Bit-Interleaved Coded Modulation Revisited:
A Mismatched Decoding Perspective

Alfonso Martinez, *Member, IEEE*, Albert Guillén i Fàbregas, *Senior Member, IEEE*, Giuseppe Caire, *Fellow, IEEE*, and Frans M. J. Willems, *Fellow, IEEE*

Abstract—We revisit the information-theoretic analysis of bit-interleaved coded modulation (BICM) by modeling the BICM decoder as a mismatched decoder. The mismatched decoding model is well defined for finite, yet arbitrary, block lengths, and naturally captures the channel memory among the bits belonging to the same symbol. We give two independent proofs of the achievability of the BICM capacity calculated by Caire *et al.*, where BICM was modeled as a set of independent parallel binary-input channels whose output is the bitwise log-likelihood ratio. Our first achievability proof uses typical sequences, and shows that due to the random coding construction, the interleaver is not required. The second proof is based on the random coding error exponents with mismatched decoding, where the largest achievable rate is the generalized mutual information. Moreover, the generalized mutual information of the mismatched decoder coincides with the infinite-interleaver BICM capacity. We show that the error exponent—and hence the cutoff rate—of the BICM mismatched decoder is upper-bounded by that of coded modulation and may thus be lower than in the infinite-interleaved model; for binary reflected Gray mapping in Gaussian channels the loss in error exponent is small. We also consider the mutual information appearing in the analysis of iterative decoding of BICM with extrinsic information transfer (EXIT) charts: if the symbol metric has knowledge of the transmitted symbol, EXIT mutual information admits a representation as a pseudo-generalized mutual information, which is in general not achievable. A different symbol decoding metric, for which the extrinsic side information refers to the hypothesized symbol, induces a generalized mutual information lower than the coded modulation capacity. In this case, perfect extrinsic side information turns the mismatched-decoder error exponent into that of coded modulation.

Index Terms—Bit-interleaved coded modulation, coded modulation, cutoff rate, error exponents, generalized mutual information, iterative decoding, maximum-likelihood decoding, mismatched decoding.

I. INTRODUCTION

IN the classical bit-interleaved coded modulation (BICM) scheme proposed by Zehavi in [1], the channel observation is used to generate decoding metrics for each of the bits of a symbol, rather than the symbol metrics used in Ungerböck's coded modulation (CM) [2]. This decoder is suboptimal and noniterative, but offers very good performance and is interesting from a practical perspective due to its low implementation complexity. In parallel, iterative decoders have also received much attention [3]–[7] thanks to their improved performance.

Caire *et al.* [8] further elaborated on Zehavi's decoder and, under the assumption of an infinite-length interleaver, presented and analyzed a BICM model as a set of parallel independent binary-input output symmetric channels. Based on the data processing theorem [9], Caire *et al.* showed that the BICM mutual information cannot be larger than that of CM. However, and rather surprisingly *a priori*, they found that the cutoff rate of BICM might exceed that of CM [10]. The error exponents for the parallel-channel model were studied by Wachsmann *et al.* [11].

In this paper, we take a closer look to the classical BICM decoder proposed by Zehavi and cast it as a mismatched decoder [12]–[14]. The observation that the classical BICM decoder treats the different bits in a given symbol as independent, even if they are clearly not, naturally leads to a simple model of the symbol mismatched decoding metric as the product of bit decoding metrics, which are in turn related to the log-likelihood ratios. We also examine the BICM mutual information in the analysis of iterative decoding by means of EXIT (extrinsic information transfer) charts [5]–[7], where the sum of the mutual informations across the parallel subchannels is used as a figure of merit of the progress in the iterative decoding process.

This paper is organized as follows. Section II introduces the system model and notation. Section III gives a proof of achievability of the BICM capacity, derived in [8] for the independent parallel channel model, by using typical sequences. Section IV shows general results on the error exponents, including the generalized mutual information and cutoff rate as particular instances. The BICM error exponent (and in particular the cutoff rate) is always upper-bounded by that of CM, as opposed to the corresponding exponent for the independent parallel channel model [8], [11], which can sometimes be larger. In particular, Section IV-D studies the achievable rates of BICM under mismatched decoding and shows that the *generalized mutual information* [12]–[14] of the BICM mismatched decoder yields the BICM capacity. The section concludes with some numerical results, including a comparison with the parallel-channel models.

In general, the loss in error exponent is negligible for binary reflected Gray mapping in Gaussian channels. In Section V, we turn our attention to the iterative decoding of BICM. First, we review how the mutual information appearing in the analysis of iterative decoding of BICM with EXIT charts, where the symbol decoding metric has some side knowledge of the transmitted symbol, admits a representation as a pseudo-generalized mutual information. A different symbol decoding metric, for which the extrinsic side information refers to the hypothesized symbol, induces a generalized mutual information lower in general than the coded modulation capacity. Moreover, perfect extrinsic side information turns the error exponent of this mismatched decoder into that of coded modulation. Finally, Section VI gives some concluding remarks.

II. CHANNEL MODEL AND CODE ENSEMBLES

A. Channel Model

We consider the transmission of information by means of a block code $\mathcal{M}$ of length N. At the transmitter, a message m is mapped onto a codeword $\boldsymbol{x} = (x_1, \ldots, x_N)$, according to one of the design options described later, in Section II-B. We denote this encoding function by ϕ. Each of the symbols x are drawn from a discrete modulation alphabet $\mathcal{X} = \{x_1, \ldots, x_M\}$, with $M \triangleq |\mathcal{X}|$ and $m = \log_2 M$ being the number of bits required to index a symbol.

We denote the output alphabet by $\mathcal{Y}$ and the channel output by $\boldsymbol{y} \triangleq (y_1, \ldots, y_N)$, with $y_k \in \mathcal{Y}$. With no loss of generality, we assume the output is continuous,[1] so that the channel output $\boldsymbol{y}$ related to the codeword $\boldsymbol{x}$ through a conditional probability density function $p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$. Further, we consider memoryless channels, for which

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

where $p_{Y|X}(y|x)$ is the channel symbol transition probability. Henceforth, we drop the words density function in our references of $p_{Y|X}(y|x)$. We denote by X, Y the underlying random variables. Similarly, the corresponding random vectors are

$$\boldsymbol{X} \triangleq \underbrace{(X, \ldots, X)}_{N \text{ times}} \quad \text{and} \quad \boldsymbol{Y} \triangleq \underbrace{(Y, \ldots, Y)}_{N \text{ times}}$$

respectively, drawn from the sets $\boldsymbol{\mathcal{X}} \triangleq \mathcal{X}^N, \boldsymbol{\mathcal{Y}} \triangleq \mathcal{Y}^N$.

A particularly interesting, yet simple, case is that of complex-plane signal sets in additive white Gaussian noise (AWGN) with fully interleaved fading where $\mathcal{Y} = \mathbb{C}$ and

$$y_k = h_k \sqrt{\mathsf{snr}} x_k + z_k, \quad k = 1, \ldots, N \qquad (2)$$

where h_k are fading coefficients with average unit energy, z_k are the complex zero-mean unit-variance AWGN samples, and snr is the signal-to-noise ratio (SNR). The decoder outputs an estimate of the message $\hat{\mathsf{m}}$ according to a given codeword *decoding metric*, which we denote by $q(\boldsymbol{x}, \boldsymbol{y})$ as

$$\hat{\mathsf{m}} = \arg \max_{\mathsf{m}} q(\boldsymbol{x}_{\mathsf{m}}, \boldsymbol{y}). \qquad (3)$$

[1]Our results are directly applicable to discrete output alphabets, by appropriately replacing integrals by sums.

The codeword metrics we consider are the product of symbol decoding metrics $q(x, y)$, for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, namely

$$q(\boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{N} q(x_k, y_k). \qquad (4)$$

Assuming that the codewords have equal probability, this decoder finds the most likely codeword as long as $q(x, y) = f(p_{Y|X}(y|x))$, where $f(\cdot)$ is a one-to-one increasing function, i.e., as long as the decoding metric is a one-to-one increasing mapping of the transition probability of the memoryless channel. If the decoding metric $q(x, y)$ is not an increasing one-to-one function of the channel transition probability, the decoder defined by (3) is a *mismatched decoder* [12]–[14].

B. Code Ensembles

1) Coded Modulation: In a *coded modulation* (CM) scheme $\mathcal{M}$, the encoder ϕ selects a codeword of N modulation symbols, $\boldsymbol{x}_{\mathsf{m}} = (x_1, \ldots, x_N)$ according to the information message m. The code is in general nonbinary, as symbols are chosen according to a probability law $P_X(x)$. Representing the information message set $\{1, \ldots, |\mathcal{M}|\}$, the rate R of this scheme in bits per channel use is given by $R = \frac{K}{N}$, where $K \triangleq \log_2 |\mathcal{M}|$ denotes the number of bits needed to represent every information message. At the receiver, a maximum metric decoder ϕ (as in (3)) acts on the received sequence $\boldsymbol{y}$ to generate an estimate of the transmitted message, $\varphi(\boldsymbol{y}) = \hat{\mathsf{m}}$. In CM constructions, such as Ungerböck's [2], the symbol decoding metric is proportional to the channel transition probability, that is, $q(x, y) \propto p_{Y|X}(y|x)$; the value of proportionality constant is irrelevant, as long as it is not zero. Reliable communication is possible at rates lower than *coded modulation capacity* or CM capacity, denoted by $\mathsf{C}_{\mathcal{X}}^{\mathrm{cm}}$ and given by

$$\mathsf{C}_{\mathcal{X}}^{\mathrm{cm}} \triangleq \mathbb{E}\left[\log \frac{p_{Y|X}(Y|X)}{\sum_{x' \in \mathcal{X}} P_X(x') p_{Y|X}(Y|x')}\right]. \qquad (5)$$

The expectation is done according to $P_X(x) p_{Y|X}(y|x)$. We often consider a uniform input distribution $P_X(x) = 2^{-m}$.

2) Bit-Interleaved Coded Modulation: In a *bit-interleaved coded modulation* (BICM) scheme $\mathcal{M}$, the encoder is restricted to be the serial concatenation of a binary code $\mathcal{C}$ of length $n \triangleq mN$ and rate $r = \frac{R}{m}$ (R being defined as for CM), a bit interleaver, and a binary labeling function $\mu : \{0,1\}^m \rightarrow \mathcal{X}$ which maps blocks of m bits to signal constellation symbols. The codewords of $\mathcal{C}$ are denoted by $\boldsymbol{b} = (b_1, \ldots, b_{mN})$. The portions of codeword allocated to the jth bit of the label are denoted by $\boldsymbol{b}_j \triangleq (b_j, b_{m+j}, \ldots, b_{m(N-1)+j})$. We denote the inverse mapping function for labeling position j as $b_j : \mathcal{X} \rightarrow \{0,1\}$, that is, $b_j(x)$ is the jth bit of symbol x. Accordingly, we now define the sets $\mathcal{X}_b^j \triangleq \{x \in \mathcal{X} : b_j(x) = b\}$ as the set of signal constellation points x whose binary label has value $b \in \{0,1\}$ in its jth position. With some abuse of notation, we will denote $B_1, \ldots, B_m$ and $b_1, \ldots, b_m$ the random variables and their corresponding realizations of the bits in a given label position $j = 1, \ldots, m$.

The classical BICM decoder [1], [18] treats each of the m bits in a symbol as independent and uses a symbol decoding
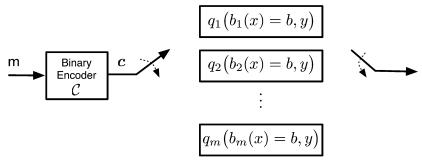
Fig. 1.   Parallel channel model of BICM.

metric proportional to the product of the *a posteriori* marginals $P_{B_j \mid Y}(b_j = b \mid y)$. More specifically, we have that

$$q(x, y) = \prod_{j=1}^{m} q_j(b_j(x), y) \qquad (6)$$

where the jth bit decoding metric $q_j(b, y)$ is given by

$$q_j(b_j(x) = b, y) = \sum_{x' \in \mathcal{X}_b^j} p_{Y \mid X}(y \mid x'). \qquad (7)$$

This metric is proportional to the transition probability of the output y given the bit b at position j, which we denote for later use by $p_j(y \mid b)$

$$p_j(y \mid b) \triangleq \frac{1}{|\mathcal{X}_b^j|} \sum_{x' \in \mathcal{X}_b^j} p_{Y \mid X}(y \mid x'). \qquad (8)$$

The set of m probabilities $p_j(y \mid b)$ can be used as a departure point to define an equivalent BICM channel model. Accordingly, Caire *et al.* defined a BICM channel [8] as the set of m parallel channels having bit $b_j(x_k)$ as input and the bit log-metric (log-likelihood) ratio for the kth symbol

$$\Xi_{m(k-1)+j} = \log \frac{q_j(b_j(x_k) = 1, y)}{q_j(b_j(x_k) = 0, y)} \qquad (9)$$

as output, for $j = 1, \ldots, m$ and $k = 1, \ldots, N$. This channel model is schematically depicted in Fig. 1. With infinite-length interleaving, the m parallel channels were assumed to be independent in [8], [11], or in other words, the correlations among the different subchannels are neglected. For this model, Caire *et al.* defined a *BICM capacity* $\mathsf{C}_{\mathcal{X}}^{\text{bicm}}$, given by

$$\mathsf{C}_{\mathcal{X}}^{\text{bicm}} \triangleq \sum_{j=1}^{m} I(B_j; Y) \qquad (10)$$

$$= \sum_{j=1}^{m} \mathbb{E} \left[\log \frac{\sum_{x' \in \mathcal{X}_B^j} p_{Y \mid X}(Y \mid x')}{\frac{1}{2} \sum_{x' \in \mathcal{X}} p_{Y \mid X}(Y \mid x')} \right] \qquad (11)$$

where the expectation is taken according to $p_j(y \mid b) P_B(b)$, for $b \in \{0, 1\}$ and $P_B(b) = \frac{1}{2}$.

In practice, due to complexity limitations, one might be interested in the following lower complexity version of (7):

$$q_j(b, y) = \max_{x \in \mathcal{X}_b^j} p_{Y \mid X}(y \mid x). \qquad (12)$$

In the log-domain this is known as the max-log approximation.

Summarizing, the BICM decoder of $\mathcal{C}$ uses a mismatched metric of the form given in (4) where the decoder of $\mathcal{C}$ outputs a binary codeword $\hat{\boldsymbol{b}}$ according to

$$\hat{\boldsymbol{b}} = \arg \max_{\boldsymbol{b} \in \mathcal{C}} \prod_{k=1}^{N} \prod_{j=1}^{m} q_j(b_j(x_k), y_k). \qquad (13)$$

III. ACHIEVABILITY OF THE BICM CAPACITY: TYPICAL SEQUENCES

In this section, we provide an achievability proof for the BICM capacity based on typical sequences. The proof is based on the usual random coding arguments [9] with typical sequences, with a slight modification to account for the mismatched decoding metric. This result is obtained without recurring to an infinite interleaver to remove the correlation among the parallel subchannels of the classical BICM model. We first introduce some notation needed for the proof.

We say that a rate R is achievable if, for every $\epsilon > 0$ and for N sufficiently large, there exists an encoder, a demapper, and a decoder such that $\frac{1}{N} \log |\mathcal{M}| \geq R - \epsilon$ and $\Pr(\hat{\mathsf{m}} \neq \mathsf{m}) \leq \epsilon$. We define the joint probability of the channel output y and the corresponding input bits $(b_1, \ldots, b_m)$ as

$$p(b_1, \ldots, b_m, y)$$
$$\triangleq \Pr(B_1 = b_1, \ldots, B_m = b_m, y \leq Y < y + dy) \quad (14)$$

for all $b_j \in \{0, 1\}, y$ and infinitely small dy. We denote the derived marginals by $P_j(b_j)$, for $j = 1, \ldots, m$, and $p_Y(y)$. The marginal distributions with respect to bit B_j and Y are special, and are denoted by $p_j(b_j, y)$. We have then the following theorem.

Theorem 1: The BICM capacity $\mathsf{C}_{\mathcal{X}}^{\text{bicm}}$ is achievable.

Proof: Fix an $\epsilon > 0$. For each $\mathsf{m} \in \mathcal{M}$ we generate a binary codeword $\boldsymbol{b}_1(\mathsf{m}) \ldots, \boldsymbol{b}_m(\mathsf{m})$ with probabilities $P_j(\boldsymbol{b}_j)$. The codebook is the set of all codewords generated with this method.

We consider a threshold decoder, which outputs $\hat{\mathsf{m}}$ only if $\hat{\mathsf{m}}$ is the unique integer satisfying

$$(\boldsymbol{b}_1(\hat{\mathsf{m}}), \ldots, \boldsymbol{b}_m(\hat{\mathsf{m}}), \boldsymbol{y}) \in \mathcal{B}_\epsilon \qquad (15)$$

where $\mathcal{B}_\epsilon$ is a set defined as

$$\mathcal{B}_\epsilon \triangleq \left\{ (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m, \boldsymbol{y}) : \frac{1}{N} \sum_{j=1}^{m} \log \frac{p_j(\boldsymbol{b}_j(\hat{\mathsf{m}}), \boldsymbol{y})}{P_j(\boldsymbol{b}_j) p(\boldsymbol{y})} \geq \Delta_\epsilon \right\}$$
$$(16)$$

for $\Delta_\epsilon \triangleq \sum_{j=1}^{m} I(B_j; Y) - 3m\epsilon$. Otherwise, the decoder outputs an error flag.

The usual random coding argument [9] shows that the error probability, averaged over the ensemble of randomly generated codes, $\bar{P}_e$, is upper-bounded by

$$\bar{P}_e \leq P_1 + (|\mathcal{M}| - 1)P_2 \tag{17}$$

where P_1 is the probability that the received $\boldsymbol{y}$ does not belong to the set $\mathcal{B}_\epsilon$

$$P_1 \triangleq \sum_{(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y}) \notin \mathcal{B}_\epsilon} p(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y}), \tag{18}$$

and P_2 is the probability that another randomly chosen codeword would be (wrongly) decoded, that is,

$$P_2 \triangleq \sum_{(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y}) \in \mathcal{B}_\epsilon} p(\boldsymbol{y}) \prod_{j=1}^{m} P_j(\boldsymbol{b}_j). \tag{19}$$

First, we prove that $\mathcal{B}_\epsilon \supseteq \mathcal{A}_\epsilon(B_1,\ldots,B_m,Y)$, where $\mathcal{A}_\epsilon$ is the corresponding jointly typical set [9]. By definition, the sequences $(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y})$ in the typical set satisfy (among other constraints) the following:

$$-\log p(\boldsymbol{y}) > N(H(Y) - \epsilon) \tag{20}$$
$$-\log P_j(\boldsymbol{b}_j) > N(H(B_j) - \epsilon), \quad j = 1,\ldots,m \tag{21}$$
$$-\log p(\boldsymbol{b}_j,\boldsymbol{y}) < N(H(B_j,Y) + \epsilon), \quad j = 1,\ldots,m. \tag{22}$$

Here $H(\cdot)$ are the entropies of the corresponding random variables. Multiplying the last equation by (-1), and summing them, we have

$$\log \frac{p_j(\boldsymbol{b}_j(\hat{\mathsf{m}}),\boldsymbol{y})}{P_j(\boldsymbol{b}_j)p(\boldsymbol{y})} \geq N(H(B_j) + H(Y) - H(B_j,Y) - 3\epsilon) \tag{23}$$
$$= N(I(B_j;Y) - 3\epsilon) \tag{24}$$

where $I(B_j;Y)$ is the corresponding mutual information. Now, summing over $j = 1,\ldots,m$ we obtain

$$\sum_{j=1}^{m} \log \frac{p_j(\boldsymbol{b}_j(\hat{\mathsf{m}}),\boldsymbol{y})}{P_j(\boldsymbol{b}_j)p(\boldsymbol{y})} \geq N\left(\sum_{j=1}^{m} I(B_j;Y) - 3\epsilon\right) = N\Delta_\epsilon. \tag{25}$$

Hence, all typical sequences belong to the set $\mathcal{B}_\epsilon$, that is, $\mathcal{A}_\epsilon \subseteq \mathcal{B}_\epsilon$. This implies that $\mathcal{B}_\epsilon^c \subseteq \mathcal{A}_\epsilon^c$ and, therefore, the probability P_1 in (18) can be upper-bounded as

$$P_1 \leq \sum_{(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y}) \notin \mathcal{A}_\epsilon} p(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y})$$
$$\leq \epsilon \tag{26}$$

for N sufficiently large. The last inequality follows from the definition of the typical set.

We now move on to P_2. For $(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y}) \in \mathcal{B}_\epsilon$, and from the definition of $\mathcal{B}_\epsilon$, we have that

$$2^{N\Delta_\epsilon} \leq \prod_{j=1}^{m} \frac{p_j(\boldsymbol{b}_j,\boldsymbol{y})}{P_j(\boldsymbol{b}_j)p(\boldsymbol{y})} = \prod_{j=1}^{m} \frac{p_j(\boldsymbol{b}_j \mid \boldsymbol{y})}{P_j(\boldsymbol{b}_j)}. \tag{27}$$

Rearranging terms we have

$$\prod_{j=1}^{m} P_j(\boldsymbol{b}_j) \leq \frac{1}{2^{N\Delta_\epsilon}} \prod_{j=1}^{m} P_j(\boldsymbol{b}_j \mid \boldsymbol{y}). \tag{28}$$

Therefore, the probability P_2 in (19) can be upper-bounded

$$P_2 \leq \frac{1}{2^{N\Delta_\epsilon}} \sum_{(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y}) \in \mathcal{B}_\epsilon} p(\boldsymbol{y}) \prod_{j=1}^{m} P_j(\boldsymbol{b}_j \mid \boldsymbol{y}) \tag{29}$$

$$\leq \frac{1}{2^{N\Delta_\epsilon}} \sum_{(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y})} p(\boldsymbol{y}) \prod_{j=1}^{m} P_j(\boldsymbol{b}_j \mid \boldsymbol{y}) \tag{30}$$

$$= \frac{1}{2^{N\Delta_\epsilon}} \sum_{(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y})} p(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_m,\boldsymbol{y}) \tag{31}$$

$$= \frac{1}{2^{N\Delta_\epsilon}}. \tag{32}$$

Now we can write for $\bar{P}_e$

$$\bar{P}_e \leq P_1 + (|\mathcal{M}| - 1)P_2$$
$$\leq \epsilon + |\mathcal{M}|2^{-N\Delta_\epsilon}$$
$$\leq 2\epsilon \tag{33}$$

for $|\mathcal{M}| = 2^{N(\Delta_\epsilon - \epsilon)}$ and large enough N. We conclude that for large enough N there exist codes such that

$$\frac{1}{N} \log_2 |\mathcal{W}| \geq \Delta_\epsilon - \epsilon = \sum_{j=1}^{m} I(B_j;Y) - (3m+1)\epsilon \tag{34}$$

and $\Pr(\hat{\mathsf{m}} \neq \mathsf{m}) \leq \epsilon$. The rate $\sum_{j=1}^{m} I(B_j;Y)$ is thus achievable.

To conclude, we verify that the BICM decoder is able to determine the probabilities required for the decoding rule defining $\mathcal{B}_\epsilon$ in (16). Since the BICM decoder uses the metric $q_j(b_j,y) \propto p_j(y \mid b_j)$, the log-metric-ratio, or equivalently the *a posteriori* bit probabilities $p_j(b_j \mid y)$, it can also compute

$$\frac{p_j(b_j,y)}{P_j(b_j)p_Y(y)} = \frac{P_j(b_j \mid y)}{P_j(b_j)} \tag{35}$$

where the bit probabilities are known, $P_j(1) = P_j(0) = \frac{1}{2}$. $\square$

IV. ACHIEVABILITY OF THE BICM CAPACITY: ERROR EXPONENTS, GENERALIZED MUTUAL INFORMATION AND CUTOFF RATE

A. Random Coding Exponent

The behavior of the average error probability of a family of randomly generated codes, decoded with a maximum-likelihood decoder, i.e., with metric $q(x,y) = p_{Y\mid X}(y \mid x)$, was studied by Gallager in [15]. In particular, Gallager showed the error probability decreases exponentially with the block length N according to a parameter called the *error exponent*, provided that the code rate R is below the channel capacity C. For memoryless channels, the average error probability over the random coding ensemble [15] can be bounded as

$$\bar{P}_e \leq \exp(-N(E_0(\rho) - \rho R)) \tag{36}$$

where $E_0(\rho)$ is the Gallager function, given by

$$E_0(\rho) \triangleq -\log\left(\int_y \left(\sum_x P_X(x)p_{Y\,|\,X}(y\,|\,x)^{\frac{1}{1+\rho}}\right)^{1+\rho} dy\right)$$
(37)

and $0 \leq \rho \leq 1$ is a free parameter. For a particular input distribution $P_X(x)$, the tightest bound is obtained by optimizing over ρ, which determines the *random coding exponent*

$$E_\mathrm{r}(R) = \max_{0 \leq \rho \leq 1} E_0(\rho) - \rho R.$$
(38)

For uniform input distribution, we define the *coded modulation exponent* $E_0^\mathrm{cm}(\rho)$ as the exponent of a decoder which uses metrics $q(x,y) = p_{Y\,|\,X}(y\,|\,x)$, namely

$$E_0^\mathrm{cm}(\rho) = -\log \mathbb{E}\left[\left(\frac{1}{2^m}\sum_{x'}\left(\frac{p_{Y\,|\,X}(Y\,|\,x')}{p_{Y\,|\,X}(Y\,|\,X)}\right)^{\frac{1}{1+\rho}}\right)^{\rho}\right].$$
(39)

The expectation is carried out according to the joint probability $p_{Y\,|\,X}(y\,|\,x)P_X(x)$.

Gallager's derivation can easily be extended to memoryless channels with generic codeword metrics decomposable as product of symbols metrics, that is, $q(\boldsymbol{x},\boldsymbol{y}) = \prod_{n=1}^{N} q(x_n,y_n)$. Details can be found in [13]. The error probability is upper-bounded by the expression

$$\bar{P}_e \leq \exp\left(-N\left(E_0^q(\rho,s) - \rho R\right)\right)$$
(40)

where the generalized Gallager function $E_0^q(\rho,s)$ is given by

$$E_0^q(\rho,s) = -\log \mathbb{E}\left[\left(\sum_{x'} P_X(x')\frac{q(x',Y)^s}{q(X,Y)^s}\right)^{\rho}\right].$$
(41)

For an input distribution $P_X(x)$, the random coding error exponent $E_\mathrm{r}^q(R)$ is then given by [13]

$$E_\mathrm{r}^q(R) = \max_{0 \leq \rho \leq 1} \max_{s > 0} E_0^q(\rho,s) - \rho R.$$
(42)

For the specific case of BICM, assuming uniformly distributed inputs and a generic bit metric $q_j(b,y)$, we have that Gallager's generalized function $E_0^\mathrm{bicm}(\rho,s)$ is given by

$$E_0^\mathrm{bicm}(\rho,s) = -\log \mathbb{E}\left[\left(\frac{1}{2^m}\sum_{x'}\prod_{j=1}^m \frac{q_j(b_j(x'),Y)^s}{q_j(b_j(X),Y)^s}\right)^{\rho}\right].$$
(43)

Since the cutoff rate is given by $R_0 = E_0(1)$, we analogously define the generalized cutoff rate as

$$R_0^q \triangleq E_\mathrm{r}^q(R=0) = \max_{s > 0} E_0^q(1,s).$$
(44)

B. Data Processing Inequality for Error Exponents

In [13], it was proved that the data-processing inequality holds for error exponents, in the sense that for a given input distribution we have that $E_0^q(\rho,s) \leq E_0(\rho)$ for any $s > 0$.

Next, we rederive this result by extending Gallager's reasoning in [15] to mismatched decoding.

The generalized Gallager function $E_0^q(\rho,s)$ in (41) can be expressed as

$$E_0^q(\rho,s) = -\log\left(\int_y \sum_x P_X(x)p_{Y\,|\,X}(y\,|\,x)\right.$$
$$\left. \times \left(\sum_{x'} P_X(x')\left(\frac{q(x',y)}{q(x,y)}\right)^s\right)^{\rho}\right) dy.$$
(45)

As long as the metric does not depend on the transmitted symbol x, the function inside the logarithm can be rewritten as

$$\int_y \left(\sum_x P_X(x)p_{Y\,|\,X}(y\,|\,x)q(x,y)^{-s\rho}\right)$$
$$\times \left(\sum_{x'} P_X(x')q(x',y)^s\right)^{\rho} dy.$$
(46)

For a fixed channel observation y, the integrand is reminiscent of the right-hand side of Hölder's inequality (see of [15, Exercise 4.15]), which can be expressed as

$$\left(\sum_i a_i b_i\right)^{1+\rho} \leq \left(\sum_i a_i^{1+\rho}\right)\left(\sum_i b_i^{\frac{1+\rho}{\rho}}\right)^{\rho}.$$
(47)

Hence, with the identifications

$$a_i = P_X(x)^{\frac{1}{1+\rho}} p_{Y\,|\,X}(y\,|\,x)^{\frac{1}{1+\rho}} q(x,y)^{\frac{-s\rho}{1+\rho}}$$
(48)
$$b_i = P_X(x)^{\frac{\rho}{1+\rho}} q(x,y)^{\frac{s\rho}{1+\rho}},$$
(49)

we can lower-bound (46) by the quantity

$$\int_y \left(\sum_x P_X(x)p_{Y\,|\,X}(y\,|\,x)^{\frac{1}{1+\rho}} dy\right)^{1+\rho}.$$
(50)

Recovering the logarithm in (45), for a general mismatched decoder, arbitrary $s > 0$, and any input distribution, we obtain that

$$E_0(\rho) \geq E_0^q(\rho,s).$$
(51)

Note that the expression in (50) is independent of s and of the specific decoding metric $q(x,y)$. Nevertheless, evaluation of Gallager's generalized function for the specific choices $s = \frac{1}{1+\rho}$ and $q(x,y) \propto p_{Y\,|\,X}(y\,|\,x)$ attains the lower bound, which is also (39).

Equality holds in Hölder's inequality if and only if for all i and some positive constant c, $a_i^{\frac{\rho}{1+\rho}} = c\,b_i^{\frac{1}{1+\rho}}$ (see [15, Exercise 4.15]). In our context, simple algebraic manipulations show that the necessary condition for equality to hold is that

$$p_{Y\,|\,X}(y\,|\,x) = c'q(x,y)^{s'}, \quad \text{for all } x \in \mathcal{X}$$
(52)

for some constants c' and s'. In other words, the metric $q(x,y)$ must be proportional to a power of the channel transition probability $p_{Y\,|\,X}(y\,|\,x)$, for the bound (51) to be tight, and therefore, to achieve the coded modulation error exponent.

C. Error Exponent for BICM With the Parallel-Channel Model

In their analysis of multilevel coding and successive decoding, Wachsmann *et al.* provided the error exponents of BICM modeled as a set of parallel channels [11]. The corresponding Gallager's function, which we denote by $E_0^{\mathrm{ind}}(\rho)$, is given by

$$
E_0^{\mathrm{ind}}(\rho) = -\sum_{j=1}^{m} \log \int_y \sum_{b_j=0}^{1} P_j(b_j) p_j(y \mid b_j)
$$
$$
\times \left(\sum_{b_j'=0}^{1} P_j(b_j') \frac{q_j(b_j', y)^{\frac{1}{1+\rho}}}{q_j(b_j, y)^{\frac{1}{1+\rho}}} \right)^{\rho} dy, \quad (53)
$$

which corresponds to a binary-input channel with input b_j, output y, and bit metric matched to the transition probability $p_j(y \mid b_j)$.

This equation can be rearranged into a form similar to the one given in previous sections. First, we insert the summation in the logarithm

$$
E_0^{\mathrm{ind}}(\rho) = -\log \left(\prod_{j=1}^{m} \int_y \sum_{b_j=0}^{1} P_j(b_j) p_j(y \mid b_j) \right.
$$
$$
\left. \times \left(\sum_{b_j'=0}^{1} P_j(b_j') \frac{q_j(b_j', y)^{\frac{1}{1+\rho}}}{q_j(b_j, y)^{\frac{1}{1+\rho}}} \right)^{\rho} dy \right). \quad (54)
$$

Then, we notice that the output variables y are dummy variables which possibly vary for each value of j. Let us denote the dummy variable in the jth subchannel by y_j'. We have then

$$
E_0^{\mathrm{ind}}(\rho) = -\log \left(\prod_{j=1}^{m} \int_{y_j'} \sum_{b_j=0}^{1} P_j(b_j) p(y_j' \mid b_j) \right.
$$
$$
\left. \times \left(\sum_{b_j'=0}^{1} p_j(b_j') \frac{q_j(b_j', y_j')^{\frac{1}{1+\rho}}}{q_j(b_j, y_j')^{\frac{1}{1+\rho}}} \right)^{\rho} dy_j' \right) \quad (55)
$$
$$
= -\log \left(\int_{\boldsymbol{y}'} \sum_{x} P_X(x) p_{\boldsymbol{Y} \mid X}(\boldsymbol{y}' \mid x) \right.
$$
$$
\left. \times \left(\sum_{x'} P_X(x') \frac{q(x', \boldsymbol{y}')^{\frac{1}{1+\rho}}}{q(x, \boldsymbol{y}')^{\frac{1}{1+\rho}}} \right)^{\rho} d\boldsymbol{y}' \right). \quad (56)
$$

Here we carried out the multiplications, defined the vector $\boldsymbol{y}'$ to be the collection of the m channel outputs, and denoted by $x = \mu(b_1, \ldots, b_m)$ and $x' = \mu(b_1', \ldots, b_m')$ the symbols selected by the bit sequences. This equation is the Gallager function of a mismatched decoder for a channel output $\boldsymbol{y}'$, such that for each of the m subchannels sees a statistically independent channel realization from the others. In general, since the original channel cannot be decomposed into parallel, conditionally independent subchannels, this parallel-channel model fails to capture the statistics of the channel.

The cutoff rate with the parallel-channel model is given by

$$
R_0^{\mathrm{ind}} = -\sum_{j=1}^{m} \log \int_y \left(\sum_{b_j=0}^{1} P_j(b_j) p_j(y \mid b_j)^{\frac{1}{2}} \right)^2 dy. \quad (57)
$$

The cutoff rate was given in [8] as m times the cutoff rate of an averaged channel

$$
R_0^{\mathrm{av}} \triangleq m \left[\log 2 \right.
$$
$$
\left. -\log \left(1 + \frac{1}{m} \sum_{j=1}^{m} \mathbb{E} \left[\sqrt{\frac{\sum_{x' \in \mathcal{X}_B^j} p_{Y \mid X}(Y \mid x')}{\sum_{x' \in \mathcal{X}_B^j} p_{Y \mid X}(Y \mid x')}} \right] \right) \right]. \quad (58)
$$

From Jensen's inequality one easily obtains that $R_0^{\mathrm{av}} \leq R_0^{\mathrm{ind}}$.

D. Generalized Mutual Information for BICM

The largest achievable rate with mismatched decoding is not known in general. Perhaps the easiest candidate to deal with is the GMI [12]–[14], given by

$$
I_{\mathrm{gmi}} \triangleq \sup_{s>0} I_{\mathrm{gmi}}(s) \quad (59)
$$

where

$$
I_{\mathrm{gmi}}(s) \triangleq \mathbb{E} \left[\log \frac{q(X, Y)^s}{\sum_{x' \in \mathcal{X}} P_X(x') q(x', Y)^s} \right]. \quad (60)
$$

As in the case of matched decoding, this definition can be recovered from the error exponent

$$
I_{\mathrm{gmi}}(s) = \frac{dE_0^q(\rho, s)}{d\rho} \bigg|_{\rho=0} = \lim_{\rho \to 0} \frac{E_0^q(\rho, s)}{\rho}. \quad (61)
$$

We next see that the generalized mutual information is equal to the BICM capacity of [8] when the metric (7) is used. Similarly to Section III, the result does not require the presence of an interleaver of infinite length. Further, the interleaver is actually not necessary for the random coding arguments. First, we have the following,

Theorem 2: The generalized mutual information of the BICM mismatched decoder is equal to the sum of the generalized mutual informations of the independent binary-input parallel channel model of BICM

$$
I_{\mathrm{gmi}} = \sup_{s>0} \sum_{j=1}^{m} \mathbb{E} \left[\log \frac{q_j(b_j, Y)^s}{\frac{1}{2} \sum_{b'=0}^{1} q_j(b_j', Y)^s} \right]. \quad (62)
$$

The expectation is carried out according to the joint distribution $P_j(b_j) p_j(y \mid b_j)$, with $P_j(b_j) = \frac{1}{2}$.

Proof: For a given s, and uniform inputs, i.e., $P_X(x) = \frac{1}{2^m}$, (60) gives

$$
I_{\mathrm{gmi}}(s) = \mathbb{E} \left[\log \frac{\prod_{j=1}^{m} q_j(b_j(X), Y)^s}{\sum_{x'} \frac{1}{2^m} \prod_{j=1}^{m} q_j(b_j(x'), Y)^s} \right]. \quad (63)
$$

We now have a closer look at the denominator in the logarithm of (63). The key observation here is that the sum over the constellation points of the product over the binary label positions can be expressed as the product over the label position is the sum of the probabilities of the bits being zero and one, i.e.,

$$
\sum_{x'} \frac{1}{2^m} \prod_{j=1}^{m} q_j(b_j(x'), Y)^s
$$
$$
= \frac{1}{2^m} \prod_{j=1}^{m} (q_j(b_j' = 0, Y)^s + q_j(b_j' = 1, Y)^s) \quad (64)
$$

$$= \prod_{j=1}^{m} \left(\frac{1}{2} (q_j(b'_j = 0, Y)^s + q_j(b'_j = 1, Y)^s) \right). \quad (65)$$

Rearranging terms in (63) we obtain

$$I_{\text{gmi}}(s) = \mathbb{E} \left[\sum_{j=1}^{m} \log \frac{q_j(b_j(x), Y)^s}{\frac{1}{2} (q_j(b'_j = 0, Y)^s + q_j(b'_j = 1, Y)^s)} \right] \quad (66)$$

$$= \sum_{j=1}^{m} \frac{1}{2} \sum_{b=0}^{1} \frac{1}{2^{m-1}} \sum_{x \in \mathcal{X}_b^j} \mathbb{E} \left[\log \frac{q_j(b_j(x), Y)^s}{\frac{1}{2} \sum_{b'=0}^{1} q_j(b'_j, Y)^s} \right] \quad (67)$$

the expectation being done according to the joint distribution $P_j(b_j) p_j(y \mid b_j)$, with $P_j(b_j) = \frac{1}{2}$. □

There are a number of interesting particular cases of the above theorem.

Corollary 1: For the classical BICM decoder with metric in (7)

$$I_{\text{gmi}} = \mathsf{C}_{\mathcal{X}}^{\text{bicm}}. \quad (68)$$

Proof: Since the metric $q_j(b_j, y)$ is proportional to $p_j(y \mid b_j)$, we can identify the quantity

$$\mathbb{E} \left[\log \frac{q_j(B_j, Y)^s}{\frac{1}{2} \sum_{b'_j=0}^{1} q_j(b'_j, Y)^s} \right] \quad (69)$$

as the generalized mutual information of a *matched* binary-input channel with transitions $p_j(y \mid b_j)$. Then, the supremum over s is achieved at $s = 1$ and we get the desired result. □

Corollary 2: For the max-log metric in (12)

$$I_{\text{gmi}} = \sup_{s>0} \sum_{j=1}^{m} \mathbb{E} \left[\log \frac{\left(\max_{x \in \mathcal{X}_j^B} p_{Y \mid X}(y \mid x) \right)^s}{\frac{1}{2} \sum_{b=0}^{1} \left(\max_{x' \in \mathcal{X}_b^j} p_{Y \mid X}(y \mid x') \right)^s} \right]. \quad (70)$$

Szczecinski *et al.* studied the mutual information with this decoder [16], using this formula for $s = 1$. Clearly, the optimization over s may induce a larger achievable rate, as we see in the next section. More generally, as we shall see later, letting $s = 1/(1+\rho)$ in the mismatched error exponent can yield some degradation.

E. BICM With Mismatched Decoding: Numerical Results

The data-processing inequality for error exponents yields $E_0^{\text{bicm}}(\rho, s) \leq E_0^{\text{cm}}(\rho)$, where the quantity in the right-hand side is the coded modulation exponent. On the other hand, no general relationship holds between $E_0^{\text{ind}}(\rho)$ and $E_0^{\text{cm}}(\rho)$. As it will be illustrated in the following examples, there are cases for which $E_0^{\text{cm}}(\rho)$ can be larger than $E_0^{\text{ind}}(\rho)$, and *vice versa*.

Figs. 2, 3, and 4 show the error exponents for coded modulation (solid), BICM with independent parallel channels (dashed), BICM using mismatched metric (7) (dash-dotted), and BICM using mismatched metric (12) (dotted) for 16-QAM with Gray mapping, Rayleigh fading, and snr = 5,15, -25 dB, respectively. Dotted lines labeled with $s = \frac{1}{1+\rho}$ correspond to the error exponent of BICM using mismatched metric (12) letting
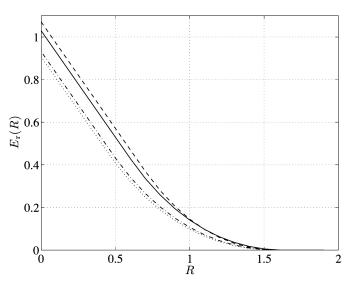

Fig. 2. Error exponents for coded modulation (solid), BICM with independent parallel channels (dashed), BICM using mismatched metric (7) (dash-dotted), and BICM using mismatched metric (12) (dotted) for 16-QAM with Gray mapping, Rayleigh fading, and snr = 5 dB.
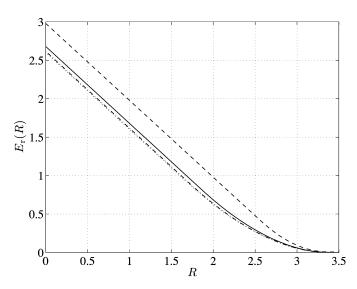

Fig. 3. Error exponents for coded modulation (solid), BICM with independent parallel channels (dashed), BICM using mismatched metric (7) (dash-dotted), and BICM using mismatched metric (12) (dotted) for 16-QAM with Gray mapping, Rayleigh fading, and snr = 15 dB.

$s = \frac{1}{1+\rho}$. The parallel-channel model gives a larger exponent than the coded modulation, in agreement with the cutoff rate results of [8]. In contrast, the mismatched-decoding analysis yields a lower exponent than coded modulation. As mentioned in the previous section, both BICM models yield the same capacity.

In most cases, BICM with a max-log metric (12) incurs in a marginal loss in the exponent for mid-to-large SNR. In this SNR range, the optimized exponent and that with $s = \frac{1}{1+\rho}$ are almost equal. For low SNR, the parallel-channel model and the mismatched-metric model with (7) have the same exponent, while we observe a larger penalty when metrics (12) are used. As we observe, some penalty is incurred at low SNR for not optimizing over s. We denote with crosses the corresponding achievable information rates.
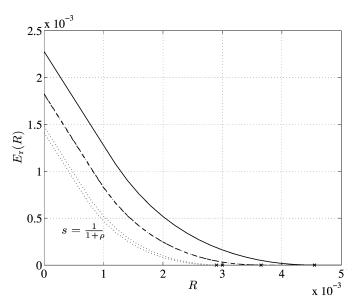
Fig. 4. Error exponents for coded modulation (solid), BICM with independent parallel channels (dashed), BICM using mismatched metric (7) (dash-dotted), and BICM using mismatched metric (12) (dotted) for 16-QAM with Gray mapping, Rayleigh fading, and $\mathsf{snr} = -25$ dB. Crosses correspond to (from right to left) coded modulation, BICM with metric (7), BICM with metric (12), and BICM with metric (12) with $s = 1$.
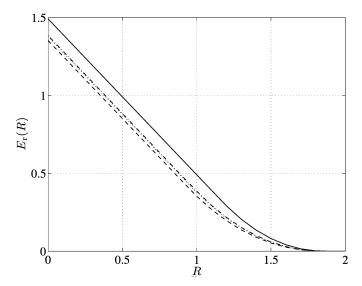

Fig. 5. Error exponents for coded modulation (solid), BICM with independent parallel channels (dashed), BICM using mismatched metric (7) (dash-dotted), and BICM using mismatched metric (12) (dotted) for 8-PSK with Gray mapping, AWGN and $\mathsf{snr} = 5$ dB.

An interesting question is whether the error exponent of the parallel-channel model is always larger than that of the mismatched decoding model. The answer is negative, as illustrated in Fig. 5, which shows the error exponents for coded modulation (solid), BICM with independent parallel channels (dashed), BICM using mismatched metric (7) (dash-dotted), and BICM using mismatched metric (12) (dotted) for 8-PSK with Gray mapping in the unfaded AWGN channel.

V. EXTRINSIC SIDE INFORMATION

Next to the classical decoder described in Section II, iterative decoders have also received much attention [3]–[7] due to their

improved performance. Iterative decoders can also be modeled as mismatched decoders, where the bit decoding metric is now of the form

$$q_j(b, y) = \sum_{x' \in \mathcal{X}_b^j} p_{Y \mid X}(y \mid x') \prod_{j' \neq j} \mathrm{ext}_{j'}(b_{j'}(x')), \qquad (71)$$

where we denote by $\mathrm{ext}_j(b)$ the extrinsic information, i.e., the "*a priori*" probability that the jth bit takes the value b. Extrinsic information is commonly generated by the decoder of the binary code $\mathcal{C}$. Clearly, we have that $\mathrm{ext}_j(0) + \mathrm{ext}_j(1) = 1$, and $0 \leq \mathrm{ext}_j(0), \mathrm{ext}_j(1) \leq 1$. Without extrinsic information, we take $\mathrm{ext}_j(0) = \mathrm{ext}_j(1) = \frac{1}{2}$, and the metric is given by (7).

In the analysis of iterative decoding, extrinsic information is often modeled as a set of random variables $\mathrm{EXT}_j(0)$, where we have defined without loss of generality the variables with respect to the all-zero symbol. We denote the joint density function by $p(\mathrm{ext}_1(0), \ldots, \mathrm{ext}_m(0)) = \prod_{j=1}^m p(\mathrm{ext}_j(0))$. We discuss later how to map the actual extrinsic information generated in the decoding process onto this density. The mismatched decoding error exponent $E_0^q(\rho, s)$ for metric (71) is given by (43), where the expectation is now carried out according to the joint density $P_X(x)p_{Y \mid X}(y \mid x)p(\mathrm{ext}_1(0)) \cdots p(\mathrm{ext}_m(0))$. Similarly, the generalized mutual information is again obtained as

$$I_{\mathrm{gmi}} = \max_s \lim_{\rho \to 0} \frac{E_0^q(\rho, s)}{\rho}.$$

It is often assumed [5] that the decoding metric acquires knowledge on the symbol x effectively transmitted, in the sense that for any symbol $x' \in \mathcal{X}$, the jth bit decoding metric is given by

$$q_j(b_j(x') = b, y)$$
$$= \sum_{x'' \in \mathcal{X}_b^j} p_{Y \mid X}(y \mid x'') \prod_{j' \neq j} \mathrm{ext}_{j'}(b_{j'}(x'') \oplus b_{j'}(x)) \quad (72)$$

where $\oplus$ denotes the binary addition. Observe that extrinsic information is defined relative to the transmitted symbol x, rather than relative to the all-zero symbol. If the jth bit of the symbols x'' and x coincide, the extrinsic information for bit zero $\mathrm{ext}_j(0)$ is selected, otherwise the extrinsic information $\mathrm{ext}_j(1)$ is used.

For the metric in (72), the proof presented in Section IV-B of the data processing inequality fails because the integrand in (46) cannot be decomposed into a product of separate terms, respectively, depending on x and x', the reason being that the metric $q(x', y)$ varies with x.

On the other hand, since the symbol metric $q(x', y)$ is the same for all symbols x', the decomposition of the generalized mutual information as a sum of generalized mutual informations across the m bit labels in Theorem 2 remains valid, and we have therefore

$$I_{\mathrm{gmi}} = \sum_{j=1}^m \mathbb{E}\left[\log \frac{q_j(B_j, Y)}{\frac{1}{2}\sum_{b'=0}^1 q_j(b_j = b', Y)}\right]. \qquad (73)$$

This expectation is carried out according to

$$P(b_j)p_j(y \mid b_j)p(\mathrm{ext}_1(0)) \cdots p(\mathrm{ext}_m(0))$$

with $P(b_j) = \frac{1}{2}$. Each of the summands can be interpreted as the mutual information achieved by nonuniform signaling in the constellation set $\mathcal{X}$, where the probabilities according to which

the symbols are drawn are a function of the extrinsic informations $\text{ext}_j(\,\cdot\,)$. The value of I_{gmi} may exceed the channel capacity [5], so this quantity is a pseudo-generalized mutual information, with the same functional form but lacking operational meaning as an achievable rate by the decoder.

Alternatively, the metric in (71) may depend on the hypothesized symbol x', that is,

$$
\begin{aligned}
&q_j(b_j(x') = b, y) \\
&= \sum_{x'' \in \mathcal{X}_b^j} p_{Y\,|\,X}(y\,|\,x'') \prod_{j' \neq j} \text{ext}_{j'}(b_{j'}(x'') \oplus b_{j'}(x')). \quad (74)
\end{aligned}
$$

Differently from (72), the bit metric varies with the hypothesized symbol x' and not with the transmitted symbol x. Therefore, Theorem 2 cannot be applied and the generalized mutual information cannot be expressed as a sum of mutual informations across the bit labels. On the other hand, the data processing inequality holds and, in particular, the error exponent and the generalized mutual information are upper-bounded by that of coded modulation. Moreover, we have the following result.

Theorem 3: In the presence of perfect extrinsic side information, the error exponent with metric (74) coincides with that of coded modulation.

Proof: With perfect extrinsic side information, all the bits $j' \neq j$ are known, and then

$$
\prod_{j' \neq j} \text{ext}_{j'}(b_{j'}(x'') \oplus b_{j'}(x')) = \begin{cases} 1, & \text{when } x'' = x' \\ 0, & \text{otherwise} \end{cases} \quad (75)
$$

which guarantees that only the symbol $x'' = x'$ is selected. Then, $q_j(b_j(x') = b, y) = p_{Y\,|\,X}(y\,|\,x')$ and the symbol metric becomes $q(x', y) = p_{Y\,|\,X}(y\,|\,x')^m$ for all $x' \in \mathcal{X}$. As we showed in (52), this is precisely the condition under which the error exponent (and the capacity) with mismatched decoding coincides that of coded modulation.

The above result suggests that with perfect extrinsic side information, the gap between the error exponent (and mutual information) of BICM and that of coded modulation can be closed if one could provide perfect side information to the decoder. A direct consequence of this result is that the generalized mutual information with BICM metric (71) and perfect extrinsic side information is equal to the mutual information of coded modulation. An indirect consequence of this result is that the multistage decoding [17], [11] does not attain the exponent of coded modulation, even though its corresponding achievable rate is the same. The reason is that the decoding metric is not of the form $c\,p_{Y\,|\,X}(y\,|\,x)^s$, for some constant c and s, except for the last bit in the decoding sequence. We hasten to remark that the above rate in presence of perfect extrinsic side information need not be achievable, in the sense that there may not exist a mechanism for accurately feeding the quantities $\text{ext}_j(b)$ to the demapper. The actual link to the iterative decoding process is open for future research.

VI. Conclusion

We have presented a mismatched-decoding analysis of BICM, which is valid for arbitrary finite-length interleavers.

We have proved that the corresponding generalized mutual information coincides with the BICM capacity originally given by Caire *et al.* modeling BICM as a set of independent parallel channels. More generally, we have seen that the error exponent cannot be larger than that of coded modulation, contrary to the analysis of BICM as a set of parallel channels. For Gaussian channels with binary reflected Gray mapping, the gap between the BICM and CM error exponents is small, as found by Caire *et al.* for the capacity. We have also seen that the mutual information appearing in the analysis of iterative decoding of BICM via EXIT charts admits a representation as a form of generalized mutual information. However, since this quantity may exceed the capacity, its operational meaning as an achievable rate is unclear. We have modified the extrinsic side information available to the decoder, to make it dependent on the hypothesized symbol rather than on the transmitted one, and shown that the corresponding error exponent is always lower-bounded by that of coded modulation. In presence of perfect extrisinc side information, both error exponents coincide. The precise connection with iterative decoding is open for future research.

References

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

[2] G. Ungerböck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 1, pp. 55–66, Jan. 1982.

[3] X. Li and J. A. Ritcey, "Bit-interleaved coded modulation with iterative decoding using soft feedback," *Electron. Lett.*, vol. 34, no. 10, pp. 942–943, 1998.

[4] X. Li and J. A. Ritcey, "Trellis-coded modulation with bit interleaving and iterative decoding," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 715–724, Apr. 1999.

[5] S. ten Brink, "Designing iterative decoding schemes with the extrinsic information transfer chart," *AEU Int. J. Electron. Commun.*, vol. 54, no. 6, pp. 389–398, 2000.

[6] S. ten Brink, J. Speidel, and R. H. Han, "Iterative demapping for QPSK modulation," *Electron. Lett.*, vol. 34, no. 15, pp. 1459–1460, 1998.

[7] S. ten Brink, "Convergence of iterative decoding," *Electron. Lett.*, vol. 35, p. 806, 1999.

[8] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 927–946, 2 1998.

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

[10] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[11] 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.

[12] 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, Nov. 1994.

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

[14] A. Ganti, A. Lapidoth, and İ. 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, Nov. 2000.

[15] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[16] L. Szczecinski, A. Alvarado, R. Feick, and L. Ahumada, "On the distribution of L-values in gray-mapped M^2-QAM signals: Exact expressions and simple approximations," in *Proc. IEEE Global Communication Conf. (GLOBECOM)*, Washington, DC, Nov. 2007, pp. 1759–1764.

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

[18] A. Guillén i Fàbregas, A. Martinez, and G. Caire, "Bit-interleaved coded modulation," *Found. Trends Commun. Inf. Theory*, vol. 5, no. 1–2, pp. 1–153, 2008.

Alfonso Martinez (S'97–M'04) was born in Zaragoza, Spain, in October 1973. He graduated in electrical engineering from the University of Zaragoza, in 1997. He received the Ph.D. degree in electrical engineering from Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in January 2008.

From 1998 to 2002, he worked at the research centre of the European Space Agency (ESTEC/ESA) in Noordwijk, The Netherlands. He conducted research on coded modulation for satellite systems. From 2003 to 2008, he was a Research and Teaching Assistant at Technische Universiteit Eindhoven, where he was involved in research on optical communication theory. In September 2005 he visited the Institute for Telecommunications Research, University of South Australia, Mawson Lakes, Australia. From August to December 2007, he was a Visiting Scientist at the Department of Engineering, University of Cambridge, Cambridge, U.K. He is currently working as a Research Associate at CWI, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands. His main research interests lie in optical communications and in coding theory, including digital modulation, and signal processing techniques.

Albert Guillén i Fàbregas (S'01–M'05–SM'09)) was born in Barcelona, Catalunya, Spain, in 1974. He received the telecommunication engineering degree and the electronics engineering degree from Universitat Politècnica de Catalunya, Barcelona, Catalunya, Spain, and the Politecnico di Torino, Torino, Italy, respectively, in 1999, and the Ph.D. degree in communication systems from Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2004.

From August 1998 to March 1999, he conducted his Final Research Project at the New Jersey Institute of Technology, Newark, NJ. He was with Telecom Italia Laboratories, Torino, Italy, from November 1999 to June 2000 and with the European Space Agency (ESA), Noordwijk, The Netherlands, from September 2000 to May 2001. During his doctoral studies, from 2001 to 2004, he was a Research and Teaching Assistant at the Institut Eurécom, Sophia-Antipolis, France. From June 2003 to July 2004, he was a Visiting Scholar at the Communications Theory Lab at EPFL. From September 2004 to December 2006, he was a Research Fellow at the Institute for Telecommunications Research, University of South Australia, Mawson Lakes, Australia. Since 2007, he has been a Lecturer in the Department of Engineering, University of Cambridge, Cambridge, U.K., where he is also a Fellow of Trinity Hall. He held visiting appointments at Ecole Nationale Supèrieure des Télécommunications, Paris, France (June–July 2005 and June 2006), Universitat Pompeu Fabra, Barcelona, Spain (September 2008), and at the Institute for Telecommunications Research, University of South Australia (December 2009). His specific research interests are in the area of communication theory, information theory, coding theory, digital modulation, and signal processing techniques, particularly with wireless terrestrial and satellite applications.

Dr. Guillén i Fàbregas is currently an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He received a predoctoral Research Fellowship of the Spanish Ministry of Education to join ESA. He received the Young Authors Award of the 2004 European Signal Processing Conference EUSIPCO 2004, Vienna, Austria, and the 2004 Nokia Best Doctoral Thesis Award in Mobile Internet and 3rd Generation Mobile Solutions from the Spanish Institution of Telecommunications Engineers. He is also a member of the ARC Communications Research Network (ACoRN) and a Junior Member of the Isaac Newton Institute for Mathematical Sciences.

Giuseppe Caire (S'92–M'94–SM'03–F'05) was born in Torino, Italy, in 1965. He received the B.Sc. degree in electrical engineering from Politecnico di Torino, Torino, Italy, in 1990, the M.Sc. degree in electrical engineering from Princeton University, Princeton, NJ, in 1992, and the Ph.D. degree from Politecnico di Torino in 1994.

He visited Princeton University in Summer 1997 and Sydney University, Sydney, Australia, in Summer 2000. He has been an Assistant Professor in Telecommunications at the Politecnico di Torino, an Associate Professor at the University of Parma, Parma, Italy, a Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, and is now a Professor in the Electrical Engineering Department of the Viterbi School of Engineering, University of Southern California, Los Angeles. He has been with the European Space Agency (ESTEC, Noordwijk, The Netherlands) from May 1994 to February 1995. His current research interests are in the field of communications theory, information theory, and coding theory with particular focus on wireless applications.

Prof. Caire served as Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS during 1998–2001 and as Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY during 2001–2003. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, and the Joint IT/COM Societies Best Paper Award in 2004. Since November 2004, he has been a Member of the Board of Governors of the IEEE Information Theory Society. He was a recipient of the AEI G. Someda Scholarship in 1991, was a recipient of the COTRAO Scholarship in 1996, and of a CNR Scholarship in 1997.

Frans M. J. Willems (F'05) was born in Stein, The Netherlands, on June 26, 1954. He received the M.Sc. degree in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands, and the Ph.D. degree from the Catholic University of Louvain, Louvain, Belgium, in 1979 and 1982, respectively.

From 1979 to 1982, he was a Research Assistant at the Catholic University of Louvain. In 1982, he became a staff member at the Electrical Engineering Department of Eindhoven University of Technology. His research contributions are in the areas of multiuser information theory and noiseless source coding. Since 1999, he has been an advisor for Philips Research Laboratories for subjects related to information theory.

Dr. Willems received the Marconi Young Scientist Award in 1982. From 1988 to 1990, he served as Associate Editor for Shannon Theory for the IEEE TRANSACTIONS ON INFORMATION THEORY. From 2002 to 2006, he was Associate Editor for Information Theory for the *European Transactions on Telecommunications*. He is corecipient of the 1996 IEEE Information Theory Society Paper Award. From 1998 to 2000, he was a member of the Board of Governors of the IEEE Information Theory Society.