

Universal Decoding for Asynchronous Slepian-Wolf Encoding

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E-mail: merhav@ee.technion.ac.il

Abstract

We consider the problem of (almost) lossless source coding of two correlated memoryless sources using separate encoders and a joint decoder, that is, Slepian-Wolf (S-W) coding. In our setting, the encoding and decoding are asynchronous, i.e., there is a certain relative delay between the two sources. Neither the source parameters nor the relative delay are known to the encoders and the decoder. Since we assume that both encoders implement standard random binning, which does not require such knowledge anyway, the focus of this work is on the decoder. Our main contribution is in proposing a universal decoder, that independent of the unknown source parameters and the relative delay, and at the same time, is asymptotically as good as the optimal maximum a posteriori probability (MAP) decoder in the sense of the random coding error exponent achieved. Consequently, the achievable rate region is also the same as if the source parameters and the delay were known to the decoder.

Index Terms: Slepian-Wolf coding, universal decoding, error exponent, asynchronous coding, delay.

1 Introduction

The problem of separate encodings and joint decoding of correlated sources, i.e., the well known Slepian–Wolf (S-W) coding problem, has received a vast level of attention ever since the landmark paper by Slepian and Wolf [18] was published, nearly five decades ago. Much less attention, however, was given to the asynchronous version of this problem, where there is a relative delay between the two correlated sources, see, e.g., [15], [11], [16], [19], [20]. The motivation for the asynchronous setting is thoroughly discussed in [11]. For memoryless correlated sources, Willems [20] assumed that the relative delay is unknown to the encoders, but known to the decoder, and proved that the achievable rate region is the same as in synchronous S-W coding. Under similar assumptions, Rimoldi and Urbanke [16], as well as Sun, Tian, Chen and Wong [19], have proposed S-W data compression schemes that are based on the notion of source splitting. In all these studies, it was assumed that the decoder has the option to postpone the actual decoding until after having received all codewords associated with the data to be decoded. Such an assumption essentially neutralizes the negative effect of the relative delay because the encoders and the decoder can still exploit the correlations between the two sources. As explained, however, by Matsuta and Uyematsu in their recent paper [11], this setup might be somewhat problematic, in practice, especially when the relative delay is very large.

The main result provided by Matsuta and Uyematsu in [11] (see also [9] and [10]) is a worst-case result in spirit. They assumed that: (i) the joint probability distribution, P_{XY} , of the two corresponding correlated random variables, X and Y , one from each source to be compressed, is only known to belong to a subset \mathcal{S} of joint probability distributions, (ii) the relative delay between the sources, d , is unknown, but known to be bounded between two limits, and (iii) the absolute value of the relative delay, $|d|$, is allowed to scale linearly with n , and the ratio $|d|/n$ tends to a constant, $\delta \in [0, 1]$, as $n \rightarrow \infty$, and δ is only known to be upper bounded by a given number Δ . They have proved a coding theorem asserting that the achievability rate region is as follows. A rate-pair (R_x, R_y) is achievable if and only if it satisfies the following three inequalities at the same time:

$$R_x \geq \sup_{P_{XY} \in \mathcal{S}} [H(X|Y) + \Delta I(X; Y)] \quad (1)$$

$$R_y \geq \sup_{P_{XY} \in \mathcal{S}} [H(Y|X) + \Delta I(X; Y)] \quad (2)$$

$$R_x + R_y \geq \sup_{P_{XY} \in \mathcal{S}} [H(X, Y) + \Delta I(X; Y)], \quad (3)$$

This result of [11] is very interesting, but it is also extremely pessimistic. It is overly pessimistic, not only because it is worst-case result, but even more importantly, because the above three suprema can potentially be achieved by *three different sources*, in general. Thus, for a single, given underlying joint source, $P_{XY} \in \mathcal{S}$ (with relative delay, δn), as bad as it may be, at least one of the above three inequalities can be improved, in general. Moreover, if \mathcal{S} happens to be the entire simplex of probability distributions $\{P_{XY}\}$ over the given alphabets, \mathcal{X} and \mathcal{Y} (which is a very realistic special case), these suprema are given by $\log |\mathcal{X}|$, $\log |\mathcal{Y}|$, and $\log |\mathcal{X}| + \log |\mathcal{Y}|$, respectively, rendering this coding theorem an uninteresting triviality, as it allows no compression at all. The fact of the matter is, however, that at least the weakness concerning the three different achievers of the suprema in (1)–(3) can be handled rather easily. Upon a careful inspection of the proof of the converse part in [11], one readily concludes that it actually supports an assertion that the achievable rate region is included in the following set:

$$\bigcap_{P_{XY} \in \mathcal{S}} \left\{ \begin{aligned} (R_x, R_y) : R_x &\geq H(X|Y) + \Delta I(X; Y), \\ R_y &\geq H(Y|X) + \Delta I(X; Y), \\ R_x + R_y &\geq H(X, Y) + \Delta I(X; Y) \end{aligned} \right\}. \quad (4)$$

Similar comments apply to the analysis of the error probability in [11], which is a pessimistic analysis, carried out for the worst source in \mathcal{S} and over all possible relative delay values, rather than the actual error probability associated with a given underlying source.

In this paper, we tackle the problem in a different manner. Instead of a worst-case approach, our approach is the following: for a given rate pair (R_x, R_y) , even if we knew the source and the relative delay, we could have handled only sources, $\{P_{XY}\}$, that satisfy $H(X|Y) + \delta I(X; Y) \leq R_x$, $H(Y|X) + \delta I(X; Y) \leq R_y$ and $H(X, Y) + \delta I(X; Y) \leq R_x + R_y$ (δ being the actual normalized relative delay). Now, the S-W encoders are always simple random-binning encoders, regardless of the source parameters, so every uncertainty, that is associated with the source parameters and the relative delay, is confronted, and therefore must be handled, by the decoder. Owing to the earlier results on the minimum-entropy universal decoder for the S-W encoding system (see, e.g., [1], [2], [3], [4], [5, Exercise 3.1.6], [7], [8], [12], [14], [17]), it is natural to set the goal of seeking a universal

decoder that is asymptotically as good as the optimal maximum a posteriori (MAP) decoder for the given source, in the sense that the random coding error exponent is the same, and hence so is the achievable rate region. In other words, unlike in previous works on universal S-W decoding, here universality is sought, not only with respect to (w.r.t.) the source distribution, P_{XY} , but also w.r.t. the unknown relative delay between the two parts of the source. Although it is natural to think of the relative delay as of yet another unknown parameter associated with the underlying source, it will be interesting to see that in our proposed universal decoder, the unknown delay will be handled differently than the other unknown parameters. We will elaborate on this point later on.

Our main contributions, in this work, are the following:

1. We propose a universal decoder that allows uncertainty, not only regarding the source parameters, but also the relative delay.
2. We prove that our universal decoder achieves the same error exponent as the optimal MAP decoder that is cognizant of both the source parameters and the relative delay. This will be done by showing that our upper bound on the error probability of the universal decoder is of the same exponential order as a lower bound on the error probability of the MAP decoder.
3. We provide the Lagrange–dual form of the resulting error exponent, and thereby characterize the achievable rate region for achieving a prescribed random coding error exponent, E .
4. We provide an outline for a possible extension to sources with memory.

The outline of the remaining part of this paper is as follows. In Section 2, we establish notation conventions. In Section 3, we formulate the problem and spell out the objectives of this work. In Section 4, we present the main theorem and discuss it. Finally, in Section 5, we prove this theorem.

2 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 $\mathbf{X} = (X_1, \dots, X_n)$, (n – positive integer) may take a specific vector value $\mathbf{x} = (x_1, \dots, x_n)$ in \mathcal{X}^n , the n -th order Cartesian power of \mathcal{X} , which is the alphabet of each component of this vector. Segments of vector components will be denoted by subscripts and superscripts, for example, x_i^j , $i < j$, will designate the segment $(x_i, x_{i+1}, \dots, x_j)$. When $i = 1$, the subscript will be omitted and therefore the notation will be x^j . By convention, when $i > j$, x_i^j will be understood to be the empty string, whose probability is formally defined to be unity. Sources and channels will be denoted by the letter P or Q , 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. 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 $\mathbf{E}_P\{\cdot\}$. Again, the subscript will be omitted if the underlying probability distribution is clear from the context. The entropy of a random variable (RV) X with a generic distribution Q will be denoted by $H_Q(X)$. Similarly, other information measures will be denoted using the customary notation, subscripted by the name of the underlying distribution Q . For example, for a pair of RVs, (X, Y) , distributed according to Q_{XY} (or Q , for short), $H_Q(X, Y)$, $H_Q(X|Y)$ and $I_Q(X; Y)$ will denote the joint entropy, the conditional entropy of X given Y , and the mutual information, respectively. For two positive sequences a_n and b_n , the notation $a_n \stackrel{\cdot}{=} b_n$ will stand for equality in the exponential scale, that is, $\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Similarly, $a_n \stackrel{\cdot}{\leq} b_n$ means that $\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{a_n}{b_n} \leq 0$, and so on. 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 empirical distribution of a sequence $\mathbf{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}_{\mathbf{x}}$, is the vector of relative frequencies, $\hat{P}_{\mathbf{x}}(x)$, of each symbol $x \in \mathcal{X}$ in \mathbf{x} . The type class of $\mathbf{x} \in \mathcal{X}^n$, denoted $\mathcal{T}(\mathbf{x})$, is the set of all vectors \mathbf{x}' with $\hat{P}_{\mathbf{x}'} = \hat{P}_{\mathbf{x}}$. When we wish to emphasize the dependence of the type class on the empirical distribution \hat{P} , we will denote it by $\mathcal{T}(\hat{P})$, with a slight abuse of notation. Information measures associated with empirical distributions will be denoted with ‘hats’ and will include the names of the vectors from which they are induced by parentheses. For example, the empirical entropy of \mathbf{x} , which is the entropy associated with $\hat{P}_{\mathbf{x}}$, will be denoted by $\hat{H}(\mathbf{x})$. An alternative notation, following the conventions described in the previous paragraph, is $H(\hat{P}_{\mathbf{x}})$.

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, $\hat{P}_{\mathbf{x}\mathbf{y}}$ will be the joint empirical distribution of $(\mathbf{x}, \mathbf{y}) = \{(x_i, y_i)\}_{i=1}^n$, $\mathcal{T}(\mathbf{x}, \mathbf{y})$ or $\mathcal{T}(\hat{P}_{\mathbf{x}\mathbf{y}})$ will denote the joint type class of (\mathbf{x}, \mathbf{y}) , $\mathcal{T}(\mathbf{x}|\mathbf{y})$ will stand for the conditional type class of \mathbf{x} given \mathbf{y} , $\hat{H}(\mathbf{x}, \mathbf{y})$ will designate the empirical joint entropy of \mathbf{x} and \mathbf{y} , $\hat{H}(\mathbf{x}|\mathbf{y})$ will be the empirical conditional entropy, and so on. Clearly, empirical information measures can be calculated, not only from the full vectors, but also from partial segments, like x_i^j and y_i^j . In this case, x_i^j and y_i^j will replace \mathbf{x} and \mathbf{y} in the above notations.

3 Problem Formulation

Let $\{(X_i, Y_i)\}$ be a pair of correlated discrete memoryless sources (DMSs) with a relative delay of d time units, that is, (X_i, Y_{i+d}) are jointly distributed according to a certain probability distribution, P_{XY} for every i , but the various pairs are mutually independent. In other words, the random vectors $Z_i = (X_i, Y_{i+d})$ are i.i.d. for different values of i . Neither P_{XY} and d are known to the encoders and decoder.

Similarly as in [11], the two separate encoders that compress $\{X_i\}$ and $\{Y_i\}$ both operate on successive blocks of length n , without any attempt to align them, because d is unknown and it may be arbitrarily large. These encoders are ordinary S-W encoders at rates R_x and R_y , respectively. In other words, each member \mathbf{x} (resp. \mathbf{y}) of \mathcal{X}^n (resp. \mathcal{Y}^n) is mapped into a bin $f(\mathbf{x}) \in \{1, 2, \dots, 2^{nR_x}\}$ (resp. $g(\mathbf{y}) \in \{1, 2, \dots, 2^{nR_y}\}$), which is selected independently at random for every n -vector in the respective source space. As always, the randomly selected mappings of both encoders are revealed to the decoder.

As already mentioned, both P_{XY} and d are unknown, but without essential loss of generality, it may be assumed that $0 \leq d \leq n$. For any $d \geq n$, the respective blocks concurrently encoded, \mathbf{X} and \mathbf{Y} , are statistically independent, and so, all values of d , from n and beyond, are actually equivalent from the viewpoints of the encoders and the decoder. The lower limit, $d \geq 0$, is assumed for convenience only. Negative values of d correspond to switching the roles of the two sources in the forthcoming results and discussions (see also [11]).¹ Our asymptotic regime will be defined as

¹We could have allowed negative values of d in the formal problem setup to begin with, but this would make the notation more cumbersome.

described in the Introduction: as $n \rightarrow \infty$, the relative delay d will be asymptotically proportional to n , i.e., the ratio d/n tends to a limit, δ . In view of the above discussion, δ can be assumed to take on values in the interval $[0, 1]$.

The decoder receives the bin indices, $f(\mathbf{x})$ and $g(\mathbf{y})$, of the compressed vectors, \mathbf{x} and \mathbf{y} , respectively, and it outputs a pair of estimates, $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}^n \times \mathcal{Y}^n$. The average probability of error is defined as

$$\bar{P}_e \triangleq \Pr\{(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) \neq (\mathbf{X}, \mathbf{Y})\}, \quad (5)$$

where both the randomness of (\mathbf{X}, \mathbf{Y}) and the randomness of the encoder mappings are taken into account. The respective error exponent is defined as

$$E(R_x, R_y) = \lim_{n \rightarrow \infty} \left[-\frac{\log \bar{P}_e}{n} \right], \quad (6)$$

provided that the limit exists.

The optimal MAP decoder, that is cognizant of both P_{XY} and d , is given by

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \arg \max_{\{(\mathbf{x}', \mathbf{y}'): f(\mathbf{x}')=f(\mathbf{x}), g(\mathbf{y}')=g(\mathbf{y})\}} P_d(\mathbf{x}', \mathbf{y}') \quad (7)$$

$$= \arg \min_{\{(\mathbf{x}', \mathbf{y}'): f(\mathbf{x}')=f(\mathbf{x}), g(\mathbf{y}')=g(\mathbf{y})\}} \{-\log P_d(\mathbf{x}', \mathbf{y}')\}, \quad (8)$$

where

$$P_d(\mathbf{x}, \mathbf{y}) = P(y_1^d) \cdot P(x_1^{n-d}, y_{d+1}^n) \cdot P(x_{n-d+1}^n), \quad (9)$$

where all three factors admit product forms,

$$P(y_1^d) = \prod_{i=1}^d P_Y(y_i), \quad (10)$$

$$P(x_1^{n-d}, y_{d+1}^n) = \prod_{i=1}^{n-d} P_{XY}(x_i, y_{i+d}), \quad (11)$$

$$P(x_{n-d+1}^n) = \prod_{i=n-d+1}^n P_X(x_i). \quad (12)$$

The average probability of error, associated with the MAP decoder, will be denoted by $\bar{P}_{e,\star}$ and its error exponent will be denoted by $E_\star(R_x, R_y)$. A general metric decoder is of the form

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \arg \min_{\{(\mathbf{x}', \mathbf{y}'): f(\mathbf{x}')=f(\mathbf{x}), g(\mathbf{y}')=g(\mathbf{y})\}} q(\mathbf{x}', \mathbf{y}'), \quad (13)$$

where the function q will be referred to as the *decoding metric*. The average error probability of the decoder that is based on the metric q , will be denoted by $P_{e,q}$, and its error exponent (if existent) will be denoted by $E_q(R_x, R_y)$.

In this paper, we propose a universal decoding metric q , that is independent of the unknown P_{XY} and d , yet its error exponent, $E_q(R_x, R_y)$, coincides with $E_*(R_x, R_y)$, and hence it is asymptotically optimal in the random coding error–exponent sense.

4 Main Result

We define the following functions for $0 \leq k \leq n$:

$$u_k(\mathbf{x}, \mathbf{y}) = k\hat{H}(y_1^k) + (n-k)\hat{H}(x_1^{n-k}, y_{k+1}^n) + k\hat{H}(x_{n-k+1}^n), \quad (14)$$

$$v_k(\mathbf{x}, \mathbf{y}) = (n-k)\hat{H}(x_1^{n-k}|y_{k+1}^n) + k\hat{H}(x_{n-k+1}^n), \quad (15)$$

$$w_k(\mathbf{x}, \mathbf{y}) = k\hat{H}(y_1^k) + (n-k)\hat{H}(y_{k+1}^n|x_1^{n-k}), \quad (16)$$

$$q_k(\mathbf{x}, \mathbf{y}) = \max\{u_k(\mathbf{x}, \mathbf{y}) - n(R_x + R_y), v_k(\mathbf{x}, \mathbf{y}) - nR_x, w_k(\mathbf{x}, \mathbf{y}) - nR_y\}, \quad (17)$$

and finally, the universal decoding metric, q , is defined as

$$q(\mathbf{x}, \mathbf{y}) = \min_{0 \leq k \leq n} q_k(\mathbf{x}, \mathbf{y}). \quad (18)$$

If the relative delay, d , is allowed to take on also negative values, i.e., $-n \leq d \leq n$, then the minimum in (18) should be extended to $-n \leq k \leq n$, where for $k < 0$, u_k , v_k , w_k , and q_k are defined exactly as above, except that the roles of \mathbf{x} and \mathbf{y} are interchanged (that is, \mathbf{x} will be shifted $|k|$ positions to the right relative to \mathbf{y} , instead of the above shift, which is the opposite). For a pair of finite–alphabet RVs, $(X, Y) \sim P_{XY}$, let us define the Rényi entropies of order $\theta > 0$ as

$$H_\theta(X) = \frac{1}{1-\theta} \log \left\{ \sum_{x \in \mathcal{X}} [P_X(x)]^\theta \right\} \quad (19)$$

$$H_\theta(Y) = \frac{1}{1-\theta} \log \left\{ \sum_{y \in \mathcal{Y}} [P_Y(y)]^\theta \right\} \quad (20)$$

$$H_\theta(X, Y) = \frac{1}{1-\theta} \log \left\{ \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} [P_{XY}(x, y)]^\theta \right\} \quad (21)$$

$$H_\theta(X|Y) = \frac{\theta}{1-\theta} \log \left\{ \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} [P_{XY}(x, y)]^\theta \right]^{1/\theta} \right\} \quad (22)$$

$$H_\theta(Y|X) = \frac{\theta}{1-\theta} \log \left\{ \sum_{x \in \mathcal{X}} \left[\sum_{y \in \mathcal{Y}} [P_{XY}(x, y)]^\theta \right]^{1/\theta} \right\}. \quad (23)$$

For $\theta \rightarrow 1$, these quantities tend to the respective Shannon entropies.

Our main result is the following.

Theorem 1 *Under the assumptions formalized in Section 3, the following is true:*

(a) *The error exponents, $E_*(R_x, R_y)$ and $E_q(R_x, R_y)$, both exist.*

(b) *$E_q(R_x, R_y) = E_*(R_x, R_y) = \min\{E_{x|y}(R_x), E_{y|x}(R_y), E_{xy}(R_x, R_y)\}$, where*

$$E_{x|y}(R_x) = \max_{0 \leq \rho \leq 1} \rho \cdot \left[R_x - \delta H_{1/(1+\rho)}(X) - (1-\delta) H_{1/(1+\rho)}(X|Y) \right] \quad (24)$$

$$E_{y|x}(R_y) = \max_{0 \leq \rho \leq 1} \rho \cdot \left[R_y - \delta H_{1/(1+\rho)}(Y) - (1-\delta) H_{1/(1+\rho)}(Y|X) \right] \quad (25)$$

$$E_{xy}(R_x, R_y) = \max_{0 \leq \rho \leq 1} \rho \cdot \left[R_x + R_y - \delta H_{1/(1+\rho)}(X) - \delta H_{1/(1+\rho)}(Y) - (1-\delta) H_{1/(1+\rho)}(X, Y) \right]. \quad (26)$$

Discussion. The remaining part of this section is devoted to a discussion on Theorem 1 and its significance.

Since the error exponents were defined under the condition that the certain limits exist, part (a) of the theorem establishes the basic fact that they indeed exist. Part (b) is more quantitative: it tells that the error exponents of the universal decoder and the MAP decoder are equal, thus rendering the universal decoder asymptotically optimal in the error exponent sense. Finally, part (b) provides also an exact single-letter expression of this error exponent, using a Gallager-style formula. Here, unlike in the synchronous case (of $\delta = 0$), we also see unconditional Rényi entropies (weighted by δ), which correspond to the compression of the segments, y_1^d and x_{n-d+1}^n , that are independent of each other and of all other pieces of data within the block, and hence no correlations can be exploited when compressing them. If d is fixed (or grows sub-linearly with n), the relative weight of these segments is asymptotically negligible, and there is no asymptotic loss compared to the synchronous case. The error exponent is given by the minimum among three error exponents:

$E_{x|y}(R_x)$ corresponds to errors in the decoding of \mathbf{x} while \mathbf{y} is decoded correctly, $E_{y|x}(R_y)$ designates the opposite type of error, and finally, $E_{xy}(R_x, R_y)$ stands for erroneous decoding of both \mathbf{x} and \mathbf{y} . The smallest of all three dominates the overall error exponent.

The above relation between the error exponent and the coding rates can be essentially inverted, in order to answer the following question: what is the achievable rate region, $\mathcal{R}(E)$, for achieving an error exponent at least as large as a prescribed value, E ? Using the above error exponent formula, the answer is readily found² to be the following.

$$\mathcal{R}(E) = \{(R_x, R_y) : R_x \geq \mathbf{R}_x(E), R_y \geq \mathbf{R}_y(E), R_x + R_y \geq \mathbf{R}_{xy}(E)\}, \quad (27)$$

where

$$\mathbf{R}_x(E) = \inf_{s \geq 1} \left[sE + \delta H_{s/(1+s)}(X) + (1 - \delta) H_{s/(1+s)}(X|Y) \right] \quad (28)$$

$$\mathbf{R}_y(E) = \inf_{s \geq 1} \left[sE + \delta H_{s/(1+s)}(Y) + (1 - \delta) H_{s/(1+s)}(Y|X) \right] \quad (29)$$

$$\mathbf{R}_{xy}(E) = \inf_{s \geq 1} \left[sE + \delta H_{s/(1+s)}(X) + \delta H_{s/(1+s)}(Y) + (1 - \delta) H_{s/(1+s)}(X, Y) \right]. \quad (30)$$

For $E \rightarrow 0$, which means a vanishing error probability, however slowly, the infima are approached by $s \rightarrow \infty$, which yield

$$\mathbf{R}_x(0) = \delta H(X) + (1 - \delta) H(X|Y) = H(X|Y) + \delta I(X; Y) \quad (31)$$

$$\mathbf{R}_y(0) = \delta H(Y) + (1 - \delta) H(Y|X) = H(Y|X) + \delta I(X; Y) \quad (32)$$

$$\mathbf{R}_{xy}(0) = \delta H(X) + \delta H(Y) + (1 - \delta) H(X, Y) = H(X, Y) + \delta I(X; Y), \quad (33)$$

as expected in view of the results of [11].

In order to try to understand the decoding metric (18), consider the following observations. This decoding metric is given by the maximum of three different metrics, which are all in the spirit of the minimum entropy (ME) universal decoding metric,³ but modified to address the dependence structure at hand. Each one of these metrics is ‘responsible’ to handle a different type of error: $u_k(\mathbf{x}, \mathbf{y}) - n(R_x + R_y)$ is associated with errors in decoding both \mathbf{x} and \mathbf{y} , $v_k(\mathbf{x}, \mathbf{y}) - nR_x$ is for errors in \mathbf{x} only, while \mathbf{y} is decoded correctly, and finally, $w_k(\mathbf{x}, \mathbf{y}) - nR_y$ is meant for the opposite

²See the last part of Subsection 5.3.

³The above defined function, $u_k(\mathbf{x}, \mathbf{y})$, was mentioned also in [11, Section V, second paragraph] as a possible decoding metric, but it was not the decoding metric actually analyzed there, because the authors argued that it cannot be analyzed by the standard method of types.

case, of decoding error in \mathbf{y} only. The maximum of all three metrics is meant to handle all three types of error at the same time. Every value of k corresponds to a certain hypothesis concerning the relative delay. Note that this decoding metric is different from the one in [14], which relies on an encoding scheme that provides pointers to the type classes of \mathbf{x} and \mathbf{y} , in addition to their bin indices.

Another observation is regarding the special stature of the relative delay parameter, d . On the face of it, it is natural to view d as yet another unknown parameter of the source, in addition to the other unknown parameters – those associated with the joint distribution, P_{XY} . If d was known, and only P_{XY} was unknown, we could have interpreted the empirical entropies in u_k , v_k and w_k (actually, with $k = d$) as negative logarithms of the maximum likelihood (ML) values of the various segments, or equivalently, as the minima of the negative log-likelihood values. For example, $(n - k)\hat{H}(x_1^{n-d}, y_{d+1}^n) = \min_{P_{XY}}[-\log P(x_1^{n-d}, y_{d+1}^n)]$, $(n - k)\hat{H}(x_1^{n-d}|y_{d+1}^n) = \min_{P_{XY}}[-\log P(x_1^{n-d}|y_{d+1}^n)]$, and so on. In other words, the minima over $\{P_{XY}\}$ are taken *before* (i.e., more internally to) the maximum over the three metrics. By contrast, the minimum over the hypothesized relative delay, k , is taken *after* (i.e., externally to) the maximum over the three metrics. Attempts were made to prove that minimum over k and the maximum among the three metrics can be commuted, but to no avail. Therefore, this point seems to be non-trivial.

Finally, it is in order to say a few words concerning sources with memory. Consider the case where $Z_i = (X_i, Y_{i+d})$ is a first-order Markov source. In this case, the techniques of [12, Subsection 5.1], suggest that one can prove the universal asymptotic optimality of a similar universal decoder, where $k\hat{H}(y_1^k)$, $k\hat{H}(x_{n-k+1}^n)$, $(n - k)\hat{H}(x_{n-k+1}^n|y_{k+1}^n)$, $(n - k)\hat{H}(y_{k+1}^n|x_{n-k+1}^n)$, and $(n - k)\hat{H}(x_{n-k+1}^n, y_{k+1}^n)$ are replaced by the respective length functions associated with the Lempel–Ziv algorithm (LZ78) [22] and the conditional LZ78 algorithm [21], $\text{LZ}(y_1^k)$, $\text{LZ}(x_{n-k+1}^n)$, $\text{LZ}(x_{n-k+1}^n|y_{k+1}^n)$, $\text{LZ}(y_{k+1}^n|x_{n-k+1}^n)$, and $\text{LZ}(x_{n-k+1}^n, y_{k+1}^n)$. It should be noted that y_1^k and x_{n-k+1}^n are not realizations of a Markov sequences, but they are realizations of a hidden-Markov process, as their correlated counterparts are not available. Nonetheless, hidden Markov sources can still be accommodated in this framework (see, e.g., [13]).

5 Proof of Theorem 1

The proof is based on a simple sandwich argument: we first derive an upper bound to the average error probability of the universal decoder that is based on q , and then a lower bound to the error probability of the MAP decoder. Both bounds turn out to be of the same exponential order. On the other hand, since the MAP decoder cannot be worse than the universal decoder, this exponential order must be exact for both decoders and its single-letter expression is easily derived using the method of types. This will establish both part (a) of Theorem 1 and the first equality in part (b). The second equality in part (b) will be obtained by deriving the Lagrange-dual of the original single-letter formula.

5.1 Upper Bound on the Error Probability of the Universal Decoder

The average probability of error of the proposed universal decoder, is as follows.

$$\bar{P}_{e,q} = \sum_{\mathbf{x}, \mathbf{y}} P_d(\mathbf{x}, \mathbf{y}) \bar{P}_{e,q}(\mathbf{x}, \mathbf{y}) \quad (34)$$

$$\triangleq \sum_{\mathbf{x}, \mathbf{y}} P_d(\mathbf{x}, \mathbf{y}) \Pr \{(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) \neq (\mathbf{x}, \mathbf{y})\} \quad (35)$$

$$= \sum_{\mathbf{x}, \mathbf{y}} P_d(\mathbf{x}, \mathbf{y}) \cdot \Pr \left[\bigcup_{\{(\mathbf{x}', \mathbf{y}') \neq (\mathbf{x}, \mathbf{y}) : q(\mathbf{x}', \mathbf{y}') \leq q(\mathbf{x}, \mathbf{y})\}} \{f(\mathbf{x}') = f(\mathbf{x}), g(\mathbf{y}') = g(\mathbf{y})\} \right]. \quad (36)$$

As for $\bar{P}_{e,q}(\mathbf{x}, \mathbf{y})$, we have

$$\begin{aligned} \bar{P}_{e,q}(\mathbf{x}, \mathbf{y}) &= \Pr \left[\bigcup_{\{(\mathbf{x}', \mathbf{y}') \neq (\mathbf{x}, \mathbf{y}) : q(\mathbf{x}', \mathbf{y}') \leq q(\mathbf{x}, \mathbf{y})\}} \{f(\mathbf{x}') = f(\mathbf{x}), g(\mathbf{y}') = g(\mathbf{y})\} \right] \\ &\leq \Pr \left[\bigcup_{\{\mathbf{x}' \neq \mathbf{x}, \mathbf{y}' \neq \mathbf{y} : q(\mathbf{x}', \mathbf{y}') \leq q(\mathbf{x}, \mathbf{y})\}} \{f(\mathbf{x}') = f(\mathbf{x}), g(\mathbf{y}') = g(\mathbf{y})\} \right] + \\ &\quad \Pr \left[\bigcup_{\{\mathbf{x}' \neq \mathbf{x} : q(\mathbf{x}', \mathbf{y}') \leq q(\mathbf{x}, \mathbf{y})\}} \{f(\mathbf{x}') = f(\mathbf{x})\} \right] + \\ &\quad \Pr \left[\bigcup_{\{\mathbf{y}' \neq \mathbf{y} : q(\mathbf{x}, \mathbf{y}') \leq q(\mathbf{x}, \mathbf{y})\}} \{g(\mathbf{y}') = g(\mathbf{y})\} \right] \\ &\triangleq \bar{P}_{e,q,1}(\mathbf{x}, \mathbf{y}) + \bar{P}_{e,q,2}(\mathbf{x}, \mathbf{y}) + \bar{P}_{e,q,3}(\mathbf{x}, \mathbf{y}). \end{aligned} \quad (37)$$

Now,

$$\begin{aligned}
\bar{P}_{e,q,1}(\mathbf{x}, \mathbf{y}) &= \Pr \left[\bigcup_{\{\tilde{\mathbf{x}} \neq \mathbf{x}, \tilde{\mathbf{y}} \neq \mathbf{y}: q(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q(\mathbf{x}, \mathbf{y})\}} \{f(\tilde{\mathbf{x}}) = f(\mathbf{x}), g(\tilde{\mathbf{y}}) = g(\mathbf{y})\} \right] \\
&\leq \min \left\{ 1, 2^{-n(R_x + R_y)} \left| \{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) : q(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q(\mathbf{x}, \mathbf{y})\} \right| \right\} \\
&= \min \left\{ 1, 2^{-n(R_x + R_y)} \left| \bigcup_{k=0}^{n-1} \{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q(\mathbf{x}, \mathbf{y})\} \right| \right\} \\
&\leq \min \left\{ 1, 2^{-n(R_x + R_y)} \left| \bigcup_{k=0}^{n-1} \{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\} \right| \right\} \\
&\leq \min \left\{ 1, 2^{-n(R_x + R_y)} \sum_{k=0}^{n-1} \left| \{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\} \right| \right\} \\
&\leq \min \left\{ 1, 2^{-n(R_x + R_y)} \sum_{k=0}^{n-1} \sum_{\{(\mathcal{T}(\tilde{x}_1^{n-k}, \tilde{y}_{k+1}^n), \mathcal{T}(\tilde{y}_1^k), \mathcal{T}(\tilde{x}_{n-k+1}^n) : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} \right. \\
&\quad \left. |\mathcal{T}(\tilde{x}_1^{n-k}, \tilde{y}_{k+1}^n)| \cdot |\mathcal{T}(\tilde{y}_1^k)| \cdot |\mathcal{T}(\tilde{x}_{n-k+1}^n)| \right\} \\
&\leq \min \left\{ 1, 2^{-n(R_x + R_y)} \sum_{k=0}^{n-1} \sum_{\{(\mathcal{T}(\tilde{x}_1^{n-k}, \tilde{y}_{k+1}^n), \mathcal{T}(\tilde{y}_1^k), \mathcal{T}(\tilde{x}_{n-k+1}^n) : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} \right. \\
&\quad \left. \exp_2\{(n-k)\hat{H}(\tilde{x}_1^{n-k}, \tilde{y}_{k+1}^n) + k\hat{H}(\tilde{y}_1^k) + k\hat{H}(\tilde{x}_{n-k+1}^n)\} \right\} \\
&= \min \left\{ 1, 2^{-n(R_x + R_y)} \sum_{k=0}^{n-1} \sum_{\{(\mathcal{T}(\tilde{x}_1^{n-k}, \tilde{y}_{k+1}^n), \mathcal{T}(\tilde{y}_1^k), \mathcal{T}(\tilde{x}_{n-k+1}^n) : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} 2^{u_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})} \right\} \\
&\doteq \min \left\{ 1, \max_{0 \leq k \leq n} \max_{\{\hat{P}_{\tilde{y}_1^k}, \hat{P}_{\tilde{x}_1^{n-d} \tilde{y}_{d+1}^n}, \hat{P}_{\tilde{x}_{n-d+1}^n} : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} \right. \\
&\quad \left. \exp_2[u_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - n(R_x + R_y)] \right\}. \tag{38}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\bar{P}_{e,q,2}(\mathbf{x}, \mathbf{y}) &= \Pr \left[\bigcup_{\{\tilde{\mathbf{x}} \neq \mathbf{x}: q(\tilde{\mathbf{x}}, \mathbf{y}) \leq q(\mathbf{x}, \mathbf{y})\}} \{f(\tilde{\mathbf{x}}) = f(\mathbf{x})\} \right] \\
&\leq \min \left\{ 1, 2^{-nR_x} \left| \{\tilde{\mathbf{x}} : q(\tilde{\mathbf{x}}, \mathbf{y}) \leq q(\mathbf{x}, \mathbf{y})\} \right| \right\} \\
&\leq \min \left\{ 1, 2^{-nR_x} \left| \bigcup_{k=0}^{n-1} \{\tilde{\mathbf{x}} : q_k(\tilde{\mathbf{x}}, \mathbf{y}) \leq q_d(\mathbf{x}, \mathbf{y})\} \right| \right\}
\end{aligned}$$

$$\begin{aligned}
&\leq \min \left\{ 1, 2^{-nR_x} \sum_{k=0}^{n-1} \left| \{ \tilde{\mathbf{x}} : q_k(\tilde{\mathbf{x}}, \mathbf{y}) \leq q_d(\mathbf{x}, \mathbf{y}) \} \right| \right\} \\
&\leq \min \left\{ 1, 2^{-nR_x} \sum_{k=0}^{n-1} \sum_{\{(\mathcal{T}(\tilde{x}_1^{n-k}|y_{k+1}^n), \mathcal{T}(\tilde{x}_{n-k+1}^n) : q_k(\tilde{\mathbf{x}}, \mathbf{y}) \leq q_d(\mathbf{x}, \mathbf{y}))\}} \right. \\
&\quad \left. |\mathcal{T}(\tilde{x}_1^{n-k}|y_{k+1}^n)| \cdot |\mathcal{T}(\tilde{x}_{n-k+1}^n)| \right\} \\
&\leq \min \left\{ 1, 2^{-nR_x} \sum_{k=0}^{n-1} \sum_{\{(\mathcal{T}(\tilde{x}_1^{n-k}|y_{k+1}^n), \mathcal{T}(\tilde{x}_{n-k+1}^n) : q_k(\tilde{\mathbf{x}}, \mathbf{y}) \leq q_d(\mathbf{x}, \mathbf{y}))\}} \right. \\
&\quad \left. \exp_2\{(n-k)\hat{H}(\tilde{x}_1^{n-k}|y_{k+1}^n) + k\hat{H}(\tilde{x}_{n-k+1}^n)\} \right\} \\
&= \min \left\{ 1, 2^{-nR_x} \sum_{k=0}^{n-1} \sum_{\{(\mathcal{T}(\tilde{x}_1^{n-k}|y_{k+1}^n), \mathcal{T}(\tilde{x}_{n-k+1}^n) : q_k(\tilde{\mathbf{x}}, \mathbf{y}) \leq q_d(\mathbf{x}, \mathbf{y}))\}} 2^{v_k(\tilde{\mathbf{x}}, \mathbf{y})} \right\} \\
&\doteq \min \left\{ 1, \max_{0 \leq k \leq n} \max_{\{\hat{P}_{\tilde{x}_1^{n-k}, y_{k+1}^n}, \hat{P}_{\tilde{x}_{n-k+1}^n} : q_k(\tilde{\mathbf{x}}, \mathbf{y}) \leq q_d(\mathbf{x}, \mathbf{y})\}} \exp_2[v_k(\tilde{\mathbf{x}}, \mathbf{y}) - nR_x] \right\} \\
&\leq \min \left\{ 1, \max_{0 \leq k \leq n} \max_{\{\hat{P}_{\tilde{y}_1^k}, \hat{P}_{\tilde{x}_1^{n-d}\tilde{y}_{d+1}^n}, \hat{P}_{\tilde{x}_{n-d+1}^n} : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} \right. \\
&\quad \left. \exp_2[v_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - nR_x] \right\}, \tag{39}
\end{aligned}$$

and in exactly the same manner,

$$\bar{P}_{e,q,3}(\mathbf{x}, \mathbf{y}) \leq \min \left\{ 1, \max_{0 \leq k \leq n} \max_{\{\hat{P}_{\tilde{y}_1^k}, \hat{P}_{\tilde{x}_1^{n-d}\tilde{y}_{d+1}^n}, \hat{P}_{\tilde{x}_{n-d+1}^n} : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} \exp_2[w_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - nR_y] \right\}. \tag{40}$$

We now use the following simple inequality that holds for every non-negative reals, a , b , and c :

$$\min\{1, a\} + \min\{1, b\} + \min\{1, c\} \leq 3 \max\{\min\{1, a\}, \min\{1, b\}, \min\{1, c\}\} \tag{41}$$

$$= \begin{cases} 3 & a > 1 \text{ or } b > 1 \text{ or } c > 1 \\ 3 \cdot \max\{a, b, c\} & a \leq 1 \text{ and } b \leq 1 \text{ and } c \leq 1 \end{cases} \tag{42}$$

$$= \begin{cases} 3 & \max\{a, b, c\} > 1 \\ 3 \cdot \max\{a, b, c\} & \max\{a, b, c\} \leq 1 \end{cases} \tag{43}$$

$$= 3 \cdot \min\{1, \max\{a, b, c\}\}. \tag{44}$$

It follows that

$$\bar{P}_{e,q}(\mathbf{x}, \mathbf{y}) \leq 3 \cdot \min \left\{ 1, \max_{0 \leq k \leq n} \max_{\{\hat{P}_{\tilde{y}_1^k}, \hat{P}_{\tilde{x}_1^{n-d}\tilde{y}_{d+1}^n}, \hat{P}_{\tilde{x}_{n-d+1}^n} : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} \right\}$$

$$\exp_2\{\max\{u_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - n(R_x + R_y), v_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - nR_x, w_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - nR_y\}\} \quad (45)$$

$$\doteq \min \left\{ 1, \max_{0 \leq k \leq n} \left\{ \max_{\{\hat{P}_{\tilde{\mathbf{y}}_1^k}, \hat{P}_{\tilde{\mathbf{x}}_1^{n-d}}, \hat{P}_{\tilde{\mathbf{y}}_{d+1}^n}, \hat{P}_{\tilde{\mathbf{x}}_{n-d+1}^n\}} : q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq q_d(\mathbf{x}, \mathbf{y})\}} 2^{q_k(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})} \right\} \right\} \quad (46)$$

$$\leq \min \{1, 2^{q_d(\mathbf{x}, \mathbf{y})}\}. \quad (47)$$

Finally, the overall average error probability, associated with the proposed universal decoder, is exponentially upper bounded by

$$\bar{P}_{e,q} \leq \mathbf{E} \min \{1, \exp_2[q_d(\mathbf{X}, \mathbf{Y})]\}. \quad (48)$$

5.2 Lower Bound on the Error Probability of the MAP Decoder

For the MAP decoder, the conditional average probability of error, for a given (\mathbf{x}, \mathbf{y}) , is as follows:

$$\begin{aligned} \bar{P}_{e,*}(\mathbf{x}, \mathbf{y}) &= \Pr \left[\bigcup_{\{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \neq (\mathbf{x}, \mathbf{y}) : P_d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \geq P_d(\mathbf{x}, \mathbf{y})\}} \{f(\tilde{\mathbf{x}}) = f(\mathbf{x}), g(\tilde{\mathbf{y}}) = g(\mathbf{y})\} \right] \\ &\geq \Pr \left[\bigcup_{\{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in \mathcal{S}_o(\mathbf{x}, \mathbf{y})\}} \{f(\tilde{\mathbf{x}}) = f(\mathbf{x}), g(\tilde{\mathbf{y}}) = g(\mathbf{y})\} \right], \end{aligned} \quad (49)$$

where

$$\mathcal{S}_o(\mathbf{x}, \mathbf{y}) = \{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \neq (\mathbf{x}, \mathbf{y}) : \tilde{y}_1^d \in \mathcal{T}(y_1^d), (\tilde{x}_1^{n-d}, \tilde{y}_{d+1}^n) \in \mathcal{T}(x_1^{n-d}, y_{d+1}^n), \tilde{x}_{n-d+1}^n \in \mathcal{T}(x_{n-d+1}^n)\}. \quad (50)$$

To further lower bound the conditional average probability of error, associated with the MAP decoder, we need the following lemma, whose proof is deferred to the appendix.

Lemma 1 *Let \mathcal{S}_o be a set of pairs of integers, $\{(i, j)\}$, with the following properties.*

1. *The pair $(0, 0)$ is a not member of \mathcal{S}_o .*
2. *For a given i , let $\mathcal{S}_{i*} = \{j : (i, j) \in \mathcal{S}_o\}$ and $\mathcal{S}_{*j} = \{i : (i, j) \in \mathcal{S}_o\}$. We assume that $|\mathcal{S}_{i*}| = \ell$ for all i such that $(i, j) \in \mathcal{S}_o$ for some j . Likewise, $|\mathcal{S}_{*j}| = k$ for all j such that $(i, j) \in \mathcal{S}_o$ for some i . Here, ℓ and k are fixed positive integers.*
3. *For every $(i, j) \in \mathcal{S}_o$, there is an event \mathcal{C}_{ij} , defined as $\mathcal{C}_{ij} = \mathcal{A}_i \cap \mathcal{B}_j$, where $\{\mathcal{A}_i\}$ and $\{\mathcal{B}_j\}$ are sequences of mutually independent events.*

4. The probabilities of $\{\mathcal{A}_i\}$ are given by $P[\mathcal{A}_0] = 1$ and $P[\mathcal{A}_i] = \alpha$ for all $i \neq 0$. Here, α is a fixed number in $[0, 1]$.

5. The probabilities of $\{\mathcal{B}_j\}$ are given by $P[\mathcal{B}_0] = 1$ and $P[\mathcal{B}_j] = \beta$ for all $j \neq 0$. Here, β is a fixed number in $[0, 1]$.

Then, under conditions 1–5,

$$P \left[\bigcup_{(i,j) \in \mathcal{S}_o} \mathcal{C}_{ij} \right] \geq \frac{1}{4} \cdot \min \{1, \max \{k\alpha, \ell\beta, (M - k - \ell)\alpha\beta\}\}, \quad (51)$$

where $M = |\mathcal{S}_o|$.

We apply Lemma 1 using the following assignments: $i = \tilde{\mathbf{x}}$, $j = \tilde{\mathbf{y}}$, and so, $i = 0$ and $j = 0$ correspond to $\tilde{\mathbf{x}} = \mathbf{x}$ and $\tilde{\mathbf{y}} = \mathbf{y}$, respectively. The event \mathcal{A}_i is $\{f(\tilde{\mathbf{x}}) = f(\mathbf{x})\}$, where \mathbf{x} and $f(\mathbf{x})$ are given. Thus, obviously $P[\mathcal{A}_0] = 1$. Likewise, the event \mathcal{B}_j is $\{g(\tilde{\mathbf{y}}) = g(\mathbf{y})\}$ for a given \mathbf{y} and $g(\mathbf{y})$ and $P[\mathcal{B}_0] = 1$. It follows then that $\alpha = 2^{-nR_x}$, $\beta = 2^{-nR_y}$, and $\mathcal{S}_o = \mathcal{S}_o(\mathbf{x}, \mathbf{y})$. Here,

$$M \doteq \exp_2\{(n-d)\hat{H}(x_1^{n-d}, y_{d+1}^n) + d\hat{H}(x_{n-d+1}^n) + d\hat{H}(y_1^d)\} = 2^{u_d(\mathbf{x}, \mathbf{y})} \quad (52)$$

$$k \doteq \exp_2\{(n-d)\hat{H}(x_1^{n-d}|y_{d+1}^n) + d\hat{H}(x_{n-d+1}^n)\} = 2^{v_d(\mathbf{x}, \mathbf{y})} \quad (53)$$

$$\ell \doteq \exp_2\{d\hat{H}(y_1^d) + (n-d)\hat{H}(y_{d+1}^n|x_1^{n-d})\} = 2^{w_d(\mathbf{x}, \mathbf{y})}. \quad (54)$$

Thus, according to Lemma 1, we obtain the matching lower bound,

$$\bar{P}_{e, \star}(\mathbf{x}, \mathbf{y}) \geq \min \left\{ 1, \max \left\{ 2^{-nR_x} \cdot 2^{v_d(\mathbf{x}, \mathbf{y})}, 2^{-nR_y} \cdot 2^{w_d(\mathbf{x}, \mathbf{y})}, 2^{-n(R_x+R_y)} \cdot 2^{u_d(\mathbf{x}, \mathbf{y})} \right\} \right\} \quad (55)$$

$$\doteq \min \{1, \exp_2[q_d(\mathbf{x}, \mathbf{y})]\}, \quad (56)$$

and so, the overall average error probability of the MAP decoder is exponentially lower bounded by

$$\bar{P}_{e, \star} \geq \mathbf{E} \min \{1, \exp_2[q_d(\mathbf{X}, \mathbf{Y})]\}, \quad (57)$$

which matches the upper bound of the universal decoder in eq. (48), as far as the exponential order goes.

5.3 The Error Exponent Formula

From the previous subsections, we learn that both decoders have an average error probability of the exponential order of $E \min\{1, \exp_2\{q_d(\mathbf{X}, \mathbf{Y})\}\}$. Standard analysis of this quantity, using the well known method of types [5], yields the following single-letter expression:

$$E_*(R_x, R_y) = E_q(R_x, R_y) = \min\{E_{xy}(R_x, R_y), E_{x|y}(R_x), E_{y|x}(R_y)\}, \quad (58)$$

where

$$\begin{aligned} E_{xy}(R_x, R_y) &= \min_{Q_{X'}, Q_{Y'}, Q_{XY}} \left\{ \delta D(Q_{X'} \| P_X) + \delta D(Q_{Y'} \| P_Y) + (1 - \delta) D(Q_{X'Y'} \| P_{XY}) + \right. \\ &\quad \left. [R_x + R_y - \delta H_Q(X') - \delta H_Q(Y') - (1 - \delta) H_Q(X, Y)]_+ \right\} \\ E_{x|y}(R_x) &= \min_{Q_{X'}, Q_{XY}} \left\{ \delta D(Q_{X'} \| P_X) + (1 - \delta) D(Q_{X'Y'} \| P_{XY}) + \right. \\ &\quad \left. [R_x - \delta H_Q(X') - (1 - \delta) H_Q(X|Y)]_+ \right\} \end{aligned} \quad (59)$$

$$\begin{aligned} E_{y|x}(R_y) &= \min_{Q_{Y'}, Q_{XY}} \left\{ \delta D(Q_{Y'} \| P_Y) + (1 - \delta) D(Q_{X'Y'} \| P_{XY}) + \right. \\ &\quad \left. [R_y - \delta H_Q(Y') - (1 - \delta) H_Q(Y|X)]_+ \right\}. \end{aligned} \quad (60)$$

To find the Lagrange-dual of $E_{xy}(R_x, R_y)$, we proceed as follows.

$$\begin{aligned} E_{xy}(R_x, R_y) &= \min_{Q_{X'}, Q_{Y'}, Q_{XY}} \left\{ \delta D(Q_{X'} \| P_X) + \delta D(Q_{Y'} \| P_Y) + (1 - \delta) D(Q_{X'Y'} \| P_{XY}) + \right. \\ &\quad \left. [R_x + R_y - \delta H_Q(X') - \delta H_Q(Y') - (1 - \delta) H_Q(X, Y)]_+ \right\} \\ &= \min_{Q_{X'}, Q_{Y'}, Q_{XY}} \max_{0 \leq \rho \leq 1} \left\{ \delta D(Q_{X'} \| P_X) + \delta D(Q_{Y'} \| P_Y) + (1 - \delta) D(Q_{X'Y'} \| P_{XY}) + \right. \\ &\quad \left. \rho [R_x + R_y - \delta H_Q(X') - \delta H_Q(Y') - (1 - \delta) H_Q(X, Y)] \right\} \\ &\stackrel{(a)}{=} \max_{0 \leq \rho \leq 1} \left[\rho (R_x + R_y) + \min_{Q_{X'}, Q_{Y'}, Q_{XY}} \left\{ \delta D(Q_{X'} \| P_X) + \right. \right. \\ &\quad \left. \delta D(Q_{Y'} \| P_Y) + (1 - \delta) D(Q_{X'Y'} \| P_{XY}) + \right. \\ &\quad \left. \rho [R_x + R_y - \delta H_Q(X') - \delta H_Q(Y') - (1 - \delta) H_Q(X, Y)] \right\} \Big] \\ &= \max_{0 \leq \rho \leq 1} \left[\rho (R_x + R_y) + \left\{ \delta \cdot \min_{Q_{X'}} [D(Q_{X'} \| P_X) - \rho H_Q(X')] + \right. \right. \\ &\quad \left. \delta \cdot \min_{Q_{Y'}} [D(Q_{Y'} \| P_Y) - \rho H_Q(Y')] + \right. \\ &\quad \left. (1 - \delta) \cdot \min_{Q_{XY}} [D(Q_{XY} \| P_{XY}) - \rho H_Q(X, Y)] \right\} \Big] \end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{=} \max_{0 \leq \rho \leq 1} \left\{ \rho(R_x + R_y) - \delta \log \left(\sum_x P_X(x)^{1/(1+\rho)} \right)^{1+\rho} - \right. \\
&\quad \left. \delta \log \left(\sum_y P_Y(y)^{1/(1+\rho)} \right)^{1+\rho} - (1-\delta) \log \left(\sum_{x,y} P_{XY}(x,y)^{1/(1+\rho)} \right)^{1+\rho} \right\} \\
&= \max_{0 \leq \rho \leq 1} \rho \cdot \left[R_x + R_y - \delta H_{1/(1+\rho)}(X) - \right. \\
&\quad \left. \delta H_{1/(1+\rho)}(Y) - (1-\delta) H_{1/(1+\rho)}(X, Y) \right], \tag{61}
\end{aligned}$$

where in (a) we invoked the minimax theorem for convex–concave functions, and in (b) we carried out the minimizations using standard methods. The Lagrange–duals of $E_{x|y}(R_x)$ and $E_{y|x}(R_y)$ are obtained in a similar fashion.

Finally, for an error exponent level, E , to be achievable by the random code, $E_{xy}(R_x, R_y)$, $E_{x|y}(R_x)$ and $E_{y|x}(R_y)$ must all be at least as large as E at the same time. The condition $E_{xy}(R_x, R_y) \geq E$ is equivalent to the condition

$$\exists 0 \leq \rho \leq 1 \quad \rho \left[(R_x + R_y) - \delta H_{1/(1+\rho)}(X) - \delta H_{1/(1+\rho)}(Y) - (1-\delta) H_{1/(1+\rho)}(X, Y) \right] \geq E \tag{62}$$

or, equivalently,

$$\exists 0 \leq \rho \leq 1 \quad R_x + R_y \geq \frac{E}{\rho} + \delta H_{1/(1+\rho)}(X) + \delta H_{1/(1+\rho)}(Y) + (1-\delta) H_{1/(1+\rho)}(X, Y), \tag{63}$$

which is the same⁴ as

$$\exists s \geq 1 \quad R_x + R_y \geq sE + \delta H_{s/(1+s)}(X) + \delta H_{s/(1+s)}(Y) + (1-\delta) H_{s/(1+s)}(X, Y), \tag{64}$$

or

$$R_x + R_y \geq \inf_{s \geq 1} \left[sE + \delta H_{s/(1+s)}(X) + \delta H_{s/(1+s)}(Y) + (1-\delta) H_{s/(1+s)}(X, Y) \right]. \tag{65}$$

Similarly, the requirements that $E_{x|y}(R_x) \geq E$ and $E_{y|x}(R_y) \geq E$ yield the individual lower bounds on R_x and R_y , that together form the achievable rate region, $\mathbf{R}(E)$, as defined.

Appendix - Proof of Lemma 1

Let us partition \mathcal{S}_o into three disjoint subsets whose union is equal to \mathcal{S}_o . These subsets are

$$\mathcal{S}_{0*} = \{(0, j) \in \mathcal{S}_o\} \tag{A.1}$$

⁴Change the variable ρ to $s = 1/\rho$.

$$\mathcal{S}_{*0} = \{(i, 0) \in \mathcal{S}_o\} \quad (\text{A.2})$$

$$\mathcal{S} = \{(i, j) \in \mathcal{S}_o : i \neq 0, j \neq 0\}, \quad (\text{A.3})$$

Clearly,

$$P \left[\bigcup_{(i,j) \in \mathcal{S}_o} \mathcal{C}_{ij} \right] \geq \max \left\{ P \left[\bigcup_{(i,j) \in \mathcal{S}_{*0}} \mathcal{C}_{ij} \right], P \left[\bigcup_{(i,j) \in \mathcal{S}_{0*}} \mathcal{C}_{ij} \right], P \left[\bigcup_{(i,j) \in \mathcal{S}} \mathcal{C}_{ij} \right] \right\}. \quad (\text{A.4})$$

As for the first term on the r.h.s., we have

$$P \left[\bigcup_{(i,j) \in \mathcal{S}_{*0}} \mathcal{C}_{ij} \right] = P \left[\bigcup_i \mathcal{A}_i \cap \mathcal{B}_0 \right] \quad (\text{A.5})$$

$$= P \left[\mathcal{B}_0 \cap \left(\bigcup_i \mathcal{A}_i \right) \right] \quad (\text{A.6})$$

$$= P[\mathcal{B}_0] \cdot P \left[\bigcup_i \mathcal{A}_i \right] \quad (\text{A.7})$$

$$= P \left[\bigcup_i \mathcal{A}_i \right] \quad (\text{A.8})$$

$$\geq \frac{1}{2} \cdot \min\{1, k\alpha\}, \quad (\text{A.9})$$

where the last step follows from de Caen's lower bound [6] on the probability of a union of a finite set of events, $\{\mathcal{E}_i, i \in \mathcal{I}\}$:

$$P \left[\bigcup_{i \in \mathcal{I}} \mathcal{E}_i \right] \geq \sum_{i \in \mathcal{I}} \frac{P^2[\mathcal{E}_i]}{\sum_{j \in \mathcal{I}} P[\mathcal{E}_i \cap \mathcal{E}_j]}, \quad (\text{A.10})$$

which in the case of pairwise independent events, simplifies to

$$\begin{aligned} P \left[\bigcup_{i \in \mathcal{I}} \mathcal{E}_i \right] &\geq \sum_{i \in \mathcal{I}} \frac{P^2[\mathcal{E}_i]}{P[\mathcal{E}_i] + \sum_{j \in \mathcal{I}} P[\mathcal{E}_i] \cdot P[\mathcal{E}_j]} \\ &= \sum_{i \in \mathcal{I}} \frac{P[\mathcal{E}_i]}{1 + \sum_{j \in \mathcal{I}} P[\mathcal{E}_j]} \\ &= \frac{\sum_{i \in \mathcal{I}} P[\mathcal{E}_i]}{1 + \sum_{i \in \mathcal{I}} P[\mathcal{E}_i]} \\ &\geq \frac{\sum_{i \in \mathcal{I}} P[\mathcal{E}_i]}{2 \cdot \max\{1, \sum_{i \in \mathcal{I}} P[\mathcal{E}_i]\}} \\ &= \frac{1}{2} \cdot \min \left\{ 1, \sum_{i \in \mathcal{I}} P[\mathcal{E}_i] \right\}. \end{aligned} \quad (\text{A.11})$$

Similarly,

$$P \left[\bigcup_{(i,j) \in \mathcal{S}_{*0}} \mathcal{C}_{ij} \right] \geq \frac{1}{2} \cdot \min\{1, \ell\beta\}. \quad (\text{A.12})$$

Moving on to the union over \mathcal{S} , let us denote $\mathcal{U} = \{i : \exists j (i, j) \in \mathcal{S}\}$ and $\mathcal{V} = \{j : \exists i (i, j) \in \mathcal{S}\}$, with cardinalities K and L , respectively. Then, by property 2 of \mathcal{S}_\circ , $M' \triangleq |\mathcal{S}| = K\ell = kL$.⁵ Applying again de Caen's lower bound, this time, to the union over \mathcal{S} , we have

$$\begin{aligned}
P \left[\bigcup_{(i,j) \in \mathcal{S}} \mathcal{C}_{ij} \right] &\geq \sum_{(i,j) \in \mathcal{S}} \frac{P^2(\mathcal{C}_{ij})}{P(\mathcal{C}_{ij}) + \sum_{(i',j') \in \mathcal{S} \setminus \{(i,j)\}} P(\mathcal{C}_{ij} \cap \mathcal{C}_{i'j'})} \\
&= \sum_{(i,j) \in \mathcal{S}} P^2(\mathcal{A}_i) P^2(\mathcal{B}_j) \cdot \left(P(\mathcal{A}_i) P(\mathcal{B}_j) + \sum_{i' \in \mathcal{S}_j \setminus \{i\}} P[\mathcal{A}_i \cap \mathcal{A}_{i'} \cap \mathcal{B}_j] + \right. \\
&\quad \left. \sum_{j' \in \mathcal{S}_i \setminus \{j\}} P[\mathcal{A}_i \cap \mathcal{B}_j \cap \mathcal{B}_{j'}] + \sum_{\{(i',j') \in \mathcal{S} : i' \neq i, j' \neq j\}} P[\mathcal{A}_i \cap \mathcal{A}_{i'} \cap \mathcal{B}_j \cap \mathcal{B}_{j'}] \right)^{-1} \\
&\geq \sum_{(i,j) \in \mathcal{S}} \frac{\alpha^2 \beta^2}{\alpha\beta + k\alpha^2\beta + \ell\alpha\beta^2 + M'\alpha^2\beta^2} \\
&= \sum_{(i,j) \in \mathcal{S}} \frac{\alpha\beta}{1 + k\alpha + \ell\beta + M'\alpha\beta} \\
&= \frac{M'\alpha\beta}{1 + k\alpha + \ell\beta + M'\alpha\beta} \\
&\geq \frac{M'\alpha\beta}{4 \cdot \max\{1, k\alpha, \ell\beta, M'\alpha\beta\}} \\
&= \frac{1}{4} \cdot \min \left\{ 1, \frac{M'\alpha}{\ell}, \frac{M'\beta}{k}, M'\alpha\beta \right\} \\
&= \frac{1}{4} \cdot \min \{1, K\alpha, L\beta, M'\alpha\beta\}. \tag{A.13}
\end{aligned}$$

Thus, overall we have

$$P \left[\bigcup_{(i,j) \in \mathcal{S}_\circ} \mathcal{C}_{ij} \right] \geq \frac{1}{4} \cdot \max \{ \min\{1, k\alpha\}, \min\{1, \ell\beta\}, \min\{1, K\alpha, L\beta, M'\alpha\beta\} \}. \tag{A.14}$$

Now, consider the following line of thought: if $M'\alpha\beta \geq K\alpha$, which is equivalent to $\ell\beta \geq 1$, then the lower bound is at least as large as $1/4$. Similarly, if $M'\alpha\beta \geq L\beta$, which is equivalent to $k\alpha \geq 1$, then again the lower bound is at least $1/4$. Thus,

$$P \left[\bigcup_{(i,j) \in \mathcal{S}_\circ} \mathcal{C}_{ij} \right] \geq \frac{1}{4} \cdot \begin{cases} 1 & k\alpha \geq 1 \text{ or } \ell\beta \geq 1 \\ \max\{k\alpha, \ell\beta, \min\{1, M'\alpha\beta\}\} & \text{otherwise} \end{cases} \tag{A.15}$$

$$= \frac{1}{4} \cdot \begin{cases} 1 & k\alpha \geq 1 \text{ or } \ell\beta \geq 1 \text{ or } M'\alpha\beta \geq 1 \\ \max\{k\alpha, \ell\beta, M'\alpha\beta\} & \text{otherwise} \end{cases} \tag{A.16}$$

⁵Note that in general, $M' \leq KL$. For example, if $\mathcal{S} = \{(1, 1), (2, 1), (2, 2), (3, 2), (3, 3), (1, 3)\}$, then $M' = |\mathcal{S}| = 6$, $k = \ell = 2$, and $\mathcal{U} = \mathcal{V} = \{1, 2, 3\}$, so $K = L = 3$. As another example: let (i, j) designate indexes of finite-alphabet n -sequences and let \mathcal{S} be a joint type, $\mathcal{T}(i, j)$, then $\mathcal{U} = \mathcal{T}(i)$, $\mathcal{S}_i = \mathcal{T}(j|i)$, $\mathcal{V} = \mathcal{T}(j)$ and $\mathcal{S}_j = \mathcal{T}(i|j)$, but $\mathcal{T}(i, j)$ is, in general, only a subset of $\mathcal{T}(i) \times \mathcal{T}(j)$.

$$= \frac{1}{4} \cdot \begin{cases} 1 & \max\{k\alpha, \ell\beta, M'\alpha\beta\} \geq 1 \\ \max\{k\alpha, \ell\beta, M'\alpha\beta\} & \max\{k\alpha, \ell\beta, M'\alpha\beta\} < 1 \end{cases} \quad (\text{A.17})$$

$$= \frac{1}{4} \cdot \min \{1, \max\{k\alpha, \ell\beta, M'\alpha\beta\}\}, \quad (\text{A.18})$$

and the proof of the lemma is completed upon observing that $M' = M - k - \ell$.

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, “Towards a general theory of source networks,” *IEEE Trans. Inform. Theory*, vol. IT-26, no. 2, pp. 155–165, March 1980.
- [5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981. Second Edition: Cambridge University Press, New York, 2011.
- [6] D. de Caen, “A lower bound on the probability of a union,” *Discrete Mathematics*, vol. 169, pp. 217–220, 1997.
- [7] S. C. Draper, “Universal incremental Slepian–Wolf coding,” *Proc. 42nd Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, USA, October 2004.
- [8] J. C. Kieffer, “Some universal noiseless multiterminal source coding theorems,” *Information and Control*, vol. 46, pp. 93–107, 1980.
- [9] T. Matsuta and T. Uyematsu, “Universal coding for asynchronous Slepian–Wolf coding systems,” *IEICE Tech. Rep.*, vol. 112, no. 460, pp. 1–6, March 2013.
- [10] T. Matsuta and T. Uyematsu, “Achievable rate regions for asynchronous Slepian–Wolf coding systems,” *Proc. 2015 IEEE Workshop on Information Theory*, pp. 312–316, October 2015.
- [11] T. Matsuta and T. Uyematsu, “Coding theorems for asynchronous Slepian–Wolf coding systems,” *IEEE Trans. Inform. Theory*, vol. 66, no. 8, pp. 4774–4795, August 2020.

- [12] N. Merhav, “Universal decoding for source–channel coding with side information,” *Communications in Information and Systems*, vol. 16, no. 1, pp. 17–58, 2016.
- [13] N. Merhav, “Universal decoding using a noisy codebook,” *IEEE Trans. Inform. Theory*, vol. 64, part 1, no. 4, pp. 2231–2239, April 2018.
- [14] 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.
- [15] N. Oki and Y. Oohama, “Coding for the asynchronous Slepian–Wolf data compression,” *Proc. 20th Symposium on Information Theory and its Applications (SITA ‘97)*, pp. 89–92, December 1997.
- [16] B. Rimoldi and R. Urbanke, “Asynchronous Slepian–Wolf coding via source–splitting,” *Proc. 1997 IEEE International Symposium on Information Theory (ISIT ‘97)*, p. 271, Ulm, Germany, June 1997.
- [17] 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.
- [18] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inform. Theory*, vol. IT-19, no. 4, pp. 471–480, July 1973.
- [19] Z. Sun, C. Tian, J. Chen, and K. M. Wong, “LDPC code design for asynchronous Slepian–Wolf coding,” *IEEE Trans. Commun.*, vol. 58, no. 2, pp. 511–520, February 2010.
- [20] F. M. J. Willems, “Totally asynchronous Slepian–Wolf data compression,” *IEEE Trans. Inform. Theory*, vol. 34, no. 1, pp. 35–44, January 1988.
- [21] J. Ziv, “Universal decoding for finite-state channels,” *IEEE Trans. Inform. Theory*, vol. IT-31, no. 4, pp. 453–460, July 1985.
- [22] 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.