Soft Output Detection for Multiple Antennas: Accelerated Sphere Decoding and Shifted List Enumeration

Joseph BOUTROS[1], Nicolas GRESSET[1,2], Loïc BRUNEL[2], Marc FOSSORIER[1,3]

[1] ENST, 46 Rue Barrault, 75013 Paris, France

[2] Mitsubishi Electric ITE-TCL, 1 allée de Beaulieu, 35708 Rennes, France

[3] Electrical Engineering Dept., University of Hawaii, Honolulu HI 96822, USA

boutros@enst.fr, gresset@enst.fr, brunel@tcl.ite.mee.com, marc@aravis.eng.hawaii.edu

Résumé – Nous proposons un détecteur APP de faible complexité capable de démoduler les constellations MAQ émises sur un canal à antennes multiples (jusqu'à 16 antennes). Le détecteur APP recherche le point de vraisemblance maximale (VM) en appliquant un décodage par sphères accéléré. A l'aide d'une énumération de Pohst à double récursion, il construit ensuite une liste centrée sur le point VM afin d'évaluer les probabilités a posteriori. Le rayon de la liste est choisi de façon à stabiliser la taille de la liste et tenir compte des frontières de la constellation finie. En utilisant une modulation MAQ-16 sur un canal symétrique à 4 antennes, un turbo code de rendement 1/2 et des blocs de longueur 20000 bits, nous observons une différence de 1,56 dB seulement entre les performances simulées et la capacité du canal au sens de Shannon.

Abstract – We propose a low complexity APP detector for demodulating QAM constellations transmitted on a MIMO channel (up to 16 antennas). The APP detector starts by applying an accelerated sphere decoder to find the maximum likelihood (ML) point. Then, using a double recursion Pohst enumerator, a shifted list is built around the ML point to evaluate the output APP. The list radius is selected in order to control the list size and to cope with the boundaries of the finite multiple antenna constellation. With a rate 1/2 turbo code and a blocklength equal to 20000 bits, we achieved a bit error rate at 1.56 dB from Shannon capacity limit while transmitting a 16-QAM on a 4x4 multiple antenna channel.

1 Introduction and system model

The growing importance of iterative processing in communication systems [2] during the last decade has permitted to attain exceptional performance on different kind of data transmission channels, e.g., bit-interleaved coded modulations (BICM) [12][4] on multiple antenna (MIMO) channel combined with joint detection and decoding at the receiver side [3][7]. When the channel number of dimensions and the number of points per dimension increase, the classical exhaustive soft output channel detector becomes intractable. We present a new non exhaustive spherical list centered on the maximum likelihood (ML) point to compute weighted channel likelihoods. For example, our APP detector achieves a bit error rate at 1.56 dB from Shannon capacity limit while transmitting a 16-QAM on a 4x4 multiple antenna channel.

Information bits are protected by an error-correcting code and interleaved by a pseudo-random permutation. The coded bits are then mapped into points of a 16-QAM constellation. Symbols are conveyed on a multiple antenna or multiple input multiple output (MIMO) channel with n_t transmit antennas and n_r receive antennas. Let $z \in \mathbb{C}^{n_t}$ denote the MIMO channel input and $y \in \mathbb{C}^{n_r}$ the MIMO channel output. Channel input and output are linked via the non-selective Rayleigh fading model:

$$y = zH + \nu = x + \nu \tag{1}$$

where $H = [h_{ij}]$ is an $n_t \times n_r$ complex matrix. The entries h_{ij} of the channel matrix are complex random variables with a Gaussian probability distribution of zero mean and unity variance. This complex system can easily be converted into a real system and viewed as a lattice Λ in $\mathbb{R}^{n_s}$ [5] perturbed by additive noise. The real space dimension is $n_s = 2n_t$. The equality $x = zH$ is now extended to the real space $\mathbb{R}^{n_s}$ to get

$$x = zM, \quad x \in \mathbb{R}^{n_s}, \quad z \in \mathbb{Z}^{n_s} \tag{2}$$

A lattice Λ is a discrete subgroup of $\mathbb{R}^{n_s}$ [5], i.e., it is a $\mathbb{Z}$-module of rank n_s. In (2), the lattice Λ is generated by the $n_s \times n_s$ real matrix $M = [M_{ij}]$ which is derived from the channel matrix H by the following simple relation

$$M_{ij} = \begin{pmatrix} \Re h_{ij} & \Im h_{ij} \\ -\Im h_{ij} & \Re h_{ij} \end{pmatrix} \tag{3}$$

where $\Re h_{ij}$ and $\Im h_{ij}$ denote the real and imaginary part of h_{ij}, respectively. The matrix M is called *lattice generator matrix*. Geometrically, the point x belongs to a discrete infinite set of points satisfying a group structure. When z is restricted to a finite QAM integer constellation, then x belongs to a finite lattice constellation denoted by Ω. With the above notations and a 2^m-QAM modulation, the cardinality of Ω is 2^{mn_t}.

Some evident capacity optimization arguments led us to choose $n_t = n_r$. The lattice representation of the MIMO channel allows us to use related theory and decoding al-

gorithms adapted to digital communications and particularly to finite constellations problems.

2 Our APP detector for multiple antennas

The direct method for establishing a non exhaustive APP detector is to build a spherical list of lattice points around y. Such lattice point enumeration is achieved in polynomial time via Pohst enumeration [8][6] inside a sphere of squared radius R^2. The number N_p of lattice points inside the sphere can be well approximated by

$$N_p \approx \frac{V_{n_s} \times R^{n_s}}{vol(\Lambda)} \tag{4}$$

where V_{n_s} is the volume of a unit radius sphere in $\mathbb{R}^{n_s}$ and $vol(\Lambda) = |det(M)|$ is the lattice fundamental volume. Such a method applied in [7] is still too complex: the lattice constellation Ω to be decoded is finite. Centering the list around y makes N_p completely unstable. The instability is due to the additive noise and to the exact position of the transmitted point within Ω. Thus, there is no evident relation similar to (4) between R^2 and N_e, the effective number of constellation points inside the spherical boundaries.

We propose a spherical list centered around the ML point x_{ML}. Since x_{ML} belongs to $\Omega \subset \Lambda$, the fraction N_e/N_p can be well selected depending on the position of x_{ML} and the shape of Ω. The fraction N_e/N_p should be controlled by taking into account the influence of these two parameters. As an example, consider a cubic integer constellation Ω in $\mathbb{R}^{n_s}$. We can evaluate the average number of points in the list by the simple hypothesis that it is divided by 2 for each dimension where the ML point is on the edge of the constellation. For a 16-QAM with $n_t = n_s/2$ transmit antennas, we have $C_{n_s}^i \times 2^{n_s}$ points with i components on the edge of the constellation. This leads to an average number of points for 16-QAM equal to

$$E[N_e] = \sum_{i=0}^{n_s} C_{n_s}^i \cdot \frac{2^{n_s}}{4^{n_s} \cdot 2^i} N_p = \left(\frac{3}{4}\right)^{n_s} N_p \tag{5}$$

When $n_t = 4$, the average reduction factor is about 1/10. Hence, in the general case of a random (non-cubic) constellation Ω given by the MIMO channel, we can adjust the sphere radius by taking into account the number of hyperplanes n_{hyp} at the constellation boundaries passing through the ML point. The number of expected points N_p is multiplied by $\alpha[n_{hyp}]$, an expansion factor of the list size which depends on n_{hyp}. Indeed, the higher the number of hyperplanes the ML point belongs to, the less the points in the list. For the special case of MIMO channels family, the empirical choice $\alpha[i] = \lfloor i/2 \rfloor + 1$ yields good results.

The number N_e is also influenced by the shape of Ω. A non-cubic shape with an acute corner attenuates the fraction N_e/N_p. We selected the normalized distance $\gamma(G)$ as a figure of merit for the shape of Ω:

$$\gamma(G) = \frac{d_{Emin}^2(G)}{vol(\Lambda)^{2/n_s}} \tag{6}$$

where $d_{Emin}^2(G)$ is the minimum diagonal element of the Gram matrix $G = MM^t$. Then, we introduce an additional expansion factor $\mu_\gamma > 1$ and apply the simple rule:

$$\gamma(G) > \gamma_i \Rightarrow \mu_\gamma = \mu_i \tag{7}$$

In the 8-dimensional real space, one may take $\gamma_1 = 3dB$, $\gamma_2 = 6dB$, $\mu_1 = 4$, and $\mu_2 = 16$. Finally, the radius R of the shifted spherical list $\mathfrak{L}$ should be computed according to

$$R = \left(\frac{\alpha[n_{hyp}] \times \mu_\gamma \times N_p \times vol(\Lambda)}{V_{n_s}} \right)^{\frac{1}{n_s}} \tag{8}$$

Our APP detector starts by applying an accelerated sphere decoder to find x_{ML} and then it builds the list using a double Pohst recursion for channel likelihoods evaluation. The final APP is determined by mixing the likelihoods and the a priori information via a sum-product formula. It is worth to note that, except for the final sum-product formula, the main processing done by our APP detector (accelerated sphere decoder+shifted spherical list) is outside the iterative detection/decoding loop.

2.1 Accelerated sphere decoding algorithm

A very efficient algorithm to find the closest point in a lattice when observing any point in the real space is the sphere decoder [10][11]. The main idea of this algorithm is to enumerate the lattice points that belong to a sphere centered on y and to calculate the corresponding Euclidean distances. The point that minimizes the distance is called the closest point (x_{ML}). If no point is found, the radius of the sphere should be enlarged. Each time a point is found, the radius of the sphere can be reduced to the distance of this new point, which limits the number of points enumerated but still ensures the closest point criterion.

There are two main strategies for point enumeration. The first was proposed by Pohst [8][6] and applied by Viterbo et al (VB) [10][11] to digital communications. The second was proposed by Schnorr and Euchner [9] and applied by Agrell et al (AEVZ) in [1]. On multiple antenna channels, VB and AEVZ complexities are similar at moderate and high signal-to-noise ratios. At low signal-to-noise ratios, AEVZ may show a speed improvement with respect to VB by a factor varying from 1 up to 4. We modified AEVZ sphere decoder in order to take into account the QAM constellation boundaries. We did not apply basis reduction (LLL or KZ) to the accelerated sphere decoder because basis reduction is incompatible with a fast checking of the QAM constellation boundaries. The finite constellation nature of the system allows to significantly reduce the complexity of the sphere decoder by dynamically modifying the bounds of research depending on those of the constellation. The accelerated sphere decoder is capable of ML performance on MIMO channels (up to 16 antennas) with a reasonable complexity (see Figure 1).

We give below the complete steps of the modified AEVZ sphere decoder.

Accelerated Sphere Decoder: applying Schnorr-Euchner strategy, and taking into account the boundaries of the finite QAM constellation

Input. A received point y, the generator matrix $M(n_s \times n_s)$ of the lattice, the radius R of the sphere, and the bounds z_{min} and z_{max} of the constellation. You can set the radius R to $+\infty$. A slight gain in speed of at most 30% can be obtained if R is linked to the Gaussian noise variance $\sigma^2 = N_0$ or to the minimum distance $d_{Emin}(\Lambda)$

Output. The ML point z_{ML} belonging to the constellation and its squared Euclidean distance to y

Step 1. (Pre-processing) Compute the Gram matrix $G = MM^t$ and do a Cholesky decomposition $G = VV^t$, where V is lower-triangular. Compute the inverse $V^I = V^{-1}$

Step 2. (Initialization) Set $bestdist \leftarrow R^2$, $k \leftarrow n_s$, $dist_k \leftarrow 0$, $e_k \leftarrow yV^I$, $z_k \leftarrow [e_{kk}]$, $z_k \leftarrow max(z_k, z_{min})$, $z_k \leftarrow min(z_k, z_{max})$, compute $\rho = (e_{kk} - z_k)/(V^I_{kk})$, $step_k \leftarrow sign(\rho)$

Step 3. Compute $newdist \leftarrow dist_k + \rho^2$. If $newdist < bestdist$ and $k \neq 1$ then go to 4 else go to 5 endif

Step 4. Compute for $i = 1, ..., k-1$ $e_{k-1,i} \leftarrow e_{k,i} - \rho V^I_{ki}$, decrement k, set $dist_k \leftarrow newdist$, $z_k \leftarrow [e_{kk}]$, $z_k \leftarrow max(z_k, z_{min})$, $z_k \leftarrow min(z_k, z_{max})$, $\rho = (e_{kk} - z_k)/(V^I_{kk})$, $step_k \leftarrow sign(\rho)$, go to 3

Step 5. If $newdist < bestdist$ then set $\hat{z} \leftarrow z$, $bestdist \leftarrow newdist$, else if $k = n$ then return $\hat{z}$ and terminate, else increment k, endif. Compute $z_k \leftarrow z_k + step_k$, if $z_k < z_{min}$ or $z_k > z_{max}$ then $step_k \leftarrow -step_k - sign(step_k)$, $z_k \leftarrow z_k + step_k$ endif. If $z_k < z_{min}$ or $z_k > z_{max}$ then go to 5, endif. $\rho \leftarrow (e_{kk} - z_k)/V^I_{kk}$, $step_k \leftarrow -step_k - sign(step_k)$, go to 3

2.2 APP evaluation based on a shifted spherical list

Given a list $\mathfrak{L}$ of constellation points in $\mathbb{R}^{n_s}$, the approximated extrinsic probability $\xi(c_j)$ of a coded bit c_j is given by the following normalized marginalization:

$$\tilde{\xi}(c_j) = \frac{\sum_{z' \in \Omega(c_j=1) \cap \mathfrak{L}} \left[\left(e^{-\frac{\|y - z'M\|^2}{2\sigma^2}} \right) \prod_{r \neq j} \pi(c_r) \right]}{\sum_{z \in \Omega \cap \mathfrak{L}} \left[\left(e^{-\frac{\|y - zM\|^2}{2\sigma^2}} \right) \prod_{r \neq j} \pi(c_r) \right]} \quad (9)$$

The subset $\Omega(c_j = 1)$ represents the set of points belonging to Ω with j-th bit equal to 1. The a priori probabilities $\pi(c_j)$ of the coded bits are fed back from a soft-input soft-output (SISO) decoder of the error-correcting code included in the BICM. Our spherical list $\mathfrak{L}$ is centered around the closest point x_{ML} found by the accelerated sphere decoder. The squared radius R^2 has been determined in a way to guarantee a moderate value for $N_e = |\mathfrak{L}|$ (e.g., 1000 points). The list size should not be too small (e.g., 10 points!), in order to guarantee an APP quality as if $\mathfrak{L} = \Omega$. On the contrary, the list size should not be too large (e.g., close to $|\Omega| = M^{n_t}$) in order to limit the detector complexity. The ML point and its neighbors in Ω yield the dominant likelihoods in the extrinsic probability generated by the APP detector.

The evaluation of the Euclidean distances between y and the points in $\mathfrak{L}$ has been optimized by applying a double Pohst recursion while enumerating the lattice points. Indeed, the first classical recursion is needed to check all constellation points at a squared distance less than R^2 from the center x_{ML}. We added a parallel second recursion centered on y to reduce the number of mathematical operations that compute the Euclidean distances $||y - zM||$ required in the extrinsic probability formula.

We give below the complete steps of the double recursion point enumeration inside $\mathfrak{L}$.

Soft Output Sphere Decoder: shifted spherical list enumeration with a double Pohst recursion

Input. A received point y, a constellation point x_{ML}, the generator matrix $M(n_s \times n_s)$ of the lattice, the radius R of the sphere according to (8) below, and the bounds z_{min} and z_{max} of the constellation

Output. A list $\mathfrak{L}$ of constellation points inside the sphere, a list of squared Euclidean distances between y and each point in the list

Step 1. (Pre-processing) Compute the Gram matrix $G = MM^T$ and do a Cholesky decomposition $G = VV^t$, V is lower-triangular. Cholesky decomposition produces an upper-triangular matrix $Q = [q_{ij}]$, $q_{ii} = V^2_{ii}$, and $q_{ij} = V_{ji}/V_{ii}$ for $i = 1 \ldots n_s$ and $j = i + 1 \ldots n_s$. Compute the inverse M^{-1}, $u = z_{ML} = x_{ML}M^{-1}$ and $\rho = yM^{-1}$. Notice that z_{ML} can be directly offered by the accelerated sphere decoder (section 2.1)

Step 2. (Initialization) Set $d^2 \leftarrow R^2$, $T_{n_s} \leftarrow R^2$, $T^d_{n_s} \leftarrow R^2$. For $j = 1 \ldots n_s$ set $S_j \leftarrow u_j$, $S^d_j \leftarrow \rho_j$, $i \leftarrow n_s$

Step 3. Compute $L_i \leftarrow min\left(\left\lfloor \sqrt{T_i/q_{ii}} + S_i \right\rfloor, z_{max}\right)$ and $z_i \leftarrow max\left(\left\lceil -\sqrt{T_i/q_{ii}} + S_i \right\rceil, z_{min}\right) - 1$

Step 4. Increment z_i. If $z_i > L_i$ if $i > 1$ compute $\xi_i \leftarrow u_i - z_i$ and $\xi^d_i \leftarrow \rho_i - z_i$, compute $T_{i-1} \leftarrow T_i - q_{ii}(S_i - z_i)^2$ and $T^d_{i-1} \leftarrow T^d_i - q_{ii}(S^d_i - z_i)^2$, compute $S_{i-1} \leftarrow u_{i-1} + \sum_{j=i}^{n_s} q_{i-1,j}\xi_j$ and $S^d_{i-1} \leftarrow \rho_{i-1} + \sum_{j=i}^{n_s} q_{i-1,j}\xi^d_j$, decrement i and go to 3, else compute $\hat{d}^2 \leftarrow R^2 - T^d_i + q_{11}(S^d_1 - z_1)^2$, store z and $\hat{d}$ in $\mathfrak{L}$, go to 4, endif, else if $i = n_s$ terminate else increment i and go to 4, endif, endif

3 Computer simulations and numerical results

The new soft output sphere decoder is applied to detect a 16-QAM modulation transmitted on a 4×4 ergodic MIMO channel. The error-correcting code is a rate 1/2 parallel Turbo code. The RSC constituent is the classical 4-state (1,5/7) code. The BICM interleaver size is 20000 bits. On Fig. 2, we can see that the above coded system achieves a distance 1.56 dB from Shannon capacity limit, under the constraint of a finite input 16-QAM alphabet. A supplementary signal-to-noise ratio gain of 0.30dB can be obtained with a large BICM interleaver of size 100000 at the expense of a greater latency.

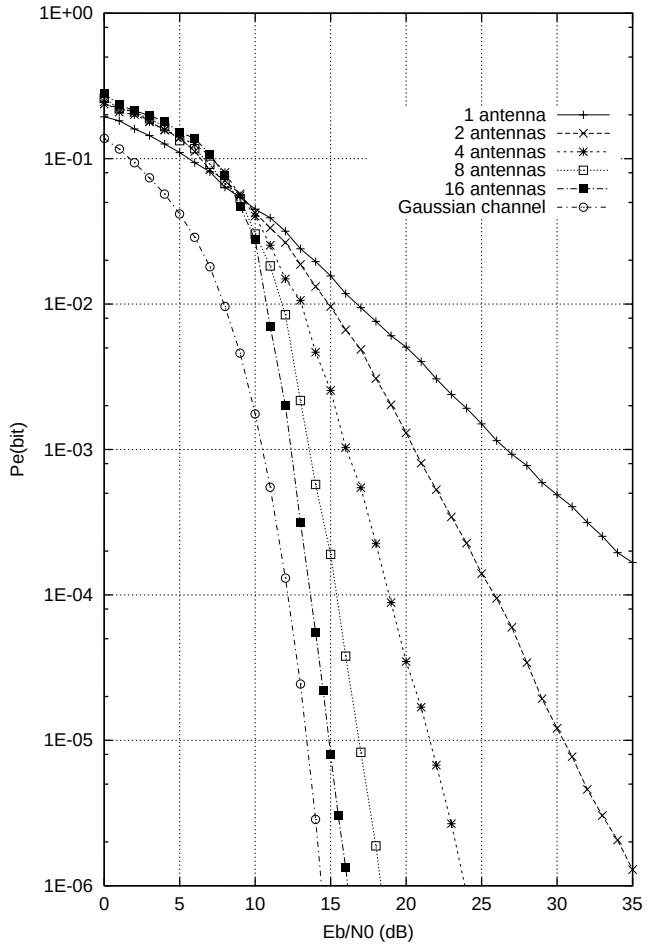

FIG. 1: Performance of the accelerated sphere decoder (ML), uncoded 16-QAM, ergodic Rayleigh MIMO channel, $n_t = n_r$ =1, 2, 4, 8 and 16 antennas.

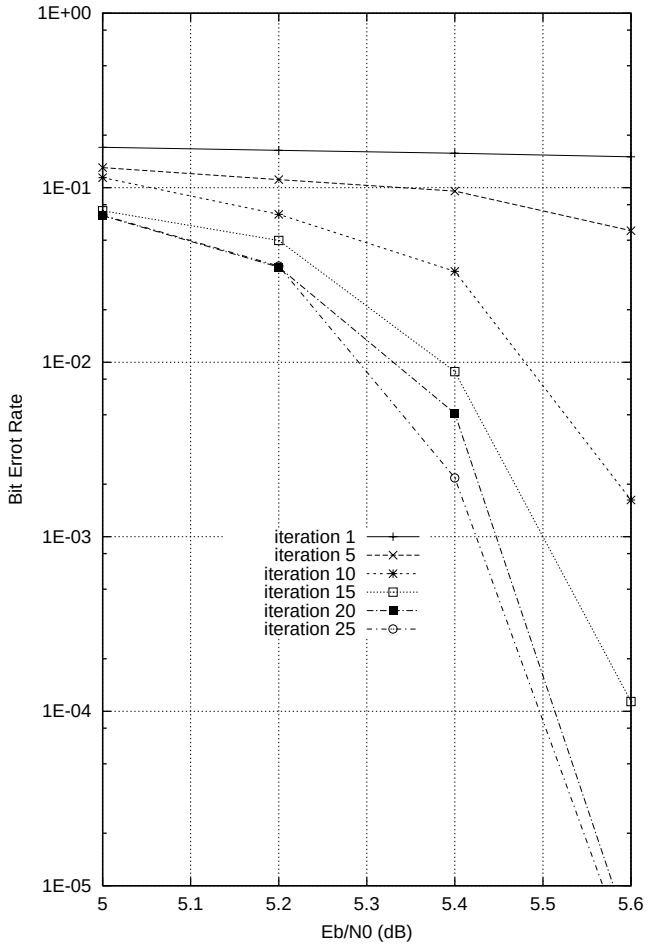

FIG. 2: Performance of the soft output sphere decoder, rate 1/2 Turbo coded 16-QAM, ergodic Rayleigh MIMO 4×4 channel, Shannon limit is at $E_b/N_0 = 4.0$ dB.

References

[1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. on Information Theory*, pp. 2201-2214, Aug 2002.

[2] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes," Proceedings of ICC'93, Geneva, pp. 1064-1070, May 1993.

[3] J. Boutros, F. Boixadera, and C. Lamy, "Bit-interleaved coded modulations for multiple-input multiple-output channels," proceedings of the *IEEE 6th International Symposium on Spread Spectrum Techniques & Applications*, New Jersey, September 2000.

[4] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. on Inf. Theory*, vol. 44, no. 3, May 1998.

[5] J.H. Conway and N.J. Sloane, *Sphere packings, lattices and groups*, 3rd ed., 1998, Springer-Verlag, New York.

[6] 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, pp. 463-471, April 1985.

[7] B. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," submitted to the *IEEE Transactions on Communications*, July 2001.

[8] M. Pohst, "On the computation of lattice vectors of minimal length, successive minima, reduced bases with applications," *ACM SIGSAM Bull.*, vol. 15, pp. 37-44, Feb 1981.

[9] C.P. Schnorr and M. Euchner, "Lattice basis reduction: improved practical algorithms and solving subset sum problems," *Mathematical Programming*, vol. 66, pp. 181-191, 1994.

[10] E. Viterbo and E. Biglieri, "A universal lattice decoder," 14^{eme} Colloque GRETSI, Juan-les-Pins, pp. 611-614, Sept. 1993.

[11] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. on Information Theory*, pp. 1639-1642, July 1999.

[12] E. Zehavi, "8-PSK trellis codes for a Rayleigh channel," *IEEE Transactions on Communications*, vol. 40, pp. 873-884, May 1992.