SPHERE DECODING FOR RETRANSMISSION DIVERSITY IN MIMO FLAT-FADING CHANNELS

Harvind Samra and Zhi Ding
Department of Electrical and Computer Engineering
University of California, Davis, CA 95616
Email: hssamra,zding@ece.ucdavis.edu

ABSTRACT

In this paper, we define an ARQ protocol for MIMO flat-fading channels which varies the bit-to-symbol mapping per retransmission. We begin by defining a model for distinctly mapped transmissions through MIMO channels, and the effect this mapping diversity has on a sphere decoder receiver. Varying the symbol mapping complicates the sphere decoding process, particularly for the enumeration of candidate solutions within the sphere. A technique that promises quick candidate enumeration is suggested, borrowing concepts from existing closest point search schemes. The value of mapping diversity, in reducing BER and reducing computational complexity, is analyzed.

1. INTRODUCTION

Critical figures of merit for a communication system often include low frame error rates (FER) and high data throughputs to its end users. Effective handling and reduction of packet retransmissions are vital in improving performance. Generally, if errors remain (possibly after error correction) in reception of a transmitted data packet, a request for retransmssion is made to the transmitter. As a result, the development and exploitation of Automatic Repeat reQuest (ARQ) protocols has been the subject of much research at both the network and physical layers. At the physical layer, various approaches have been proposed for both packet combining and improving diversity among these retransmissions. Chase developed a maximum likelihood combining scheme for an arbitrary number of coded packets, concatenating M copies of a codeword into a single codeword [1]. Harvey and Wicker proposed several ARQ strategies, including an approach where soft-decoded codewords from multiple packet transmissions are combined into a single soft codeword [2]. Stuber and Narayanan developed an ARQ receiver using error correcting codes where the extrinsic information from the decoding of previous packets is reused [3]. Recently, an approach involving adapting the bit-to-symbol mapping for each retransmission was developed, providing significant reduction in BER while requiring minimal increase in system complexity [4, 5].

This work studies the effectiveness of packet retransmissions in multiple-input, multiple-output (MIMO) systems, where multiple antennas may be present at the transmitter and/or receiver. The spatial diversity provided by multiple antennas is known to dramatically increase system capacity [6]. As a result, MIMO systems have become important and popular research subjects. In particular, the combination of MIMO spatial diversity with temporal diversity has led to widespread development of space-time coding schemes [7, 8]. As packet retransmissions are a form of temporal diversity, this paper overlaps considerably with space-time coding. However, in ARQ we are limited to using incremental redundancy in order to minimize the number of retransmissions. Others have recently studied packet retransmissions in MIMO systems. Onggosanusi *et al.* introduced methods for combining packet transmissions using zero-forcing and MMSE receivers [9]. Ding and Rice proposed a hybrid ARQ protocol involving spatio-temporal vector coding and multi-dimensional TCM [10]. Nguyen and Ingram investigated hybrid ARQ protocols for systems that use recursive space-time codes [11].

In this paper, we define an ARQ protocol for MIMO flat-fading channels where the bit-to-symbol mapping is varied per retransmission (mapping diversity). We begin with defining a model for distinctly mapped transmissions through MIMO channels, and the impact mapping diversity has on a sphere decoder receiver. Varying the symbol mapping complicates the sphere decoding process, mostly in the enumeration of candidates within the sphere. We propose a method that promises fast candidate enumeration, using a technique developed for various computational vision problems. A brief review of mapping diversity is provided, with some insight into its effect in reducing BER and computational complexity. We conclude with simulation results to validate the protocol.

2. SYSTEM MODEL

We begin with a set $\mathcal{C}$ of real or complex numbers that represent the points of a signal constellation, e.g. 16QAM. Given a packet of bits, consecutive groups of $\log_2 |\mathcal{C}|$ bits (s represents the decimal equivalent of these bits) are assigned to symbols in $\mathcal{C}$ via a symbol mapping function $\psi : \{0, 1, \ldots, |\mathcal{C}| - 1\} \rightarrow \mathcal{C}$. We consider M transmissions of this packet, illustrated in Fig. 1, with a transmitter with N antennas and a receiver with K antennas. More specifically, we narrow our focus on blocks of the packet, with each block $\mathbf{s} = [s_1, \ldots, s_N]^T$ containing N user labels, that are transmitted M times. A $K \times N$ matrix $\mathbf{H}_m$ ($K \geq N$) represents the m^{th} transmission channel, with $h_{m,ij}$ indicating the fading coefficient between transmit antenna i and receive antenna j. This fading gain is Rayleigh distributed, so that $h_{m,ij}$ is a complex-valued Gaussian variable of zero mean and unit variance. For the m^{th} transmission, the receiver obtains

$$\mathbf{y}_m = \begin{bmatrix} y_{m,1} \\ \vdots \\ y_{m,K} \end{bmatrix} = \mathbf{H}_m \begin{bmatrix} \psi_m[s_1] \\ \vdots \\ \psi_m[s_N] \end{bmatrix} + \begin{bmatrix} w_{m,1} \\ \vdots \\ w_{m,K} \end{bmatrix} \quad (1)$$

$$= \mathbf{H_m}\vec{\psi}_m[\mathbf{s}] + \mathbf{w}_m.$$

Supported in part by NSF grant ECS-0121469.

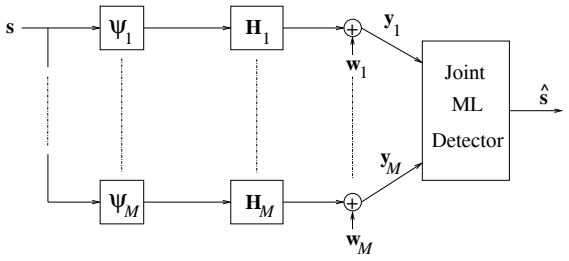

Fig. 1. Block diagram of multiple MIMO transmissions of a packet with mapping diversity.

The noise vector $\mathbf{w}_m$ is Gaussian with zero mean and variance $\sigma_w^2 \mathbf{I}_K$. Note that each label s_n is distinctly mapped via M mapping functions $\psi_1, \ldots, \psi_M$. The impact of varying the mapping is discussed later; it has been shown to greatly reduce BER for single-antenna systems [4, 5].

The receiver employs a joint ML decoder that incorporates $\mathbf{y}_1, \ldots, \mathbf{y}_M$ to produces estimates $\hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_N]^T$. Assuming that the channels are perfectly known, the ML decoding rule is

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \sum_{m=1}^{M} ||\mathbf{y}_m - \mathbf{H}_m \vec{\psi}_m[\mathbf{s}]||^2. \tag{2}$$

At first glance, minimizing the metric in (2) appears to require an exhaustive search over all $|\mathcal{C}|^N$ candidates. Fortunately, through some algebraic manipulations, the metric can be expressed so that an exhaustive search is avoided. By defining $\mathbf{p}_m = \mathbf{H}_m^{\dagger}\mathbf{y}_m$, the metric is

$$\sum_{m=1}^{M} (\vec{\psi}_m[\mathbf{s}] - \mathbf{p}_m)^H \mathbf{H}_m^H \mathbf{H}_m (\vec{\psi}_m[\mathbf{s}] - \mathbf{p}_m).$$

Let $\mathbf{U}_m$ be an upper-triangular matrix that satisfies $\mathbf{U}_m^H \mathbf{U}_m = \mathbf{H}_m^H \mathbf{H}_m$. This is typically accomplished with a Cholesky decomposition. The elements of $\mathbf{U}_m$ are indicated using $u_{m,ij}$, and the metric becomes

$$\sum_{n=1}^{N}\sum_{m=1}^{M} u_{m,nn}^2 \left| \psi_m[s_n] - p_{m,n} + \sum_{j=n+1}^{N} \frac{u_{m,nj}}{u_{m,nn}} (\psi_m[s_j] - p_{m,j}) \right|^2 \tag{3}$$

This metric bears many similarities with the metric that is minimized using sphere decoding. The primary difference involves the inclusion of multiple transmissions and their different mappings. We now discuss the application of sphere decoding to minimize our metric while maintaining a low computational burden.

3. APPLYING SPHERE DECODING

In ML detection, sphere decoding has become a low cost alternative to exhaustive, brute-force searches [12, 13]. The primary strategy in search reduction is the removal of all points outside of a hypersphere centered about the received data point $\mathbf{p}$. Beginning with s_N, labels whose mapped symbols fall inside the hypersphere are sequentially selected to comprise a data estimate $\hat{\mathbf{s}}$. Several techniques exist that quickly enumerate all symbols/labels within the hypersphere [13, 14]. This estimate is saved, and the radius of the hypersphere is decreased to the distance between the symbols of $\hat{\mathbf{s}}$ and $\mathbf{p}$. This process is repeated until no candidates exist within the hypersphere, and the most recent data estimate $\mathbf{s}$ becomes the ML estimate. The expected complexity of sphere decoding is cubic (N^3) in most instances [12].

To apply sphere decoding to our multiple-transmission scenario, we first define a hypersphere of radius r_N, with a modification of the rule proposed by Hochwald and ten Brink [14]

$$r_N^2 = \sum_{m=1}^{M} 2\sigma_v^2 \gamma K - \mathbf{y}_m^H (I_K - \mathbf{H}_m (\mathbf{H}_m^H \mathbf{H}_m)^{\dagger} \mathbf{H}_m^H) \mathbf{y}_m.$$

The parameter γ is chosen to ensure that the hypersphere contains some candidates. From the metric (3), we focus on the label s_N by looking only at the $n = N$ term of the summation. Thus, we seek to choose a label s_N that satisfies the inequality

$$\sum_{m=1}^{M} u_{m,NN}^2 |\psi_m[s_N] - p_{m,N}|^2 < r_N^2. \tag{4}$$

The set of all candidate labels that satisfy (4) is denoted by $\mathcal{S}_N$. The issue of quickly enumerating these labels is discussed momentarily.

Given a method for fast candidate enumeration, the sphere decoding algorithm is easily applied. With the set $\mathcal{S}_N$, a candidate label $\hat{s}_N$ is chosen and removed from $\mathcal{S}_N$, and we proceed in finding a candidate for s_{N-1}. The process described above is repeated, where the $n = N-1$ term of the summation in (3) is isolated, and $\hat{\mathbf{s}}$ is substituted into the $n = N$ term. A candidate $\hat{s}_{N-1}$ is chosen and removed from $\mathcal{S}_{N-1}$, and we proceed to find $\mathcal{S}_{N-2}$ and $\hat{s}_{N-2}$, etc. In constructing the candidate set $\mathcal{S}_n$, the inequality of interest becomes

$$\sum_{m=1}^{M} u_{m,nn}^2 |\psi_m[s_n] - a_m|^2 < r_n^2, \tag{5}$$

with

$$a_m = p_{m,n} - \sum_{j=n+1}^{N} \frac{u_{m,nj}}{u_{m,nn}} (\psi_m[\hat{s}_j] - p_{m,j}),$$

$$b_m = \sum_{j=n+1}^{N} u_{m,jj}^2 \left| \psi_m[\hat{s}_j] - p_{m,j} + \sum_{t=j+1}^{N} \frac{u_{m,jt}}{u_{m,jj}} (\psi_m[\hat{s}_t] - p_{m,t}) \right|^2.$$

and radius

$$r_n^2 = r_N^2 - \sum_{m=1}^{M} b_m.$$

The variable a_m is the center of the region defined by (5), and b_m adjusts the radius of this region by accounting for the $n+1$ through N terms of the summation in (3).

When the decoder finishes with the $n = 1$ stage, we have an estimate $\hat{\mathbf{s}}$. This estimate is not necessarily the true ML estimate as other points may still lie with the hypersphere. The radius r_N is updated to the distance between $\hat{\mathbf{s}}$ and $\mathbf{p}$, which is computed using (3). The decoding process starts over using this updated radius to determine if a better estimate exists.

When the set $\mathcal{S}_n$ is empty, then at least one of the previously chosen candidates $\hat{s}_{n+1}, \ldots, \hat{s}_N$ is incorrect and should be changed. The decoder moves sequentially through the existing candidate sets starting from $\mathcal{S}_{n+1}$ to $\mathcal{S}_N$ until a non-empty set $\mathcal{S}_e$ is encountered. It then chooses and removes a new candidate $\hat{s}_e$ from $\mathcal{S}_e$ and continues the decoding process with the creation of a new $\mathcal{S}_{e-1}$

and $\hat{s}_{e-1}$, etc. When the sets $\mathcal{S}_{n+1}, \ldots, \mathcal{S}_N$ are all empty, there are two possible actions. If estimates for $\hat{\mathbf{s}}$ already exist, then the most recent estimate is the true ML estimate. Otherwise, the initial radius r_N is too small and is increased; the decoding process starts over with this new radius.

4. CANDIDATE ENUMERATION

When $M = 1$, the problem of enumerating points in (5) simplifies to normal sphere decoding, and as previously mentioned, techniques exist to quickly find a satisfactory s_n. Unfortunately, these techniques do not extend feasibly to cases where $M > 1$. One solution is to directly compute (5) for all $|\mathcal{C}|$ possible labels, and sort these in ascending order to produce a list of viable candidates for s_n.

We propose an enumeration method that incorporates concepts applied in many computational vision problems that require pattern matching of eigen-decompositions. These problems frequently require closest point searches to find the point q, in a fixed set of points $\mathcal{Q}$ in a d-dimensional space, that is closest in Euclidean distance to a received data point x. To accomplish this search efficiently, an algorithm was introduced where the points of $\mathcal{Q}$ are sorted in each dimension, creating d sorted lists [15]. Given a point x, a hypercube $\mathcal{D}$ is defined with center x, and points $q \in \mathcal{Q}$ within the hypercube are quickly identified from the sorted lists.

The region specified by (5) can be rewritten, separating the real and imaginary components, as

$$\sum_{m=1}^{M} \left(\frac{\Re\{\psi_m[s_n]\} - \Re\{a_m\}}{r_n/u_{m,nn}} \right)^2 + \left(\frac{\Im\{\psi_m[s_n]\} - \Im\{a_m\}}{r_n/u_{m,nn}} \right)^2 < 1.$$

This form follows that of a $2M$-dimensional hyper-ellipsoid region $\mathcal{E}_n$ with axes of length $r_n/u_{m,nn}$ along the m^{th} real and imaginary components, and center

$$c_n = (\Re\{a_1\}, \ldots, \Re\{a_M\}, \Im\{a_1\}, \ldots, \Im\{a_M\}).$$

To find the candidate values of s_n inside $\mathcal{E}_n$, a hyper-box region $\mathcal{D}_n$ is defined that has center c_n and edge length $2r_n/u_{m,nn}$ along the m^{th} real and imaginary components. Since $\mathcal{D}_n$ tightly bounds $\mathcal{E}_n$, all points in $\mathcal{E}_n$ will also be in $\mathcal{D}_n$. Assuming that all symbols $\psi_m[s_n] \in \mathcal{C}$ are sorted along the M real and M imaginary dimensions, a set of labels $\mathcal{S}_n^{'}$ inside $\mathcal{D}_n$ is easily created. The set $\mathcal{S}_n$ of candidate labels that fall in $\mathcal{E}_n$ are discovered by evaluating (5) for all labels in $\mathcal{S}_n^{'}$. The details of using sorted lists to create $\mathcal{S}_n^{'}$ are found in the work of Nene and Nayar, who indicate that the computational complexity of this search algorithm is generally independent of the problem dimensionality [15]. We note that the sorted lists can be created off-line and stored in memory at the receiver.

5. APPLYING MAPPING DIVERSITY

To transmit the same label M times via M distinct mappings, effective mapping functions $\psi_1, \ldots, \psi_M$ should be selected. Previous works studied systems where the transmitter and receiver each had a single antenna [4, 5]. Thus, any transmitted label was considered individually, without any interference from other labels within the packet. With multiple antennas, clearly this is not the case, as all N labels simultaneously interfere with each other. However, as illustrated in (5), there are advantages to varying the mappings. In the initial mapping $\psi_1[s_n]$, consider any pair of labels that are mapped to a pair of closely spaced symbols in $\mathcal{C}$. Additional mappings $\psi_2, \ldots, \psi_M$ can increase the Euclidean distance between this pair of labels by mapping them to symbols in $\mathcal{C}$ much further apart. This obviously reduces the probability of label misdetection. It also distributes the points $\psi_1[s_n], \ldots, \psi_M[s_n]$ more evenly, so that fewer candidates will fall inside $\mathcal{E}_n$ (and $\mathcal{D}_n$), thereby reducing the computational complexity.

With these points in mind, we consider each transmitted label individually in choosing effective mappings. In this context, the variables $u_{1,nn}, \ldots, u_{M,nn}$ roughly act as fading gains on M transmissions of the label s_n. These transmissions take the form of $z_m = u_m \psi_m[s] + w_m$, with w_m being white, Gaussian noise as before. We now review existing techniques to develop mappings in flat-fading channels [5].

For finding mappings, optimality is defined as minimizing the bit error rate (BER). A generic BER upper bound was developed for M transmissions under symbol mapping diversity:

$$\sum_{s=0}^{|\mathcal{C}|-1} \sum_{\substack{v=0 \\ v \neq s}}^{|\mathcal{C}|-1} \Pr\{s\} B[s,v] \Pr\{\alpha_M[v] < \alpha_M[s] \mid s\}, \qquad (6)$$

where $\Pr\{s\}$ is the probability of transmitting s (generally $1/|\mathcal{C}|$),

$$B[s,v] = \frac{(\text{number of differing bits between } s \text{ and } v)}{\log_2 |\mathcal{C}|}$$

is defined to account for the number of bit errors that result from a label misdetection, and $\Pr\{\alpha_M[v] < \alpha_M[s] \mid s\}$ is the pairwise error probability (PEP) that v is more likely to be detected than s given that s is transmitted. Here, the metric $\alpha_M[s]$ is defined as

$$\sum_{m=1}^{M} |z_m - u_m \psi_m[s]|^2.$$

For ARQ, a simpler, and probably sub-optimal, iterative solution is performed by computing the $M-$th mapping given the previous $M-1$ mappings, usually starting with any Gray mapping for $M = 1$. In this iterative, best effort approach, future mappings are justifiably ignored since additional packet transmissions are undesirable. Our optimization problem then simplifies to

$$\min_{\psi_M \in \Psi} \sum_{s=0}^{|\mathcal{C}|-1} \sum_{\substack{t=0 \\ t \neq s}}^{|\mathcal{C}|-1} g\left[s, \psi_M[s], t, \psi_M[t]\right], \qquad (7)$$

where Ψ is the set of all possible mappings and $g[s,a,t,b]$ is the pairwise BER that results from mapping label s to symbol $a \in \mathcal{C}$ and label t to symbol $b \in \mathcal{C}$ in the M^{th} mapping,

$$g[s,a,t,b] = \Pr\{s\} B[s,t] Pr\{\delta < 0\}.$$

Though still computationally difficult, (7) falls into a category of combinatorial optimization problems referred to as the Quadratic Assignment Problem (QAP), and exact and approximate solvers are available. For the optimization posed in (7),

$$\delta = (\alpha_{M-1}[k] - \alpha_{M-1}[s]) + |z_M - u_M b|^2 - |z_M - u_M a|^2,$$

leading to a PEP $\Pr\{\delta < 0\}$ of

$$E_{\mathbf{u}} \left\{ Q\left(\sqrt{\frac{1}{2\sigma_w^2} \left(\sum_{m=1}^{M-1} |u_m|^2 |d_m[s,t]|^2 + |u_M|^2 |a-b|^2 \right)} \right) \right\}.$$

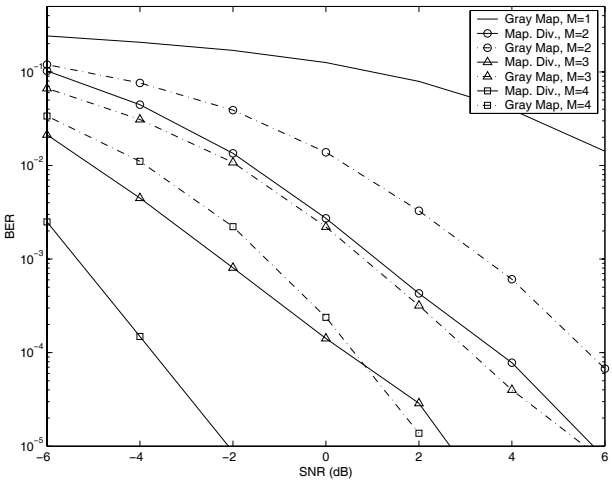

Fig. 2. BER of $M = 2, 3, 4$ transmissions through a 4×4 MIMO channel using 16QAM constellations.

with $d_m[s,t] = \psi_m[s] - \psi_m[t]$ and $\mathbf{u} = \{u_1, \ldots, u_M\}$. This PEP is numerically computed to provide the function $g[s, a, k, b]$ to solve (7), producing optimal mappings for fading channels [5].

6. SIMULATION RESULTS

We briefly present simulation results that demonstrate the effectiveness of mapping diversity in MIMO environments. The modulation type is 16QAM in a system with $N = 4$ transmit and $K = 4$ receiver antennas. With a packet size of 800 bits (50 blocks), we compare retransmissions made using mapping diversity against those using identical Gray mappings, for $M = 2, 3, 4$ transmissions. The channels $\mathbf{H}_1, \ldots, \mathbf{H}_4$ are independent, and remain constant through the (re)transmission of the packet.

Fig. 2 contains BER results obtained via Monte Carlo simulation of 2000 packets, with SNR defined as $E\{|s_n|^2\}/\sigma_w^2$. Clearly, mapping diversity provides significant gains in BER, with about a 2 dB improvement for $M = 2$ and over a 4 dB gain for $M = 4$. A useful observation is the inherent gain achieved from the channel variation between retransmissions, regardless of the mapping strategy employed. In cases where channels do not vary, block precoding or permutation of retransmissions should sufficiently vary the effective transmission channel.

7. CONCLUSION

An ARQ protocol for MIMO flat-fading channels was proposed that incorporates mapping diversity. By combining distinctly mapped transmissions through MIMO channels, significant gains are achieved. The joint receiver involves the application of sphere decoding, with a significant modification required for the enumeration of candidates within the sphere. Borrowing concepts from an existing closest point search technique employed for pattern matching applications, a fast enumeration method is readily available.

Extensions of the work include the use of precoding for channels that do not vary significantly between retransmissions, and a more detailed complexity analysis of the enumeration method. We note that the ARQ mechanism described here may double as a flexible, yet effective space-time coding scheme, particularly with precoding for non-varying channels.

8. REFERENCES

[1] D. Chase, "Code combining–a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. Commun.*, vol. 33, pp. 385–393, May 1985.

[2] B. Harvey and S. Wicker, "Packet combining systems based on the Viterbi decoder," *IEEE Trans. Commun.*, vol. 42, no. 2/3/4, pp. 1544–1557, Feb/Mar/Apr 1994.

[3] K. Narayanan and G. Stuber, "A novel ARQ technique using the turbo coding principle," *IEEE Commun. Lett.*, vol. 1, no. 2, pp. 49–51, Mar. 1997.

[4] H. Samra, Z. Ding, and P. M. Hahn, "Optimal symbol mapping diversity for multiple packet transmissions." Hong Kong, China: ICASSP, 2003.

[5] H. Samra and Z. Ding, "Symbol mapping diversity design for packet retransmissions through fading channels." San Francisco, CA: Global Telecommunications Conference, 2003.

[6] G. J. Foschini and M. J. Gans, "On limits of wireless communication in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, no. 3, pp. 311–315, Mar. 1998.

[7] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communications: Performance criterion and code construction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 744–765, Mar. 1998.

[8] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999.

[9] E. N. Onggosanusi, A. G. Dabak, Y. Hui, and G. Jeong, "Hybrid ARQ transmission and combining for MIMO systems," in *Proceedings of the IEEE Intl. Conf. on Commun.*, vol. 5, May 2003, pp. 3205–3209.

[10] Z. Ding and M. Rice, "Type-I hybrid-ARQ using MTCM spatio-temporal vector coding for MIMO systems," in *Proceedings of the IEEE Intl. Conf. on Commun.*, vol. 4, May 2003, pp. 2758–2762.

[11] A. V. Nguyen and M. A. Ingram, "Hybrid ARQ protocols using space-time codes," in *Proceedings of the 54th IEEE Veh. Tech. Conf.*, vol. 4, Oct. 2001, pp. 2364–2368.

[12] B. Hassibi and H. Vikalo, "On the expected complexity of integer least-squares problems," in *Proceedings of the IEEE ICASSP*, vol. 2, May 2002, pp. 1497–1500.

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

[14] 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.

[15] S. A. Nene and S. K. Nayar, "A simple algorithm for nearest neighbor search in high dimensions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 9, pp. 989–1003, Sept. 1997.