IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Universal Decoding for Source–Channel Coding with Side Information

Neri Merhav

CCIT Report #885
July 2015

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT Report #885 July 2015

Universal Decoding for Source–Channel Coding with Side Information *

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

Abstract

We consider a setting of Slepian–Wolf coding, where the random bin of the source vector undergoes channel coding, and then decoded at the receiver, based on additional side information, correlated to the source. For a given distribution of the randomly selected channel codewords, we propose a universal decoder that depends on the statistics of neither the correlated sources nor the channel, assuming first that they are both memoryless. Exact analysis of the random–binning/random–coding error exponent of this universal decoder shows that it is the same as the one achieved by the optimal maximum a–posteriori (MAP) decoder. Previously known results on universal Slepian–Wolf source decoding, universal channel decoding, and universal source–channel decoding, are all obtained as special cases of this result. Subsequently, we outline further generalizations of our results in several directions, including: (i) finite–state sources and finite–state channels, along with a universal decoding metric that is based on Lempel–Ziv parsing, (ii) arbitrary sources and channels, where the universal decoding is with respect to a given class of decoding metrics, and (iii) full (symmetric) Slepian–Wolf coding, where both source streams are separately fed into random–binning source encoders, followed by random channel encoders, which are then jointly decoded by a universal decoder.

*This research was partially supported by the Israel Science Foundation (ISF), grant no. 412/12.

1 Introduction

Universal decoding for unknown channels is a topic that attracted considerable attention throughout the last four decades. In [9], Goppa was the first to offer the *maximum mutual information* (MMI) decoder, which decodes the message as the one whose codeword has the largest empirical mutual information with the channel output sequence. Goppa proved that for discrete memoryless channels (DMC's), MMI decoding attains capacity. Csiszár and Körner [4, Theorem 5.2] have further showed that the random coding error exponent of the MMI decoder, pertaining to the ensemble of the uniform random coding distribution over a certain type class, achieves the same random coding error exponent as the optimum, maximum likelihood (ML) decoder. Ever since these early works on universal channel decoding, a considerably large volume of research work has been done, see, e.g., [3], [7], [8], [11], [13], [14], [16], [25], for a non–exhaustive list of works on memoryless channels, as well as more general classes of channels.

At the same time, considering the analogy between channel coding and Slepian–Wolf (SW) source coding, it is not surprising that universal schemes for SW decoding, like the *minimum entropy* (ME) decoder, have also been derived, first, by Csiszár and Körner [4, Exercise 3.1.6], and later further developed by others in various directions, see, e.g., [1], [6], [10], [26], [27], [31].

Much less attention, however, has been devoted to universal decoding for joint source–channel coding, where both the source and the channel are unknown to the decoder, Csiszár [2] was the first to propose such a universal decoder, which he referred to as the *generalized MMI* decoder. The generalized MMI decoding metric, to be maximized among all messages, is essentially[1] given by the difference between the empirical input–output mutual information of the channel and the empirical entropy of the source. In a way, it naturally combines the concepts of MMI channel decoding and ME source decoding. But the emphasis in [2] was inclined much more towards upper and lower bounds on the reliability functions, whereas the universality of the decoder was quite a secondary issue. Consequently, later articles that refer to [2] also focus, first and foremost, on the joint source–channel reliability function and not really on universal decoding. We are not aware of subsequent works on universal source–channel decoding other than [15], which concerns a completely different setting, of zero–delay coding.

In this work, we consider universal joint source–channel decoding in a setting that is more general than that of [2], and we also further extend it in several directions. In particular, we consider the communication system depicted in Fig. 1, which is described as follows: A source vector $\boldsymbol{u}$, emerging from a discrete memoryless source (DMS), undergoes Slepian–Wolf encoding (random binning) at rate R, followed by channel coding (random coding). The discrete memoryless channel (DMC) output $\boldsymbol{y}$ is fed into the decoder, along with a side information (SI) vector $\boldsymbol{v}$, correlated to the source $\boldsymbol{u}$, and the output of the decoder, $\hat{\boldsymbol{u}}$, should agree with $\boldsymbol{u}$ with probability as high as possible. Our first goal is to characterize the exact exponential rate of the expected error probability, associated with the optimum MAP decoder, where the expectation is over the joint ensemble of the random

[1] Strictly speaking, Csiszźar's decoding metric is slightly different, but is asymptotically equivalent to this definition.

binning encoder and the random channel code. We refer to this exponential rate as the *random–binning/random–coding error exponent*. Our second goal is to show that this error exponent is also achieved by a universal decoder, that depends neither on the statistics of the source, nor of the channel, and which is similar to Csiszár's generalized MMI decoder. Beyond the fact this model is more general than the one in [2] (in the sense of including the random binning component as well as decoder SI), the assertion of the universal optimality of the generalized MMI decoder is somewhat stronger here than in [2]. In [2] the performance of the generalized MMI decoder is compared directly to an upper bound on the joint source–channel reliability function, and the claim on the optimality of this decoder is asserted only in the range where this bound is tight. Here, on the other hand (similarly as in earlier works on universal pure channel decoding), we argue that the generalized MMI decoder is always asymptotically as good as the optimal MAP decoder, in the error exponent sense, no matter whether or not there is a gap between the achievable exponent and the upper bound on the reliability function.

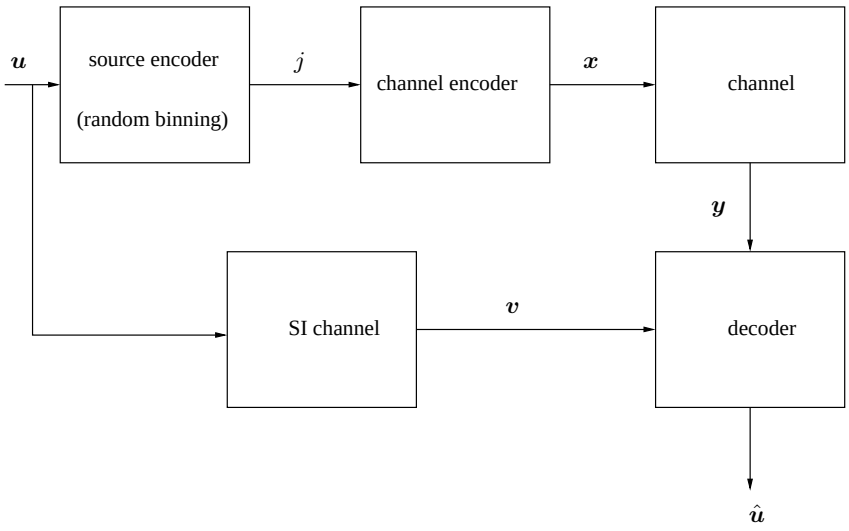

Figure 1: Slepian–Wolf source coding, followed by channel coding. The source $\boldsymbol{u}$ is source–channel encoded, whereas the correlated SI $\boldsymbol{v}$ (described as being generated by a DMC fed by $\boldsymbol{u}$) is available at the decoder.

One of the motivations for studying this model is that it captures, in a unified framework, several important special cases of communication systems, from the perspective of universal decoding.

1. Separate source coding and channel coding without SI: letting $\boldsymbol{v}$ be degenerate.
2. Pure SW source coding: letting the channel be clean ($\boldsymbol{y} \equiv \boldsymbol{x}$) and assuming the channel alphabet to be very large (so that the probability for two or more identical codewords would be negligible).

3. Pure channel coding: letting the source be binary and symmetric, and the SI be degenerate.

4. Joint source–channel coding with and without SI: letting the binning rate R be sufficiently high, so that probability of ambiguous binning (i.e., when two or more source vectors are mapped into the same bin) is negligible. In this case, the mapping between source vectors and channel input vectors is one–to–one with high probability, and therefore, this is a joint source–channel code.

5. Systematic coding: letting the SI channel (from $\boldsymbol{u}$ to $\boldsymbol{v}$) in Fig. 1 be identical to the main channel (from $\boldsymbol{x}$ to $\boldsymbol{y}$), and then the SI channel may represent transmission of the systematic (uncoded) part of the code (see discussions on this point of view also in [18] and [29]).

In addition to studying the above described model, we also outline further extensions of our setting in several directions. These include:

1. Extending the scope from memoryless sources and channels to finite–state sources and finite–state channels. Here, the universal joint source-channel decoding metric is based on Lempel–Ziv (LZ) parsing, with the inspiration of [32].

2. Further extending the scope to arbitrary sources and channels, but allowing a given, limited class of reference decoding metrics. We propose a universal joint source–channel decoder with the property that, no matter what the underlying source and channel may be, this universal decoder is asymptotically as good as the best decoder in the class for these source and channel. This extends the recent study in [24], from pure channel coding to joint source–channel coding.

3. Generalizing to the model to separate encodings (source binning followed by channel coding) and joint decoding of two correlated sources (see Fig. 2). Here the universal decoder must handle several types of error events (more details in the sequel).

Finally, a few words are in order concerning the error exponent analysis. The ensemble of codes in our setting combines random binning (for the source coding part) and random coding (for the channel coding part), which is somewhat more involved than ordinary error exponent analyses that involve either one but not both. This requires quite a careful analysis, in two steps, where in the first, we take the average probability of error over the ensemble of random binning codes, for a given channel code, and at the second step, we average over the ensemble of channel codes. The latter employs the type class enumeration method [19, Chap. 6], which has already proved rather useful as a tool for obtaining exponentially tight random coding bounds in various contexts (see, e.g., [20], [21], [22], [23] for a sample), and this work is no exception in that respect, as the resulting error exponents are tight for the average code.

The remaining part of the paper is organized as follows. In Section 2, we establish notation conventions, formalize the model and the problem, and finally, review some preliminaries. Section 3 provides the main result along with some discussion. The proof of this result appears in Section 4, and finally, Section 5 is devoted for outlining the various extensions described above.

2 Notation Conventions, Problem Setting and Preliminaries

2.1 Notation Conventions

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted, respectively, by capital letters and the corresponding lower case letters, both in the bold face font. Their alphabets will be superscripted by their dimensions. For example, the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, the n–th order Cartesian power of $\mathcal{X}$, which is the alphabet of each component of this vector. Sources and channels will be denoted by the letter P, Q, or W, subscripted by the names of the relevant random variables/vectors and their conditionings, if applicable, following the standard notation conventions, e.g., Q_X, $P_{Y|X}$, and so on. When there is no room for ambiguity, these subscripts will be omitted. To avoid cumbersome notation, the various probability distributions will be denoted as above, no matter whether probabilities of single symbols or n–vectors are addressed. Thus, for example, $P_U(u)$ (or $P(u)$) will denote the probability of a single symbol $u \in \mathcal{U}$, whereas $P_U(\boldsymbol{u})$ (or $P(\boldsymbol{u})$) will stand for the probability of the n–vector $\boldsymbol{u} \in \mathcal{U}^n$. The probability of an event $\mathcal{E}$ will be denoted by $\Pr\{\mathcal{E}\}$, and the expectation operator with respect to (w.r.t.) a probability distribution P will be denoted by $\boldsymbol{E}\{\cdot\}$. The entropy of a generic distribution Q on $\mathcal{X}$ will be denoted by $H(Q)$. For two positive sequences a_n and b_n, the notation $a_n \doteq b_n$ will stand for equality in the exponential scale, that is, $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Accordingly, the notation $a_n \doteq 2^{-n\infty}$ means that a_n decays at a super–exponential rate (e.g., double–exponentially). Unless specified otherwise, logarithms and exponents, throughout this paper, should be understood to be taken to the base 2. The indicator function of an event $\mathcal{E}$ will be denoted by $\mathcal{I}\{E\}$. The notation $[x]_+$ will stand for $\max\{0, x\}$. The minimum between two reals, a and b, will frequently be denoted by $a \wedge b$. The cardinality of a finite set, say $\mathcal{A}$, will be denoted by $|\mathcal{A}|$.

The empirical distribution of a sequence $\boldsymbol{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}_{\boldsymbol{x}}$, is the vector of relative frequencies $\hat{P}_{\boldsymbol{x}}(x)$ of each symbol $x \in \mathcal{X}$ in $\boldsymbol{x}$. The type class of $\boldsymbol{x} \in \mathcal{X}^n$, denoted $\mathcal{T}(\boldsymbol{x})$, is the set of all vectors $\boldsymbol{x}'$ with $\hat{P}_{\boldsymbol{x}'} = \hat{P}_{\boldsymbol{x}}$. When we wish to emphasize the dependence of the type class on the empirical distribution, say Q, we will denote it by $\mathcal{T}(Q)$. Information measures associated with empirical distributions will be denoted with 'hats' and will be subscripted by the sequences from which they are induced. For example, the entropy associated with $\hat{P}_{\boldsymbol{x}}$, which is the empirical entropy of $\boldsymbol{x}$, will be denoted by $\hat{H}_{\boldsymbol{x}}(X)$. Again, the subscript will be omitted whenever it is clear from the context what sequence the empirical distribution was extracted from. Similar conventions will apply to the joint empirical distribution, the joint type class, the conditional empirical distributions and the conditional type classes associated with pairs of sequences of length n. Accordingly, $\hat{P}_{\boldsymbol{xy}}$ would be the joint empirical distribution of $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i)\}_{i=1}^n$, $\mathcal{T}(\boldsymbol{x}, \boldsymbol{y})$ or $\mathcal{T}(\hat{P}_{\boldsymbol{xy}})$ will denote the joint type class of $(\boldsymbol{x}, \boldsymbol{y})$, $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ will stand for the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$, $\hat{H}_{\boldsymbol{xy}}(X, Y)$ will designate the empirical joint entropy of $\boldsymbol{x}$ and $\boldsymbol{y}$, $\hat{H}_{\boldsymbol{xy}}(X|Y)$ will be

the empirical conditional entropy, $\hat{I}_{\boldsymbol{xy}}(X;Y)$ will denote empirical mutual information, and so on.

2.2 Problem Setting

Let $(\boldsymbol{U},\boldsymbol{V})=\{(U_t,V_t)\}_{t=1}^n$ be n independent copies of a pair of random variables, $(U,V)\sim P_{UV}$, taking on values in finite alphabets, $\mathcal{U}$ and $\mathcal{V}$, respectively. The vector $\boldsymbol{U}$ will designate the source vector to be encoded, whereas the vector $\boldsymbol{V}$ will serve as correlated SI, available to the decoder. Let W designate a DMC, with single–letter, input–output transition probabilities $W(y|x)$, $x\in\mathcal{X}$, $y\in\mathcal{Y}$, $\mathcal{X}$ and $\mathcal{Y}$ being finite input and output alphabets, respectively. When the channel is fed by an input vector $\boldsymbol{x}\in\mathcal{X}^n$, it produces[2] a channel output vector $\boldsymbol{y}\in\mathcal{Y}^n$, according to

$$W(\boldsymbol{y}|\boldsymbol{x})=\prod_{t=1}^n W(y_t|x_t). \tag{1}$$

Consider the communication system depicted in Fig. 1. When a given realization $\boldsymbol{u}=(u_1,\ldots,u_n)$, of the source vector $\boldsymbol{U}$, is fed into the system, it is encoded into one out of $M=2^{nR}$ bins, selected independently at random for every member of $\mathcal{U}^n$. Here, $R>0$ is referred to as the *binning rate*. The bin index $j=f(\boldsymbol{u})$ is mapped into a channel input vector $\boldsymbol{x}(j)\in\mathcal{X}^n$, which in turn is transmitted across the channel W. The various codewords $\{\boldsymbol{x}(j)\}_{j=1}^M$ are selected independently at random under the uniform distribution across a given type class $\mathcal{T}(Q)$, Q being a given probability distribution over $\mathcal{X}$. The randomly chosen codebook $\{\boldsymbol{x}(1),\boldsymbol{x}(2),\ldots,\boldsymbol{x}(M)\}$ will be denoted by $\mathcal{C}$. Both the channel encoder, $\mathcal{C}$, and the realization of the random binning source encoder, f, are revealed to the decoder as well. With a slight abuse of notation, we will sometimes denote $\boldsymbol{x}(j)=\boldsymbol{x}[f(\boldsymbol{u})]$ by $\boldsymbol{x}[\boldsymbol{u}]$. The optimal (MAP) decoder estimates $\boldsymbol{u}$, using the channel output $\boldsymbol{y}=(y_1,\ldots,y_n)$ and the SI vector $\boldsymbol{v}=(v_1,\ldots,v_n)$, according to:

$$\hat{\boldsymbol{u}}=\arg\max_{\boldsymbol{u}} P(\boldsymbol{u},\boldsymbol{v})W(\boldsymbol{y}|\boldsymbol{x}[\boldsymbol{u}]). \tag{2}$$

The average probability of error, $\overline{P_{\rm e}}$, is the probability of the event $\{\hat{\boldsymbol{U}}\neq\boldsymbol{U}\}$, where in addition to the randomness of $(\boldsymbol{U},\boldsymbol{V})$ and the channel output $\boldsymbol{Y}$, the randomness of the source binning code and the channel code are also taken into account. The random–binning/random–coding error exponent, associated with the optimal, MAP decoder, is defined as

$$E(R,Q)=\lim_{n\to\infty}\left[-\frac{\log\overline{P_{\rm e}}}{n}\right], \tag{3}$$

provided that the limit exists (a fact that will become evident from the analysis in the sequel).

Our first objective is to derive a single–letter expression for the exact random–binning/random–coding error exponent $E(R,Q)$. While the MAP decoder depends on the source P and the channel

[2] Without essential loss of generality, and similarly as in [2], we assume that the source and the channel operate at the same rate, so that while the source emits the n–vector $(\boldsymbol{U},\boldsymbol{V})$, the channel is used n times exactly, transforming $\boldsymbol{x}\in\mathcal{X}^n$ to $\boldsymbol{y}\in\mathcal{Y}^n$. The extension to the case where operation rates are different (bandwidth expansion factor different from 1) is straightforward but is avoided here, in the quest of keeping notation and expressions less cumbersome.

W, our second (and main) goal is to propose a universal decoder, independent of P and W, whose average error probability decays exponentially at the same rate, $E(R, Q)$. Finally, we aim to extend the scope beyond memoryless systems, as well as to the setting where the role of $\boldsymbol{V}$ is no longer merely to serve as SI at the decoder, but rather as another source vector, encoded similarly, but separately from $\boldsymbol{U}$ (see Fig. 2).

2.3 Preliminaries – the Joint Source–Channel Error Exponent

In [2], Csiszár derived upper and lower bounds on the reliability function of lossless joint source–channel coding (without SI). According to Csiszár's model, the source vector $\boldsymbol{u}$, drawn from a DMS P, is mapped directly into a channel input vector $\boldsymbol{x}[\boldsymbol{u}]$, which in turn is transmitted over a DMC W, whose output $\boldsymbol{y}$ is the input to a decoder, that produces the estimate $\hat{\boldsymbol{u}}$, and the goal was to find exponential error bounds to the best achievable probability of the error. Csiszár has shown in that paper that the reliability function, E_{jsc}, of lossless joint source–channel coding is upper bounded by

$$E^{\text{jsc}} \leq \min_{R}[E^{\text{s}}(R) + E^{\text{c}}(R)] \tag{4}$$

where[3] $E^{\text{s}}(R)$ is the source reliability function, given by

$$E^{\text{s}}(R) = \min_{\{P':\ H(P')\geq R\}} D(P'\|P), \tag{5}$$

$D(P'\|P)$ being the Kullback–Leibler divergence between P' and P, and $E^{\text{c}}(R)$ is the channel reliability function, for which there is a closed–form expression available only at rate zero (the zero–rate expurgated exponent) and at rates above the critical rate (the sphere–packing exponent). The lower bound in [2] is given by

$$E^{\text{jsc}} \geq \min_{R}[E^{\text{s}}(R) + E^{\text{c}}_{\text{r}}(R)], \tag{6}$$

where $E^{\text{c}}_{\text{r}}(R)$ is the random coding error exponent of the channel W. The upper and the lower bounds coincide (and hence provide the exact reliability function) whenever the minimizing R, of the upper bound (4), exceeds the critical rate of W. An expression equivalent to (6) is given by

$$E^{\text{jsc}} \geq \max_{Q} \min_{P',W'} \{D(P'\|P) + [I(X;Y') - H(U')]_+\} \tag{7}$$

where U' is an auxiliary random variable drawn by P', X is governed by Q, and Y' designates the output of an auxiliary channel W' fed by X. Here, the term $D(P'\|P)$ is parallel to the source coding exponent, $E^{\text{s}}(R)$, whereas the other term can be referred to the channel coding exponent $E^{\text{c}}_{\text{r}}(R)$ (see [2]). This is true since the minimization over P' in (7) can be carried out in two steps, where in the first, one minimizes over all $\{P'\}$ with a given entropy $H(U') = R$ (thus giving rise to $E^{\text{s}}(R)$ according to (5)), and then minimizes over R.

[3] Note that here, R is just an auxiliary variable over which the minimization of (4) is carried out, and it should not be confused with the binning rate R in the other parts of this paper.

In a nutshell, the idea behind the converse part in [2] is that each type class P', of source vectors, can be thought of as being mapped by the encoder into a separate channel subcode at rate $R = H(U')$, and then the probability of error is lower bounded by the contribution of the worst subcode. This is to say, that for the purpose of the lower bound, only decoding errors *within* each subcode are counted, whereas errors caused by confusing two source vectors that belong two different subcodes, are ignored. An interesting point, in this context, is that whenever the upper and the lower bound coincide (in the exponential scale), this means that confusions *within* the subcodes dominate the error probability, at least as far as error exponents are concerned, whereas errors of confusing codewords from different subcodes are essentially immaterial. We will witness the same phenomenon from a different perspective, in the sequel.

For the achievability part of [2], Csiszár analyzes the performance of a universal decoder, that is asymptotically equivalent to the following:

$$\hat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} [\hat{I}_{\boldsymbol{x}[\boldsymbol{u}]\boldsymbol{y}}(X;Y) - \hat{H}_{\boldsymbol{u}}(U)]. \tag{8}$$

As mentioned earlier, Csiszár refers to his decoder as the generalized MMI decoder.

3 Main Results

Our problem setting and results are more general than those of [2] from the following aspects: (i) we include side information $\boldsymbol{V}$, (ii) we include a cascade of random binning encoder and a channel encoder (separate source– and channel coding), and (iii) we compare the performance of the universal decoder to that of the MAP decoder (2) and show that they *always* (i.e., even when the random coding ensemble is not good enough to achieve the reliability function) have the same error exponent, whereas Csiszár compares the performance of (8) to the upper bound (4) and thus may conclude for asymptotic optimality of the decoder (together the encoder) only when the exact joint source–channel reliability function is known.

Concerning (ii), one may wonder what is the motivation for separate source– and channel coding, because joint source–channel coding is always at least as good. The answer is two–fold: (a) in some applications, system constraints dictate separate source– and channel coding, for example, when the two encodings are performed at different units/locations or when general engineering considerations (like modularity) dictate separation, and (b) the joint source–channel setting can always be obtained as a special case, by choosing the binning rate R sufficiently high, since then the binning encoder is a one–to–one mapping with an overwhelmingly high probability.

Our main result is given by the following theorem.

Theorem 1 *Consider the problem setting defined in Subsection 2.2.*

(a) The random–binning/random–coding error exponent of the MAP decoder is given by

$$E(R,Q) = \min_{P_{U'V'},W'} \{D(P_{U'V'}\|P_{UV}) + D(W'\|W|Q) + [R \wedge I(X;Y') - H(U'|V')]_+\} \tag{9}$$

where $D(W'\|W|Q)$ is Kullback–Leibler divergence between an auxiliary channel W' and the channel W, weighted by Q, $(U',V')\in\mathcal{U}\times\mathcal{V}$ are auxiliary random variables jointly distributed according to $P_{U'V'}$, and $Y'\in\mathcal{Y}$ is another auxiliary random variable that designates the output of channel W' when fed by $X\sim Q$.

(b) The universal decoders,

$$\hat{\boldsymbol{u}}=\arg\max_{\boldsymbol{u}}[\hat{I}_{\boldsymbol{x}[\boldsymbol{u}]\boldsymbol{y}}(X;Y)-\hat{H}_{\boldsymbol{u}\boldsymbol{v}}(U|V)] \tag{10}$$

and

$$\hat{\boldsymbol{u}}=\arg\max_{\boldsymbol{u}}[R\wedge\hat{I}_{\boldsymbol{x}[\boldsymbol{u}]\boldsymbol{y}}(X;Y)-\hat{H}_{\boldsymbol{u}\boldsymbol{v}}(U|V)], \tag{11}$$

both achieve $E(R,Q)$.

Decoder (10) is, of course, a natural extension of (8) to our setting. As for (11), while it offers no apparent advantage over (10), it is given here as an alternative decoder for future reference. It will turn out later that conceptually, (11) lends itself more naturally to the extension that deals with separate encodings and joint decoding of two correlated sources, where in the extended version of (11), it will not be obvious (at least not to the author) that the operator $R\wedge(\cdot)$ is neutral (i.e., an expression like $R\wedge x$ can be simply replaced by x, as is indeed suggested here by the equivalence between (10) and (11)). Another interesting point concerning (11), is that it appears more clearly as a joint extension of the MMI decoder of pure channel decoding and the ME decoder of pure source coding. When R dominates the term $R\wedge\hat{I}_{\boldsymbol{x}[\boldsymbol{u}]\boldsymbol{y}}(X;Y)$, the source coding component of the problem is more prominent and (11) is essentially equivalent to the ME decoder. Otherwise, it is essentially equivalent to (10).

As can be seen, $E(R,Q)$ is monotonically non–decreasing in R,[4] but when R is sufficiently large, the term $R\wedge I(X;Y')$ is dominated by $I(X;Y')$, which yields saturation of $E(R,Q)$ to the level of the joint source–channel random coding exponent (for a given Q), similarly as in (7), except that here, the entropy $H(U')$ is replaced by the conditional entropy $H(U'|V')$, due to the SI. For another extreme case, if the channel is clean, Q is uniform, and the channel alphabet is very large, then $\hat{I}(X;Y')=\hat{H}(X)=\log|\mathcal{X}|$ is large as well, and then $R\wedge\hat{I}(X';Y)$ is dominated by R. In this case, we recover the SW random binning error exponent (see, e.g., [31] and references therein).

4 Proof of Theorem 1

The outline of the proof is as follows. We begin by showing that $E(R,Q)$ is an upper bound on the error exponent associated with the MAP decoder, and then we show that both universal decoders (10) and (11) attain $E(R,Q)$. The combination of these two facts will prove both parts of Theorem 1 at the same time.

[4] This fact is not completely trivial, since an increase in R improves the source binning part, but one may expect that it harms the channel coding part. Nonetheless, as will become apparent in the sequel (see footnote 5), the combined effect of source binning and channel coding gives a non–decreasing exponent as a function of R.

As a first step, let the codebook $\mathcal{C}$, as well as the vectors $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{x}$ and $\boldsymbol{y}$ be given, and let $\overline{P}_e(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C})$ be the average error probability given $(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C})$, where the averaging is w.r.t. the ensemble of random binning codes. For a given $\boldsymbol{u}' \neq \boldsymbol{u}$, let us define the set

$$\mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}) = \mathcal{T}(Q) \bigcap \left\{\boldsymbol{x}' :\ W(\boldsymbol{y}|\boldsymbol{x}')P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})W(\boldsymbol{y}|\boldsymbol{x})\right\}. \tag{12}$$

The conditional error event, given $(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C})$, is given by

$$\begin{aligned}\mathcal{E}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C}) &= \bigcup_{\boldsymbol{u}'\neq\boldsymbol{u}} \left\{P(\boldsymbol{u}',\boldsymbol{v})W(\boldsymbol{y}|\boldsymbol{x}[\boldsymbol{u}']) \geq P(\boldsymbol{u},\boldsymbol{v})W(\boldsymbol{y}|\boldsymbol{x}[\boldsymbol{u}])\right\} \\ &\triangleq \bigcup_{\boldsymbol{u}'\neq\boldsymbol{u}} \mathcal{E}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C}) \end{aligned} \tag{13}$$

The probability of the pairwise error event, $\mathcal{E}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C})$ (again, w.r.t. the randomness of the bin assignment), is given by:

$$\overline{\Pr}\{\mathcal{E}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C})\} = 2^{-nR}\left|\mathcal{C}\bigcap\mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\right| + 2^{-nR}\mathcal{I}\{P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})\}, \tag{14}$$

where the first term accounts for errors pertaining to $\boldsymbol{u}'$–vectors that are assigned to bins other than $f(\boldsymbol{u})$, and the second term is associated with errors related to $\boldsymbol{u}'$–vectors that belong to the same bin. Now,

$$\begin{aligned} & \overline{P}_e(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C}) \\ = \ & \overline{\Pr}\{\mathcal{E}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C})\} \\ \doteq \ & \sum_{\{\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})\}} \overline{\Pr}\left\{\bigcup_{\boldsymbol{u}''\in\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})} \mathcal{E}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\mathcal{C})\}\right\} \\ \doteq \ & \sum_{\{\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})\}} \min\left\{1, |\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})|\cdot 2^{-nR}\left[\left|\mathcal{C}\bigcap\mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\right| + \mathcal{I}\{P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})\}\right]\right\} \end{aligned} \tag{15}$$

where the exponential tightness of the truncated union bound (for pairwise independent events) in the last expression is known from [30, Lemma A.2, p. 109] and it can also be readily deduced from de Caen's lower bound on the probability of a union of events [5]. The next step is to average over the randomness of $\mathcal{C}$ (except the codeword for the bin of the actual source vector $\boldsymbol{u}$, which is still given to be $\boldsymbol{x}$):

$$\begin{aligned}\overline{P}_e(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}) \ \doteq \ & \sum_{\{\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})\}} \boldsymbol{E}\left(\min\left\{1, |\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})|\cdot 2^{-nR}\left[\left|\mathcal{C}\bigcap\mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\right| + \right.\right.\right. \\ & \left.\left.\left. \mathcal{I}\{P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})\}\right]\right\}\right). \end{aligned} \tag{16}$$

Now, using the identity $\boldsymbol{E}\{Z\} = \int_0^\infty \Pr\{Z \geq t\}\mathrm{d}t$, which is valid for any non–negative random variable Z, we have

$$\boldsymbol{E}\left(\min\left\{1, |\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})|\cdot 2^{-nR}\left[\left|\mathcal{C}\bigcap\mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\right| + \mathcal{I}\{P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})\}\right]\right\}\right)$$

$$
\begin{aligned}
&= \int_0^1 \mathrm{d}t \cdot \Pr\left\{ |\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})| \cdot 2^{-nR} \left[\left| \mathcal{C} \bigcap \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}) \right| + \mathcal{I}\{P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})\} \right] \geq t \right\} \\
&= \int_0^1 \mathrm{d}t \cdot \Pr\left\{ \mathcal{I}\{P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})\} + \sum_i \mathcal{I}[\boldsymbol{X}(i) \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})] \geq \frac{t \cdot 2^{nR}}{|\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})|} \right\} \\
&\doteq n \ln 2 \cdot \int_0^\infty \mathrm{d}\theta \cdot 2^{-n\theta} \Pr\left\{ \mathcal{I}\{P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})\} + \right. \\
&\quad \left. \sum_i \mathcal{I}[\boldsymbol{X}(i) \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})] \geq 2^{n[R-\theta-\hat{H}(U'|V)]} \right\},
\end{aligned}
$$

where in the last passage, we have changed the integration variable from t to θ, according to the relation $t = 2^{-n\theta}$.

Consider first the case where $P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})$. Then, the integrand is given by

$$
\Pr\left\{ 1 + \sum_i \mathcal{I}[\boldsymbol{X}(i) \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})] \geq 2^{n[R-\theta-\hat{H}(U'|V)]} \right\} \tag{17}
$$

which is obviously equal to unity for all $\theta \geq [R - \hat{H}(U'|V)]_+$. Thus, the tail of the integral (17) is given by

$$
n \ln 2 \cdot \int_{[R-\hat{H}(U'|V)]_+}^\infty \mathrm{d}\theta \cdot 2^{-n\theta} = 2^{-n[R-\hat{H}(U'|V)]_+}. \tag{18}
$$

For $\theta < [R - \hat{H}(U'|V)]_+$, the unity term in (17) can be safely neglected, and

$$
\Pr\left\{ \sum_i \mathcal{I}[\boldsymbol{X}(i) \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})] \geq 2^{n[R-\theta-\hat{H}(U'|V)]} \right\} \tag{19}
$$

is the probability of a large deviations event associated with a binomial random variable with 2^{nR} trials and probability of success of the exponential order of 2^{-nJ}, with J being defined as

$$
J \triangleq \min\left\{ \hat{I}(X';Y) : \ \hat{P}_{\boldsymbol{x}'\boldsymbol{y}} \text{ is such that } \boldsymbol{x}' \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}) \right\}, \tag{20}
$$

where $\hat{I}(X';Y)$ is shorthand notation for $\hat{I}_{\boldsymbol{x}'\boldsymbol{y}}(X;Y)$ and where it should be kept in mind that J depends on $\hat{P}_{\boldsymbol{u}\boldsymbol{v}}$, $\hat{P}_{\boldsymbol{u}'\boldsymbol{v}}$, and $\hat{P}_{\boldsymbol{x}\boldsymbol{y}}$. According to [19, Chap. 6], the large deviations behavior is as follows:

$$
\begin{aligned}
&\Pr\left\{ \sum_i \mathcal{I}[\boldsymbol{X}(i) \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})] \geq 2^{n[R-\theta-\hat{H}(U'|V)]} \right\} \\
&\doteq \begin{cases} 2^{-n[J-R]_+} & R - \theta - \hat{H}(U'|V) \leq [R-J]_+ \\ 2^{-n\infty} & R - \theta - \hat{H}(U'|V) > [R-J]_+ \end{cases} \\
&= \begin{cases} 2^{-n[J-R]_+} & \theta \geq [R - \hat{H}(U'|V) - [R-J]_+]_+ \\ 2^{-n\infty} & \text{elsewhere} \end{cases}
\end{aligned} \tag{21}
$$

Thus, the other contribution to (17) is given by

$$\begin{aligned}
& n\ln 2 \cdot \int_{[R-\hat{H}(U'|V)-[R-J]_+]_+}^{[R-\hat{H}(U'|V)]_+} \mathrm{d}\theta \cdot 2^{-n\theta} \cdot 2^{-n[J-R]_+} \\
\dot{=} \;& \exp_2\{-n([R-\hat{H}(U'|V)-[R-J]_+]_+ + [J-R]_+)\} \\
= \;& \exp_2\{-n([R\wedge J-\hat{H}(U'|V)]_+ + [J-R]_+)\} \\
= \;& \exp_2\{-n([R\wedge J-\hat{H}(U'|V)]_+ - R\wedge J + R\wedge J + [J-R]_+)\} \\
= \;& \exp_2\{-n[-R\wedge J\wedge \hat{H}(U'|V)+J]\} \\
= \;& \exp_2\{-n[J-R\wedge \hat{H}(U'|V)]_+\}, \qquad (22)
\end{aligned}$$

where we have repeatedly used the identity $a-[a-b]_+ = a\wedge b$. Thus, the total conditional error exponent, for the case $P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})$, is given by

$$\begin{aligned}
& \min\{[R-\hat{H}(U'|V)]_+, [J-R\wedge\hat{H}(U'|V)]_+\} \\
= \;& \min\{R-R\wedge\hat{H}(U'|V), J-R\wedge J\wedge\hat{H}(U'|V)\} \\
= \;& \left[R\wedge J-\hat{H}(U'|V)\right]_+, \qquad (23)
\end{aligned}$$

where the last line follows from the following consideration:[5] If $\hat{H}(U'|V) > J$, then all three expressions obviously vanish and then the equality is trivially met. Otherwise, $\hat{H}(U'|V) \leq J$ implies that the term $R\wedge\hat{H}(U'|V)$, in the second line, can safely be replaced by $R\wedge J\wedge\hat{H}(U'|V)$, which makes the second line identical to $R\wedge J - R\wedge J\wedge\hat{H}(U'|V) = \left[R\wedge J-\hat{H}(U'|V)\right]_+$. In the case, $P(\boldsymbol{u}',\boldsymbol{v}) < P(\boldsymbol{u},\boldsymbol{v})$, the conditional error exponent is just $[J-R\wedge\hat{H}(U'|V)]_+$.

Let $E_0(\hat{P}_{\boldsymbol{uv}},\hat{P}_{\boldsymbol{u'v}},\hat{P}_{\boldsymbol{xy}})$ denote the overall conditional error exponent given $(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})$, i.e.,

$$E_0(\hat{P}_{\boldsymbol{uv}},\hat{P}_{\boldsymbol{u'v}},\hat{P}_{\boldsymbol{xy}}) \;=\; \begin{cases} \left[R\wedge J-\hat{H}(U'|V)\right]_+ & P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v}) \\ \left[J-R\wedge\hat{H}(U'|V)\right]_+ & \text{otherwise} \end{cases} \qquad (24)$$

Then, we finally have:

$$E(R,Q) = \lim_{n\to\infty} \min_{\hat{P}_{\boldsymbol{uv}},\hat{P}_{\boldsymbol{xy}}} [D(\hat{P}_{\boldsymbol{uv}}\|P_{UV}) + D(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}\|W|Q) + E_1(\hat{P}_{\boldsymbol{uv}},\hat{P}_{\boldsymbol{xy}})], \qquad (25)$$

where

$$E_1(\hat{P}_{\boldsymbol{uv}},\hat{P}_{\boldsymbol{xy}}) = \min_{\hat{P}_{\boldsymbol{u'v}}} E_0(\hat{P}_{\boldsymbol{uv}},\hat{P}_{\boldsymbol{u'v}},\hat{P}_{\boldsymbol{xy}}). \qquad (26)$$

[5] The first line of (23) corresponds to the worst between the source coding exponent, $[R-\hat{H}(U'|V)]_+$, and the channel coding exponent, $[J-R\wedge\hat{H}(U'|V)]_+$, which is to be expected in separate source– and channel coding. While the former is non-decreasing in R, the latter is non-increasing. From the last line of (23), we learn that the overall exponent is non-decreasing in R.

An obvious upper bound[6] is obtained by

$$\begin{aligned} E_1(\hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{xy}}) &\leq E_0(\hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{xy}}) \\ &\leq \Big[R \wedge \hat{I}(X;Y) - \hat{H}(U|V)\Big]_+ \\ &\stackrel{\Delta}{=} E_1^*(\hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{xy}}), \end{aligned} \tag{27}$$

where we have used the fact that $\boldsymbol{x} \in \mathcal{A}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$ and so, for $\hat{P}_{\boldsymbol{u'v}} = \hat{P}_{\boldsymbol{uv}}$, one has $J \leq \hat{I}(X;Y)$. Thus,

$$\begin{aligned} E(R,C) &\leq \min_{P_{U'V'}, W'} \big[D(P_{U'V'} \| P_{UV}) + D(W' \| W | Q) + E_1^*(P_{U'V'}, Q \times W')\big] \\ &= \min_{P_{U'V'}, W'} \Big\{D(P_{U'V'} \| P_{UV}) + D(W' \| W | Q) + \big[R \wedge I(X;Y') - H(U'|V')\big]_+\Big\} \\ &\stackrel{\Delta}{=} E_U(R,Q). \end{aligned} \tag{28}$$

We next argue that the universal decoder (10) achieves $E_U(R,Q)$ and hence $E(R,Q) = E_U(R,Q)$. To see why this is true, one repeats exactly the same derivation, with the following two simple modifications:

1. $\mathcal{A}(\boldsymbol{u}, \boldsymbol{u}', \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$ is replaced by

$$\tilde{\mathcal{A}}(\boldsymbol{u}, \boldsymbol{u}', \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) = \mathcal{T}(Q) \cap \{\boldsymbol{x}' : \ \hat{I}(X';Y) - \hat{H}(U'|V) \geq \hat{I}(X;Y) - \hat{H}(U|V)\} \tag{29}$$

   and accordingly, J is replaced by

$$\tilde{J} = \min\{\hat{I}(X';Y) : \ \hat{I}(X';Y) - \hat{H}(U'|V) \geq \hat{I}(X;Y) - \hat{H}(U|V)\}. \tag{30}$$

2. The indicator function $\mathcal{I}\{P(\boldsymbol{u}', \boldsymbol{v}) \geq P(\boldsymbol{u}, \boldsymbol{v})\}$ is replaced by $\mathcal{I}\{\hat{H}(U'|V) \leq \hat{H}(U|V)\}$.

The result is then similar except that $E_0(\hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{u'v}}, \hat{P}_{\boldsymbol{xy}})$ is replaced by

$$\tilde{E}_0(\hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{u'v}}, \hat{P}_{\boldsymbol{xy}}) = \begin{cases} [R \wedge \tilde{J} - \hat{H}(U'|V)]_+ & \hat{H}(U'|V) \leq \hat{H}(U|V) \\ \Big[\tilde{J} - R \wedge \hat{H}(U'|V)\Big]_+ & \text{otherwise} \end{cases} \tag{31}$$

Now, observe that for the first line of (31),

$$\begin{aligned} & R \wedge \tilde{J} - \hat{H}(U'|V) \\ \geq \ & R \wedge [\hat{I}(X;Y) - \hat{H}(U|V) + \hat{H}(U'|V)] - \hat{H}(U'|V) \end{aligned}$$

[6] We are upper bounding the minimum of E_1 over $\{\hat{P}_{\boldsymbol{u'v}}\}$ by the value of E_0 where $\hat{P}_{\boldsymbol{u'v}} = \hat{P}_{\boldsymbol{uv}}$, and will shortly see that this bound is actually tight. This means that the error exponent is dominated by erroneous vectors $\{\boldsymbol{u}'\}$ that are within the same conditional type (given $\boldsymbol{v}$) as the correct source vector $\boldsymbol{u}$. This is coherent with the observation discussed in Subsection 2.3, that errors within the subcode pertaining to the same type class dominate the error exponent.

$$
\begin{aligned}
&= [R - \hat{H}(U'|V)] \wedge [\hat{I}(X;Y) - \hat{H}(U|V)] \\
&\geq [R - \hat{H}(U|V)] \wedge [\hat{I}(X;Y) - \hat{H}(U|V)] \quad \text{since } \hat{H}(U'|V) \leq \hat{H}(U|V) \\
&= R \wedge \hat{I}(X;Y) - \hat{H}(U|V), \qquad (32)
\end{aligned}
$$

where the first line follows from the definition of $\tilde{J}$. As for the second line,

$$
\begin{aligned}
\tilde{J} - R \wedge \hat{H}(U'|V) &\geq \hat{I}(X;Y) - \hat{H}(U|V) + \hat{H}(U'|V) - R \wedge \hat{H}(U'|V) \\
&= \hat{I}(X;Y) - \hat{H}(U|V) + [\hat{H}(U'|V) - R]_+ \\
&\geq \hat{I}(X;Y) - \hat{H}(U|V) + [\hat{H}(U|V) - R]_+ \quad \text{since } \hat{H}(U'|V) > \hat{H}(U|V) \\
&= \hat{I}(X;Y) - R \wedge \hat{H}(U|V) \\
&\geq R \wedge \hat{I}(X;Y) - \hat{H}(U|V). \qquad (33)
\end{aligned}
$$

We conclude then that, no matter whether $\hat{H}(U'|V) \leq \hat{H}(U|V)$ or $\hat{H}(U'|V) > \hat{H}(U|V)$, we always have:

$$
\tilde{E}_0(\hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{u'v}}, \hat{P}_{\boldsymbol{xy}}) \geq \Big[R \wedge \hat{I}(X;Y)\} - \hat{H}(U|V)\Big]_+ = E_1^*(\hat{P}_{\boldsymbol{uv}}, \hat{P}_{\boldsymbol{xy}}), \qquad (34)
$$

and so, the overall exponent $E_U(R)$ is achieved by (10).

As for the alternative universal decoding metric (11), the derivation is, once again, the very same, with the pairwise error event $\mathcal{A}(\boldsymbol{u}, \boldsymbol{u'}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$ and the variable $\tilde{J}$ redefined accordingly as the minimum of $\hat{I}(X';Y)$ s.t. $R \wedge \hat{I}(X';Y) - \hat{H}(U'|V) \geq R \wedge \hat{I}(X;Y) - \hat{H}(U|V)$, and then the last two inequalities are modified as follows: Instead of (32), we have

$$
\begin{aligned}
& R \wedge \tilde{J} - \hat{H}(U'|V) \\
= \;& R \wedge R \wedge \hat{I}(X';Y) - \hat{H}(U'|V) \\
\geq \;& R \wedge [R \wedge \hat{I}(X;Y) - \hat{H}(U|V) + \hat{H}(U'|V)] - \hat{H}(U'|V) \\
= \;& [R - \hat{H}(U'|V)] \wedge \{[R \wedge \hat{I}(X;Y)] - \hat{H}(U|V)\} \\
\geq \;& [R - \hat{H}(U|V)] \wedge \{[R \wedge \hat{I}(X;Y)] - \hat{H}(U|V)\} \quad \text{since } \hat{H}(U'|V) \leq \hat{H}(U|V) \\
= \;& R \wedge \hat{I}(X;Y) - \hat{H}(U|V). \qquad (35)
\end{aligned}
$$

and instead of (33):

$$
\begin{aligned}
\tilde{J} - R \wedge \hat{H}(U'|V) &\geq R \wedge \tilde{J} - R \wedge \hat{H}(U'|V) \\
&\geq R \wedge \hat{I}(X;Y) - \hat{H}(U|V) + \hat{H}(U'|V) - R \wedge \hat{H}(U'|V) \\
&= R \wedge \hat{I}(X;Y) - \hat{H}(U|V) + [\hat{H}(U'|V) - R]_+ \\
&\geq R \wedge \hat{I}(X;Y) - \hat{H}(U|V) + [\hat{H}(U|V) - R]_+ \quad \text{since } \hat{H}(U'|V) > \hat{H}(U|V) \\
&= R \wedge \hat{I}(X;Y) - R \wedge \hat{H}(U|V) \\
&\geq R \wedge \hat{I}(X;Y) - \hat{H}(U|V). \qquad (36)
\end{aligned}
$$

This completes the proof of Theorem 1.

5 Extensions

As mentioned in the Introduction, in this section, we outline extensions of the above results in several directions, including: (i) finite–state sources and channels with LZ universal decoding metrics, (ii) arbitrary sources and channels with universal decoding w.r.t. a given class of metric decoders, and (iii) separate source–channel encodings and joint universal decoding of correlated source.

While in (i) and (ii) we no longer expect to have single–letter formulas for the error exponent, we will still be able to propose asymptotically optimum universal decoding metrics in the error exponent sense. Since the analysis techniques are very similar to these of the proof of Theorem 1, we will only describe the differences and the modifications relative to that proof.

5.1 Finite–State Sources/Channels and a Universal LZ Decoding Metric

In [32], Ziv considered the class of finite–state channels and proposed a universal decoding metric that is based on conditional LZ parsing. Here, we discuss a similar model with a suitable extension of Ziv's decoding metric in the spirit of the generalized MMI decoder.

Consider a sequence of pairs of random variables $\{(U_i, V_i)\}_{i=1}^n$, drawn from a finite–alphabet, finite–state source, defined according to

$$P(\boldsymbol{u}, \boldsymbol{v}) = \prod_{t=1}^{n} P(u_t, v_t|s_t) \tag{37}$$

where s_t is the joint state of the two sources at time t, which evolves according to

$$s_t = g(s_{t-1}, u_{t-1}, v_{t-1}), \tag{38}$$

with $g : \mathcal{S} \times \mathcal{U} \times \mathcal{V} \to \mathcal{S}$ being the source next–state function, and $\mathcal{S}$ being a finite set of states. The initial state, s_1, is assumed to be an arbitrary fixed member of $\mathcal{S}$. By the same token, the channel is also assumed to be finite–state (as in [32]), i.e.,

$$W(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{n} W(y_t|x_t, z_t), \qquad z_t = h(z_{t-1}, x_{t-1}, y_{t-1}), \tag{39}$$

where z_t is the channel state at time t, taking on values in a finite set $\mathcal{Z}$ and $h : \mathcal{Z} \times \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is the channel next–state function. Once again, the initial state, z_1, is an arbitrary member of $\mathcal{Z}$.

The remaining details of the communication system are the same as described in Subsection 2.2, with the exception that the random coding distribution, now denoted by $Q(\boldsymbol{x})$, is allowed here to be more general than a uniform distribution across a type class (or the uniform distribution across $\mathcal{X}^n$, as assumed in [32]). Similarly as in [24], we assume that Q may be any exchangeable probability distribution (i.e., $\boldsymbol{x}'$ is a permutation of $\boldsymbol{x}$ implies $Q(\boldsymbol{x}') = Q(\boldsymbol{x})$), and that, moreover, if the state variable z_t includes a component, say, σ_t, that is fed merely by $\{x_t\}$ (but not $\{y_t\}$), then it is enough that Q would be invariant within conditional types of $\boldsymbol{x}$ given $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$.

Let $\hat{H}_{\rm LZ}(\boldsymbol{x}|\boldsymbol{y})$ denote the normalized conditional LZ compressibility of $\boldsymbol{x}$ given $\boldsymbol{y}$, as defined in [32, eq. (20)] (and denoted by $u(\boldsymbol{x},\boldsymbol{y})$ therein).[7] Next define

$$\hat{I}_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) = -\frac{\log Q(\boldsymbol{x})}{n} - \hat{H}_{\rm LZ}(\boldsymbol{x}|\boldsymbol{y}), \tag{40}$$

and finally, define the universal decoder

$$\tilde{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} \left[\hat{I}_{\rm LZ}(\boldsymbol{x}[\boldsymbol{u}];\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v})\right]. \tag{41}$$

Note that the first term on the r.h.s. of (40) plays a role analogous to that of the unconditional empirical entropy, $\hat{H}_{\boldsymbol{x}}(X)$, of the memoryless case (and indeed, at least for the uniform distribution over a type class, as assumed in the previous sections, it is asymptotically equivalent), and so, the difference in (40) makes sense as an extension of the empirical mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$.

As in Theorem 2, part b (and as an extension to [32]), we now argue that the universal decoder (41) achieves an average error probability that is, within a sub-exponential function of n, the same as the average error probability of the MAP decoder for the given source (37) and channel (39). Thus, whenever the latter has an exponentially decaying average error probability, then so does (41), and with the same exponential rate. The proof of this claim goes largely in the footsteps of the proof of Theorem 1. Below we outline the main steps, highlighting the main modifications required.

1. The conditional type class of $\boldsymbol{u}$ given $\boldsymbol{v}$, $\mathcal{T}(\boldsymbol{u}|\boldsymbol{v})$, is redefined as the set $\{\boldsymbol{u}' :\ P(\boldsymbol{u}',\boldsymbol{v}) = P(\boldsymbol{u},\boldsymbol{v})\}$, where P is given as in (37).[8] Obviously, for every given $\boldsymbol{v}$, the various 'types' $\{\mathcal{T}(\boldsymbol{u}|\boldsymbol{v})\}$ are equivalence classes, and hence form a partition of $\mathcal{U}^n$. One important property that would essential for the proof is that the number $K_n(\boldsymbol{v})$ of distinct types, $\{\mathcal{T}(\boldsymbol{u}|\boldsymbol{v})\}$, under this definition, for a given $\boldsymbol{v}$, grows sub–exponentially in n (just like in the case of ordinary types). This guarantees that the probability of a union of events, over $\{\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})\}$, is of the same exponential order as the maximum term, as was the case in the proof of Theorem 1. Interestingly, this can easily be proved using the theory of LZ data compression:

$$K_n(\boldsymbol{v}) = \sum_{\boldsymbol{u}\in\mathcal{U}^n} \frac{1}{|\mathcal{T}(\boldsymbol{u}|\boldsymbol{v})|} \le \sum_{\boldsymbol{u}\in\mathcal{U}^n} 2^{-n\hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v})+no(n)} \le 2^{no(n)}, \tag{42}$$

where $o(n)$ stands for a vanishing term (uniformly in both $\boldsymbol{u}$ and $\boldsymbol{v}$), the first inequality is by [32, Lemma 1, p. 459][9] and the second inequality is due to the fact that $n\hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v})$ is (within negligible

[7] This means that $n\hat{H}_{\rm LZ}(\boldsymbol{x}|\boldsymbol{y})$ is the length of the conditional Lempel–Ziv code for $\boldsymbol{x}$, where $\boldsymbol{y}$ serves as SI available to both the encoder and decoder, which is based on joint incremental parsing of the sequence pair $(\boldsymbol{x},\boldsymbol{y})$ (see also [17]). Here, we are deliberately using a somewhat different notation than the usual, which hopefully makes the analogy to the memoryless case self–evident.

[8] Note that the requirement $P(\boldsymbol{u}',\boldsymbol{v}) = P(\boldsymbol{u},\boldsymbol{v})$ is imposed here only for the given P, not even for every finite–state source in the class.

[9] Not to be confused with the lemma on page 456 of [32], which is also referred to as Lemma 1.

terms) a legitimate length function for lossless compression of $\boldsymbol{u}$ (with SI $\boldsymbol{v}$) (see [32, Lemma 2] and [17]) and hence must satisfy the Kraft inequality for every given $\boldsymbol{v}$.

2. The quantity $\hat{H}(U'|V)$, in the proof of Theorem 1, is replaced by $\frac{1}{n}\log|\mathcal{T}(\boldsymbol{u}'|\boldsymbol{v})|$ with the above modified definition of the conditional type.

3. The definition of J is changed to

$$J = \min\left\{-\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}'|\boldsymbol{y})]:\ \boldsymbol{x}' \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\right\}, \tag{43}$$

where $\mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})$ is the pairwise error event pertaining to the MAP decoder (for the lower bound) or to the universal decoder (41) (for the upper bound). By our assumptions, Q assigns the same probability to all members of $\mathcal{T}(\boldsymbol{x}'|\boldsymbol{y})$, thus

$$\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}'|\boldsymbol{y})] = \frac{\log Q(\boldsymbol{x}')}{n} + \frac{\log|\mathcal{T}(\boldsymbol{x}'|\boldsymbol{y})|}{n}. \tag{44}$$

4. Using the above, and following the same steps as in the proof of Theorem 1, the conditional average error probability, given $(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})$, associated with the MAP decoder, can be shown to be lower bounded by an expression of the exponential order of $\exp\{-nE_0(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}$, where

$$\begin{aligned}
E_0(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}) &\leq \left[\min\left\{R, -\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]\right\} - \frac{1}{n}\log|\mathcal{T}(\boldsymbol{u}|\boldsymbol{v})|\right]_+ \\
&= \left[\min\left\{R, -\frac{1}{n}\log Q(\boldsymbol{x}) - \frac{1}{n}\log|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|\right\} - \frac{1}{n}\log|\mathcal{T}(\boldsymbol{u}|\boldsymbol{v})|\right]_+ \\
&\leq \left[\min\left\{R, -\frac{1}{n}\log Q(\boldsymbol{x}) - \hat{H}_{\mathrm{LZ}}(\boldsymbol{x}|\boldsymbol{y})\right\} - \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}|\boldsymbol{v})\right]_+ + o(n) \\
&= \left[R \wedge \hat{I}_{\mathrm{LZ}}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}|\boldsymbol{v})\right]_+ + o(n) \\
&\stackrel{\Delta}{=} E_1^*(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}),
\end{aligned} \tag{45}$$

and where we have used twice Lemma 1 of [32, p. 459] and the fact that $\boldsymbol{x} \in \mathcal{A}(\boldsymbol{u},\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})$ and so, for $P(\boldsymbol{u}',\boldsymbol{v}) = P(\boldsymbol{u},\boldsymbol{v})$, one has $J \leq -\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]$.

5. For the upper bound on the error probability of (41), $\mathcal{A}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})$ is replaced by

$$\tilde{\mathcal{A}}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}) = \{\boldsymbol{x}':\ I_{\mathrm{LZ}}(\boldsymbol{x}';\boldsymbol{y}) - \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}'|\boldsymbol{v}) \geq I_{\mathrm{LZ}}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}|\boldsymbol{v})\} \tag{46}$$

and accordingly, J is replaced by

$$\begin{aligned}
\tilde{J} &= \min\{I_{\mathrm{LZ}}(\boldsymbol{x}';\boldsymbol{y}):\ I_{\mathrm{LZ}}(\boldsymbol{x}';\boldsymbol{y}) - \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}'|\boldsymbol{v}) \geq I_{\mathrm{LZ}}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}|\boldsymbol{v})\} \\
&= I_{\mathrm{LZ}}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}'|\boldsymbol{v}).
\end{aligned} \tag{47}$$

6. The indicator function $\mathcal{I}[P(\boldsymbol{u}',\boldsymbol{v}) \geq P(\boldsymbol{u},\boldsymbol{v})]$ is replaced by $\mathcal{I}[\hat{H}_{\mathrm{LZ}}(\boldsymbol{u}'|\boldsymbol{v}) \leq \hat{H}_{\mathrm{LZ}}(\boldsymbol{u}|\boldsymbol{v})]$.

7. For the error probability analysis of the universal decoder (41), the union over erroneous source vectors $\{\boldsymbol{u}'\}$ is partitioned into (a sub–exponential number of) 'types' of the form

$$\mathcal{T}_\ell(\boldsymbol{u}'|\boldsymbol{v}) = \{\tilde{\boldsymbol{u}}:\ P(\tilde{\boldsymbol{u}},\boldsymbol{v}) = P(\boldsymbol{u}',\boldsymbol{v}),\ n\hat{H}_{\rm LZ}(\tilde{\boldsymbol{u}}|\boldsymbol{v}) = \ell\}, \tag{48}$$

for $\ell = 1, 2, \ldots$, and one uses the fact that $|\mathcal{T}_\ell(\boldsymbol{u}'|\boldsymbol{v})| \leq 2^\ell$, as $n\hat{H}_{\rm LZ}(\cdot|\boldsymbol{v})$ is a length function of a lossless data compression algorithm.

8. It is observed that $\sum_i \mathcal{I}[\boldsymbol{X}(i) \in \tilde{\mathcal{A}}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})]$ is a binomial random variables with 2^{nR} trials and probability of success of the exponential order of $2^{-n\tilde{J}}$. To see why the latter is true, consider the following:

$$\begin{aligned}
& Q\left\{\boldsymbol{X}' \in \tilde{\mathcal{A}}(\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\right\} \\
= \ & Q\left\{I_{\rm LZ}(\boldsymbol{X}';\boldsymbol{y}) \geq I_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\rm LZ}(\boldsymbol{u}'|\boldsymbol{v})\right\} \\
= \ & \sum_{\{\boldsymbol{x}':\ I_{\rm LZ}(\boldsymbol{x}';\boldsymbol{y}) \geq I_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\rm LZ}(\boldsymbol{u}'|\boldsymbol{v})\}} Q(\boldsymbol{x}') \\
= \ & \sum_{\{\boldsymbol{x}':\ I_{\rm LZ}(\boldsymbol{x}';\boldsymbol{y}) \geq I_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\rm LZ}(\boldsymbol{u}'|\boldsymbol{v})\}} Q(\boldsymbol{x}')2^{n\hat{H}_{\rm LZ}(\boldsymbol{x}'|\boldsymbol{y})} \cdot 2^{-n\hat{H}_{\rm LZ}(\boldsymbol{x}'|\boldsymbol{y})} \\
= \ & \sum_{\{\boldsymbol{x}':\ I_{\rm LZ}(\boldsymbol{x}';\boldsymbol{y}) \geq I_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\rm LZ}(\boldsymbol{u}'|\boldsymbol{v})\}} \exp_2\{-nI_{\rm LZ}(\boldsymbol{x}';\boldsymbol{y})\} \cdot 2^{-n\hat{H}_{\rm LZ}(\boldsymbol{x}'|\boldsymbol{y})} \\
\leq \ & \sum_{\boldsymbol{x}' \in \mathcal{X}^n} \exp_2\{-n[I_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\rm LZ}(\boldsymbol{u}'|\boldsymbol{v})]\} \cdot 2^{-n\hat{H}_{\rm LZ}(\boldsymbol{x}'|\boldsymbol{y})} \\
\leq \ & \exp_2\{-n[I_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\rm LZ}(\boldsymbol{u}'|\boldsymbol{v})]\} \sum_{\boldsymbol{x}' \in \mathcal{X}^n} 2^{-n\hat{H}_{\rm LZ}(\boldsymbol{x}'|\boldsymbol{y})} \\
\leq \ & \exp_2\{-n[I_{\rm LZ}(\boldsymbol{x};\boldsymbol{y}) - \hat{H}_{\rm LZ}(\boldsymbol{u}|\boldsymbol{v}) + \hat{H}_{\rm LZ}(\boldsymbol{u}'|\boldsymbol{v})] + o(n)\},
\end{aligned} \tag{49}$$

where in the last step, we have used again Kraft's inequality.

9. Using exactly the same method as in the proof of Theorem 1, one can show that that conditional error probability of the universal decoder (41) is upper bounded by an expression whose exponential order is lower bounded by $E_1^*(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})$.

It should be noted that these results continue to apply for *arbitrary* sources and channels (even deterministic ones), where the assertion would be that the decoder (41) competes favorably (in the error exponent sense) relative to any decoding metric of the form

$$\sum_{t=1}^n m_{\rm s}(u_t, v_t, s_t) + \sum_{t=1}^n m_{\rm c}(x_t, y_t, z_t), \tag{50}$$

where s_t and z_t evolve according to next–state functions h and g, as defined above. This follows from the observation that the assumption on underlying finite–state sources and finite–state channels was actually used merely in the assumed structure of the MAP decoding metric, with which decoder

(41) competes. The fact that the overall probability of error is eventually averaged over all source vectors and channel noise realizations pertaining to finite–state probability distributions, was not really used here, since we compared the conditional error probabilities given $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$. The same observation has been exploited also in [24] for universal pure channel coding, and it will be further developed in the next subsection.

5.2 Arbitrary Sources and Channels With a Given Class of Metric Decoders

In [24], the following setting of universal channel decoding was studied: Given a random coding distribution Q, for independent random selection of 2^{nR} codewords $\{\boldsymbol{x}_i\}$, and given a limited class of reference decoders, defined by a family of decoding metrics $\{m_\theta(\boldsymbol{x}, \boldsymbol{y}),\ \theta \in \Theta\}$ (θ being an index or a parameter), find a decoding metric that is universal in the sense of achieving an average error probability that is, within a sub–exponential function of n, as good as the best decoder in the class, no matter what the underlying channel, $W(\boldsymbol{y}|\boldsymbol{x})$, may be. The following decoder was shown in [24] to possess this property under a certain condition that will be specified shortly: estimate the message i as the one that minimizes $u(\boldsymbol{x}_i, \boldsymbol{y}) = \log Q[\mathcal{T}(\boldsymbol{x}_i|\boldsymbol{y})]$, where $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ designates a notion of a "type" induced by the family of decoding metrics (rather than by channels), namely,

$$\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) = \left\{\boldsymbol{x}' :\ m_\theta(\boldsymbol{x}', \boldsymbol{y}) = m_\theta(\boldsymbol{x}, \boldsymbol{y})\ \forall\ \theta \in \Theta\right\}. \tag{51}$$

As $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ are equivalence classes, they form a partition of $\mathcal{X}^n$ for every given $\boldsymbol{y}$. The condition required for the universality of this decoding metric is that the number of distinct 'types' $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ would grow sub–exponentially with n.

A similar approach can be taken in the present problem setting. Given a family of decoding metrics of the form[10]

$$m_\theta(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) = m_{\mathrm{s},\theta}(\boldsymbol{u}, \boldsymbol{v}) + m_{\mathrm{c},\theta}(\boldsymbol{x}, \boldsymbol{y}), \qquad \theta \in \Theta, \tag{52}$$

let us define

$$\mathcal{T}_{\mathrm{s}}(\boldsymbol{u}|\boldsymbol{v}) = \{\boldsymbol{u}' :\ m_{\mathrm{s},\theta}(\boldsymbol{u}', \boldsymbol{v}) = m_{\mathrm{s},\theta}(\boldsymbol{u}, \boldsymbol{v})\ \forall\ \theta \in \Theta\} \tag{53}$$

$$\mathcal{T}_{\mathrm{s}}(\boldsymbol{x}|\boldsymbol{y}) = \{\boldsymbol{x}' :\ m_{\mathrm{c},\theta}(\boldsymbol{x}', \boldsymbol{y}) = m_{\mathrm{c},\theta}(\boldsymbol{x}, \boldsymbol{y})\ \forall\ \theta \in \Theta\}, \tag{54}$$

and assume, as before, that the numbers of distinct 'types', $\{\mathcal{T}_{\mathrm{s}}(\boldsymbol{u}|\boldsymbol{v})\}$ and $\{\mathcal{T}_{\mathrm{c}}(\boldsymbol{x}|\boldsymbol{y})\}$, both grow sub–exponentially with n. Then, the universal decoder

$$\hat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \left\{\log |\mathcal{T}_{\mathrm{s}}(\boldsymbol{u}|\boldsymbol{v})| + \log Q[\mathcal{T}_{\mathrm{c}}(\boldsymbol{x}[\boldsymbol{u}]|\boldsymbol{y})]\right\} \tag{55}$$

competes favorably with all metrics in the above family, no matter what the underlying source and the underlying channel may be. The proof combines the ideas of the proof of Theorem 1 above with those of [24], with the proper adjustments, of course, but it is otherwise straightforward. Here,

[10] This additive structure can be justified by the fact that the MAP decoding metric is also additive, as it maximizes $\log P(\boldsymbol{u}, \boldsymbol{v}) + \log W(\boldsymbol{y}|\boldsymbol{x}[\boldsymbol{u}])$.

the term $\log|\mathcal{T}_{\rm s}(\boldsymbol{u}|\boldsymbol{v})|$ is the analogue of the conditional empirical entropy pertaining to the source part, whereas the term $\log Q[\mathcal{T}_{\rm c}(\boldsymbol{x}[\boldsymbol{u}]|\boldsymbol{y})$ plays the role of the negative empirical mutual information between $\boldsymbol{x}[\boldsymbol{u}]$ and $\boldsymbol{y}$. Therefore if, for example,

$$m_{{\rm c},\theta}(\boldsymbol{x},\boldsymbol{y}) = \sum_{t=1}^{n} m_{{\rm c},\theta}(x_t,y_t) \quad \text{and} \tag{56}$$

$$m_{{\rm s},\theta}(\boldsymbol{u},\boldsymbol{v}) = \sum_{t=1}^{n} m_{{\rm s},\theta}(u_t,v_t), \tag{57}$$

as is the case when the sources and the channel are memoryless, then $\{\mathcal{T}_{\rm c}(\boldsymbol{x}|\boldsymbol{y})\}$ and $\{\mathcal{T}_{\rm s}(\boldsymbol{u}|\boldsymbol{v})\}$ become conditional type classes in the usual sense, and we are back to the generalized MMI decoder of Section 3, provided that Q is, again, the uniform distribution within a single type class. As a final note, in this context, we mention that in this setting, the input and the output alphabets of the channel may also be continuous, see, e.g., [24, p. 5575, Example 3].

5.3 Separate Encodings and Universal Joint Decoding of Correlated Sources

Consider the system depicted in Fig. 2, which illustrates a scenario of separate source–channel encodings and joint decoding of two correlated sources, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. For the sake of simplicity of the presentation, we return to the assumption of memoryless systems, as in Section 3.

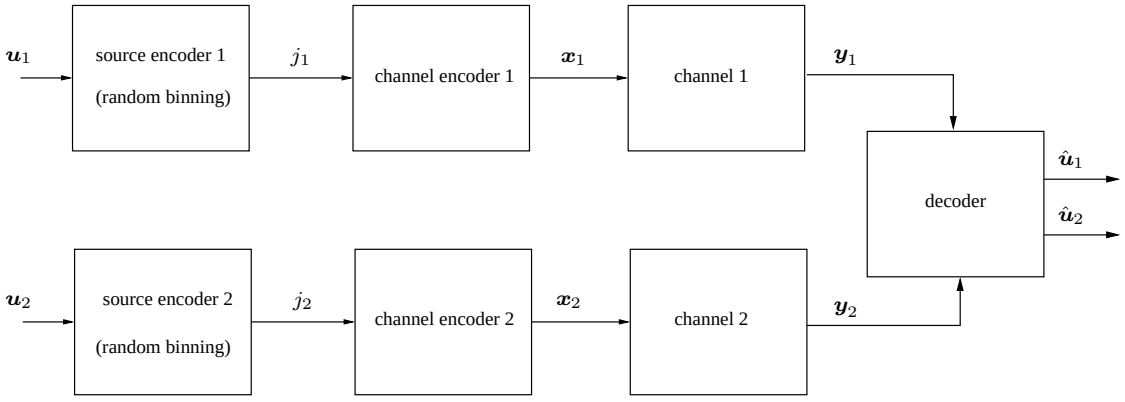

Figure 2: Separate source–channel encodings and joint decoding of two correlated sources.

Consider n independent copies $\{(U_{1,i},U_{2,i})\}_{i=1}^{n}$ of a finite–alphabet pair of random variables $(U_1,U_2) \sim P_{U_1U_2}$, as well as n uses of two independent finite–alphabet DMC's $W_1(\boldsymbol{y}_1|\boldsymbol{x}_1) = \prod_{t=1}^{n} W(y_{1,t}|x_{1,t})$ and $W_2(\boldsymbol{y}_2|\boldsymbol{x}_2) = \prod_{t=1}^{n} W(y_{2,t}|x_{2,t})$. For $k=1,2$, consider the following mechanism: The source vector $\boldsymbol{u}_k = (u_{k,1},\ldots,u_{k,n})$ is encoded into one out of $M_k = 2^{nR_k}$ bins, selected independently at random for every member of $\mathcal{U}_k^n$. The bin index $j_k = f_k(\boldsymbol{u}_k)$ is in turn mapped into a channel input vector $\boldsymbol{x}_k(i) \in \mathcal{X}_1^n$, which is transmitted across the channel W_k. The various codewords $\{\boldsymbol{x}_k(i)\}_{i=1}^{M_k}$ are selected independently at random under the uniform distribution within

given type classes $\mathcal{T}(Q_k)$, where Q_k is a given distribution across $\mathcal{X}_k$. The randomly chosen codebook $\{\boldsymbol{x}_k(1), \boldsymbol{x}_k(2), \ldots, \boldsymbol{x}_k(M_k)\}$ will be denoted by $\mathcal{C}_k$. Similarly, as before, we will sometimes denote $\boldsymbol{x}_k(j_k) = \boldsymbol{x}_k[f_k(\boldsymbol{u}_k)]$ by $\boldsymbol{x}_k[\boldsymbol{u}_k]$. The optimal (MAP) decoder estimates $(\boldsymbol{u}_1, \boldsymbol{u}_2)$, using the channel outputs $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, according to

$$(\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{u}}_2) = \arg \max_{\boldsymbol{u}_1, \boldsymbol{u}_2} P(\boldsymbol{u}_1, \boldsymbol{u}_2) W_1(\boldsymbol{y}_1|\boldsymbol{x}_1[\boldsymbol{u}_1]) W_2(\boldsymbol{y}_2|\boldsymbol{x}_2[\boldsymbol{u}_2]). \tag{58}$$

The main structure of the analysis continues to be essentially the same as in Section 4. The situation here, however, is significantly more involved, because five different types of pairwise error events $\{(\boldsymbol{u}_1, \boldsymbol{u}_2) \to (\boldsymbol{u}_1', \boldsymbol{u}_2')\}$ should be carefully handled:

1. $\boldsymbol{u}_1' \neq \boldsymbol{u}_1$ and $\boldsymbol{u}_2' = \boldsymbol{u}_2$.
2. $\boldsymbol{u}_2' \neq \boldsymbol{u}_2$ and $\boldsymbol{u}_1' = \boldsymbol{u}_1$.
3. Both $\boldsymbol{u}_1' \neq \boldsymbol{u}_1$ and $\boldsymbol{u}_2' \neq \boldsymbol{u}_2$, but (at least) $\boldsymbol{u}_2'$ is mapped into the same bin as $\boldsymbol{u}_2$.
4. Both $\boldsymbol{u}_1' \neq \boldsymbol{u}_1$ and $\boldsymbol{u}_2' \neq \boldsymbol{u}_2$, but (at least)[11] $\boldsymbol{u}_1'$ is mapped into the same bin as $\boldsymbol{u}_1$.
5. Both $\boldsymbol{u}_1' \neq \boldsymbol{u}_1$ and $\boldsymbol{u}_2' \neq \boldsymbol{u}_2$, and neither $\boldsymbol{u}_1'$ nor $\boldsymbol{u}_2'$ belongs to the same bin as the respective true source vector.

Errors of types 1 and 2 are of the same nature as in Section 3, where the source that is estimated correctly, is actually in the role of SI at the decoder. Following (11), the respective metrics are[12]

$$f_1(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2) = R_1 \wedge \hat{I}(X_1; Y_1) - \hat{H}(U_1|U_2) \tag{59}$$

$$f_2(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2) = R_2 \wedge \hat{I}(X_2; Y_2) - \hat{H}(U_2|U_1). \tag{60}$$

where $\hat{I}(X_1; Y_1)$ and $\hat{H}(U_1|U_2)$ are shorthand notations for $\hat{I}_{\boldsymbol{x}_1\boldsymbol{x}_2}(X_1; Y_1)$ and $\hat{H}_{\boldsymbol{u}_1\boldsymbol{u}_2}(U_1|U_2)$, respectively, and so on. Errors of types 3 and 4 will turn out to be addressed by metrics of the form

$$f_3(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2) = R_1 \wedge \hat{I}(X_1; Y_1) + R_2 - \hat{H}(U_1, U_2) \tag{61}$$

$$f_4(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2) = R_2 \wedge \hat{I}(X_2; Y_2) + R_1 - \hat{H}(U_1, U_2). \tag{62}$$

Finally, error of type 5 is accommodated by

$$\begin{aligned} f_5(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2) &= \hat{I}(X_1; Y_1) + \hat{I}(X_2; Y_2) - \\ &\quad \min\{\hat{H}(U_1, U_2), R_1 \wedge \hat{I}(X_1; Y_1) + R_2 \wedge \hat{I}(X_2; Y_2)\} \\ &\equiv [R_1 \wedge \hat{I}(X_1; Y_1) + R_2 \wedge \hat{I}(X_2; Y_2) - \hat{H}(U_1, U_2)]_+ + \end{aligned}$$

[11] Here, we are counting twice the case "$\boldsymbol{u}_1' \neq \boldsymbol{u}_1$ and $\boldsymbol{u}_2' \neq \boldsymbol{u}_2$ and both estimates are in the bins of their respective true source vectors." This is done simply for symmetry the structure above, without affecting the error exponent.

[12] Note that f_1 does not really depend on $(\boldsymbol{x}_2, \boldsymbol{y}_2)$, and similarly, f_2 does not depend on $(\boldsymbol{x}_1, \boldsymbol{y}_1)$. Nonetheless, we deliberately adopt this uniform notation for convenience later on.

$$[\hat{I}(X_1;Y_1) - R_1]_+ + [\hat{I}(X_2;Y_2) - R_2]_+ \tag{63}$$

But we need a *single* universal decoding metric that copes with all five types of errors at the same time.

Similarly as in [24, eqs. (57)-(60)], this objective is accomplished by a metric which is given by the minimum among all five metrics above, i.e., we define our decoding metric as

$$f_0(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2) = \min_{1 \le i \le 5} f_i(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2), \tag{64}$$

meaning that the proposed universal decoder is given by

$$(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2) = \arg \max_{\boldsymbol{u}_1, \boldsymbol{u}_2} f_0(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1[\boldsymbol{u}_1], \boldsymbol{x}_2[\boldsymbol{u}_2], \boldsymbol{y}_1, \boldsymbol{y}_2). \tag{65}$$

The conditional probability of error given $(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2)$, for both the MAP decoder and the universal decoder, can be shown to be of the exponential order of $\exp_2\{-n[f_0(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2)]_+\}$.

To show this, the analysis of the probability of error, for both the MAP decoder and the universal decoder, should be divided into several parts, according to the various types of error events. Errors of types 1 and 2 are addressed exactly as in Section 4. The more complicated part of the analysis is due to errors of types 3–5, where both competing source vectors are in error. However, this analysis too follows the same basic ideas. Here we will outline only the main ingredients that are different from those of the proof of Theorem 1.

For a given $\boldsymbol{u}_1' \neq \boldsymbol{u}_1$ and $\boldsymbol{u}_2' \neq \boldsymbol{u}_2$ (errors of types 3–5), let us define the pairwise error event

$$\begin{aligned} & \mathcal{A}(\boldsymbol{u}_1, \boldsymbol{u}_1', \boldsymbol{u}_2, \boldsymbol{u}_2', \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2) \\ = \; & [\mathcal{T}(Q_1) \times \mathcal{T}(Q_2)] \bigcap \\ & \left\{(\boldsymbol{x}_1', \boldsymbol{x}_2') : \; P(\boldsymbol{u}_1', \boldsymbol{u}_2') W_1(\boldsymbol{y}_1|\boldsymbol{x}_1') W_2(\boldsymbol{y}_2|\boldsymbol{x}_2') \ge P(\boldsymbol{u}_1, \boldsymbol{u}_2) W_1(\boldsymbol{y}_1|\boldsymbol{x}_1) W_2(\boldsymbol{y}_2|\boldsymbol{x}_2)\right\}. \end{aligned}$$

The conditional error event, given $(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2, \mathcal{C}_1, \mathcal{C}_2)$, is given by

$$\begin{aligned} & \mathcal{E}(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2, \mathcal{C}_1, \mathcal{C}_2) \\ = \; & \bigcup_{\boldsymbol{u}_1' \neq \boldsymbol{u}_1, \; \boldsymbol{u}_2' \neq \boldsymbol{u}_2} \left\{P(\boldsymbol{u}_1', \boldsymbol{u}_2') W_1(\boldsymbol{y}_1|\boldsymbol{x}_1[\boldsymbol{u}_1']) W_2(\boldsymbol{y}_2|\boldsymbol{x}_2[\boldsymbol{u}_2']) \ge \right. \\ & \left. P(\boldsymbol{u}_1, \boldsymbol{u}_2) W_1(\boldsymbol{y}_1|\boldsymbol{x}_1[\boldsymbol{u}_1]) W_2(\boldsymbol{y}_2|\boldsymbol{x}_2[\boldsymbol{u}_2])\right\} \\ \stackrel{\Delta}{=} \; & \bigcup_{\boldsymbol{u}_1' \neq \boldsymbol{u}_1, \; \boldsymbol{u}_2' \neq \boldsymbol{u}_2} \mathcal{E}(\boldsymbol{u}_1, \boldsymbol{u}_1', \boldsymbol{u}_2, \boldsymbol{u}_2', \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2, \mathcal{C}_1, \mathcal{C}_2) \end{aligned} \tag{66}$$

Here too, the exponential tightness of the truncated union bound for two dimensional unions of events with independence structure as above can be established using de Caen's lower bound [5] (see [28]). For errors of types 3 and 4, let us define

$$\begin{aligned} & \mathcal{A}_1(\boldsymbol{u}_1, \boldsymbol{u}_1', \boldsymbol{u}_2, \boldsymbol{u}_2', \boldsymbol{x}_1, \boldsymbol{y}_1) \\ = \; & \mathcal{T}(Q_1) \bigcap \left\{\boldsymbol{x}_1' : \; P(\boldsymbol{u}_1', \boldsymbol{u}_2') W_1(\boldsymbol{y}_1|\boldsymbol{x}_1') \ge P(\boldsymbol{u}_1, \boldsymbol{u}_2) W_1(\boldsymbol{y}_1|\boldsymbol{x}_1)\right\} \end{aligned} \tag{67}$$

and

$$\begin{aligned}
&\mathcal{A}_2(\boldsymbol{u}_1,\boldsymbol{u}_1',\boldsymbol{u}_2,\boldsymbol{u}_2',\boldsymbol{x}_2,\boldsymbol{y}_2)\\
=\;&\mathcal{T}(Q_2)\bigcap\{\boldsymbol{x}_2':\ P(\boldsymbol{u}_1',\boldsymbol{u}_2')W_2(\boldsymbol{y}_2|\boldsymbol{x}_2')\geq P(\boldsymbol{u}_1,\boldsymbol{u}_2)W_2(\boldsymbol{y}_2|\boldsymbol{x}_2).\}
\end{aligned}\tag{68}$$

The probability of $\mathcal{E}(\boldsymbol{u}_1,\boldsymbol{u}_1',\boldsymbol{u}_2,\boldsymbol{u}_2',\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}_1,\boldsymbol{y}_2,\mathcal{C}_1,\mathcal{C}_2)$ (w.r.t. the randomness of the bin assignment) is given by:

$$\begin{aligned}
&\overline{\Pr}\{\mathcal{E}(\boldsymbol{u}_1,\boldsymbol{u}_1',\boldsymbol{u}_2,\boldsymbol{u}_2',\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}_1,\boldsymbol{y}_2,\mathcal{C}_1,\mathcal{C}_2)\}\\
=\;&2^{-n(R_1+R_2)}\left|[\mathcal{C}_1\times\mathcal{C}_2]\bigcap\mathcal{A}(\boldsymbol{u}_1,\boldsymbol{u}_1',\boldsymbol{u}_2,\boldsymbol{u}_2',\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}_1,\boldsymbol{y}_2)\right|+\\
&2^{-n(R_1+R_2)}\left|\mathcal{C}_1\bigcap\mathcal{A}_1(\boldsymbol{u}_1,\boldsymbol{u}_1',\boldsymbol{u}_2,\boldsymbol{u}_2',\boldsymbol{x}_1,\boldsymbol{y}_1)\right|+\\
&2^{-n(R_1+R_2)}\left|\mathcal{C}_2\bigcap\mathcal{A}_2(\boldsymbol{u}_1,\boldsymbol{u}_1',\boldsymbol{u}_2,\boldsymbol{u}_2',\boldsymbol{x}_2,\boldsymbol{y}_2)\right|+\\
&+2^{-n(R_1+R_2)}\mathcal{I}\{P(\boldsymbol{u}_1',\boldsymbol{u}_2')\geq P(\boldsymbol{u}_1,\boldsymbol{u}_2)\},
\end{aligned}\tag{69}$$

where the first term stands for errors of type 5, the second and third terms represent errors of types 3 and 4, and the last term is associated with an error where both $\boldsymbol{u}_1'\neq\boldsymbol{u}_1$ and $\boldsymbol{u}_2'\neq\boldsymbol{u}_1$, but the respective bins both coincide. Passing temporarily to shorthand notation, let us denote

$$N\stackrel{\Delta}{=}\left|[\mathcal{C}_1\times\mathcal{C}_2]\bigcap\mathcal{A}\right|+\left|\mathcal{C}_1\bigcap\mathcal{A}_1\right|+\left|\mathcal{C}_2\bigcap\mathcal{A}_2\right|+\mathcal{I}\{P(\boldsymbol{u}_1',\boldsymbol{u}_2')\geq P(\boldsymbol{u}_1,\boldsymbol{u}_2)\}\stackrel{\Delta}{=}N_{12}+N_1+N_2+I.\tag{70}$$

The next step, as before, is to average over the randomness of all codewords in $\mathcal{C}_1$ and $\mathcal{C}_2$, To analyze the large deviations behavior of $N_{12}+N_1+N_2$, the contributions of the individual random variables can be handled separately, since $\Pr\{N_{12}+N_1+N_2>\text{threshold}\}$ is of the same exponential order of the sum

$$\Pr\{N_{12}>\text{threshold}\}+\Pr\{N_1>\text{threshold}\}+\Pr\{N_2>\text{threshold}\}.$$

Now, N_1 and N_2 are binomial random variables whose numbers of trials are 2^{nR_1} and 2^{nR_2}, respectively, and whose probabilities of success decay exponentially according to the relevant channel mutual informations, similarly as before. So their contributions are again analyzed with great similarly to those of type 1 and type 2 errors.

Finally, it remains to handle N_{12}, which not a binomial random variables, but it can be decomposed as the sum (over combinations of conditional types of $\boldsymbol{x}_1'$ given $\boldsymbol{y}_1$ and of $\boldsymbol{x}_2'$ given $\boldsymbol{y}_2$) of products of independent binomial random variables, for which we reuse the notations N_1 and N_2 (for a given combination of such types). Using the same techniques as in [19, Chap. 6], one can easily obtain the following generic result concerning the large deviations behavior of $N_1\cdot N_2$: If N_1 is a binomial random variable with 2^{nA_1} trials and probability of success 2^{-nB_1} and N_2 is an independent binomial random variable with 2^{nA_2} trials and probability of success 2^{-nB_2}, then

$$\Pr\{N_1\cdot N_2\geq 2^{nC}\}\ \doteq\ \max_{0\leq\alpha\leq C}\Pr\{N_1\geq 2^{n\alpha}\}\cdot\Pr\{N_2\geq 2^{n(C-\alpha)}\}$$

$$\dot{=} \quad 2^{-nE} \tag{71}$$

with

$$E = \begin{cases} [B_1 - A_1]_+ + [B_2 - A_2]_+ & C \leq [A_1 - B_1]_+ + [A_2 - B_2]_+ \\ \infty & C > [A_1 - B_1]_+ + [A_2 - B_2]_+ \end{cases} \tag{72}$$

Using this fact, it is possible to obtain the contribution of the type 5 error event.

Upon carrying out the analysis along these lines, the state of affairs turns out to be as described next. In the analysis of the conditional probability of error, the contribution of a given type class, $\mathcal{T}(\boldsymbol{u}_1', \boldsymbol{u}_2')$, of competing source vectors, which are encoded into $\boldsymbol{x}_1'$ and $\boldsymbol{x}_2'$ (from given conditional type classes given $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively) is the following: the probability of error of type i is of the exponential order of $\exp\{-n[f_i(\boldsymbol{u}_1', \boldsymbol{u}_2', \boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{y}_1, \boldsymbol{y}_2)]_+\}$, $i = 1, \ldots, 5$. Thus, the total conditional error probability contributed by this combination of types is of the exponential order of

$$\begin{aligned} \sum_{i=1}^{5} \exp\{-n[f_i(\boldsymbol{u}_1', \boldsymbol{u}_2', \boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{y}_1, \boldsymbol{y}_2)]_+\} &\dot{=} \exp\{-n \min_i [f_i(\boldsymbol{u}_1', \boldsymbol{u}_2', \boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{y}_1, \boldsymbol{y}_2)]_+\} \\ &= \exp\{-n[f_0(\boldsymbol{u}_1', \boldsymbol{u}_2', \boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{y}_1, \boldsymbol{y}_2)]_+\}. \end{aligned} \tag{73}$$

For the total contribution of all type classes, the exponent $[f_0(\boldsymbol{u}_1', \boldsymbol{u}_2', \boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{y}_1, \boldsymbol{y}_2)]_+$ should be minimized over all such combinations of types (that yield the relevant pairwise error event). An upper bound on this exponent is obtained by selecting the same combination of types as those of the correct source vectors (instead of taking this minimum), namely, the conditional error probaiblity of the MAP decoder is simply lower bounded by the exponential order of $\exp\{-n[f_0(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2)]_+\}$. As for the universal decoder, one should minimize the exponent $[f_0(\boldsymbol{u}_1', \boldsymbol{u}_2', \boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{y}_1, \boldsymbol{y}_2)]_+$ as well, but only over the combinations of type classes that are associated with the pairwise error event of this decoder, namely, those for which $f_0(\boldsymbol{u}_1', \boldsymbol{u}_2', \boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{y}_1, \boldsymbol{y}_2) \geq f_0(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2)$. However, this minimum is exactly $[f_0(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2)]_+$, which agrees with that of the upper bound associated with the MAP decoder.

References

[1] J. Chen, D.-k. He, A. Jagmohan, and L. A. Lastras–Montaño, " On universal variable–rate Slepian–Wolf coding," *Proc. 2008 IEEE International Conference on Communications (ICC 2008)*, pp. 1426–1430, 2008.

[2] I. Csiszár, "Joint source–channel error exponent," *Problems of Control and Information Theory*, vol. 9, no. 5, pp. 315–328, 1980.

[3] I. Csiszár, "Linear codes for sources and source networks: error exponents, universal coding," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 4, pp. 585–592, July 1982.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981.

[5] D. de Caen, "A lower bound on the probability of a union," *Discrete Math.*, vol. 169, pp. 217–220, 1997.

[6] S. C. Draper, "Universal incremental Slepian–Wolf coding," *Proc. 42nd Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, USA, October 2004.

[7] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1726–1745, September 1998.

[8] M. Feder and N. Merhav, "Universal composite hypothesis testing: a competitive minimax approach," *IEEE Trans. Inform. Theory*, special issue in memory of Aaron D. Wyner, vol. 48, no. 6, pp. 1504–1517, June 2002.

[9] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Cont. Information Theory*, vol. 4, pp. 97–102, 1975.

[10] J. C. Kieffer, "Some universal noiseless multiterminal source coding theorems," *Information and Control*, vol. 46, pp. 93–107, 1980.

[11] A. Lapidoth and J. Ziv, "On the universality of the LZ–based noisy channels decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1746–1755, September 1998.

[12] Y.-S. Liu and B. L. Hughes, "A new universal random coding bound for the multiple access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 376–386, March 1996.

[13] Y. Lomnitz and M. Feder, "Communication over individual channels – a general framework," arXiv:1023.1406v1 [cs.IT] 7 Mar 2012.

[14] Y. Lomnitz and M. Feder, "Universal communication over modulo–additive channels with an individual noise sequence," arXiv:1012.2751v2 [cs.IT] 7 May 2012.

[15] S. Matloub and T. Weissman, "Universal zero–delay joint source–channel coding," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5240–5250, December 2006.

[16] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1261–1269, July 1993.

[17] N. Merhav, "Universal detection of messages via finite–state channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242–2246, September 2000.

[18] N. Merhav, "Shannon's secrecy system with informed receivers and its application to systematic coding for wiretapped channels," *IEEE Trans. Inform. Theory*, special issue on *Information–Theoretic Security*, vol. 54, no. 6, pp. 2723–2734, June 2008.

[19] N. Merhav, "Statistical physics and information theory," *Foundations and Trends in Communications and Information Theory*, vol. 6, nos. 1–2, pp. 1–212, 2009.

[20] N. Merhav, "Relations between random coding exponents and the statistical physics of random codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 1, pp. 83–92, January 2009.

[21] N. Merhav, "Erasure/list exponents for Slepian–Wolf decoding," *IEEE Trans. Inform. Theory*, vol. 60, no. 8, pp. 4463–4471, August 2014.

[22] N. Merhav, "Exact random coding exponents of optimal bin index decoding," *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 6024–6031, October 2014.

[23] N. Merhav, "Erasure/list exponents for Slepian–Wolf decoding," *IEEE Trans. Inform. Theory*, vol. 60, no. 8, pp. 4463–4471, August 2014.

[24] N. Merhav, "Universal decoding for arbitrary channels relative to a given family of decoding metrics," *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5566–5576, September 2013.

[25] V. Misra and T. Weissman, "The porosity of additive noise sequences," arXiv:1025.6974v1 [cs.IT] 31 May 2012.

[26] Y. Oohama and T. S. Han, "Universal coding for the Slepian–Wolf data compression system and the strong converse theorem," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1908–1919, November 1994.

[27] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Variable–rate universal Slepian–Wolf coding with feedback," *Proc. 39th Asilomar Conference on Signals, Systems and Computers*, pp. 8–12, November 2005.

[28] J. Scarlett, A. Martinéz, and A. i. Fábregas, "Multiuser techniques for mismatched decoding," submitted to *IEEE Trans. Inform. Theory*, November 2013. `arxiv.org/pdf/1311.6635`

[29] S. Shamai (Shitz), S. Verdú and R. Zamir, "Systematic lossy source/ channel coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564–579, March 1998.

[30] N. Shulman, *Communication over an Unknown Channel via Common Broadcasting*, Ph.D. dissertation, Department of Electrical Engineering – Systems, Tel Aviv University, July 2003.

[31] N. Weinberger and N. Merhav, "Optimum trade-off between the error exponent and the excess–rate exponent of variable–rate Slepian–Wolf coding," *IEEE Trans. Inform. Theory*, vol. 61, no. 4, pp. 2165–2190, April 2015.

[32] J. Ziv, "Universal decoding for finite–state channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.

[33] J. Ziv and A. Lempel, "Compression of individual sequences via variable–rate coding," *IEEE Trans. Inform. Theory*, vol. IT–24, no. 5, pp. 530–536, September 1978.