{E2,1}$
$\boldsymbol{\lambda}_{A1,1}$
$\boldsymbol{\lambda}_{A2,1}$
$\boldsymbol{\lambda}_{E1,1}$
Π^{-1}
soft sphere decoder
$\tilde{\mathbf{y}}_1$
$\tilde{\mathbf{H}}_1$
interfer. est. and prewhitten
$\boldsymbol{\lambda}_{E2,M_b}$
$\boldsymbol{\lambda}_{A1,M_b}$
$\boldsymbol{\lambda}_{A2,M_b}$
$\boldsymbol{\lambda}_{E1,M_b}$
$\tilde{\mathbf{y}}_{M_b}$
$\tilde{\mathbf{H}}_{M_b}$

Fig. 2. Real equivalent bit-level multistream system model for QAM signaling.

$+C_{\tilde{\mathbf{y}}}(M) + C_{\text{HSD}}(M) + MC_{\text{HSD}}(M-1)$. When applied to iterative decoding, our SoD-HSD has performed at each iteration. The overall average complexity of iterative SoD-HSD is, therefore, N_{iter} times that of one iteration, where N_{iter} denotes the number of iterations. The list search in LSD entails higher complexity than HSD. However, LSD requires only one QR decomposition of $\mathbf{H}$ and one list search for all bits. LSD is also advantageous during the iterative decoding process. The list can be saved once for all iterations, if memory allows. Taking all aspects into account, it is hard to assess whether LSD or SoD-HSD has lower complexity. We will illustrate in Section IV the running-time comparison between LSD and SoD-HSD with the same setup and comparable programming. We will observe that exact SoD-HSD exhibits a slight edge relative to LSD as the constellation size and/or the block size increase. Approximate SoD-HSD, on the other hand, can be up to 40 times faster than LSD for 64-QAM in a 12×12 setup.

III. Bit-Level Multistream Coded V-Blast System

The SoD-HSD schemes in Section II were derived when $\mathbf{s} = \mathbf{x}$ in (2). However, SoD-HSD cannot be applied directly to QAM signaling. For QAM signaling, after the complex-to-real conversion, $\mathbf{s}$ is a $2N_t \times 1$ vector composed of pulse amplitude modulation (PAM) symbols with natural bit mapping, i.e., $\mathbf{s} = \sum_{i=1}^{M_b} 2^{i-1}\mathbf{x}_i$. Because different $x_{i,k}$ bits in s_k will be received with generally different SNRs, we adopt a bit-level multistream coded LST transmission for QAM signaling, as depicted in Fig. 2. At the transmitter, the stream of information bits $\mathbf{b}$ is first divided into M_b substreams $\{\mathbf{b}_i\}_{i=1}^{M_b}$. Each substream is coded with an error-control code (ECC) and scrambled through a random interleaver. From each substream, we take one interleaved bit to form a PAM symbol consisting of M_b bits. Two PAM symbols are further combined to form a QAM symbol. The QAM symbol vectors are then transmitted using V-BLAST. At the receiver end, iterative decoding is performed in a layered fashion per bit level. In the MIMO channel decoding module, the complex MIMO block model is first converted to the real block model by separating the real and imaginary parts. When decoding one bit level, interference from other bit levels will be treated as Gaussian noise, and based on the *a priori* information from the levels with higher power, the mean and covariance matrix of the equivalent noise will be estimated as in [12]. This reduces decoding of each bit level to an equivalent quaternary phase-shift keying (QPSK) decoding problem in the presence of colored Gaussian noise. After prewhitening the noise, our SoD-HSD scheme can be readily applied with the extrinsic information exchanged through the interleaver/deinterleaver between the ECC decoding module and the MIMO channel decoding module. The detailed decoding process is omitted here due to space limitations.

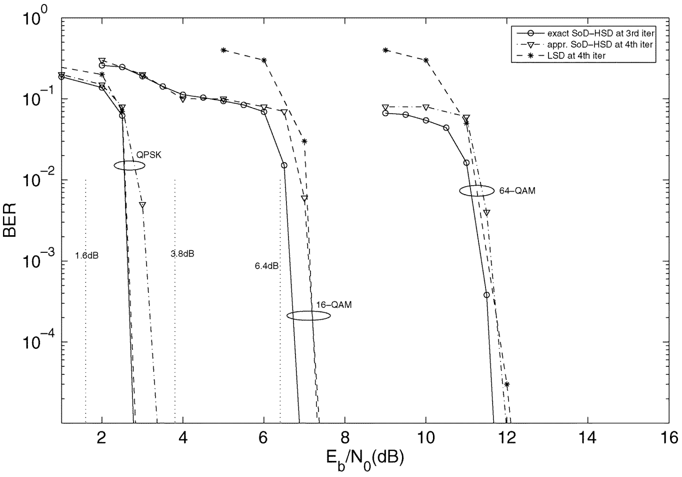

Fig. 3. Exact SoD-HSD, approximate SoD-HSD, and LSD in an 8×8 setup.

IV. Simulations

In this section, we present simulations using the same parallel concatenated (turbo) codes as in [2]. To maintain comparable ECC decoding complexity with [2], for QPSK signaling, we choose the interleaver size to be 9000, and the number of inner iterations for the ECC decoding module to be 10. For 16-QAM signaling, we choose the interleaver size of each bit level to be 4, 500, and the number of inner iterations for the ECC decoding module of bit level 2 and bit level 1 to be 5 and 15, respectively. For 64-QAM signaling, we choose the interleaver size of each bit level to be 3000, and the number of inner iterations for the ECC decoding module of bit level 3, 2, and 1 to be 4, 8, and 18, respectively.

Simulation 1: Fig. 3 depicts average BER performance in an 8×8 transmit/receive antenna setup when using LSD, exact SoD-HSD, and the approximate SoD-HSD in (5). The limits (e.g., 1.6, 3.8, and 6.4 dB) shown in the figures are the lowest possible SNR values to achieve capacity for QPSK, 16-QAM, and 64-QAM, respectively. We performed three and four outer iterations for the exact and approximate SoD-HSD, respectively. Compared with [2], our exact SoD-HSD achieves almost identical performance for QPSK signaling. For 16-QAM and

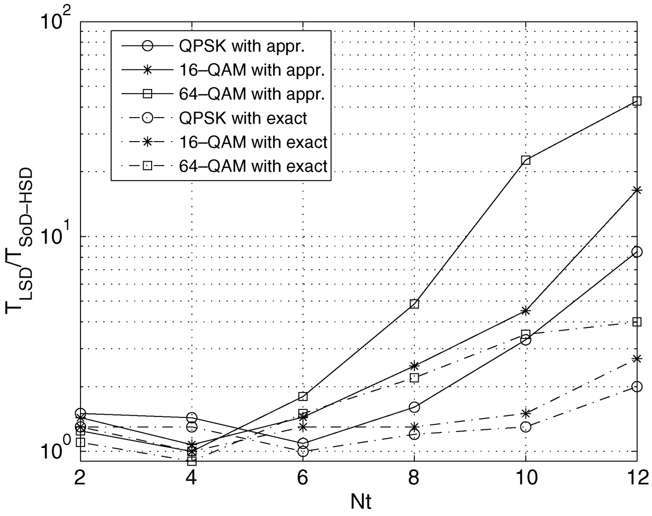

Fig. 4. Average time ratio of LSD and SoD-HSD.

64-QAM, our scheme outperforms [2] by about 0.5 dB. We observe that with the approximate scheme, we have about 0.5 dB loss relative to our exact scheme. However, we can still achieve the same performance as [2] for 16-QAM and 64-QAM.

Simulation 2: Fig. 4 depicts running-time comparison between our exact as well as approximate SoD-HSD and LSD in different setups. The y-axis depicts the average running-time ratio of LSD over SoD-HSD per ECC frame per iteration. Only the time spent on soft MIMO decoding, excluding ECC decoding, was counted. In LSD, we save the list for later iterations. Fig. 4 shows that for a small number of antennas and/or constellation size, LSD and SoD-HSD have comparable running time. But SoD-HSD becomes faster as the constellation size and/or the number of antennas increase. Our approximate SoD-HSD can be up to 40 times faster than LSD for 64-QAM in a 12×12 setup.

Acknowledgment

The authors would like to thank Dr. W. Zhao of Qualcomm for helpful discussions on this and related topics. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

References

[1] P. Robertson, E. Villebrun, and P. Hoeher, "A comparison of optimal and suboptimal MAP decoding algorithms operating in the log domain," in *Proc. Int. Conf. Commun.*, Seattle, WA, Jun. 1995, pp. 1009–1013.

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

[3] A. B. Reid, A. J. Grant, and A. P. Kind, "Low complexity list detection for high-rate MIMO channels," in *Proc. 4th Australian Commun. Theory Workshop*, Melbourne, Australia, Feb. 2003, pp. 66–69.

[4] W. K. Ma, T. N. Davidson, K. M. Wong, Z. Q. Luo, and P. C. Ching, "Quasi-maximum-likelihood multiuser detection using semidefinite relaxation with application to synchronous CDMA," *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 912–922, Apr. 2002.

[5] C. Schlegel*,* *Trellis Coding*. Piscataway, NJ: IEEE Press, 1997.

[6] M. Pohst, "On the computation of lattice vectors of minimal length, successive minima and reduced bases with applications," in *Proc. ACM SIGSAM*, 1981, pp. 37–44.

[7] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, pp. 463–471, Apr. 1985.

[8] B. Hassibi and H. Vikalo, "On the expected complexity of sphere decoding," in *Proc. 35th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2001, pp. 1051–1055.

[9] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[10] A. Chan and I. Lee, "A new reduced-complexity sphere decoder for multiple antenna systems," in *Proc. Int. Conf. Commun.*, New York, NY, Apr. 28–May 2 2002, vol. 1, pp. 460–464.

[11] W. Zhao and G. Giannakis, "Reduced complexity closest point decoding for random lattices," in *Proc. 41st Annu. Allerton Conf. Commun. Control, Comput.*, Monticello, IL, Oct. 2003.

[12] J. Luo, K. R. Pattipati, P. K. Willett, and F. Hasegawa, "Near-optimal multiuser detection in synchronous CDMA using probabilistic data association," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 361–363, Sep. 2001.