Factorization properties of the optimal signaling distribution of multi-dimensional QAM constellations

Yankov, Metodi Plamenov; Forchhammer, Søren; Larsen, Knud J.; Christensen, Lars P. B.

Published in:
Proceedings of ISCCSP 2014

Link to article, DOI:
10.1109/ISCCSP.2014.6877894

Publication date:
2014

Document Version
Peer reviewed version

FACTORIZATION PROPERTIES OF THE OPTIMAL SIGNALING DISTRIBUTION OF MULTI-DIMENSIONAL QAM CONSTELLATIONS

Metodi P. Yankov, Søren Forchhammer, Knud J. Larsen

Technical University of Denmark
Department of Photonics Engineering
2800 Kgs Lyngby, Denmark, meya@fotonik.dtu.dk

Lars P. B. Christensen

Fingerprint Cards
Lyskær 3CD, 2730 Herlev, Denmark

ABSTRACT

In this work we study the properties of the optimal Probability Mass Function (PMF) of a discrete input to a general Multiple Input Multiple Output (MIMO) channel. We prove that when the input constellation is constructed as a Cartesian product of 1-dimensional constellations, the optimal PMF factorizes into the product of the marginal 1D PMFs. This confirms the conjecture made in [1], which allows for optimizing the input PMF efficiently when the rank of the MIMO channel grows. The proof is built upon the iterative Blahut-Arimoto algorithm. We show that if the initial PMF is factorized, the PMF on each successive step is also factorized. Since the algorithm converges to the optimal PMF, it must therefore also be factorized.

Index Terms— MIMO, QAM, Constellation shaping

I. INTRODUCTION

As the data rate demand increases, the physical links in band limited scenarios are pushed to operate at high SNR and with high order of modulation in order to achieve high spectral efficiency. Furthermore, it is well known that at high SNR, uniformly distributed signaling achieves the Shannon capacity rate for 1.53dB more energy, called the *shaping gap* or *shaping gain*. In order to close that gap, continuously distributed Gaussian signaling is required [2]. While it is clear that such signals are completely described by their mean and variance, nothing explicit can be said about the shape of the optimal *Probability Mass Function* (PMF) of a given discrete signaling set.

We consider a standard *Multiple Input Multiple Output* (MIMO) channel model:

$$Y = HX + W \tag{1}$$

where X is $2M$ dimensional vector random variable, taking values from the discrete, real-valued set $\mathcal{X}^{2M}$, H is the $2Nx2M$ real-valued equivalent of the NxM complex channel matrix, W is the usual $2N$ dimensional AWGN and Y is the $2N$ dimensional channel output. The set $\mathcal{X}$ is the basic PAM constellation set. The channel is assumed to be perfectly known at the receiver unless otherwise stated.

The algorithm for finding the optimal PMF input to an AWGN channel was derived independently by Blahut [3] and Arimoto [4]. It uses *Expectation-Maximization* (EM) type update rules, sequentially increasing the concave cost function (in this case the *Mutual Information* (MI) as a function of the input distribution) subject to average power constraint $\sum_{i=1}^{|\mathcal{X}^{2M}|} p(\mathbf{x_i})|\mathbf{x_i}|^2 \leq P_{av}$, and normalization constraint $\sum_{i=1}^{|\mathcal{X}^{2M}|} p(\mathbf{x_i}) = 1$, where $\mathbf{x_i}$ is the $i'th$ element of the signaling set and $P(X = \mathbf{x_i}) = p(\mathbf{x_i})$ is its probability.

In [5], the authors replace the power constraint with equality, and then sweep all possible scaled versions of the signaling set, i.e. $\hat{\mathcal{X}^{2M}} = \alpha\mathcal{X}^{2M}$. The EM algorithm is then run for each of these sets, and the one achieving maximum MI is chosen as optimal. The optimal PMF of $\mathcal{X}^{2M}$ is the PMF of the optimal set, and its respective MI is the channel capacity. This algorithm was later extended to cover MIMO scenarios in [1]. Our work is largely based on the derivations made in [1], and so we briefly introduce the mathematical notation of the algorithm used there.

I-A. EM for finding the optimal distribution on MIMO channel with average power constraint

The channel capacity when signaling with $\mathcal{X}^{2M}$ and averaging among the possible channel realizations can be expressed as [1]:

$$C = \max_{p(X)} \mathcal{I}(X;Y|H) =$$
$$= \max_{p(X)} \sum_{i=1:|\mathcal{X}^{2M}|} p(\mathbf{x_i}) \left(\log_2 \frac{1}{p(\mathbf{x_i})} + T_i \right) \tag{2}$$

where,:

$$T_i = \int_{\mathbf{H}} p(\mathbf{H}) \int_{\mathbf{y}} p(\mathbf{y}|\mathbf{x_i},\mathbf{H}) \log_2 p(\mathbf{x_i}|\mathbf{y},\mathbf{H}) d\mathbf{y} d\mathbf{H} \tag{3}$$

We have replaced $p(H = \mathbf{H})$ with $p(\mathbf{H})$ for simplicity and $\mathcal{I}$ is the MI. The EM algorithm is as follows:

Sweep $\alpha \in [\alpha_{min}; \alpha_{max}]$

- **Initialization:**
  Initialize $p(\mathbf{x})$, so that it satisfies $\sum_{i=1}^{|\mathcal{X}^{2M}|} p(\mathbf{x_i}) = 1$, and $\sum_{i=1}^{|\mathcal{X}^{2M}|} p(\mathbf{x_i})|\alpha\mathbf{x_i}|^2 = P_{av}$

- **E:** For each point i and fixed $p(\mathbf{x_i})$, compute T_i.
- **M:** Given the set of fixed T_i values, maximize $\mathcal{I}(X;Y|H)$ w.r.t. the probability of each point. This is a strictly concave optimization problem, solved using Lagrange multipliers.
- **Go to 'E' until convergence**

As the order of modulation and number of dimensions of the signal grow, this algorithm is impractical even when performed offline due to the exponential increase in the number of parameters to be optimized. In [1] the authors conjecture, that the optimal PMF factorizes into the PMFs of each dimension, however, no theoretical proof has been provided. In this work we prove the conjecture. We also show how the Blahut-Arimoto (BA) algorithm can be used to solve the power-allocation problem for discrete signaling sets.

I-B. Notation

In the rest of the paper we use the following notation: x_i denotes the $i'th$ value from the set $\mathcal{X}$. The one-dimensional variable X_i represents dimension i, and $p(X_i = x_j) = p(\mathbf{x_{ij}})$ is the probability of that variable taking value x_j. Capacity achieving PMF will be denoted as $p(X)^*$, and $p(X)^n$ will be the PMF at the beginning of the **E** step on the $n'th$ iteration. The entropy function is denoted as $\mathcal{H}(\cdot)$. Dependency on the noise variance is omitted in the expressions for probability and entropy for brevity.

II. FACTORIZING THE OPTIMAL QAM DISTRIBUTION

In the following, without loss of optimality of the EM algorithm, we assume it has been initialized with a distribution, which is symmetric around 0, and which factorizes as $p(X)^1 = \prod_{k=1:2M} p(X_k)^1$.

Theorem 1. *If $p(X)^n = \prod_{k=1:2M} p(X_k)^n$, then at step n, for fixed $p(X|Y,H)$, the conditional entropy $\mathcal{H}(X_k|X_{\{1:2M\}\setminus k}, Y, H = \mathbf{H})$ is independent of $p\left(X_{\{1:2M\}\setminus k}\right)^{n+1}$*

Proof. See Appendix A □

Theorem 2. *If $p(X)^n = \prod_{k=1:2M} p(X_k)^n$, then $p(X)^{n+1} = \prod_{k=1:2M} p(X_k)^{n+1}$, and therefore $p(X)^* = \prod_{i=1:2M} p(X_k)^*$*

Proof. The MI after step n for real-valued 2x2 MIMO channel can be written as:

$$\mathcal{I}(X;Y|H) = \mathcal{H}(X|H) - \mathcal{H}(X|Y,H) = \mathcal{H}(X_1) + \mathcal{H}(X_2|X_1) - \mathcal{H}(X_1|Y,H) - \mathcal{H}(X_2|X_1,Y,H) \tag{4}$$

where we have used the chain rule for entropy and conditional entropy. It is clear that $\mathcal{H}(X_1|Y,H)$ is independent of $p(X_2)^{n+1}$. After Theorem I, $\mathcal{H}(X_2|X_1,Y,H)$ is also independent of $p(X_1)^{n+1}$, or the conditional entropy $\mathcal{H}(X|Y,H)$ is separated into functions of the marginal PMFs. Let's assume that the optimal PMF at this step was found as $p(X)^*$, and its respective marginals are $p(X_k)^* = \sum_{X_{\{1:2M\}\setminus k}} p(X)^*$. We then consider the PMF, obtained as product of those marginals $p(X)^{\tilde{}} = \prod_{k=1:2M} p(X_k)^*$. The entropy of X as a function of this PMF is:

$$\mathcal{H}(p(X)^{\tilde{}}) = \mathcal{H}(X_1) + \mathcal{H}(X_2) \geq \mathcal{H}(p(X)^*) \tag{5}$$

Then for the MI as a function of the PMF we have:

$$\mathcal{I}(p(X)^{\tilde{}}) = \mathcal{H}(p(X)^{\tilde{}}) - \mathcal{H}(X|Y,H) \geq \mathcal{H}(p(X)^*) - \mathcal{H}(X|Y,H) = \mathcal{I}(p(X)^*) \tag{6}$$

However, $\mathcal{I}(p(X)^*) \geq \mathcal{I}(p(X))$ for any $p(X)$, and therefore $\mathcal{I}(p(X)^*) = \mathcal{I}(p(X)^{\tilde{}})$. In [6] the authors prove, that the MI is strictly concave in $p(X)$, and therefore the optimal distribution is unique. The optimal distribution therefore must be the same as the product of its marginals. The extension to $M > 2$ is straight-forward. The MI after step n can be expressed as:

$$\mathcal{I}(p(X)^{n+1}) = \mathcal{H}(X) - \sum_{k=1:2M} \mathcal{H}(X_k|X_{\{1:k-1\}}, Y, H) \tag{7}$$

where each element in the sum only depends on its respective marginal PMF. Then if $\mathcal{I}(p(X)^{n+1}) = \max \mathcal{I}(p(X)^{n+1}) \Rightarrow \mathcal{H}(X_m|\{X_{1:2M}\}\setminus X_m) = \max \mathcal{H}(X_m|\{X_{1:2M}\}\setminus X_m)$, and therefore $p(X_m|\{X_{1:2M}\}\setminus X_m)^{n+1} = p(X_m)^{n+1}$. Since $p(X)^{n+1} = \prod_{i=1:M} p(X_i)^{n+1}$, the theorem is proven. □

II-A. Modified BA algorithm

The channel capacity of the simple 2x2 real-valued channel can now be expressed as:

$$C = \max_{p(X)} \mathcal{I}(X;Y|H) = \max_{p(X_1)} (\mathcal{I}(X_1;Y|H)) + \max_{p(X_2)} (\mathcal{H}(X_2) - \mathcal{H}(X_2|X_1,Y,H)) \tag{8}$$

The maximization problem is separated into 2 maximization problems of much lower dimensionality, and power constrain $P_{av}^k = \sum_{l=1:|\mathcal{X}|} p(x_l)\alpha_k |x_l|^2$, where α_k is the scaling coefficient for the $k'th$ dimension. The degrees of freedom in the maximization problem are reduced from $|\mathcal{X}|^{2M}$ to only $|\mathcal{X}|$. If we assume symmetric distribution on H, it is clear that $p(X_1)^* = p(X_2)^* = \cdots = p(X_{2M})^*$, and $P_{av}^1 = P_{av}^2 = \cdots = P_{av}^{2M} = P_{av}/2M$. In that case the modified BA algorithm from Section I takes the form:

Sweep $\alpha \in [\alpha_{min}; \alpha_{max}]$

- **Initialization:**
  Initialize $p(X_1)$, so that it satisfies $\sum_{i=1}^{|\mathcal{X}|} p(\mathbf{x_{1i}}) = 1$, and $\sum_{i=1}^{|\mathcal{X}|} p(\mathbf{x_{1i}})|\alpha x_i|^2 = \frac{P_{av}}{2M}$

- **E:** For each point $x_i \in \mathcal{X}$ and fixed $p(x_i)$, compute $\hat{T}_i$:

$$\hat{T}_i = \int_{\mathbf{H}} p(\mathbf{H}) \int_{\mathbf{y}} p(\mathbf{y}|\mathbf{x_{1i}}, \mathbf{H}) \log_2 p(\mathbf{x_{1i}}|\mathbf{y}, \mathbf{H}) \quad (9)$$

- **M:** Given the set of fixed $\hat{T}_i$ values, find $p(X_1)$ which maximizes $I(X_1; Y|H)$ as:

$$p(X_1) = \arg\max \mathcal{I}(X_1; Y|H) =$$
$$\arg\max \sum_{i=1:|\mathcal{X}|} p(\mathbf{x_{1i}}) \left(\log_2 \frac{1}{p(\mathbf{x_{1i}})} + \hat{T}_i \right) \quad (10)$$

- **Go to 'E' until convergence**

III. FURTHER DISCUSSION

III-A. Orthogonal channel

Let us consider the case, when $\mathbf{H}$ is diagonal. When the noise is i.i.d. Gaussian, the likelihood on this channel can be expressed as a product of the likelihoods on each dimension:

$$p(Y|X, \mathbf{H}) = \prod_{k=1:2M} p(Y_k|X_k, \mathbf{H_{kk}}) \quad (11)$$

where $\mathbf{H_{kk}}$ is the element on the $k'th$ row and $k'th$ column (we assume the channel has full rank). In this case it is straightforward to prove that the entropy $\mathcal{H}(X|Y, H)$ can be expressed as a sum of functions of the marginal PMFs, namely $\mathcal{H}(X|Y, H) = \sum_{k=1:2M} \mathcal{H}(X_m|Y_m, H)$. It suffices to show that at step n, the conditional distribution $p(X|Y, H)$ factorizes as:

$$p(X|Y, H) = \frac{p(Y|X, \mathbf{H}) p(X)^n}{\sum_{k=1:|\mathcal{X}^{2M}|} p(Y|\mathbf{x}_k, \mathbf{H}) p(\mathbf{x}_k)^n} =$$
$$\frac{\prod_{k=1:2M} p(Y_k|X_k, \mathbf{H_{kk}}) p(X_k)}{\prod_{k=1:2M} \sum_{i=1:|\mathcal{X}|} p(Y_k|\mathbf{x_{ki}}, \mathbf{H}) p(\mathbf{x_{ki}})^n} =$$
$$\prod_{k=1:2M} p(X_k|Y_k, \mathbf{H}) \quad (12)$$

III-B. Non-symmetric channel matrix distribution

When the channel distribution is not symmetric, the power constraint on each maximization problem is not necessarily the same. In order to find these constraints, a power allocation solution is needed. However, this can be circumvented if the scaling coefficient α can be chosen differently for each dimension, and the maximization problem is run on the full-rank system for each allowed $\alpha = [\alpha_1, ...\alpha_{2M}]^T$ and the set $\mathcal{X}^{2M} = \prod_{k=1:2M} \alpha_k \mathcal{X}$. As usual, the set, achieving the maximum MI will be chosen as optimal. Since for fixed $p(X)$, α_k directly gives the power, allocated to dimension k, it is clear that the power allocation is obtained by the modified Blahut-Arimoto algorithm. When channel state information at the transmitter is not available or is imperfect, and symmetry in the channel distribution cannot be assumed, the optimization problem cannot be separated into problems of reduced dimensionality, because the optimal power allocation is not known a-priori, and only the total power constraint:

$$P_{av} = \sum_{k=1:|\mathcal{X}^{2M}|} p(\mathbf{x_k}) [\alpha_1, ...\alpha_{2M}] [|x_k^1|^2, ...|x_k^{2M}|^2]^T \quad (13)$$

can be considered (here $\mathbf{x_k} = [x_k^1, ...x_k^{2M}]^T$). In this case the degrees of freedom in the optimization process for each α are reduced from $|\mathcal{X}|^{2M}$ to $2M|\mathcal{X}|$, but the number of values of the vector α that need to be swept grows exponentially with M.

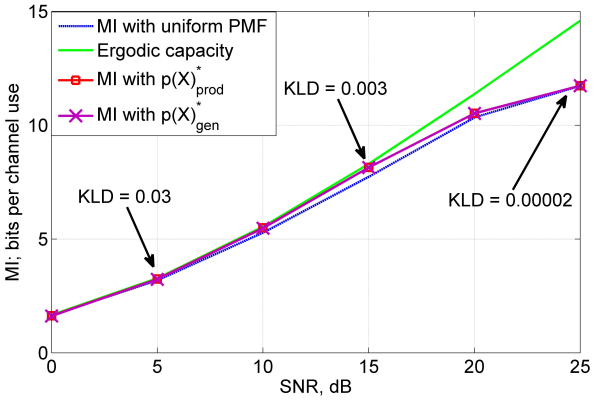

Fig. 1. MI with uniform PMF, the optimal PMFs obtained from the algorithms in Sections I and II, and the ergodic capacity. The KLD between the two optimal PMFs is indicated for selected SNR values

IV. SOME RESULTS ON THE OPTIMAL PMFS

We compare the PMF, obtained by the general algorithm from Section I, with the PMF, obtained by the modified algorithm from Section II by means of the *Kullback-Leibler Distance* (KLD). If $p(X)^*_{gen}$ is the former, and $p(X)^*_{prod}$ is the latter PMF, the KLD is defined as:

$$D(p(X)^*_{gen} || p(X)^*_{prod}) = \sum_{i=1:|\mathcal{X}^{2M}|} p(\mathbf{x_i})^*_{gen} \log_2 \frac{p(\mathbf{x_i})^*_{gen}}{p(\mathbf{x_i})^*_{prod}}$$

In Fig. 1 the MI is given for both PMFs, together with the MI with uniform PMF and the ergodic capacity [7]. The MI is practically the same for both. The KLD is also indicated in the figure for selected SNR values. We see that it is practically zero (we attribute any error to numerical inaccuracy) at the low SNR (where the shaping gain is very small), the medium SNR (where the largest shaping gain can be expected) and at high SNR (where the uniform PMF is near-optimal).

V. CONCLUSION

In this paper the factorization properties of the optimal PMF input to a MIMO channel were considered. It was proven, that the optimal PMF factorizes into the product of its marginal PMFs, confirming the conjecture, made in [1]. The proof relies on the iterative BA algorithm, showing

that if the initial PMF factorizes, it stays factorized on each subsequent iteration, evidently reaching the unique optimum, which must also be factorized. Using the factorization property, it was also shown how the power allocation problem can be solved by the BA algorithm.

VI. APPENDIX A

We prove Theorem 1 for $M = 2$ and real-valued channel for simpler notation, and then show that it can be extended to higher order MIMO and complex-valued channel. Let us denote $f(p(X_1)^{n+1}, p(X_2)^{n+1}) = -\mathcal{H}(X_2|X_1, Y, H = \mathbf{H})$ at iteration n for fixed $p(X|Y, \mathbf{H})$:

$$f = \sum_{k=1:|\mathcal{X}|} p(\mathbf{x_{1k}})^{n+1} \int_{\mathbf{y}} p(\mathbf{y}|\mathbf{H}, \mathbf{x_{1k}}) \cdot$$

$$\sum_{j=1:|\mathcal{X}|} p(\mathbf{x_{2j}}|\mathbf{x_{1k}}, \mathbf{y}, \mathbf{H}) \log_2 p(\mathbf{x_{2j}}|\mathbf{x_{1k}}, \mathbf{y}, \mathbf{H}) d\mathbf{y} =$$

$$= \sum_{k=1:|\mathcal{X}|} p(\mathbf{x_{1k}})^{n+1} \sum_{j=1:|\mathcal{X}|} p(\mathbf{x_{2j}}|\mathbf{x_{1k}})^{n+1} \cdot g_j(\mathbf{x_{1k}}) \quad (14)$$

where the function $g_j(X_1)$ is defined as:

$$g_j(X_1) = \int_{\mathbf{y}} p(\mathbf{y}|\mathbf{H}, \mathbf{x_{2j}}, X_1).$$

$$\log_2 \frac{p(\mathbf{y}|\mathbf{H}, \mathbf{x_{2j}}, X_1) p(\mathbf{x_{2j}})^n}{\sum_{i=1:|\mathcal{X}|} p(\mathbf{y}|\mathbf{H}, \mathbf{x_{2i}}, X_1) p(\mathbf{x_{2i}})^n} d\mathbf{y} =$$

$$= \log_2 p(\mathbf{x_{2j}})^n + D(p(Y|\mathbf{H}, \mathbf{x_{2j}}, X_1)||p(Y|\mathbf{H}, X_1)) \quad (15)$$

In (14) the Bayes theorem is used to express the conditional distribution $p(X_2|X_1, \mathbf{y}, \mathbf{H})$, we remove the dependency on H where it is not relevant, and we have used the fact that $p(\mathbf{x_{2j}}|\mathbf{x_{1k}})^n = p(\mathbf{x_{2j}})^n$.

Let us examine the PDFs in the KLD. The first one is a Gaussian with mean $\mathbf{H}\,[X_1, \mathbf{x_{2j}}]^T$ and covariance matrix given by the noise. The second is a *Mixture of Gaussians* (MoG) with the same covariance matrices, and mixing coefficients given by the marginal PMF $p(X_2)^n$. For different values of X_1, the shape of both PDFs is unchanged, only the mass is linearly shifted. We note that this shift is the same for both PDFs. This means that the relative offset between them is the same regardless of X_1, or in other words - the KLD is unchanged for different X_1. This property is illustrated in Fig. 2. We plot the likelihood and the marginal PDF for a random $\mathbf{H}$, 2 randomly chosen values of $X_1 : X_1 = x_k$ and $X_1 = x_m$, fixed $X_2 = x_j$ and a random valid PMF $p(X_2)$. The likelihood (the white Gaussian curve) sits on top of one of the components of the MoG (plotted in gray). The shift is along the line, defined by the points $\mathbf{H}[X_1, x_j]^T$.

The logarithm in (15) is clearly independent of X_1, therefore g_j is independent of X_1. Equation (14) can now be rewritten as:

$$f = \sum_{j=1:|\mathcal{X}|} p(\mathbf{x_{2j}}) \sum_{k=1:|\mathcal{X}|} p(\mathbf{x_{1k}}|\mathbf{x_{2j}}).g_j = \sum_{j=1:|\mathcal{X}|} p(\mathbf{x_{2j}}).g_j$$

which proves the theorem.

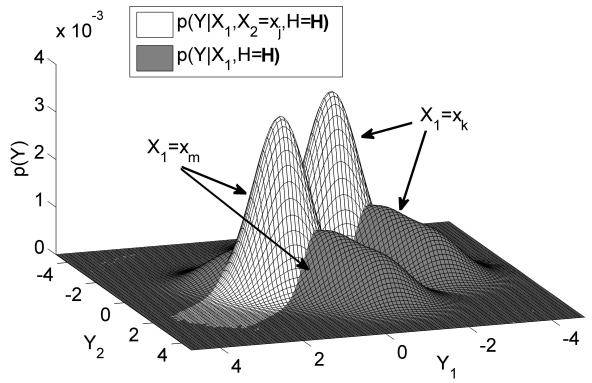

Fig. 2. Illustration of the KLD independence of X_1. The masses of the likelihood and the marginal PDF are shifted by the same factor, retaining their relative offset, and therefore the KLD

It is straight-forward to extend the proof to $M > 2$ and complex-valued input and channel. The upper-mentioned linear shift will be across multiple dimensions, instead of just 1, still keeping the KLD $D(p(Y|\mathbf{H}, \mathbf{x_{2j}}, X_{\{1:2M\}\setminus 2})||p(Y|\mathbf{H}, X_{\{1:2M\}\setminus 2}))$ independent of $X_{\{1:2M\}\setminus 2}$. Then $g_j(X_1, X_3, ...X_{2M})$ only depends on $p(\mathbf{x_{2j}})$, and $f = \sum_{j=1:|\mathcal{X}|} p(\mathbf{x_{2j}}).g_j$.

VII. REFERENCES

[1] J. Bellorado, S. Ghassemzadeh and A. Kavcic, "Approaching the capacity of the MIMO Rayleigh flat-fading channel with QAM constellations, independent across antennas and dimensions", *IEEE Trans. Commun.*, vol. 5, no. 6, pp. 1322-1332, 2006

[2] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, New York, Wiley, 1991.

[3] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460-473, 1972.

[4] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14-20, 1972.

[5] N. Varnica, X. Ma and A. Kavcic, "Capacity of power constrained memoryless AWGN channels with fixed input constellations," *GLOBECOM*, pp. 1339-1343 vol.2, 2002.

[6] I. C. Abou-Faycal, M. D. Trott and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1290-1301, 2001.

[7] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj and H. V. Poor, *MIMO Wireless Communications*, Cambridge University Press 2007