Bit-Interleaved Coded Modulation with Shaping

Albert Guillén i Fàbregas
Department of Engineering
University of Cambridge, UK
guillen@ieee.org

Alfonso Martinez
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
alfonso.martinez@ieee.org

***Abstract*— The performance of bit-interleaved coded modulation (BICM) with shaping (i.e., non-equiprobable bit probabilities) is studied. For the AWGN channel, the rates achievable with BICM and shaping are practically identical to those of coded modulation or multilevel coding, virtually closing the gap that made BICM suboptimal in terms of information rates.**

I. Introduction

For non-binary transmission in the AWGN channel, three main coding constructions that achieve information rates close to the channel capacity are known: coded modulation (CM), bit-interleaved coded modulation (BICM), and multilevel coding (MLC). CM dates back to the pioneering work of Ungerböck [1], and merges coding and modulation in a single entity. In contrast, BICM separates them, and maps a simple binary code onto a non-binary modulation [2], [3], [4]. MLC makes use of multiple binary codes, one for each bit in the binary label of the modulation symbol [5], [6].

CM yields the highest information rates. It is closely followed by MLC (for equiprobable modulation symbols the rates coincide) and, with a larger loss, by BICM. In terms of error exponents, the situation is somewhat reversed, with CM again the best, but now BICM beats MLC at low rates. Whereas previous analyses of BICM in the literature assume that the modulation symbols are used equiprobably, in this work we lift this assumption and consider shaping, whereby the bit or symbol probabilities are arbitrary. We will see that BICM with shaping achieves both information rates and error exponents very close to those of CM, thus closing the gap which made MLC better in terms of information rates. Practical coding schemes based on BICM with shaping have been studied in [7], illustrating that the gains predicted by our theoretical analysis are achievable in practice.

II. Preliminaries

Consider a memoryless channel with input X and output Y, respectively belonging to the sets $\mathcal{X}$ and $\mathcal{Y}$. A block code $\mathcal{M} \subseteq \mathcal{X}^N$ is a set of $|\mathcal{M}|$ vectors (or codewords) $\boldsymbol{x}$ of length N (the number of channel uses), i. e. $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$. Let $M \stackrel{\Delta}{=} |\mathcal{X}|$ denote the cardinality of $\mathcal{X}$ and $m \stackrel{\Delta}{=} \log_2 M$ the number of bits required to index a symbol. The output $\boldsymbol{y} \stackrel{\Delta}{=} (y_1, \ldots, y_N)$ is a random transformation of the input with transition probability distribution $P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$. For memoryless channels $P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ admits the decomposition

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^{N} P_{Y|X}(y_k|x_k) \tag{1}$$

With no loss of generality, we limit our attention to continuous output and identify $P_{Y|X}(y|x)$ as a probability density function. We adopt the convention that capital letters represent random variables, while the corresponding small letters correspond to realizations of the variables.

At the source, a message m drawn with equal probability from a message set $\{1, \ldots, |\mathcal{M}|\}$ is mapped onto a codeword $\boldsymbol{x}$. We denote this encoding function by ϕ, i. e. $\phi(\mathsf{m}) = \boldsymbol{x}$. The corresponding transmission rate R is given by $R \stackrel{\Delta}{=} \frac{1}{N} \log |\mathcal{M}|$. At the receiver, the decoder determines the codeword *decoding metric*, denoted by $q(\boldsymbol{x}, \boldsymbol{y})$, for all codewords, and outputs the message $\widehat{\mathsf{m}}$ whose metric is largest,

$$\widehat{\mathsf{m}} = \underset{\mathsf{m} \in \{1, \ldots, |\mathcal{M}|\}}{\arg\max}\ q(\phi(\mathsf{m}), \boldsymbol{y}). \tag{2}$$

The metrics we consider are products of symbol decoding metrics $q(x, y)$, namely (with some abuse of notation)

$$q(\boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{N} q(x_k, y_k). \tag{3}$$

For maximum a posteriori (MAP) decoders, the decoding metric is given by $q(x, y) = P_{Y|X}(y|x)P_X(x)$. More generally, a decoder finds the most likely codeword as long as the metric $q(x, y)$ is a strictly increasing bijective function of $P_{Y|X}(y|x)P_X(x)$. Instead, if the metric $q(x, y)$ is not a bijective function of $P_{Y|X}(y|x)P_X(x)$, we have a *mismatched decoder* [8], [9].

A. Coded Modualtion

The random coding ensemble corresponding to CM has channel inputs selected i.i.d. from $\mathcal{X}$ according to a probability distribution $P_X(x)$. The CM decoder is MAP. The largest information rate that can be achieved with CM under the constraint $x \in \mathcal{X}$ is

$$\mathsf{C}^{\mathrm{cm}} = \sup_{P_X(X)} I(X; Y). \tag{4}$$

Also, for any input distribution $P_X(X)$, the error probability averaged over the random coding ensemble satisfies [10]

$$\bar{P}_e \leq \mathrm{e}^{-NE_{\mathrm{r}}(R)} \tag{5}$$

[1]This work has been supported by the International Joint Project 2008/R2 of the Royal Society.

where $E_{\rm r}(R) = \sup_{0\le\rho\le 1} E_0(\rho) - \rho R$ and

$$E_0(\rho) \triangleq -\log \mathbb{E}\left[\left(\sum_{x'} P_X(x')\left(\frac{P_{Y|X}(Y|x')}{P_{Y|X}(Y|X)}\right)^{\frac{1}{1+\rho}}\right)^{\rho}\right]. \tag{6}$$

The expectation is carried out according to the joint distribution $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x)$.

B. Bit-Interleaved Coded Modulation

In practical CM schemes, since the codewords are selected elements of $\mathcal{X}^N$ and the alphabet $\mathcal{X}$ has typically more than 2 elements, the corresponding codes are non-binary. BICM is a different construction where the underlying code is binary. Originally analyzed in [3] under the assumption of infinite-depth interleaving, this restriction was recently lifted in [4], [11], where it was shown that BICM has a natural description in terms of mismatched decoding.

The BICM encoder consists of a binary code $\mathcal{C}$ that generates a codeword of mN bits, $\boldsymbol{b} = (b_1, \ldots, b_{mN})$. This codeword is interleaved and mapped onto a vector of N modulation symbols according to a labeling rule $\mu : \mathbb{F}_2^m \to \mathcal{X}$, such that

$$x_k = \mu\big(b_{(k-1)m+1}, \ldots, b_{km}\big), \quad k = 1, \ldots, N. \tag{7}$$

Thus, $\phi(\mathsf{m}) = \big(\mu(b_1, \ldots, b_m), \ldots, \mu(b_{(N-1)m+1}, \ldots, b_{mN})\big)$. Analogously, we denote the inverse labeling function by b_j, so that $b_j(x)$ is the j-th bit in the binary label of modulation symbol x, for $j = 1, \ldots, m$. We define also the sets $\mathcal{X}_b^j$ as those elements of $\mathcal{X}$ having bit b in the j-th label position, i.e., $\mathcal{X}_b^j \triangleq \{x \in \mathcal{X} : b_j(x) = b\}$. By construction, the modulation symbols x are used with probabilities

$$P_X^{\rm bicm}(x) = \prod_{j=1}^{m} P_{B_j}\big(b_j(x)\big) \tag{8}$$

where $P_{B_j}\big(b\big)$ is the probability of the j-th bit. Fig. 1 shows the random generation of the binary codebook of $\mathcal{C}$. The labeling rule μ modulates the resulting binary codebook column-wise. Note that the interleaver which gives the name to BICM has been absorbed in the description of the random coding ensemble. In practice [7], different bit-probability assignments can easily be implemented by employing an interleaver that independently scrambles the coded bits assigned to each of position of the label. In other words, according to Fig. 1, we would need m interleavers operating on a row basis. This is similar to how coded bits are assigned to multiple fading blocks in block-fading channels [12].

In addition to the different code construction, BICM also differs from CM at the receiver side. The BICM symbol metric treats the m bits in a symbol as if they were independent, and is given by (see [4], [11] for more details)

$$q^{\rm bicm}(x,y) = \prod_{j=1}^{m} q_j\big(b_j(x), y\big) \tag{9}$$

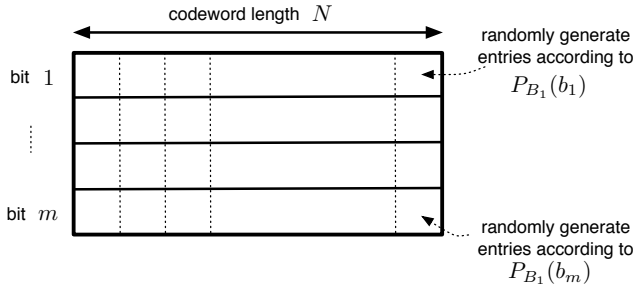

Fig. 1. Binary andom coding ensemble for BICM with shaping.

where $q_j(b_j, y)$ is the j-th bit metric given by

$$q_j(b_j(x) = b, y) = \sum_{x' \in \mathcal{X}_b^j} P_{Y|X}(y|x') P_X^{\rm bicm}(x'). \tag{10}$$

C. Multilevel Coding

Multilevel codes (MLC) combined with multistage decoding (MSD) have been proposed [5], [6] as an efficient method to attain the channel capacity by using binary codes. For BICM, a single binary code $\mathcal{C}$ is used to generate a binary codeword, which is used to select modulation symbols by means of a binary labeling function μ. In MLC, the input binary code $\mathcal{C}$ is the Cartesian product of m binary codes of length N, one per modulation level, i. e. $\mathcal{C} = \mathcal{C}_1 \times \ldots \times \mathcal{C}_m$, and the input distribution for the symbol $x(b_1, \ldots, b_j)$ has the form

$$P_X^{\rm mlc}(x) = P_{B_1,\ldots,B_M}(b_1, \ldots, b_m) = \prod_{j=1}^{m} P_{B_j}(b_j). \tag{11}$$

For a fixed input distribution on the bits, MLC achieve the mutual information [5], [6] both with MAP joint decoding and with multistage decoding. The largest information rate that can be achieved with MLC under the constraint $x \in \mathcal{X}$ is

$$\mathsf{C}^{\rm mlc} = \sup_{P_{B_1}(B_1),\ldots,P_{B_m}(B_m)} I(X;Y). \tag{12}$$

The error exponents of MLC with multistage decoding were derived in [13], [14], [4], [15], where it was also shown the error exponent is upper bounded by one, making the MLC exponent much worse than that of CM.

III. Achievable Rates with BICM

References [4], [11] show that the rate

$$I_{\rm gmi} = \sup_{s>0} I_{\rm gmi}(s), \tag{13}$$

where

$$I_{\rm gmi}(s) = \mathbb{E}\left[\log \frac{\prod_{j=1}^{m} q_j\big(b_j(X), Y\big)^s}{\frac{1}{M}\sum_{x'}\prod_{j=1}^{m} q_j\big(b_j(x'), Y\big)^s}\right], \tag{14}$$

also named generalized mutual information (GMI), is achievable for BICM with equiprobable bits, $P_{B_j}(b) = \frac{1}{2}$. The proof is based on a simple extension of Gallager's analysis of ML decoding in terms of error exponents to mismatched decoding [16], [4]. References [4], [11] also show that the above rate

may be decomposed as the sum of m bit GMI terms (with s fixed), and that it coincides with the BICM capacity defined in [3]. We next generalize this result for arbitrary bit probabilities.

Theorem 1: The GMI of the BICM mismatched decoder is equal to $I_{\rm gmi} = \sup_{s>0} I_{\rm gmi}(s)$, where

$$I_{\rm gmi}(s) = \sum_{j=1}^{m} \mathbb{E}\left[\log \frac{q_j(B_j,Y)^s}{\sum_{b'_j=0}^{1} q_j(b'_j,Y)^s P_{B_j}(b'_j)}\right]. \quad (15)$$

is the sum of the GMIs (with fixed s) of m binary-input channels. The expectation is over the joint distribution $P_{B_j}(b_j)P_j(y|b_j)$, with

$$P_j(y|b) \triangleq \sum_{x\in\mathcal{X}_b^j} \frac{P_{Y|X}(y|x)P_X^{\rm bicm}(x)}{\sum_{x'\in\mathcal{X}_b^j} P_X^{\rm bicm}(x')}. \quad (16)$$

An alternative expression is

$$I_{\rm gmi}(s) = \sum_{j=1}^{m} \mathbb{E}\left[\log \frac{q_j\big(b_j(X),Y\big)^s}{\sum_{b'_j=0}^{1} q_j(b'_j,Y)^s P_{B_j}(b'_j)}\right], \quad (17)$$

where the expectation is over the joint distribution $P_X^{\rm bicm}(x)P_{Y|X}(y|x)$.

Proof: See Appendix. ∎

In the remainder of the paper, for the sake of simplicity and without loss of generality, we focus on the classical BICM metric given in Eq. (10). Note that Theorem 1 generalizes to other metrics, in which case, s should also be optimized.

Corollary 1: For the classical BICM decoder with metric in Eq. (10) the supremum over s is achieved at $s = 1$, and $I_{\rm gmi} = \sum_{j=1}^{m} I(B_j;Y)$.

Proof: Since the metric $q_j(b_j, y)$ is proportional to $P_j(y|b_j)$, we can identify the quantity

$$\mathbb{E}\left[\log \frac{q_j\big(B_j,Y\big)^s}{\sum_{b'_j=0}^{1} q_j(b'_j,Y)^s P_{B_j}(b'_j)}\right] \quad (18)$$

as the GMI of a *matched* binary-input channel with transitions $P_j(y|b_j)$. Then, the supremum over s is achieved at $s = 1$ (that is, the mutual information $I(B_j;Y)$) and we get the result. ∎

The above results suggest that we can chose the input bit distribution that yields the largest GMI, i.e., effectively implying shaping the bit probabilities in BICM as

$$\mathsf{C}^{\rm bicm} = \sup_{P_{B_1}(B_1),\ldots,P_{B_m}(B_m)} \sum_{j=1}^{m} I(B_j;Y). \quad (19)$$

For i.i.d. codebooks, $\mathsf{C}^{\rm bicm}$ is also the largest rate that can be transmitted with vanishing error probability [17]. This capacity should be compared with the equivalent quantities on CM and MLC, given in Eqs. (4) and (12) respectively,

$$\mathsf{C}^{\rm cm} = \sup_{P_X(X)} I(X;Y), \quad (20)$$

$$\mathsf{C}^{\rm mlc} = \sup_{P_{B_1}(B_1),\ldots,P_{B_m}(B_m)} I(X;Y). \quad (21)$$

Recall that BICM differs from CM at the transmitter, where the modulation symbol probabilities have the specific form $P_X^{\rm bicm}(x) = \prod_{j=1}^{m} P_{B_j}\big(b_j(x)\big)$, and at the receiver, where the bit metrics in Eq. (10) are used for decoding.

In terms of the random coding error exponent, the analysis in [4], [11] can be merged with the previous proof to show that for any input distribution $P_X^{\rm bicm}(X) = \prod_{j=1}^{m} P_{B_j}\big(b_j(X)\big)$, the error probability averaged over the ensemble of random codes is upper bounded by

$$\bar{P}_e \le {\rm e}^{-NE_{\rm r}^q(R)} \quad (22)$$

where $E_{\rm r}^q(R) = \sup_{\substack{0\le\rho\le 1\\ s>0}} E_0^q(\rho,s) - \rho R$, and

$$E_0^q(\rho,s) \triangleq -\log \mathbb{E}\left[\left(\sum_{x'} P_X^{\rm bicm}(x')\left(\frac{q^{\rm bicm}(x',Y)}{q^{\rm bicm}(X,Y)}\right)^s\right)^{\rho}\right]. \quad (23)$$

is the generalized Gallager function. The expectation is over the joint distribution $P_{Y|X}(y|x)P_X^{\rm bicm}(x)$.

IV. Shaping for the Gaussian Channel

A. Channel Model

We consider transmission using complex-plane signal sets ($\mathcal{X}\subset\mathbb{C}$, $\mathcal{Y}=\mathbb{C}$) in the AWGN channel such that

$$Y = \sqrt{\mathsf{snr}}\,X + Z \quad (24)$$

where $Z\sim\mathcal{N}_{\mathbb{C}}(0,1)$ and snr is the signal-to-noise ratio (SNR). We wish to solve the optimization problems in Eqs. (4), (12) and (19) with the additional constraints that $x\in\mathcal{X}$, $\mathbb{E}[X]=0$, and $\mathbb{E}[|X|^2]=1$.

B. Examples

In this section, we show some examples using binary reflected Gray mapping[2]. For shaping, 2^m-QAM signal sets are of special interest; this constellation is the Cartesian product of two $2^{\frac{m}{2}}$-PAM constellations, one for each of the in-phase and quadrature components of the channel. Since the optimum input distribution is known to be Gaussian, a good input distribution over the set $\mathcal{X}$ should approach in some sense a Gaussian density. Symmetry between the in-phase and quadrature components and along the zero axis (positive and negative planes have equal probability) dictates that the optimization problems in Eqs. (4) and (12) respectively have

- $2^{\frac{m}{2}-1}-1$ free parameters for CM, and
- $\frac{m}{2}-1$ free parameters for BICM and MLC.

For BICM we used the symmetries of binary reflected Gray mapping and the fact that the most significant bit selects the positive or negative half-plane, and always has probability $\frac{1}{2}$.

Note that the CM optimization problem does not restrict the input distribution to be $P_X(x) = \prod_{j=1}^{m} P_{B_j}(b_j(x))$, hence

[2]Recall that the binary reflected Gray mapping for m bits may be generated recursively from the mapping for $m-1$ bits by prefixing a binary 0 to the mapping for $m-1$ bits, then prefixing a binary 1 to the reflected (i. e. listed in reverse order) mapping for $m-1$ bits. For QAM modulations in the Gaussian channel, the symbol mapping is the Cartesian product of Gray mappings over the in-phase and quadrature PAM components.

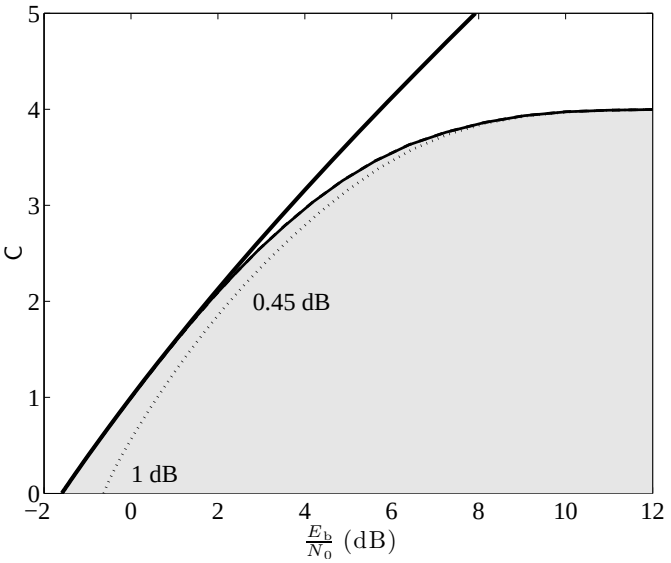

Fig. 2. Capacities for Gaussian inputs (thick solid), CM/MLC with shaping (thin solid), BICM with shaping (dashed) and BICM with equiprobable inputs (dotted line) for 16-QAM with Gray mapping and shaping as a function of $\frac{E_b}{N_0}$ (dB).

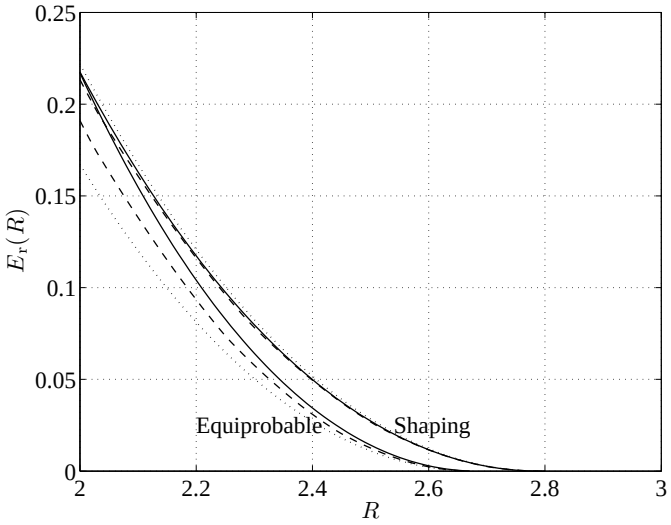

Fig. 3. Error exponent zoom at capacity for 16-QAM with and without shaping for CM (solid) and BICM (dashed) at $\mathsf{snr} = 8$ dB. The error exponent of the BICM independent parallel channel model is shown for comparison (dotted). When shaping is used, the input distribution is the corresponding optimal capacity achieving distribution.

being able to achieve potentially larger rates. As we shall see, the resulting difference in information rates is however marginal. Moreover, note that there is an exponential relationship between the number of free parameters for BICM and CM, which can induce rather large computational savings for large signal sets. For example, since for 16-QAM there is only one free parameter for MLC and CM, the optimization will result in the best performance, i.e., MLC is optimal and BICM, as we shall see, is very close. However, for $m > 4$ this is no longer true and the optimization over symbol probabilities without restriction $P_X(x)$ to be the product of bit probabilities could potentially yield larger rates.

Figure 2 shows the improvement in BICM capacity derived from shaping for 16-QAM with binary reflected Gray mapping. As we observe $\mathsf{C}^{\mathrm{bicm}}$ (dashed) is almost indistinguishable from C^{cm} or $\mathsf{C}^{\mathrm{mlc}}$ (thin solid) or channel capacity itself (thick solid). This shows that shaping for BICM can recover the BICM capacity loss for equiprobable bits and effectively close the gap with CM and MLC. Recall that shaping gains are achieved with a one-shot non-iterative demodulator. In general, the decoding complexity of BICM is larger than that of MLC, since the codes of MLC are shorter. In practice, however, if the decoding complexity grows linearly with the number of bits in a codeword, e. g. with LDPC or turbo codes, the overall complexity of BICM becomes comparable to that of MLC.

Figure 3 shows the error exponents for CM and BICM, with and without shaping, for 16-QAM at $\mathsf{snr} = 8$ dB. When shaping is used, the input distribution is the corresponding capacity-achieving distribution. We observe that when shaping is used, in the region near capacity, the overall BICM error exponent is very close to that of CM, while when equiprobable bits are used, the exponent deviates from that of CM. Remark that, according to [4], [11] the BICM error exponent cannot be larger than that of CM, as opposed to that of the independent parallel channel model [3]. Note that the error exponent of BICM is much larger than that of MLC (always being given by the minimum of the error exponents of the various levels, which results in an error exponent smaller than 1) [13], [14], [4], [15]. Therefore, BICM outperforms MLC in terms of error probability and yields nearly the same rates with shaping.

C. Wideband Regime

The gain from shaping in BICM is especially significant at low snr. In this wideband regime [18] one considers a Taylor series in terms of snr,

$$R(\mathsf{snr}) = c_1\mathsf{snr} + c_2\mathsf{snr}^2 + \mathrm{o}\big(\mathsf{snr}^2\big). \tag{25}$$

where the A scheme is said to be first- and second-order optimal if $c_1 = 1$ and $c_2 = -\frac{1}{2}$ [18]. In those conditions, such a system is both power- and bandwidth-efficient. For instance, it is well known that for low snr, QPSK is both first- and second-order optimal [18].

The low-snr performance of BICM was studied in [19], where general expressions for the coefficients c_1 and c_2 were given for general mapping rules and equiprobable signaling. For the particular case of binary reflected Gray mapping with squared QAM constellations, it was found that BICM was suboptimal, in the sense that it did not achieve the optimum c_1 and c_2. References [20], [21] proposed first-order optimal mapping rules for BICM that achieve $c_1 = 1$, or equivalently $\frac{E_b}{N_0}_{\lim} \triangleq \frac{\log 2}{c_1} = -1.59$ dB, with equiprobable signaling.

Theorem 2: Shaping makes BICM transmission over QAM modulations with binary reflected Gray mapping first- and second-order optimal, i.e., $c_1 = 1$ and $c_2 = -\frac{1}{2}$.

The key fact is that at low snr the bit probabilities are such that a QPSK constellation is effectively selected. To see how, note that for $m = 2$ we have QPSK with Gray mapping. Limiting ourselves to one dimension, the binary reflected Gray

mapping for $\frac{m}{2}+1$ bits is constructed from the mapping for $\frac{m}{2}$ bits by prefixing a binary 0 to the mapping for $\frac{m}{2}-1$ bits, then prefixing a binary 1 to the reflected (i. e. listed in reverse order) mapping for $\frac{m}{2}-1$ bits. With shaping, one has the flexibility to fix the probabilities of each of the additional bits to a given value, say, to 0, so that one is effectively transmitting over a BPSK constellation (QPSK over the two axis) when the resulting constellation is normalized in mean and energy. This is property does not *necessarily* hold for other mapping rules.

APPENDIX: PROOF OF THEOREM 1

For fixed s and probabilities $P_X^{\mathrm{bicm}}(x)=\prod_{j=1}^m P_{B_j}\big(b_j(x)\big)$ the GMI can be written as

$$I_{\mathrm{gmi}}(s)=\mathbb{E}\left[\log\frac{q^{\mathrm{bicm}}(X,Y)^s}{\sum_{x'}q^{\mathrm{bicm}}(x',Y)^s P_X^{\mathrm{bicm}}(x')}\right] \tag{26}$$

$$=\mathbb{E}\left[\log\frac{\prod_{j=1}^m q_j\big(b_j(X),Y\big)^s}{\sum_{x'}\prod_{j=1}^m q_j\big(b_j(x'),Y\big)^s P_{B_j}\big(b_j(x')\big)}\right], \tag{27}$$

where the expectation is carried out according to $P_X^{\mathrm{bicm}}(x)P_{Y|X}(y|x)$.

We now have a closer look at the denominator in the logarithm of (27). The key observation is that the sum over the constellation points ($x'\in\mathcal{X}$) of the product of a function $f\big(b_j(x')\big)$ evaluated at all the binary label positions admits an alternative expression, namely

$$\sum_{x'\in\mathcal{X}}\left(\prod_{j=1}^m f\big(b_j(x')\big)\right)=\prod_{j=1}^m\left(\sum_{b_j\in\{0,1\}} f(b_j)\right). \tag{28}$$

Indeed, after carrying out the product in the right-hand side, we obtain the desired sum over all 2^m binary m-tuples $(b_1,\ldots,b_m)$ of summands of the form $f(b_1)\cdots f(b_m)$.

Therefore, for the specific choice $f\big(b_j(x')\big)=q_j\big(b_j(x'),Y\big)^s P_{B_j}\big(b_j(x')\big)$ we have the product over all label positions of the sum of the probabilities of the bit b_j being zero and one, i.e.,

$$\sum_{x'\in\mathcal{X}}\left(\prod_{j=1}^m q_j\big(b_j(x'),Y\big)^s P_{B_j}(b_j(x))\right) \tag{29}$$

$$=\prod_{j=1}^m\left(\sum_{b'_j\in\{0,1\}} q_j(b_j,Y)^s P_{B_j}(b'_j)\right). \tag{30}$$

Next, going back to (27), we obtain

$$I_{\mathrm{gmi}}(s)=\mathbb{E}\left[\log\left(\prod_{j=1}^m\frac{q_j\big(b_j(X),Y\big)^s}{\sum_{b'_j=0}^1 q_j(b'_j,Y)^s P_{B_j}(b'_j)}\right)\right], \tag{31}$$

$$=\sum_{j=1}^m\mathbb{E}\left[\log\frac{q_j\big(b_j(X),Y\big)^s}{\sum_{b'_j=0}^1 q_j(b'_j,Y)^s P_{B_j}(b'_j)}\right], \tag{32}$$

where the expectation is over the joint distribution $P_X^{\mathrm{bicm}}(x)P_{Y|X}(y|x)$. This gives Eq. (17) since the GMI is the supremum over all s [8], [9]. As for Eq. (15), we derive it by noting that, for each j, the summation over x in the expectation can be split into two parts and rearranged as follows,

$$\sum_x f(x)=\sum_{b_j\in\{0,1\}}\sum_{x\in\mathcal{X}_b^j} f(x) \tag{33}$$

$$=\sum_{b_j\in\{0,1\}} P_{B_j}(b_j)\sum_{x\in\mathcal{X}_b^j}\frac{f(x)}{P_{B_j}(b_j)}. \tag{34}$$

As $P_{B_j}(b_j)=\sum_{x'\in\mathcal{X}_b^j}P_X^{\mathrm{bicm}}(x')$ by construction, recovering the expression of $f(x)$ we obtain $P_j(y|b_j)$ in Eq. (16).

REFERENCES

[1] G. Ungerböck, "Channel Coding With Multilevel/Phase Signals.," *IEEE Trans. Inf. Theory*, vol. 28, no. 1, pp. 55–66, 1982.

[2] E. Zehavi, "8-PSK trellis codes for a Rayleigh channel," *IEEE Trans. Commun.*, vol. 40, no. 5, pp. 873–884, 1992.

[3] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 927–946, 1998.

[4] A. Guillén i Fàbregas, A. Martinez, and G. Caire, *Bit-Interleaved Coded Modulation*, vol. 5, Foundations and Trends on Communications and Information Theory, Now Publishers, 2008.

[5] H. Imai and S. Hirakawa, "A new multilevel coding method using error-correcting codes," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 371–377, May 1977.

[6] U. Wachsmann, R. F. H. Fischer, and J. B. Huber, "Multilevel codes: theoretical concepts and practical design rules," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1361–1391, Jul. 1999.

[7] M. J. Hossain, A. Alvarado, and L Szczecinski, "BICM transmission using non-uniform QAM constellations: Performance analysis and design," in *IEEE Int. Conf. Commun., Cape Town, South Africa*, 2010.

[8] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, 1994.

[9] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: general alphabets, channels with memory, and the wideband limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, 2000.

[10] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, Inc. New York, NY, USA, 1968.

[11] A. Martinez, A. Guillén i Fàbregas, G. Caire, and F. Willems, "Bit-interleaved coded modulation revisited: A mismatched decoding perspective," *IEEE Trans on Inf. Theory*, vol. 55, no. 6, pp. 2756–2765, Jun. 2009.

[12] R. Knopp and P. A. Humblet, "On coding for block fading channels," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 189–205, 2000.

[13] G. Beyer, K. Engdahl, and K. S. Zigangirov, "Asymptotical analysis and comparison of two coded modulation schemes using PSK signaling - Part I," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2782–2792, 2001.

[14] G. Beyer, K. Engdahl, and K. S. Zigangirov, "Asymptotical analysis and comparison of two coded modulation schemes using PSK signaling - Part II," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2793–2806, 2001.

[15] A. Ingber and M. Feder, "Capacity and Error Exponent Analysis of Multilevel Coding with Multistage Decoding," in *IEEE Int. Symp. Inf. Theory, Seoul, Korea*, Jul. 2009, pp. 1799–1803.

[16] G. Kaplan and S. Shamai, "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEU. Archiv für Elektronik und Übertragungstechnik*, vol. 47, no. 4, pp. 228–239, 1993.

[17] A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1520–1529, Sept. 1996.

[18] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.

[19] A. Martinez, A. Guillén i Fàbregas, G. Caire, and F. Willems, "Bit-interleaved coded modulation in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5447–5455, Dec. 2008.

[20] C. Stierstorfer and R. F. H. Fischer, "Asymptotically optimal mappings for BICM with M-PAM and M-QAM," *IET Electronics Letters*, vol. 45, no. 3, pp. 173–174, Jan. 2009.

[21] E. Agrell and A. Alvarado, "On optimal constellations for BICM at low SNR," in *IEEE Inf. Theory Workshop, Taormina, Italy*, Oct. 2009.