On More General Distributions of Random Binning for 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 separate encodings and joint decoding of two correlated discrete memoryless sources (DMSs), that is, the well–known Slepian-Wolf (S-W) source coding problem. Traditionally, ensembles of S–W codes are defined such that every bin of each n–vector of each source is randomly drawn under the uniform distribution across the sets $\{0, 1, \ldots, 2^{nR_X} - 1\}$ and $\{0, 1, \ldots, 2^{nR_Y} - 1\}$, where R_X and R_Y are the coding rates of the two sources, X and Y, respectively. In a few more recent works, where only one source, say, X, is compressed and the other one, Y, serves as side information available at the decoder, the scope is extended to variable–rate S–W (VRSW) codes, where the rate is allowed to depend on the type class of the source string, but still, the random–binning distribution is assumed uniform within the corresponding, type–dependent, bin index set.

In this expository work, we investigate the role of the uniformity of the random binning distribution from the perspective of the trade-off between the reliability (defined in terms of the error exponent) and the compression performance (measured from the viewpoint of the source coding exponent). To this end, we study a much wider class of random–binning distributions, which includes the ensemble of VRSW codes as a special case, but it also goes considerably beyond. We first show that, with the exception of some pathological cases, the smaller ensemble, of VRSW codes, is as good as the larger ensemble in terms the trade–off between the error exponent and the source coding exponent. Notwithstanding this finding, the wider class of ensembles proposed is motivated in two ways. The first is that it outperforms VRSW codes in the above–mentioned pathological cases, and the second is that it allows robustness: in the event of a system failure that causes unavailability of the compressed bit–stream from one of the sources, it still allows reconstruction of the other source within some controllable distortion.

Index Terms: Slepian–Wolf coding, random binning, universal decoding, error exponent, source coding exponent, variable–rate coding.

---

1 Introduction

The problem of (almost) loss separate source coding and joint decoding of two correlated sources, i.e., the well known Slepian–Wolf (S-W) coding problem, has received a vast level of attention ever since the celebrated article by Slepian and Wolf [11] was published, nearly five decades ago.

Ensembles of S–W codes have almost exclusively been defined by independent, random selection of a bin indices under the *uniform distribution*, for each and every possible source vector. In particular, denoting generically the two sources by X and Y, each n–vector of X (resp. Y) is mapped into a randomly selected bin in $\{0, 1, \ldots, 2^{nR_X}\}$ (resp. $\{0, 1, \ldots, 2^{nR_Y}\}$), which is described using nR_X (resp. nR_Y) bits. To the best knowledge of the author, the only exception to this rule is in the notion of variable–rate S–W (VRSW) codes, see, e.g., [1], [2], [3], [10], [12]. In those works, the simpler version of S–W setting was considered, where only one source, say, X, is compressed, whereas the other one, Y, serves as side information available at the decoder, and the coding rate for X, is allowed to vary with dependence on the type class of the source sequence, $\boldsymbol{x} = (x_1, \ldots, x_n)$, or equivalently, the empirical distribution of the source sequence, $\hat{P}_{\boldsymbol{x}}$. As was shown in [12], VRSW codes offer considerable improvement relative to their fixed–rate counterparts, in terms of achievable trade–offs between the error exponent and the source coding exponent. Moreover, the optimal rate function, $R(\hat{P}_{\boldsymbol{x}})$, was characterized in [12]. However, here too, within each and every type class of source sequences, the ensemble of codes is defined by independent random selection bin assignment under the uniform distribution over $\{0, 1, \ldots, 2^{nR(\hat{P}_{\boldsymbol{x}})} - 1\}$.

It is quite natural to ask, at this point, why is the uniform distribution being used so exclusively in the context of S–W random binning, whereas in other contexts of ensembles of codes (for both source coding and channel coding), general random coding distributions are used, which depend on the underlying source, or channel. A simple, intuitive answer could be that the bins are randomly drawn in the "compressed domain": after optimal compression, strings of bits form a binary symmetric source (or at least, nearly so), and therefore, they are distributed uniformly. On the face of it, such an answer sounds appealing and satisfactory. Indeed, this uniform distribution is good enough for proving the S–W coding theorem, which is compatible with the converse theorem. However, as will be demonstrated shortly, this kind of explanation is somewhat too simplistic when it comes to more refined figures of merit, like the error exponent and the source coding exponent.

As will be demonstrated shortly, the optimality of the uniform random–binning distribution is not quite obvious when one examines the problem through the lens of those figures of merit.

In view of this background, our aim, in this expository paper, is to study this question with some more detail. We carry out this study by investigating a considerably wider class of bin–assignment distributions from the viewpoint of achievable trade-offs between the error exponent and the source coding exponent. In particular, we consider an ensemble model where source n–vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, are mapped into 'bins' that are represented by other finite–alphabet n–vectors, $\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathcal{U}^n$ and $\boldsymbol{v} = (v_1, \ldots, v_n) \in \mathcal{V}^n$, respectively, that are drawn under conditional probability distributions, $A(\boldsymbol{u}|\boldsymbol{x})$ and $B(\boldsymbol{v}|\boldsymbol{y})$, which have a simple structure: $A(\boldsymbol{u}|\boldsymbol{x})$ (resp. $B(\boldsymbol{v}|\boldsymbol{y})$) depends on $\boldsymbol{x}$ and $\boldsymbol{u}$ (resp. $\boldsymbol{y}$ and $\boldsymbol{v}$) only via their empirical joint distribution, $\hat{P}_{\boldsymbol{ux}}$ (resp. $\hat{P}_{\boldsymbol{vy}}$). More precisely,

$$A(\boldsymbol{u}|\boldsymbol{x}) \doteq \exp\{-nF(\hat{P}_{\boldsymbol{ux}})\}, \qquad B(\boldsymbol{v}|\boldsymbol{y}) \doteq \exp\{-nG(\hat{P}_{\boldsymbol{vy}})\}, \tag{1}$$

where $\doteq$ stands for asymptotic equality in the exponential order (to be formally defined in the sequel) and F and G are non-negative functions of the respective empirical joint distributions. The outputs, $\boldsymbol{u}$ and $\boldsymbol{v}$, of these mappings, are then losslessly compressed and sent to the decoder, which in turn, first reconstructs $\boldsymbol{u}$ and $\boldsymbol{v}$, and then selects the most likely pair of source vectors among all those mapped to the given $\boldsymbol{u}$ and $\boldsymbol{v}$.

The motivation for studying the class of conditional distributions (1), henceforth referred to as *random–binning channels* (RBCs), is that, on the one hand, thanks to their structure, they are still amenable to analysis using the standard method of types [4], and on the other hand, they are general enough to cover a large variety of relevant special cases, such as memoryless sources and their mixtures, memoryless channels and their mixtures, uniform distributions within single (or several) type classes, or conditional type classes, etc. Obviously, the ordinary ensemble of S–W codes corresponds to the special case where $F(\hat{P}_{\boldsymbol{ux}}) \equiv R_X$, $G(\hat{P}_{\boldsymbol{vy}}) \equiv R_Y$, and the alphabets, $\mathcal{U}$ and $\mathcal{V}$, of each u_i and v_i, are of sizes 2^{R_X} and 2^{R_Y}, respectively.[1] VRSW codes can be viewed in this formal framework too: for given binning alphabets, $\mathcal{U}$ and $\mathcal{V}$, one may define $F(\hat{P}_{\boldsymbol{ux}}) = H(\hat{P}_{\boldsymbol{u}})$ and $G(\hat{P}_{\boldsymbol{vy}}) = H(\hat{P}_{\boldsymbol{v}})$, i.e., the entropies associated with certain empirical distributions, $\hat{P}_{\boldsymbol{u}}$ and $\hat{P}_{\boldsymbol{v}}$, that may depend on $\hat{P}_{\boldsymbol{x}}$ and $\hat{P}_{\boldsymbol{y}}$, respectively. The rate functions, $R_X(\hat{P}_{\boldsymbol{x}})$ and $R_Y(\hat{P}_{\boldsymbol{y}})$, are then given by $H(\hat{P}_{\boldsymbol{u}})$ and $H(\hat{P}_{\boldsymbol{v}})$, respectively, keeping in mind these correspondences between

---

[1] If 2^{R_X} is not integer, the rate R_X can be approximated arbitrarily closely by alphabet extensions. A similar comment applies, of course, also to R_Y.

$\hat{P}_{\boldsymbol{x}}$ and $\hat{P}_{\boldsymbol{u}}$, and between $\hat{P}_{\boldsymbol{y}}$ and $\hat{P}_{\boldsymbol{v}}$. In other words, for VRSW codes, $F(\hat{P}_{\boldsymbol{ux}})$ (resp. $G(\hat{P}_{\boldsymbol{vy}})$) depends on $\hat{P}_{\boldsymbol{xu}}$ (resp. $\hat{P}_{\boldsymbol{yv}}$) only via the marginals, $\hat{P}_{\boldsymbol{x}}$ and $\hat{P}_{\boldsymbol{u}}$ (resp. $\hat{P}_{\boldsymbol{y}}$ and $\hat{P}_{\boldsymbol{v}}$).

The class of RBCs (1) seems to offer much beyond the power of the ensembles of VRSW codes: it allows to shape the empirical joint distribution, or the coordination, between $\boldsymbol{x}$ and $\boldsymbol{u}$ (resp. $\boldsymbol{y}$ and $\boldsymbol{v}$). In fact, it is this coordination feature that makes (1) significantly more general than the ordinary ensembles of S–W coding. It is deeper and more important than the option of just selecting a non-uniform distribution under (1). But what is the motivation to shape such an empirical correlation between each source sequence and its bin sequence? The first obvious answer is that the ensemble of RBCs (1) is more general than that of variable–rate S–W codes, and therefore, one might expect that the extra degrees of freedom can potentially be used to gain improved performance in terms of the trade-off between the error exponent and the source coding exponent. A somewhat deeper answer is best presented if we look at a simple example: suppose that $\boldsymbol{y}$ has already been decoded, and it now being used to decode $\boldsymbol{x}$ as well. Suppose also that A assigns the uniform distribution over a certain conditional type class of $\boldsymbol{u}$ given $\boldsymbol{x}$, say, all $\boldsymbol{u}$–vectors that fall within Hamming distance 1 away from $\boldsymbol{x}$ (assuming, for simplicity, that both $\boldsymbol{x}$ and $\boldsymbol{u}$ are binary vectors). Under such circumstances, the decoder knows that the only candidates for guessing $\boldsymbol{x}$ based on $\boldsymbol{u}$ and $\boldsymbol{y}$ must be within Hamming distance 1 from $\boldsymbol{u}$. In other words, the number of incorrect source vectors, $\{\boldsymbol{x}'\}$, that may confuse the decoder is significantly reduced to lie within Hamming sphere of radius 1 around $\boldsymbol{u}$. It appears then that, in a way, $\boldsymbol{u}$ could serve as additional side information, beyond $\boldsymbol{y}$. Trivially, such a correlation between $\boldsymbol{x}$ and $\boldsymbol{u}$ cannot improve the achievable compression rate of the source X, given by the conditional entropy of X given Y, but at least on the face of it, one might speculate that the error exponent could be improved this way.

Our first result shows that, in spite of the above described idea of using $\boldsymbol{u}$ for reducing the uncertainty in $\boldsymbol{x}$, no coordination between $\boldsymbol{x}$ and $\boldsymbol{u}$ can possibly improve the trade-off between error exponents and source coding exponents. In other words, with the exception of a few pathological special cases (where the method of types is not applicable in a simple way), the best trade-off between those exponents, that it is achievable with the class of RBCs (1), can also be achieved using the smaller class of VRSW codes. This means that the ensemble of VRSW codes is good in a stronger sense than that was known before.[2]

---

[2]In [12], it was shown that in the simpler version of the S-W problem, where only one source is compressed and

Nonetheless, the general RBCs of (1) still have *raison d'être* for at least two reasons.

1. The first reason is that it covers the above–mentioned pathological special cases. Consider, for example, a memoryless source $\{X_i\}$ with a uniform distribution over the alphabet $\{-3, -1, +1, +3\}$, and suppose that $y_i = |x_i|$, $i = 1, 2, \ldots, n$ ($\{x_i\}$ being realizations of $\{X_i\}$). Suppose further that $\boldsymbol{y} = (y_1, \ldots, y_n)$ is compressed to its entropy, as is, which means that the RBC B is the identity channel, that assigns probability one to $\boldsymbol{v} = \boldsymbol{y}$, and that each v_i is represented by one bit. If the RBC A assigns probability one to the sequence of signs, $u_i = \text{sgn}(x_i)$, $i = 1, 2, \ldots, n$ (which is also a special case of (1), then both sources are perfectly recoverable at the decoder, as $x_i = u_i y_i$, and therefore the probability of error is strictly zero, which means an infinite error exponent. The source coding exponent, i.e., the exponential rate of the probability that both code–lengths exceed a threshold n, is also infinite, as there are exactly n (uncompressed) bits for each source. On the other hand, for ordinary S–W codes, and even for VRSW codes, the error exponent is finite even if one does not care about the source coding exponent. This is because there are in general many source sequences in each conditional type given the absolute values (all differing in permutations of signs) and some of them may be mapped into the same bin.[3]

2. Maintaining coordination between $\boldsymbol{u}$ and $\boldsymbol{x}$, and between $\boldsymbol{v}$ and $\boldsymbol{y}$, buys robustness in the following sense: imagine that due to some system failure, one of the compressed bit-streams, say, that of the Y-source, might not be available to the decoder, or that it might be severely corrupted. In this case, if a code drawn from the RBC (1) is used, one can still reconstruct $\boldsymbol{u}$ from its compressed bits, and by shaping the conditional distribution of $\boldsymbol{u}$ given $\boldsymbol{x}$, one can control the distortion and keep it within a prescribed level D away from $\boldsymbol{x}$. This is in sharp contrast to the situation with ordinary S–W codes, or even VRSW codes, where the absence of side information could be totally catastrophic.[4]

---

the other one serves as side information, the ensemble of VRSW codes with an optimized rate function, are optimal at low rates, in the sense of meeting a converse bound. However, this applies only at low enough rates, and it is based on the duality with channel coding and the sphere–packing bound. But this duality argument does not seem to be applicable to the general S-W problem, where both sources are compressed.

[3]Technically speaking, the reason that this example is exceptional is that the method of types fails, in this case, to assess the cardinality of the set of incorrect typical source vectors, $\{\boldsymbol{x'}\}$, that may confuse the decoder. While, in general, the cardinality of a typical set is assessed, exponentially tightly, in terms of the corresponding empirical conditional entropy, in this particular case, this confusion set is empty, which is equivalent to replacing the conditional empirical entropy by $-\infty$. More details can be found in the sequel.

[4]In the simpler version of the S–W problem, where only one source is compressed, and the other one serves as

The outline of the remaining part of the paper are as follows. In Section 2, we establish notation conventions. In Section 3, we formalize the setting and spell out our objectives. In Section 4, we establish the asymptotic optimality of the ensemble of VRSW codes within the class of RBC's (1). In Section 5, we consider RBCs of the form (1) and discuss the compromise between distortion, error exponent, and source coding exponent. Finally, Section 6 is devoted to proofs of our quantitative results.

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 $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – 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 capital letters, 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 $\mathbb{E}_P\{\cdot\}$. Again, the subscript will be omitted if the underlying probability distribution is clear from the context. The entropy of a generic random variable X, with a distribution Q on $\mathcal{X}$, will be denoted by $H_Q(X)$. Similar conventions will apply to other information measures. For example, if (X, Y) is a pair of random variables, jointly distributed according to Q_{XY} (or just Q, for short), $H_Q(X, Y)$, $H_Q(X|Y)$ and $I_Q(X; Y)$ will denote the induced joint entropy, conditional entropy of X given Y, and mutual information. Similar notation rules will apply to larger sets of random variables (triplets, quadruplets, etc). The Kullback–Leibler

---

side information, this scenario is a special case of a Wyner–Ziv setting where the side information may or may not be present, which was studied independently by Heegard and Berger [6] and Kaspi [7]. Narrowing our model to the compression of one source only, our scheme may not be optimal in their sense, but it is considerably simpler.

divergence between two probability distributions, Q_{XY} and P_{XY} is defined as

$$D(Q_{XY}\|P_{XY}) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} Q_{XY}(x,y)\log\frac{Q_{XY}(x,y)}{P_{XY}(x,y)}, \tag{2}$$

and the conditional weighted divergence is defined as

$$D(Q_{Y|X}\|P_{Y|X}|Q_X) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} Q_{XY}(x,y)\log\frac{Q_{Y|X}(y|x)}{P_{Y|X}(y|x)}, \tag{3}$$

where logarithms, here and throughout the sequel, are understood to be taken to the base 2, unless specified otherwise.

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, $\lim_{n\to\infty}\frac{1}{n}\log\frac{a_n}{b_n} = 0$. Similarly, $a_n \stackrel{\cdot}{\leq} b_n$ means that $\limsup_{n\to\infty}\frac{1}{n}\log\frac{a_n}{b_n} \leq 0$, 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 cardinality of a finite set, $\mathcal{A}$, 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 $\hat{P}$, we will denote it by $\mathcal{T}(\hat{P})$. 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 $\hat{H}_{\boldsymbol{x}}(X)$. Alternative notations that will be used are $H(\hat{P}_{\boldsymbol{x}})$ and $\hat{H}(\boldsymbol{x})$. 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 (and multiples) 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 $\boldsymbol{y}$, $\hat{H}_{\boldsymbol{xy}}(X|Y)$ or $\hat{H}(\boldsymbol{x}|\boldsymbol{y})$ will be the empirical conditional entropy, $\hat{I}_{\boldsymbol{xy}}(X; Y)$ will denote empirical mutual information, and so on. The same notation rules apply also to triplets and quadruplets of typical sequences, e.g., $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{x}$ and $\boldsymbol{y}$.

3 Problem Setting and Objectives

Consider a pair of sources defined by independent copies, $\{(X_i, Y_i),\ i = 1, 2, \ldots\}$, of a pair of finite–alphabet random variables, (X, Y), jointly distributed according to a probability distribution, P_{XY}. The alphabets of X and Y are denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. Both sequences are compressed separately and decoded jointly.

The encoder model is as follows. Each source vector, $\boldsymbol{x} = (x_1, \ldots, x_n)$ (resp. $\boldsymbol{y} = (y_1, \ldots, y_n)$), is mapped into a bin, $f(\boldsymbol{x}) = \boldsymbol{u} = (u_1, \ldots, u_n) \in \mathcal{U}^n$ (resp. $g(\boldsymbol{y}) = \boldsymbol{v} = (v_1, \ldots, v_n) \in \mathcal{V}^n$), where $\mathcal{U}$ (resp. $\mathcal{V}$) is a finite alphabet, and then, the vector $\boldsymbol{u}$ (resp. $\boldsymbol{v}$) is compressed losslessly into a binary string. The decoder receives the two bit–streams. It first reconstructs $\boldsymbol{u}$ and $\boldsymbol{v}$, and then uses them in order to estimate $(\boldsymbol{x}, \boldsymbol{y})$ using the MAP estimator,

$$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = h[\boldsymbol{u}, \boldsymbol{v}] = \arg \max_{\{(\boldsymbol{x}, \boldsymbol{y}):\ f(\boldsymbol{x}) = \boldsymbol{u},\ g(\boldsymbol{y}) = \boldsymbol{v}\}} P(\boldsymbol{x}, \boldsymbol{y}), \tag{4}$$

where

$$P(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} P_{XY}(x_i, y_i). \tag{5}$$

The mappings f and g, henceforth referred to as the *bin–assignment mappings* (BAMs), are generated in the following manner. For each $\boldsymbol{x} \in \mathcal{X}^n$ (resp. $\boldsymbol{y} \in \mathcal{Y}^n$), the corresponding bin $\boldsymbol{u}$ (resp. $\boldsymbol{v}$) is selected independently at random according to conditional distributions of the form (1), where $F(\cdot)$ and $G(\cdot)$ are arbitrary functionals of the respective empirical joint distributions, $\hat{P}_{\boldsymbol{xu}}$ and $\hat{P}_{\boldsymbol{yv}}$, of the pair of vectors, $(\boldsymbol{x}, \boldsymbol{u})$, and $(\boldsymbol{y}, \boldsymbol{v})$. Once the BAMs have been selected, they are revealed to both encoder and decoder.

We are interested in the trade-off between two performance metrics. The first is the average error probability with respect to (w.r.t.) the ensemble of BAMs,

$$\bar{P}_{\text{err}} = \mathbb{E}\{\Pr\{(\boldsymbol{X}, \boldsymbol{Y}) \neq h[f(\boldsymbol{X}), g(\boldsymbol{Y})]\}\} = \sum_{\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}} P(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{u}|\boldsymbol{x}) B(\boldsymbol{v}|\boldsymbol{y}) \cdot \bar{P}_{\text{err}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) \tag{6}$$

where

$$\bar{P}_{\text{err}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) \stackrel{\Delta}{=} \Pr\left[\bigcup_{\{(\boldsymbol{x}', \boldsymbol{y}') \neq (\boldsymbol{x}, \boldsymbol{y}):\ P(\boldsymbol{x}', \boldsymbol{y}') \geq P(\boldsymbol{x}, \boldsymbol{y})\}} \{f(\boldsymbol{x}') = \boldsymbol{u},\ g(\boldsymbol{y}') = \boldsymbol{v}\} \right]. \tag{7}$$

Accordingly, we define the error exponent as

$$\boldsymbol{E}_{\text{err}}(A, B) = \lim_{n \to \infty} \left[-\frac{\log \bar{P}_{\text{err}}}{n} \right], \tag{8}$$

8

provided that the limit exists. We denote the error exponent as a functional of the RBCs A and B, for a reason that will become apparent shortly.

The second performance metric is the average *excess code-length probability* (a.k.a. the "buffer overflow" probability), defined again w.r.t. the ensemble of BAMs (1). Given two prescribed threshold rate parameters, $\tilde{R}_X$ and $\tilde{R}_Y$, this probability is defined as

$$
\begin{aligned}
\bar{P}_{\text{ecl}} &= \mathbb{E}\left[\min_{L_X, L_Y} \Pr\{L_X[f(\boldsymbol{X})] \geq n\tilde{R}_X, \ L_Y[g(\boldsymbol{Y})] \geq n\tilde{R}_Y\}\right] \\
&= \mathbb{E}\left\{ \min_{L_X, L_Y} \sum_{(\boldsymbol{x}, \boldsymbol{y}): \ L_X[f(\boldsymbol{x})] \geq n\tilde{R}_X, \ L_Y[g(\boldsymbol{y})] \geq n\tilde{R}_Y\}} P(\boldsymbol{x}, \boldsymbol{y})\right\},
\end{aligned}
\tag{9}
$$

where the expectation is over the ensemble of BAMs, $L_X[\cdot]$ and $L_Y[\cdot]$ are length functions[5] of two uniquely decipherable (UD) lossless source codes, and $\Pr\{\cdot\}$ is w.r.t. the randomness of $(\boldsymbol{X}, \boldsymbol{Y})$. For given realization, f and g, of the BAMs, the optimal length function, that achieves the minimum in (9) depends on the induced probability distributions of $\boldsymbol{u}$ and $\boldsymbol{v}$,

$$
P(\boldsymbol{u}) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \cdot \mathcal{I}\{f(\boldsymbol{x}) = \boldsymbol{u}\}, \quad P(\boldsymbol{v}) = \sum_{\boldsymbol{y}} P(\boldsymbol{y}) \cdot \mathcal{I}\{g(\boldsymbol{y}) = \boldsymbol{v}\}.
\tag{10}
$$

Since f and g are random, then so are $\{P(\boldsymbol{u})\}$ and $\{P(\boldsymbol{v})\}$. The exact (asymptotic) analysis of $\bar{P}_{\text{ecl}}$ is non–trivial. We therefore replace the definition of (9) by a more convenient definition, as follows: instead of minimizing over L_X and L_Y, we take both of them to be the length function L_* of the universal lossless code of memoryless sources, which represents $\boldsymbol{u}$ (resp. $\boldsymbol{v}$) in two parts: the first is is the index of the type class of $\boldsymbol{u}$ (resp. $\boldsymbol{v}$) and the second is the index of $\boldsymbol{u}$ (resp. $\boldsymbol{v}$) within its type class. Since the number of types of n–sequences is polynomial in n, the first part requires $O(\log n)$ bits, whereas the second part requires $\log |\mathcal{T}(\boldsymbol{u})| \approx nH(\hat{P}_{\boldsymbol{u}})$ (resp. $\log |\mathcal{T}(\boldsymbol{v})| \approx nH(\hat{P}_{\boldsymbol{v}})$) bits. We now adopt the alternative definition of the excess code length probability,

$$
\tilde{P}_{\text{ecl}} = \mathbb{E}\left[\Pr\{L_*[f(\boldsymbol{X})] \geq n\tilde{R}_X, \ L_*[g(\boldsymbol{Y})] \geq n\tilde{R}_Y\}\right],
\tag{11}
$$

which is obviously an upper bound on $\bar{P}_{\text{ecl}}$. The *excess code–length exponent* is then defined by

$$
\boldsymbol{E}_{\text{ecl}}(\tilde{R}_X, \tilde{R}_Y, A, B) = \lim_{n \to \infty}\left[-\frac{\log \tilde{P}_{\text{ecl}}}{n}\right]
$$

---

[5]Recall that a length function $L[\boldsymbol{u}] = L[f(\boldsymbol{x})]$ of a source code is defined as the length (in bits) of the compressed representation of $\boldsymbol{u}$. For a length function to correspond to a UD code, it must obey the Kraft inequality, $\sum_{\boldsymbol{u}} 2^{-L(\boldsymbol{u})} \leq 1$.

$$= \lim_{n \to \infty} \left[-\frac{\log \mathbb{E}[\Pr\{H(\hat{P}_{\boldsymbol{u}}) \geq \tilde{R}_X, \ H(\hat{P}_{\boldsymbol{v}}) \geq \tilde{R}_Y\}]}{n} \right]. \tag{12}$$

For a given pair of threshold rates, $(\tilde{R}_X, \tilde{R}_Y)$, that satisfy $\tilde{R}_X > H(X|Y)$, $\tilde{R}_Y > H(Y|X)$, and $\tilde{R}_X + \tilde{R}_Y > H(X,Y)$, every pair (A, B) of RBCs, achieves a certain error exponent, $\boldsymbol{E}_{\mathrm{err}}(A, B)$, and a certain excess code–length exponent, $\boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, A, B)$, i.e., a certain point $(\boldsymbol{E}_{\mathrm{err}}(A, B), \boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, A, B))$ in the plane of the two exponents, and naturally, there is a certain tension between them. We are interested in those RBC pairs $\{(A, B)\}$ that achieve the best trade-off curve, that is,

$$\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}_X, \tilde{R}_Y) = \max\{\boldsymbol{E}_{\mathrm{err}}(A, B) : \ \boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, A, B) \geq E_{\mathrm{ecl}}\}, \tag{13}$$

where E_{ecl} designates a prescribed excess code–length exponent level,[6] that varies between zero and infinity.

4 Asymptotic Optimality of VRSW Codes

Let the ensemble of S-W codes be given by eq. (1), where the RBCs, A and B, is defined with given functions, F and G, respectively. Our first lemma provides general formulas for the error exponent and the source coding exponent.

Lemma 1 *Referring to the problem setting defined in Section 3,*

1. *The excess code–length exponent is given by*

$$\boldsymbol{E}_{ecl}(\tilde{R}_X, \tilde{R}_Y, A, B) = \min_{\{Q_{UVXY}: \ H_Q(U) \geq \tilde{R}_X, \ H_Q(V) \geq \tilde{R}_Y\}} \{D(Q_{XY} \| P_{XY}) + F(Q_{UX}) + G(Q_{VY}) - H_Q(U|X) - H_Q(V|Y)\}. \tag{14}$$

2. *Assuming that $(\boldsymbol{x}, \boldsymbol{y})$ cannot be uniquely determined from $(\boldsymbol{u}, \boldsymbol{v})$,*

$$\boldsymbol{E}_{err}(A, B) = \min\{\boldsymbol{E}_1(A), \boldsymbol{E}_2(B), \boldsymbol{E}_3(A, B)\}, \tag{15}$$

where

$$\boldsymbol{E}_1(A) = \min_{Q_{UXY}} \{D(Q_{XY} \| P_{XY}) + F(Q_{UX}) - H_Q(U|X,Y) + $$

---

[6] Note that the excess code length exponent function is denoted with a bold face $\boldsymbol{E}$, as opposed to the excess code length exponent level (which is just a given non–negative real) is denoted in an ordinary font.

$$[F(Q_{UX}) - H_Q(X|Y,U)]_+\}$$ (16)

$$\boldsymbol{E}_2(B) = \min_{Q_{VXY}} \{D(Q_{XY}\|P_{XY}) + G(Q_{VY}) - H_Q(V|X,Y) +$$

$$[G(Q_{VY}) - H_Q(Y|X,V)]_+\}$$ (17)

$$\boldsymbol{E}_3(A,B) = \min_{Q_{UVXY}} \{D(Q_{XY}\|P_{XY}) + F(Q_{UX}) + G(Q_{VY}) - H_Q(U,V|X,Y) +$$

$$[F(Q_{UX}) + G(Q_{VY}) - H_Q(X,Y|U,V)]_+\}.$$ (18)

In the second part of Lemma 1, $\boldsymbol{E}_1(A)$ is the error exponent associated with decoding errors in $\boldsymbol{x}$ only, $\boldsymbol{E}_2(B)$ is the one associated with decoding errors in $\boldsymbol{y}$ only, and $\boldsymbol{E}_3(A,B)$ stands for simultaneous decoding errors in both $\boldsymbol{x}$ and $\boldsymbol{y}$. The second part of Lemma 1 generalizes the well–known formula for of the random coding error exponent of ordinary S-W codes, where $F(Q_{UX}) = H_Q(U) = R_X$ and $G(Q_{VY}) = H_Q(V) = R_Y$. The minimization of the error exponent expression of $E_{\text{err}}(A,B)$ over Q_{UVXY} under the constraint $H_Q(U) = R_X$, with U being independent of X, causes U and V to be independent of (X,Y), and of each other, and then the error exponent formulas simplify to (see, e.g., [12]),

$$\boldsymbol{E}_1(A) = \min_{Q_{XY}} \{D(Q_{XY}\|P_{XY}) + [R_X - H_Q(X|Y)]_+\},$$ (19)

and similar comments apply to $\boldsymbol{E}_2(B)$ and $\boldsymbol{E}_3(A,B)$. It appears then that in the more general expression of $\boldsymbol{E}_1(A,B)$, $H_Q(X|Y)$ (in the square brackets) is replaced by $H_Q(X|Y,U)$. Likewise, in $\boldsymbol{E}_2(A,B)$, $H_Q(Y|X)$ is replaced by $H_Q(Y|X,V)$ and in $\boldsymbol{E}_3(A,B)$, $H_Q(X,Y)$ is replaced by $H_Q(X,Y|U,V)$. This is coherent with the insight that U and V play the role of additional side information, as described in the Introduction.

The next lemma gives a simple guideline for a good choice of the functions, F and G.

Lemma 2 *Both $\boldsymbol{E}_{ecl}(\tilde{R}_X, \tilde{R}_Y, A, B)$ and $\boldsymbol{E}_{err}(A,B)$ are maximized by functions F and G with the following property: For every Q_X, there exists $Q^*_{U|X}$, which may depend on Q_X, such that*

$$F(Q_{UX}) = \begin{cases} H_Q(U|X) & Q_{UX} = Q_X \times Q^*_{U|X} \\ \infty & elsewhere \end{cases}$$ (20)

A similar argument applies to G with Q_X, Q_{UX} and $Q^*_{U|X}$, being replaced by Q_Y, Q_{VY}, and $Q^*_{V|Y}$, respectively.

Lemma 2 tells us that the best RBCs put all their mass on a single conditional type class of $\boldsymbol{u}$ given $\boldsymbol{x}$ (resp. $\boldsymbol{v}$ given $\boldsymbol{y}$), wherein the distribution assigned is uniform, i.e.,

$$A^*(\boldsymbol{u}|\boldsymbol{x}) = \begin{cases} \frac{1}{|\mathcal{T}(\boldsymbol{u}|\boldsymbol{x})|} & \hat{P}_{\boldsymbol{u}\boldsymbol{x}} = \hat{P}_{\boldsymbol{x}} \times Q^*_{U|X} \\ 0 & \text{elsewhere} \end{cases} \tag{21}$$

and

$$B^*(\boldsymbol{v}|\boldsymbol{y}) = \begin{cases} \frac{1}{|\mathcal{T}(\boldsymbol{v}|\boldsymbol{y})|} & \hat{P}_{\boldsymbol{v}\boldsymbol{y}} = \hat{P}_{\boldsymbol{y}} \times Q^*_{V|Y} \\ 0 & \text{elsewhere} \end{cases} \tag{22}$$

This result is analogous to well known results about the optimality of ensembles of fixed composition with uniform distributions therein.

Combining Lemma 1 and Lemma 2, we now obtain the following corollary concerning the expressions for the source coding exponent and the error exponent, where with slight abuse of notation, we replace A and B, in $\boldsymbol{E}_{\text{ecl}}(\tilde{R}_X, \tilde{R}_Y, A, B)$ and $\boldsymbol{E}_{\text{err}}(A, B)$, by $Q^*_{U|X}$ and $Q^*_{V|Y}$, respectively.

Corollary 1

$$\boldsymbol{E}_{ecl}(\tilde{R}_X, \tilde{R}_Y, Q^*_{U|X}, Q^*_{V|Y}) = \min_{\{Q_{UVXY}: \ H_Q(V) \geq \tilde{R}_Y\}} D(Q_{XY} \| P_{XY}), \tag{23}$$

$$\begin{aligned}
\boldsymbol{E}_1(Q^*_{U|X}) &= \min_{Q_{UXY}} \{ D(Q_{XY} \| P_{XY}) + H_Q(U|X) - H_Q(U|X,Y) + \\
&\quad [H_Q(U|X) - H_Q(X|Y,U)]_+ \},
\end{aligned} \tag{24}$$

$$\begin{aligned}
\boldsymbol{E}_2(Q^*_{V|Y}) &= \min_{Q_{VXY}} \{ D(Q_{XY} \| P_{XY}) + H_Q(V|Y) - H_Q(V|X,Y) + \\
&\quad [H_Q(V|Y) - H_Q(Y|X,V)]_+ \},
\end{aligned} \tag{25}$$

$$\begin{aligned}
\boldsymbol{E}_3(Q^*_{U|X}, Q^*_{V|Y}) &= \min_{Q_{UVXY}} \{ D(Q_{XY} \| P_{XY}) + H_Q(U|X) + H_Q(V|Y) - H_Q(U,V|X,Y) + \\
&\quad [H_Q(U|X) + H_Q(V|Y) - H_Q(X,Y|U,V)]_+ \},
\end{aligned} \tag{26}$$

where all minimizations over Q_{UVXY} are subject to the (additional) constraints, $Q_{U|X} = Q^*_{U|X}$ (which is a function of Q_X), and $Q_{V|Y} = Q^*_{V|Y}$ (which is a function of Q_Y).

Note that for $\boldsymbol{E}_{\text{ecl}}(\tilde{R}_X, \tilde{R}_Y, Q^*_{U|X}, Q^*_{V|Y})$, the dependence on $Q^*_{U|X}$ and $Q^*_{V|Y}$ is only via the induced marginals, Q^*_U and Q^*_V. Corollary 1 exhibits the seemingly non–trivial impact of generating dependencies between X and U and between Y and V. On the one hand, it helps as it reduces the subtracted terms, $H_Q(X|U,Y)$, $H_Q(Y|X,V)$ and $H_Q(X,Y|U,V)$ (as U and V serve as side

informations), but on the other hand, it reduces also the terms $H_Q(U|X)$ and $H_Q(V|Y)$, and it is not a priori obvious what is the overall effect of such dependencies on the error exponent.

The next result tells us, however, that for the best RBCs of the class of Lemma 2, $Q_{U|X}^* = Q_U^*$ and $Q_{V|Y}^* = Q_V^*$, that is, U is independent of X, and V is independent of Y, under Q.

Theorem 1 *For a given Q_X (resp. Q_Y) and it is associated conditional distribution, $Q_{U|X}^*$ (resp. $Q_{V|Y}^*$), let Q_U^* (resp. Q_V^*) denote the induced marginal. Then,*

$$\boldsymbol{E}_{ecl}(\tilde{R}_X, \tilde{R}_Y, Q_U^*, Q_V^*) = \boldsymbol{E}_{ecl}(\tilde{R}_X, \tilde{R}_Y, Q_{U|X}^*, Q_{V|Y}^*), \tag{27}$$

$$\boldsymbol{E}_{err}(Q_U^*, Q_V^*) \geq \boldsymbol{E}_{err}(Q_{U|X}^*, Q_{V|Y}^*), \tag{28}$$

$$\boldsymbol{E}_{err}(Q_U^*, Q_V^*) = \min\{E_1(Q_U^*), E_2(Q_V^*), E_3(Q_U^*, Q_V^*)\} \tag{29}$$

$$\boldsymbol{E}_1(Q_U^*) = \min_{Q_{XY}}\{D(Q_{XY}\|P_{XY}) + [H_{Q^*}(U) - H_Q(X|Y)]_+\} \tag{30}$$

$$\boldsymbol{E}_2(Q_V^*) = \min_{Q_{XY}}\{D(Q_{XY}\|P_{XY}) + [H_{Q^*}(V) - H_Q(Y|X)]_+\} \tag{31}$$

$$\boldsymbol{E}_3(Q_U^*, Q_V^*) = \min_{Q_{XY}}\{D(Q_{XY}\|P_{XY}) + [H_{Q^*}(U) + H_{Q^*}(V) - H_Q(X,Y)]_+\}, \tag{32}$$

where $H_{Q^*}(U)$ and $H_{Q^*}(V)$ denote the entropies pertaining to Q_U^* and Q_V^*, respectively.

The last part of Theorem 1 is identified with the error exponent of VRSW codes with rate functions $H_{Q^*}(U)$ and $H_{Q^*}(V)$ that depend on Q_X and Q_Y, respectively. Indeed, as explained in the Introduction, the ensemble of VRSW codes, with rate functions $R_X(\hat{P}_{\boldsymbol{x}})$ and $R_Y(\hat{P}_{\boldsymbol{y}})$, corresponds to the special case of (1), where $F(\hat{P}_{\boldsymbol{ux}}) = H(\hat{P}_{\boldsymbol{u}})$ and $G(\hat{P}_{\boldsymbol{yv}}) = H(\hat{P}_{\boldsymbol{v}})$, with $\hat{P}_{\boldsymbol{u}}$ and $\hat{P}_{\boldsymbol{v}}$ being chosen such that $H(\hat{P}_{\boldsymbol{u}}) = R_X(\hat{P}_{\boldsymbol{x}})$ and $H(\hat{P}_{\boldsymbol{v}}) = R_Y(\hat{P}_{\boldsymbol{y}})$.[7]

5 Using General Random Binning Channels for Robustness

As mentioned in the Introduction, in spite of the conclusion of Theorem 1, there are nevertheless at least two motivations for using general RBC's that maintain some coordination between the source vectors and their associated bins. The first is that it covers pathological, degenerate cases, as demonstrated in the Introduction. The other, and perhaps more important motivation is that it

---

[7]This is a complete equivalence since the typical sequences of type $\hat{P}_{\boldsymbol{u}}$ can be mapped bijectively into binary sequences of length about $nH(\hat{P}_{\boldsymbol{u}}) = nR_X(\hat{P}_{\boldsymbol{x}})$ and a similar comment applies to $\hat{P}_{\boldsymbol{v}}$. Note that the choice of the alphabets $\mathcal{U}$ and $\mathcal{V}$ is completely immaterial in the case of VRSW codes. It is important only in the general case, where (1) depends on the full joint empirical distributions, $\hat{P}_{\boldsymbol{ux}}$ and $\hat{P}_{\boldsymbol{vy}}$.

allows more robustness to system failure: in the unfortunate event that the compressed bit stream of one of the sources, say, Y, becomes unavailable to the decoder (for whatever reason), the other source, X, can still be reconstructed from its compressed bits within some distortion: if $Q_{U|X}^*$ is chosen such that $\sum_{i=1}^n d(x_i, u_i) = n \sum_{x,u} Q_X(x) Q_{U|X}(u|x) d(x,u) \leq nD$, for some distortion measure $d(\cdot,\cdot)$ and distortion level D, then the reconstruction of $\boldsymbol{u}$ can serve as an estimator of $\boldsymbol{x}$.[8] Obviously, this robustness comes at the price of some degradation in the trade–off between the two other performance metrics: the error exponent and the excess code–length exponent.[9] In other words, we now have a more general trade-off between three, rather than two, performance metrics: the error exponent, the excess code–length exponent, and the robustness to system failure, measured in terms of the distortions in the reconstruction of each source, in the event that the other source may not be available. More specifically, we would like to derive the following quantity for a given pair of threshold rates, $(\tilde{R}_X, \tilde{R}_Y)$:

$$
\begin{aligned}
\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, D_X, D_Y) = \max \Big\{ &\boldsymbol{E}_{\mathrm{err}}(Q_{U|X}, Q_{V|Y}):\ \boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, Q_{U|X}, Q_{V|Y}) \geq E_{\mathrm{ecl}}, \\
&\sum_{x,u} Q_{UX}(u,x) d_X(x,u) \leq D_X,\ \sum_{y,v} Q_{VY}(v,y) d_Y(y,v) \leq D_Y \Big\}, \quad (33)
\end{aligned}
$$

where $d_X(\cdot,\cdot)$ and $d_Y(\cdot,\cdot)$ are distortion measures associated with the two sources, D_X and D_Y are the maximum allowable distortion levels, and where we have removed the asterisks from $Q_{U|V}^*$ and $Q_{V|Y}^*$. The requirement $\boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, Q_{U|X}, Q_{V|Y}) \geq E_{\mathrm{ecl}}$ is equivalent to the requirement that for every Q_{XY} with $D(Q_{XY} \| P_{XY}) \leq E_{\mathrm{ecl}}$, $Q_{U|X}$ and $Q_{V|Y}$ must be such that $H_Q(U) \leq \tilde{R}_X$ and $H_Q(V) \leq \tilde{R}_Y$. In addition, the two distortion constraints must be satisfied.

In principle, given the earlier results presented above, the complete derivation of this quantity can be carried out in full generality, but it is evidently extremely complicated in terms of the number of parameters to be optimized and the number of nested optimizations required. Therefore, in order to characterize some of the trade-offs under discussion, instead of the full optimization above, we will confine ourselves to the simpler case where the source $\boldsymbol{Y}$ is fully available at the decoder, and only $\boldsymbol{X}$ must be decoded. This corresponds to the following formal choices in our setting:

1. $Q_{V|Y}$ is set to be the identity channel, i.e., $\boldsymbol{V} = \boldsymbol{Y}$ with probability one.

---

[8] Slightly more generally, we could also use a certain function q for the reconstruction, i.e., $\sum_{i=1}^n d(x_i, q(u_i)) \leq nD$, but for the sake of simplicity, we will let q to be the identity function.

[9] Moreover, the distortion constraints might incur penalties in terms of the achievable rates altogether.

2. In view of item no. 1, we set $D_Y = 0$.

3. Since we have no requirements considering the excess code–length exponent of $\boldsymbol{Y}$, we formally set $\tilde{R}_Y = 0$. This degenerates the excess code–length event to be $\{L_X[f(\boldsymbol{X})] \geq n\tilde{R}_X\}$, or $\{H_Q(U) \geq \tilde{R}_X\}$ alone, without any requirement concerning the description length of $\boldsymbol{Y}$, since the event $\{L_Y[g(\boldsymbol{Y})] \geq 0$ (or $H_Q(V) \geq 0$) occurs with probability one, it is neutral in terms of intersections with other events.

Since $E_2(Q_{V|Y})$ and $E_3(Q_{U|X}, Q_{V|Y})$ are error exponents that are associated with decoding errors in $\boldsymbol{Y}$, they become irrelevant in this case, and so, only $E_1(Q_{U|V})$ plays a role here.[10]

We have therefore simplified the problem as follows.

$$
\begin{aligned}
\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D) \;\; = \;\; \max\Big\{ & \boldsymbol{E}_1(Q_{U|X}) : \;\; \boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}, Q_{U|X}) \geq E_{\mathrm{ecl}}, \\
& \sum_{x,u} Q_{UX}(u,x) d(x,u) \leq D \Big\},
\end{aligned}
\tag{34}
$$

where we have omitted the subscript X from d_X, D_X and $\tilde{R}_X$, and we have redefined $\boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, Q_{U|V})$ to be

$$
\boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, Q_{U|X}) = \min\{D(Q_X \| P_X) : \;\; H_Q(U) \geq \tilde{R}\}.
\tag{35}
$$

Before we proceed, it would be convenient to represent $E_1(Q_{U|X})$ as follows:

$$
\begin{aligned}
E_1(Q_{U|X}) \;\; = \;\; & \min_{Q_X} \min_{Q_{Y|UX}} \{D(Q_{XY} \| P_{XY}) + \\
& H_Q(U|X) - H_Q(U|X,Y) + [H_Q(U|X) - H_Q(X|U,Y)]_+\} \\
\triangleq \;\; & \min_{Q_X} \zeta(Q_{UX}),
\end{aligned}
\tag{36}
$$

where we remind that $E_1(Q_{U|X})$ is defined for a given choice of $Q_{U|X}$ for every Q_X, and where we have defined

$$
\zeta(Q_{UX}) = \min_{Q_{Y|UX}} \{D(Q_{XY} \| P_{XY}) + H_Q(U|X) - H_Q(U|X,Y) + [H_Q(U|X) - H_Q(X|U,Y)]_+\}. \tag{37}
$$

In Subsection 6.4, we prove the following alternative expression for $\zeta(Q_{UX})$.

$$
\zeta(Q_{UX}) \;\; = \;\; \inf_V \sup_{0 \leq \lambda \leq 1} \Big\{ \sum_{u,x} Q_{UX}(u,x) \log \frac{[Q_X(x)]^{1+\lambda}}{P_X(x)[Q_U(u)]^\lambda} -
$$

---

[10]Formally, we may consider $E_2(Q_{V|Y})$ and $E_3(Q_{U|X}, Q_{V|Y})$ to be infinite.

$$(1 + \lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \Big\}, \tag{38}$$

where

$$Z(u,x,V,\lambda) \triangleq \sum_y [P_{Y|X}(y|x)]^{1/(1+\lambda)} [V(y|u)]^{\lambda/(1+\lambda)}, \tag{39}$$

and where here $V = \{V(y|u),\ u \in \mathcal{U},\ y \in \mathcal{Y}\}$ (not to be confused with the random variable V, which no longer plays a role here) is a matrix of conditional probabilities of y given u, i.e., $V(y|u) \geq 0$ for all (u,y) and $\sum_y V(y|u) = 1$ for all u.

For $\boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}, Q_{U|X})$ to exceed a prescribed level E_{ecl}, every Q_X with $D(Q_X \| P_X) \leq E_{\mathrm{ecl}}$, must correspond to $Q_{U|X}$ such that the induced Q_U satisfies $\mathbb{E}_Q \log[1/M(U)] \leq \tilde{R}$ for some probability distribution M defined over $\mathcal{U}$. Accordingly, we define

$$\begin{aligned}
\mathcal{B}(Q_X, M, \tilde{R}, D, E_{\mathrm{ecl}}) &= \{Q_{U|X}:\ \mathbb{E}_Q d(X,U) \leq D,\ \text{and} \\
&\quad D(Q_X \| P_X) \leq E_{\mathrm{ecl}} \text{ implies } -\mathbb{E}_Q \log M(U) \leq \tilde{R}\}. \tag{40}
\end{aligned}$$

Comment: we could have taken $M = Q$ and thereby replace the requirement $-\mathbb{E}_Q \log M(U) \leq \tilde{R}$ by $H_Q(U) \leq \tilde{R}$, but this would make the problem non–convex. Now, we have

$$\begin{aligned}
\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D) &= \inf_{Q_X} \sup_M \sup_{Q_{U|X} \in \mathcal{B}(Q_X, M, \tilde{R}, D, E_{\mathrm{ecl}})} \zeta(Q_{UX}) \\
&= \min\{\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D), \tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)\}, \tag{41}
\end{aligned}$$

where

$$\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D) = \inf_{\{Q_X:\ D(Q_X\|P_X) \leq E_{\mathrm{ecl}}\}} \sup_M \sup_{\{Q_{U|X}:\ -\mathbb{E}_Q \log M(U) \leq \tilde{R},\ \mathbb{E}_Q d(X,U) \leq D\}} \zeta(Q_{UX}), \tag{42}$$

$$\tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D) = \inf_{\{Q_X:\ D(Q_X\|P_X) > E_{\mathrm{ecl}}\}} \sup_{\{Q_{U|X}:\ \mathbb{E}_Q d(X,U) \leq D\}} \zeta(Q_{UX}). \tag{43}$$

Note that $\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D)$ is monotonically non-increasing in E_{ecl} for fixed $(\tilde{R}, D)$, whereas $\tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)$ is monotonically non-decreasing, but the minimum between them, $\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D)$, is non–increasing in E_{ecl}, because of the following consideration: denoting

$$K(Q_X) = \sup_M \sup_{\{Q_{U|X}:\ -\mathbb{E}_Q \log M(U) \leq \tilde{R},\ \mathbb{E}_Q d(X,U) \leq D\}} \zeta(Q_{UX}) \tag{44}$$

and

$$L(Q_X) = \sup_{\{Q_{U|X}:\ \mathbb{E}_Q d(X,U) \leq D\}} \zeta(Q_{UX}), \tag{45}$$

16

we have

$$
\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D) = \inf_{Q_X} \Big\{ \mathcal{I}[Q_X : D(Q_X \| P_X) \leq E_{\mathrm{ecl}}] \cdot K(Q_X) +
$$

$$
\mathcal{I}[Q_X : D(Q_X \| P_X) > E_{\mathrm{ecl}}] \cdot L(Q_X) \Big\} \tag{46}
$$

$$
= \inf_{Q_X} \{ \mathcal{I}[Q_X : D(Q_X \| P_X) > E_{\mathrm{ecl}}] \cdot [L(Q_X) - K(Q_X)] + K(Q_X) \}, \tag{47}
$$

which is obviously monotonically non–increasing since $L(Q_X) \geq K(Q_X)$ and $\mathcal{I}[Q_X : D(Q_X \| P_X) > E_{\mathrm{ecl}}$ is non-increasing in E_{ecl}. We now argue that $\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D) \leq \tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)$ and hence $\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D) = \hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D)$ for all E_{ecl}, unless there is a range of E_{ecl} where $\tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)$ is fixed, as a function of E_{ecl} (i.e., the constraint is inactive). To see why this is true, suppose conversely, that for some $\hat{E}_{\mathrm{ecl}} > 0$, we have $\hat{\boldsymbol{E}}(\tilde{R}, \hat{E}_{\mathrm{ecl}}, D) > \tilde{\boldsymbol{E}}(\hat{E}_{\mathrm{ecl}}, D)$. Then for all $E_{\mathrm{ecl}} < \hat{E}_{\mathrm{ecl}}$, we have $\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D) \geq \hat{\boldsymbol{E}}(\tilde{R}, \hat{E}_{\mathrm{ecl}}, D) > \tilde{\boldsymbol{E}}(\hat{E}_{\mathrm{ecl}}, D) \geq \tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)$, which means that $\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D) = \tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)$ for all $E_{\mathrm{ecl}} \leq \hat{E}_{\mathrm{ecl}}$. Since $\tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)$ is monotonically non–deceasing as a function of E_{ecl}, then so is $\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D)$ in the range $E_{\mathrm{ecl}} < \hat{E}_{\mathrm{ecl}}$. But this is a contradiction to the non–increasing monotonicity of $\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D)$ unless $\tilde{\boldsymbol{E}}(E_{\mathrm{ecl}}, D)$ is fixed in the range $E_{\mathrm{ecl}} < \hat{E}_{\mathrm{ecl}}$, in which case it is equal to $\tilde{\boldsymbol{E}}(0, D)$ throughout this range. Thus, an alternative expression is

$$
\boldsymbol{E}_{\mathrm{err}}(E_{\mathrm{ecl}}, \tilde{R}, D) = \min\{ \hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D), \tilde{\boldsymbol{E}}(0, D) \}. \tag{48}
$$

We now have (see Subsection 6.5) the following lower bounds to $\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D)$ and $\tilde{\boldsymbol{E}}(0, D)$.

$$
\tilde{\boldsymbol{E}}(0, D) \geq \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{C} \Big[\zeta D -
$$

$$
(1 + \lambda) \log \Big\{ \sum_x \min_u \Big(2^{\zeta d(x,u)} \sum_y P(x, y)^{1/(1+\lambda)} C(u, y)^{\lambda/(1+\lambda)} \Big) \Big\} \Big], \tag{49}
$$

where $\{C(u, y)\}$ is an auxiliary probability distribution over $\mathcal{U} \times \mathcal{Y}$, and

$$
\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D) \geq \sup_{\theta \geq 0} \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \sup_M \inf_W \Big\{ \rho \tilde{R} + \zeta D - \theta E_{\mathrm{ecl}} -
$$

$$
(1 + \theta + \lambda) \log \Big(\sum_x \Big[\frac{P_X^{1+\theta}(x)}{S(x, M, W, V, \rho, \zeta, \lambda)} \Big]^{1/(1+\theta+\lambda)} \Big) \Big\}
$$

$$
= \sup_{\theta \geq 0} \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \Big\{ \rho \tilde{R} + \zeta D - \theta E_{\mathrm{ecl}} -
$$

$$
(1 + \theta + \lambda) \sup_V \inf_M \sup_W \log \Big(\sum_x \Big[\frac{P_X^{1+\theta}(x)}{S(x, M, W, V, \rho, \zeta, \lambda)} \Big]^{1/(1+\theta+\lambda)} \Big) \Big\}, \tag{50}
$$

17

where

$$S(x, M, W, V, \rho, \zeta, \lambda) \triangleq \max_u \left\{ \frac{M^\rho(u)}{W^\lambda(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u, x, V, \lambda)} \right\}, \tag{51}$$

$M(\cdot)$ and $W(\cdot)$ being additional auxiliary probability distributions over $\mathcal{U}$.

As we can see, in spite of the reduction to the simpler version of the S–W problem, the resulting exponent expression is still, by no means trivial, due to the large number of nested optimizations, especially in eq. (50). Because of this excessive complexity of the expressions involved, we will not carry out a full investigation of the error exponent formula. Instead, we will conclude this article by characterizing the behavior of the error exponent at two important extreme special cases: $E_{\mathrm{ecl}} = 0$ and $E_{\mathrm{ecl}} = \infty$. The rationale behind the focus on these two extremes is two–fold: first, these two extremes are relatively simple, and secondly, they are relevant if one wishes to contrast this with the behavior of the ensemble of fixed–rate S–W codes, where the excess code–length exponent is indeed always equal to either zero or infinity, depending on whether $\tilde{R}$ is below or above the actual coding rate. In both cases, we wish to characterize the trade-off between the distortion D, between U and X, and the threshold rate, $\tilde{R}$, such that the error exponent is still positive. Since we are looking at the limit where the error exponent vanishes, we will assume that $\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D)$ dominates the error exponent.

In the case $E_{\mathrm{ecl}} = 0$, the maximizing conjugate parameter, θ, tends to infinity. Now, in this limit, we have that

$$
\begin{aligned}
& \lim_{\theta \to \infty} (1 + \theta + \lambda) \log \left(\sum_x \left[\frac{P_X^{1+\theta}(x)}{S(x, M, W, V, \rho, \zeta, \lambda)} \right]^{1/(1+\theta+\lambda)} \right) \\
= {} & \lim_{\theta \to \infty} (1 + \theta + \lambda) \log \left(\sum_x P_X(x) \left[\frac{P_X^{-\lambda}(x)}{S(x, M, W, V, \rho, \zeta, \lambda)} \right]^{1/(1+\theta+\lambda)} \right) \\
= {} & \sum_x P_X(x) \log \left[\frac{P_X^{-\lambda}(x)}{S(x, M, W, V, \rho, \zeta, \lambda)} \right] \\
= {} & \lambda H(X) - \sum_x P_X(x) \log S(x, M, W, V, \rho, \zeta, \lambda) \\
\triangleq {} & T_0(M, W, V, \rho, \zeta, \lambda). \tag{52}
\end{aligned}
$$

Thus, the condition for a positive error exponent becomes

$$\exists \, 0 \le \lambda \le 1 \; \forall \, \rho \ge 0, \zeta \ge 0, \; \rho \tilde{R} + \zeta D \ge \sup_V \inf_M \sup_W T_0(M, W, V, \rho, \zeta, \lambda), \tag{53}$$

18

or, equivalently,

$$
\begin{aligned}
\tilde{R} \;\geq\;& \inf_{0\leq\lambda\leq1} \sup_{\rho\geq0} \sup_{\zeta\geq0} \frac{1}{\rho} \cdot [\sup_{V} \inf_{M} \sup_{W} T_0(M,W,V,\rho,\zeta,\lambda) - \zeta D] \\
=\;& \inf_{0\leq\lambda\leq1} \sup_{s\geq0} \sup_{\zeta\geq0} s[\sup_{V} \inf_{M} \sup_{W} T_0(M,W,V,1/s,\zeta,\lambda) - \zeta D].
\end{aligned}
\tag{54}
$$

In the case $E_{\mathrm{ecl}} = \infty$, the maximizing conjugate parameter, θ, must vanish. In this case, we get a similar lower bound to the required threshold rate, $\tilde{R}$, except that T_0 is replaced by

$$
\begin{aligned}
T_\infty(M,W,V,\rho,\zeta,\lambda) \;=\;& \lim_{\theta\to\infty} (1+\theta+\lambda)\log\left(\sum_x \left[\frac{P_X^{1+\theta}(x)}{S(x,M,W,V,\rho,\zeta,\lambda)}\right]^{1/(1+\theta+\lambda)}\right) \\
=\;& (1+\lambda)\log\left(\sum_x \left[\frac{P_X(x)}{S(x,M,W,V,\rho,\zeta,\lambda)}\right]^{1/(1+\lambda)}\right).
\end{aligned}
\tag{55}
$$

6 Proofs

6.1 Proof of Lemma 1

The first part of Lemma 1 is obtained by a simple application of the method of types [4].

$$
\begin{aligned}
\tilde{P}_{\mathrm{ecl}} \;=\;& \sum_{\boldsymbol{x},\boldsymbol{y}} \sum_{\{\boldsymbol{u}:\, H(\hat{P}_{\boldsymbol{u}})\geq\tilde{R}_X,\ H(\hat{P}_{\boldsymbol{v}})\geq\tilde{R}_Y\}} P(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{u}|\boldsymbol{x})B(\boldsymbol{v}|\boldsymbol{y}) \\
=\;& \sum_{\{\mathcal{T}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}):\, H(\hat{P}_{\boldsymbol{u}})\geq tR_X,\ H(\hat{P}_{\boldsymbol{v}})\geq tR_Y\}} |\mathcal{T}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y},\boldsymbol{u},\boldsymbol{v})| \times \\
& \exp_2\{-n[\hat{H}_{\boldsymbol{xy}}(X,Y) + D(\hat{P}_{\boldsymbol{xy}}\|P_{XY}) + F(\hat{P}_{\boldsymbol{ux}}) + G(\hat{P}_{\boldsymbol{vy}})]\} \\
\doteq\;& \max_{\{\mathcal{T}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}):\, H(\hat{P}_{\boldsymbol{u}})\geq\tilde{R}_X,\ H(\hat{P}_{\boldsymbol{v}})\geq\tilde{R}_Y\}} \exp_2\{n\hat{H}_{\boldsymbol{uvxy}}(U,V,X,Y)\} \times \\
& \exp_2\{-n[\hat{H}_{\boldsymbol{xy}}(X,Y) + D(\hat{P}_{\boldsymbol{xy}}\|P_{XY}) + F(\hat{P}_{\boldsymbol{ux}}) + G(\hat{P}_{\boldsymbol{v}})]\} \\
\doteq\;& \max_{\{Q_{UVXY}:\, H_Q(U)\geq\tilde{R}_X,\ H_Q(V)\geq\tilde{R}_Y\}} \exp_2\{n[H_Q(X,Y,U,V) - H_Q(X,Y) - \\
& D(Q_{XY}\|P_{XY}) - F(Q_{UX}) - G(Q_{VY})]\} \\
=\;& \exp_2\Bigg[-n \min_{\{Q_{UVXY}:\, H_Q(U)\geq\tilde{R}_X,\ H_Q(V)\geq\tilde{R}_Y\}} \{D(Q_{XY}\|P_{XY}) + \\
& F(Q_{UX}) + G(Q_{VY}) - H_Q(U,V|X,Y)\}\Bigg] \\
=\;& \exp_2\Bigg[-n \min_{\{Q_{XY},Q_{U|X},Q_{V|Y}:\, H_Q(U)\geq\tilde{R}_X,\ H_Q(V)\geq\tilde{R}_Y\}} \{D(Q_{XY}\|P_{XY}) + F(Q_{UX}) + \\
& G(Q_{VY}) - H_Q(U|X) - H_Q(V|Y)\}\Bigg],
\end{aligned}
\tag{56}
$$

where the last step follows since the maximum of $H_Q(U, V|X, Y)$, for given Q_{UX} and Q_{VY}, is achieved with $H_Q(U, V|X, Y) = H_Q(U|X) + H_Q(V|Y)$.

Moving on the second part of Lemma 1, we begin from an upper bound on the error probability. Consider the sub-optimal decoder,

$$(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \arg \min_{\{(\boldsymbol{x}, \boldsymbol{y}): \ f(\boldsymbol{x}) = \boldsymbol{u}, \ g(\boldsymbol{y}) = \boldsymbol{v}\}} m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}), \tag{57}$$

where the decoding metric, m, is defined by

$$m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) = \max\{m_1(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}), m_2(\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}), m_3(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})\}, \tag{58}$$

where

$$m_1(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) = \hat{H}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{u}, \boldsymbol{v}) - F(\hat{P}_{\boldsymbol{u}\boldsymbol{x}}) - G(\hat{P}_{\boldsymbol{v}\boldsymbol{y}}) \tag{59}$$

$$m_2(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) = \hat{H}(\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{y}) - F(\hat{P}_{\boldsymbol{u}\boldsymbol{x}}) \tag{60}$$

$$m_3(\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) = \hat{H}(\boldsymbol{y}|\boldsymbol{v}, \boldsymbol{x}) - G(\hat{P}_{\boldsymbol{v}\boldsymbol{y}}). \tag{61}$$

The error probability associated with this decoding metric is given by

$$\begin{aligned}
\bar{P}_{\mathrm{err}} &= \sum_{\boldsymbol{u}, \boldsymbol{v}\boldsymbol{x}, \boldsymbol{y}} P(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{u}|\boldsymbol{x}) B(\boldsymbol{v}|\boldsymbol{y}) \bar{P}_{\mathrm{err}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) \\
&= \sum_{\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}} P(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{u}|\boldsymbol{x}) B(\boldsymbol{v}|\boldsymbol{y}) \times \\
&\quad \Pr\left[\bigcup_{\{(\boldsymbol{x}', \boldsymbol{y}') \neq (\boldsymbol{x}, \boldsymbol{y}): \ m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}', \boldsymbol{y}') \leq m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})\}} \{f(\boldsymbol{x}') = f(\boldsymbol{x}), \ g(\boldsymbol{y}') = g(\boldsymbol{y})\}\right] \\
&\triangleq \sum_{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}} P(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{u}|\boldsymbol{x}) B(\boldsymbol{v}|\boldsymbol{y}) \cdot \bar{P}_{\mathrm{err}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}).
\end{aligned} \tag{62}$$

As for $\bar{P}_{\mathrm{err}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})$, we have

$$\begin{aligned}
\bar{P}_{\mathrm{err}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}) &= \Pr\left[\bigcup_{\{(\boldsymbol{x}', \boldsymbol{y}') \neq (\boldsymbol{x}, \boldsymbol{y}): \ m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}', \boldsymbol{y}') \leq m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})\}} \{f(\boldsymbol{x}') = f(\boldsymbol{x}), \ g(\boldsymbol{y}') = g(\boldsymbol{y})\}\right] \\
&\leq \Pr\left[\bigcup_{\{\boldsymbol{x}' \neq \boldsymbol{x}, \ \boldsymbol{y}' \neq \boldsymbol{y}: \ m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}', \boldsymbol{y}') \leq m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})\}} \{f(\boldsymbol{x}') = f(\boldsymbol{x}), \ g(\boldsymbol{y}') = g(\boldsymbol{y})\}\right] + \\
&\quad \Pr\left[\bigcup_{\{\boldsymbol{x}' \neq \boldsymbol{x}: \ m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}', \boldsymbol{y}) \leq m(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y})\}} \{f(\boldsymbol{x}') = f(\boldsymbol{x})\}\right] +
\end{aligned}$$

20

$$\Pr\left[\bigcup_{\{\boldsymbol{y}'\neq\boldsymbol{y}:\ m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}')\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\{g(\boldsymbol{y}')=g(\boldsymbol{y})\}\right]$$

$$\triangleq\ \bar{P}_{\mathrm{err},1}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})+\bar{P}_{\mathrm{err},2}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})+\bar{P}_{\mathrm{err},3}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}). \tag{63}$$

Now,

$$
\begin{aligned}
\bar{P}_{\mathrm{err},1}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}) &= \Pr\left[\bigcup_{\{\tilde{\boldsymbol{x}}\neq\boldsymbol{x},\ \tilde{\boldsymbol{y}}\neq\boldsymbol{y}:\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\{f(\tilde{\boldsymbol{x}})=f(\boldsymbol{x}),\ g(\tilde{\boldsymbol{y}})=g(\boldsymbol{y})\}\right] \\
&\leq \min\left\{1, A(\boldsymbol{u}|\tilde{\boldsymbol{x}})B(\boldsymbol{v}|\tilde{\boldsymbol{y}})\Big|\{(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}):\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}\Big|\right\} \\
&= \min\left\{1, \exp_2\{-n[F(\hat{P}_{\boldsymbol{u}\tilde{\boldsymbol{x}}})+G(\hat{P}_{\boldsymbol{v}\tilde{\boldsymbol{y}}})]\}\times\right. \\
&\qquad \left.\Big|\{(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}):\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}\Big|\right\} \\
&\leq \min\left\{1, \exp_2\{-n[F(\hat{P}_{\boldsymbol{u}\tilde{\boldsymbol{x}}})+G(\hat{P}_{\boldsymbol{v}\tilde{\boldsymbol{y}}})]\}\cdot \sum_{\{(\mathcal{T}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}|\boldsymbol{u},\boldsymbol{v}):\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\right. \\
&\qquad \left.|\mathcal{T}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}|\boldsymbol{u},\boldsymbol{v})|\right\} \\
&\leq \min\left\{1, \sum_{\{\mathcal{T}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}|\boldsymbol{u},\boldsymbol{v}):\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\exp_2\{n[\hat{H}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}|\boldsymbol{u},\boldsymbol{v})-\right. \\
&\qquad \left.F(\hat{P}_{\boldsymbol{u}\tilde{\boldsymbol{x}}})-G(\hat{P}_{\boldsymbol{v}\tilde{\boldsymbol{y}}})]\}\right\} \\
&= \min\left\{1, \sum_{\{\mathcal{T}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}|\boldsymbol{u},\boldsymbol{v}):\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\exp_2[nm_1(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})]\right\}. \tag{64}
\end{aligned}
$$

In the same fashion, we obtain (see also [9])

$$\bar{P}_{\mathrm{err},2} \ \dot{\leq}\ \min\left\{1, \sum_{\{\mathcal{T}(\tilde{\boldsymbol{x}}|\boldsymbol{u},\boldsymbol{y}):\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\boldsymbol{y})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\exp_2[nm_2(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\boldsymbol{y})]\right\} \tag{65}$$

$$\bar{P}_{\mathrm{err},3} \ \dot{\leq}\ \min\left\{1, \sum_{\{\mathcal{T}(\tilde{\boldsymbol{y}}|\boldsymbol{v},\boldsymbol{x}):\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\boldsymbol{y})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\exp_2[nm_3(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\tilde{\boldsymbol{y}})]\right\}, \tag{66}$$

and so,

$$
\begin{aligned}
\bar{P}_{\mathrm{err}} \ \dot{\leq}\ \min\Big\{1, \max_{\{\hat{P}_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}|\boldsymbol{u}\boldsymbol{v}}:\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}}\{\exp_2[nm_1(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})]+ \\
\exp_2[nm_2(\boldsymbol{u},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})]+\exp_2[nm_3(\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})]\}\Big\} \tag{67}
\end{aligned}
$$

$$\doteq \min\left\{1, \max_{\{\hat{P}_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}|\boldsymbol{u}\boldsymbol{v}}:\ m(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\leq m(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})\}} \exp_2[nm(\boldsymbol{u},\boldsymbol{v},\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})]\right\} \tag{68}$$

$$= \min\left\{1, \exp_2[nm(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})]\}\right\}. \tag{69}$$

Using the method of types, the resulting error exponent is

$$\begin{aligned}
\boldsymbol{E}_{\mathrm{err}}(A,B) \ \geq\ & \min_{Q_{UVXY}}\left\{D(Q_{XY}\|P_{XY}) + F(Q_{UX}) + G(Q_{VY}) - H_Q(U,V|X,Y) + \right.\\
& [\min\{F(Q_{UX}) - H_Q(X|U,Y), G(Q_{VY}) - H_Q(Y|V,X),\\
& \left. F(Q_{UX}) + G(Q_{VY}) - H_Q(X,Y|U,V)\}]_+\right\}\\
=\ & \min_{Q_{UVXY}}\left\{D(Q_{XY}\|P_{XY}) + F(Q_{UX}) + G(Q_{VY}) - H_Q(U,V|X,Y) +\right.\\
& \min\{[F(Q_{UX}) - H_Q(X|U,Y)]_+, [G(Q_{VY}) - H_Q(Y|V,X)]_+,\\
& \left. [F(Q_{UX}) + G(Q_{VY}) - H_Q(X,Y|U,V)]_+\}\right\}\\
=\ & \min\{\boldsymbol{E}_1(A), \boldsymbol{E}_2(B), \boldsymbol{E}_3(A,B)\}, \tag{70}
\end{aligned}$$

where the inequality is due to the fact that we have analyzed a sub–optimal decoder, whereas $\boldsymbol{E}_{\mathrm{err}}(A,B)$ is defined for the optimal, MAP decoder (4). A matching lower bound, of the same exponential order, is obtained by analyzing the error probability for a given $(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})$ under the MAP decoder, where the inner summation over $\{(\boldsymbol{x}',\boldsymbol{y}')\}$ is confined to the same conditional type as that of $(\boldsymbol{x},\boldsymbol{y})$ given $(\boldsymbol{u},\boldsymbol{v})$ (where $P(\boldsymbol{x}',\boldsymbol{y}') = P(\boldsymbol{x},\boldsymbol{y})$ and $A(\boldsymbol{u}|\boldsymbol{x}') = A(\boldsymbol{u}|\boldsymbol{x})$), see, [9] where the same idea was used as well. Since the sub-optimal decoding metric (58) cannot be better than the MAP decoding metric, a simple sandwich argument yields that the last inequality is, in fact, an equality:

$$\boldsymbol{E}_{\mathrm{err}}(A,B) = \min\{\boldsymbol{E}_1(A), \boldsymbol{E}_2(B), \boldsymbol{E}_3(A,B)\}. \tag{71}$$

This completes the proof of Lemma 1.

6.2 Proof of Lemma 2

First, observe that for both $E_{\mathrm{err}}(A,B)$ and $E_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, A, B)$ to improve, the larger we can make F and G – the better. But since it is associated with a conditional distribution, there is an inherent limitation, which stems from the following simple derivation:

$$1 = \sum_{\boldsymbol{u}} A(\boldsymbol{u}|\boldsymbol{x}) = \sum_{\{\mathcal{T}(\boldsymbol{u}|\boldsymbol{x})\}} |\mathcal{T}(\boldsymbol{u}|\boldsymbol{x})| \cdot 2^{-nF(\hat{P}_{\boldsymbol{u}\boldsymbol{x}})} \doteq \max_{Q_{U|X}} 2^{n[H_Q(U|X) - F(Q_{UX})]}, \tag{72}$$

or equivalently, $\max_{Q_{U|X}}[H_Q(U|X) - F(Q_{UX})] = 0$. In other words, for each Q_X, the following holds true: for every $Q_{U|X}$, $F(Q_{UX}) \geq H_Q(U|X)$ and the inequality must be met with equality for at least one conditional distribution, $Q_{U|X}$. The best one can in maximizing $F(Q_{UX})$ (and hence also both exponents) is therefore to take it to infinity for all $Q_{U|X}$, except one conditional distribution, $Q_{U|X}$, which we denote by $Q_{U|X}^*$, and which may depend on Q_X. The same line of thought applies, of course, to G.

6.3 Proof of Theorem 1

The first part, $\boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, Q_U^*, Q_V^*) = \boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, Q_{U|X}^*, Q_{V|Y}^*)$, is obvious since the dependence of $\boldsymbol{E}_{\mathrm{ecl}}(\tilde{R}_X, \tilde{R}_Y, Q_{U|X}^*, Q_{V|Y}^*)$ on $Q_{U|X}^*$ and $Q_{V|Y}^*$ is only via its unconditional marginals, Q_U^* and Q_V^*, anyway. As for the second part, consider the following chain of inequalities, for a given Q_X and $Q_{U|X}^*$:

$$\min_{Q_{Y|UX}} \{D(Q_{XY}\|P_{XY}) + H_Q(U|X) - H_Q(U|X,Y) + [H_Q(U|X) - H_Q(X|Y,U)]_+\} \tag{73}$$

$$\leq \min_{Q_{Y|UX}=Q_{Y|X}} \{D(Q_{XY}\|P_{XY}) + H_Q(U|X) - H_Q(U|X,Y) +$$
$$[H_Q(U|X) - H_Q(X|Y,U)]_+\} \tag{74}$$

$$= \min_{Q_{Y|UX}=Q_{Y|X}} \{D(Q_{XY}\|P_{XY}) + H_Q(U|X) - H_Q(U|X) +$$
$$[H_Q(U|X) - H_Q(X|Y) + I_Q(X;U|Y)]_+\} \tag{75}$$

$$= \min_{Q_{Y|UX}=Q_{Y|X}} \{D(Q_{XY}\|P_{XY}) + [H_Q(U|X) - H_Q(X|Y) + H_Q(U|Y) - H_Q(U|X,Y)]_+\} \tag{76}$$

$$= \min_{Q_{Y|UX}=Q_{Y|X}} \{D(Q_{XY}\|P_{XY}) + [H_Q(U|X) - H_Q(X|Y) + H_Q(U|Y) - H_Q(U|X)]_+\} \tag{77}$$

$$= \min_{Q_{Y|UX}=Q_{Y|X}} \{D(Q_{XY}\|P_{XY}) + [H_Q(U|Y) - H_Q(X|Y)]_+\} \tag{78}$$

$$\leq \min_{Q_{Y|UX}=Q_{Y|X}} \{D(Q_{XY}\|P_{XY}) + [H_Q(U) - H_Q(X|Y)]_+\} \tag{79}$$

$$\leq \min_{Q_{Y|X}} \{D(Q_{XY}\|P_{XY}) + [H_Q(U) - H_Q(X|Y)]_+\}. \tag{80}$$

Minimizing the first and the last expressions in the above chain over Q_X, while holding $Q_{U|X} = Q_{U|X}^*$ for each Q_X, we get

$$\boldsymbol{E}_1(Q_{U|X}^*) \leq \min_{Q_{XY}} \{D(Q_{XY}\|P_{XY}) + [H_{Q^*}(U) - H_Q(X|Y)]_+\}, \tag{81}$$

Replacing U by V and interchanging the roles of X and Y, we similarly obtain

$$\boldsymbol{E}_2(Q_{V|Y}^*) \leq \min_{Q_{XY}} \{D(Q_{XY}\|P_{XY}) + [H_{Q^*}(V) - H_Q(Y|X)]_+\}. \tag{82}$$

23

Likewise, let us define $\mathcal{S} \triangleq \{Q_{UVXY}: \ Q_{U|X} = Q^*_{U|X}, \ Q_{V|Y} = Q^*_{V|Y}\}$ and let $\mathcal{S}' = \mathcal{S} \cap \{Q_{UVXY}: \ Q_{UVXY} = Q_{XY} \times Q_{U|X} \times Q_{V|Y}\}$. Then,

$$
\begin{aligned}
\boldsymbol{E}_3(A, B) &= \min_{Q_{UVXY} \in \mathcal{S}} \{D(Q_{XY}\|P_{XY}) + H_Q(U|X) + H_Q(V|Y) - H_Q(U,V|X,Y) + \\
&\qquad [H_Q(U|X) + H_Q(V|Y) - H_Q(X,Y|U,V)]_+\} \qquad\qquad (83) \\
&\leq \min_{Q_{UVXY} \in \mathcal{S}'} \{D(Q_{XY}\|P_{XY}) + H_Q(U|X) + H_Q(V|Y) - H_Q(U,V|X,Y) + \\
&\qquad [H_Q(U|X) + H_Q(V|Y) - H_Q(X,Y) + I_Q(X,Y;U,V)]_+\} \qquad\quad (84) \\
&= \min_{Q_{UVXY} \in \mathcal{S}'} \{D(Q_{XY}\|P_{XY}) + H_Q(U|X) + H_Q(V|Y) - H_Q(U|X) - H_Q(V|Y) + \\
&\qquad [H_Q(U|X) + H_Q(V|Y) - H_Q(X,Y) + H_Q(U,V) - H_Q(U,V|X,Y)]_+\} \qquad (85) \\
&= \min_{Q_{UVXY} \in \mathcal{S}'} \{D(Q_{XY}\|P_{XY}) + \\
&\qquad [H_Q(U|X) + H_Q(V|Y) - H_Q(X,Y) + H_Q(U,V) - H_Q(U|X) - H_Q(V|Y)]_+\}\,(86) \\
&= \min_{Q_{UVXY} \in \mathcal{S}'} \{D(Q_{XY}\|P_{XY}) + [H_Q(U,V) - H_Q(X,Y)]_+\} \qquad\qquad\qquad (87) \\
&\leq \min_{Q_{UVXY} \in \mathcal{S}'} \{D(Q_{XY}\|P_{XY}) + [H_Q(U) + H_Q(V) - H_Q(X,Y)]_+\} \qquad\qquad (88) \\
&= \min_{Q_{XY}} \{D(Q_{XY}\|P_{XY}) + [H_Q(U) + H_Q(V) - H_Q(X,Y)]_+\}, \qquad\qquad\quad (89)
\end{aligned}
$$

and the proof of Theorem 1 (including the last part) is now completed once one observes that the right–hand sides of (81), (82) and (83) are exactly the expressions obtained for $\boldsymbol{E}_{\mathrm{err}}(Q^*_U, Q^*_V)$. Indeed, for $Q^*_{U|X} = Q^*_U$ and $Q^*_{V|Y} = Q^*_V$, the minimizing Q_{UVXY} in Corollary 1 is such that (X, Y), U and V are all independent, which yields $H_Q(U|X,Y) = H_Q(U|X) = H_Q(U)$, $H_Q(V|X,Y) = H_Q(V|Y) = H_Q(V)$, $H_Q(X,Y|U,V) = H_Q(X,Y)$, $H_Q(X|Y,U) = H_Q(X|Y)$, and $H_Q(Y|X,V) = H_Q(Y|X)$.

6.4 Proof of Eq. (38)

$$
\begin{aligned}
\zeta(Q_{UX}) &= \inf_{Q_{Y|UX}} \{D(Q_{XY}\|P_{XY}) + H_Q(U|X) - H_Q(U|X,Y) + \\
&\qquad [H_Q(U|X) - H_Q(X|U,Y)]_+\} \\
&= \inf_{Q_{Y|UX}} \{D(Q_{XY}\|P_{XY}) + H_Q(Y|X) - H_Q(Y|U,X) + \\
&\qquad [H_Q(U|X) - H_Q(X|U) + H_Q(Y|U) - H_Q(Y|U,X]_+\} \\
&= \inf_{Q_{Y|UX}} \{D(Q_{XY}\|P_{XY}) + H_Q(Y|X) - H_Q(Y|U,X) +
\end{aligned}
$$

$$[H_Q(U) - H_Q(X) + H_Q(Y|U) - H_Q(Y|U,X)]_+\}$$

$$= \inf_{Q_{Y|UX}} \sup_{0 \le \lambda \le 1} \{D(Q_{XY}\|P_{XY}) + H_Q(Y|X) - (1+\lambda)H_Q(Y|U,X) +$$

$$\lambda[H_Q(U) - H_Q(X) + H_Q(Y|U)]$$

$$= \inf_{Q_{Y|UX}} \sup_{0 \le \lambda \le 1} \inf_{V} \sum_{u,x,y} Q_{UXY}(u,x,y) \left[\log \frac{Q_{XY}(x,y)}{P_{XY}(x,y)} - \log Q_{Y|X}(y|x) + \right.$$

$$\left. (1+\lambda)\log Q_{Y|UX}(y|u,x) - \lambda \log Q_U(u) + \lambda \log Q_X(x) - \lambda \log V(y|u) \right]$$

$$= \inf_{Q_{Y|UX}} \sup_{0 \le \lambda \le 1} \inf_{V} \sum_{u,x,y} Q_{UXY}(u,x,y) \log \frac{Q_{XY}(x,y)[Q_{Y|UX}(y|u,x)]^{1+\lambda}[Q_X(x)]^{\lambda}}{P_{XY}(x,y)Q_{Y|X}(y|x)[Q_U(u)]^{\lambda}[V(y|u)]^{\lambda}}$$

$$= \inf_{Q_{Y|UX}} \sup_{0 \le \lambda \le 1} \inf_{V} \sum_{u,x,y} Q_{UXY}(u,x,y) \log \frac{Q_X(x)[Q_{Y|UX}(y|u,x)]^{1+\lambda}[Q_X(x)]^{\lambda}}{P_{XY}(x,y)[Q_U(u)]^{\lambda}[V(y|u)]^{\lambda}}$$

$$= \inf_{Q_{Y|UX}} \sup_{0 \le \lambda \le 1} \inf_{V} \sum_{u,x,y} Q_{UXY}(u,x,y) \log \frac{[Q_X(x)]^{1+\lambda}[Q_{Y|UX}(y|u,x)]^{1+\lambda}}{P_{XY}(x,y)[Q_U(u)]^{\lambda}[V(y|u)]^{\lambda}}$$

$$= \inf_{Q_{Y|UX}} \sup_{0 \le \lambda \le 1} \inf_{V} \left\{ \sum_{u,x} Q_{UX}(u,x) \log \frac{[Q_X(x)]^{1+\lambda}}{P_X(x)[Q_U(u)]^{\lambda}} + \right.$$

$$\left. \sum_{u,x} Q_{UX}(u,x) \sum_y Q_{Y|UX}(y|u,x) \log \frac{[Q_{Y|UX}(y|u,x)]^{1+\lambda}}{P_{Y|X}(y|x)[V(y|u)]^{\lambda}} \right\}$$

$$= \inf_{Q_{Y|UX}} \inf_{V} \sup_{0 \le \lambda \le 1} \left\{ \sum_{u,x} Q_{UX}(u,x) \log \frac{[Q_X(x)]^{1+\lambda}}{P_X(x)[Q_U(u)]^{\lambda}} + \right.$$

$$\left. (1+\lambda) \sum_{u,x} Q_{UX}(u,x) \sum_y Q_{Y|UX}(y|u,x) \log \frac{Q_{Y|UX}(y|u,x)}{[P_{Y|X}(y|x)]^{1/(1+\lambda)}[V(y|u)]^{\lambda/(1+\lambda)}} \right\}$$

$$= \inf_{V} \inf_{Q_{Y|UX}} \sup_{0 \le \lambda \le 1} \left\{ \sum_{u,x} Q_{UX}(u,x) \log \frac{[Q_X(x)]^{1+\lambda}}{P_X(x)[Q_U(u)]^{\lambda}} + \right.$$

$$\left. (1+\lambda) \sum_{u,x} Q_{UX}(u,x) \sum_y Q_{Y|UX}(y|x,u) \log \frac{Q_{Y|UX}(y|x,u)}{[P_{Y|X}(y|x)]^{1/(1+\lambda)}[V(y|u)]^{\lambda/(1+\lambda)}} \right\}$$

$$= \inf_{V} \sup_{0 \le \lambda \le 1} \inf_{Q_{Y|UX}} \left\{ \sum_{u,x} Q_{UX}(u,x) \log \frac{[Q_X(x)]^{1+\lambda}}{P_X(x)[Q_U(u)]^{\lambda}} + \right.$$

$$\left. (1+\lambda) \sum_{u,x} Q_{UX}(u,x) \sum_y Q_{Y|UX}(y|u,x) \log \frac{Q_{Y|UX}(y|u,x)}{[P_{Y|X}(y|x)]^{1/(1+\lambda)}[V(y|u)]^{\lambda/(1+\lambda)}} \right\}$$

$$= \inf_{V} \sup_{0 \le \lambda \le 1} \left\{ \sum_{u,x} Q_{UX}(u,x) \log \frac{[Q_X(x)]^{1+\lambda}}{P_X(x)[Q_U(u)]^{\lambda}} - \right.$$

$$\left. (1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log \left(\sum_y [P_{Y|X}(y|x)]^{1/(1+\lambda)}[V(y|u)]^{\lambda/(1+\lambda)} \right) \right\}. \tag{90}$$

6.5 Proof of Eqs. (49) and (50)

As for eq. (49), we have:

$$
\tilde{\boldsymbol{E}}(E_{\text{ecl}}, D) = \inf_{Q_X} \sup_{\{Q_{U|X}: \ \mathbb{E}_Q d(X,U) \leq D\}} \zeta(Q_{UX}) \tag{91}
$$

$$
= \inf_{Q_X} \sup_{Q_{U|X}} \inf_{\zeta \geq 0} \inf_V \sup_{0 \leq \lambda \leq 1} \left\{ \sum_{u,x} Q_{UX}(u,x) \log \frac{[Q_X(x)]^{1+\lambda}}{P_X(x)[Q_U(u)]^{\lambda}} + \zeta[D - \mathbb{E}_Q d(X,U)] - \right.
$$
$$
\left. (1 + \lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \right\} \tag{92}
$$

$$
= \inf_{Q_X} \sup_{Q_{U|X}} \inf_{\zeta \geq 0} \inf_V \sup_{0 \leq \lambda \leq 1} \min_W \left\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) - \right.
$$
$$
\lambda \mathbb{E}_Q \log W(U) - \zeta \mathbb{E}_Q d(X,U) -
$$
$$
\left. (1 + \lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \right\} \tag{93}
$$

$$
= \inf_{Q_X} \sup_{Q_{U|X}} \inf_{\zeta \geq 0} \inf_V \sup_{0 \leq \lambda \leq 1} \min_W \left\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) - \right.
$$
$$
\left. \sum_x Q_X(x) \sum_u Q_{U|X}(u|x) \log[W^{\lambda}(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)] \right\} \tag{94}
$$

$$
= \inf_{Q_X} \sup_{Q_{U|X}} \inf_{\zeta \geq 0} \sup_{0 \leq \lambda \leq 1} \min_{V,W} \left\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) - \right.
$$
$$
\left. \sum_x Q_X(x) \sum_u Q_{U|X}(u|x) \log[W^{\lambda}(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)] \right\} \tag{95}
$$

$$
= \inf_{Q_X} \sup_{0 \leq \lambda \leq 1} \sup_{Q_{U|X}} \inf_{\zeta \geq 0} \min_{V,W} \left\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) - \right.
$$
$$
\left. \sum_x Q_X(x) \sum_u Q_{U|X}(u|x) \log[W^{\lambda}(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)] \right\} \tag{96}
$$

$$
= \inf_{Q_X} \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \sup_{Q_{U|X}} \left\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) - \right.
$$
$$
\left. \sum_x Q_X(x) \sum_u Q_{U|X}(u|x) \log[W^{\lambda}(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)] \right\} \tag{97}
$$

$$
= \inf_{Q_X} \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \left\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) - \right.
$$
$$
\left. \sum_x Q_X(x) \min_u \log[W^{\lambda}(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)] \right\} \tag{98}
$$

$$
\overset{\Delta}{=} \inf_{Q_X} \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \left\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) - \right.
$$

$$\sum_x Q_X(x) \log T(x, \zeta, \lambda, V, W)\bigg\} \tag{99}$$

$$\geq \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \inf_{Q_X} \bigg\{ \zeta D + D(Q_X \| P_X) - \lambda H_Q(X) -$$
$$\sum_x Q_X(x) \log T(x, \zeta, \lambda, V, W)\bigg\} \tag{100}$$

$$= \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \inf_{Q_X} \bigg\{ \zeta D + \sum_x Q_X(x) \log \frac{Q_X^{1+\lambda}(x)}{P_X(x) T(x, \zeta, \lambda, V, W)}\bigg\} \tag{101}$$

$$= \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \inf_{Q_X} \bigg\{ \zeta D +$$
$$(1+\lambda) \cdot \sum_x Q_X(x) \log \frac{Q_X(x)}{[P_X(x) T(x, \zeta, \lambda, V, W)]^{1/(1+\lambda)}}\bigg\} \tag{102}$$

$$= \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \bigg[\zeta D - (1+\lambda) \log \bigg\{ \sum_x [P_X(x) T(x, \zeta, \lambda, V, W)]^{1/(1+\lambda)}\bigg\}\bigg]$$

$$= \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{V,W} \bigg[\zeta D - (1+\lambda) \log \bigg\{ \sum_x P_X(x)^{1/(1+\lambda)} \times$$
$$\min_u W^{\lambda/(1+\lambda)}(u) 2^{\zeta d(x,u)} \sum_y P(y|x)^{1/(1+\lambda)} V(y|u)^{\lambda/(1+\lambda)}\bigg\}\bigg]$$

$$= \sup_{0 \leq \lambda \leq 1} \inf_{\zeta \geq 0} \min_{C} \bigg[\zeta D -$$
$$(1+\lambda) \log \bigg\{ \sum_x \min_u 2^{\zeta d(x,u)} \sum_y P(x,y)^{1/(1+\lambda)} C(u,y)^{\lambda/(1+\lambda)}\bigg\}\bigg]. \tag{103}$$

Concerning eq. (50), we have the following chain of equalities and inequalities:

$$\hat{\boldsymbol{E}}(\tilde{R}, E_{\mathrm{ecl}}, D) = \inf_{\{Q_X: \, D(Q_X \| P_X) \leq E_{\mathrm{ecl}}\}} \sup_{M} \sup_{\{Q_{U|X}: \, -\mathbb{E}_Q \log M(U) \leq \tilde{R}, \, \mathbb{E}_Q d(X,U) \leq D\}}$$
$$\inf_V \sup_{0 \leq \lambda \leq 1} \bigg\{ D(Q_X \| P_X) - \lambda H_Q(X) +$$
$$\lambda H_Q(U) - (1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda)\bigg\}$$

$$= \inf_{\{Q_X: \, D(Q_X \| P_X) \leq E_{\mathrm{ecl}}\}} \bigg[D(Q_X \| P_X) + \sup_M \sup_{Q_{U|X}} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \sup_{0 \leq \lambda \leq 1} \bigg\{ \rho \tilde{R} + \zeta D +$$
$$\rho \mathbb{E}_Q \log M(U) - \zeta \mathbb{E}_Q d(X,U) - \lambda H_Q(X) + \lambda H_Q(U) -$$
$$(1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda)\bigg\}\bigg]$$

$$\geq \inf_{\{Q_X: \, D(Q_X \| P_X) \leq E_{\mathrm{ecl}}\}} \bigg[D(Q_X \| P_X) + \sup_M \sup_{Q_{U|X}} \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \bigg\{ \rho \tilde{R} + \zeta D +$$

27

$$\rho \mathbb{E}_Q \log M(U) - \zeta \mathbb{E}_Q d(X,U) - \lambda H_Q(X) + \lambda H_Q(U) -$$

$$(1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \Big\} \Big]$$

$$= \inf_{\{Q_X: \ D(Q_X\|P_X) \leq E_{\mathrm{ecl}}\}} \Big[D(Q_X\|P_X) + \sup_{0 \leq \lambda \leq 1} \sup_M \sup_{Q_{U|X}} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \Big\{ \rho \tilde{R} + \zeta D +$$

$$\rho \mathbb{E}_Q \log M(U) - \zeta \mathbb{E}_Q d(X,U) - \lambda H_Q(X) + \lambda H_Q(U) -$$

$$(1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \Big\} \Big]$$

$$= \inf_{\{Q_X: \ D(Q_X\|P_X) \leq E_{\mathrm{ecl}}\}} \Big[D(Q_X\|P_X) + \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \sup_M \sup_{Q_{U|X}} \Big\{ \rho \tilde{R} + \zeta D +$$

$$\rho \mathbb{E}_Q \log M(U) - \zeta \mathbb{E}_Q d(X,U) - \lambda H_Q(X) + \lambda H_Q(U) -$$

$$(1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \Big\} \Big]$$

$$= \inf_{\{Q_X: \ D(Q_X\|P_X) \leq E_{\mathrm{ecl}}\}} \Big[D(Q_X\|P_X) + \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \sup_M \sup_{Q_{U|X}} \inf_W \Big\{ \rho \tilde{R} + \zeta D +$$

$$\rho \mathbb{E}_Q \log M(U) - \zeta \mathbb{E}_Q d(X,U) - \lambda H_Q(X) - \lambda \mathbb{E}_Q \log W(U) -$$

$$(1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \Big\} \Big]$$

$$= \inf_{\{Q_X: \ D(Q_X\|P_X) \leq E_{\mathrm{ecl}}\}} \Big[D(Q_X\|P_X) + \sup_{0 \leq \lambda \leq 1} \Big(- \lambda H_Q(X) + \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \Big[\rho \tilde{R} + \zeta D +$$

$$\inf_V \sup_M \inf_W \sup_{Q_{U|X}} \Big\{$$

$$\lambda \mathbb{E}_Q \log W(U) - \zeta \mathbb{E}_Q d(X,U) - (1+\lambda) \sum_{u,x} Q_{UX}(u,x) \log Z(u,x,V,\lambda) \Big\} \Big] \Big) \Big]$$

$$= \inf_{\{Q_X: \ D(Q_X\|P_X) \leq E_{\mathrm{ecl}}\}} \Big[D(Q_X\|P_X) + \sup_{0 \leq \lambda \leq 1} \Big(- \lambda H_Q(X) +$$

$$\inf_{\rho \geq 0} \inf_{\zeta \geq 0} \Big[\rho \tilde{R} + \zeta D + \inf_V \sup_M \inf_W \sup_{Q_{U|X}} \Big\{$$

$$\sum_x Q_X(x) \sum_u Q_{U|X}(u|x) \log \frac{M^\rho(u)}{W^\lambda(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)} \Big\} \Big] \Big) \Big]$$

$$= \inf_{\{Q_X: \ D(Q_X\|P_X) \leq E_{\mathrm{ecl}}\}} \Big[D(Q_X\|P_X) + \sup_{0 \leq \lambda \leq 1} \Big(- \lambda H_Q(X) +$$

$$\inf_{\rho \geq 0} \inf_{\zeta \geq 0} \Big[\rho \tilde{R} + \zeta D +$$

$$\inf_V \sup_M \inf_W \Big\{ \sum_x Q_X(x) \max_u \log \frac{M^\rho(u)}{W^\lambda(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)} \Big\} \Big] \Big) \Big]$$

$$= \inf_{Q_X} \sup_{\theta \geq 0} \left[D(Q_X \| P_X) + \theta[D(Q_X \| P_X) - E_{\text{ecl}}] + \sup_{0 \leq \lambda \leq 1} \left(-\lambda H_Q(X) + \right. \right.$$

$$\inf_{\rho \geq 0} \inf_{\zeta \geq 0} \left[\rho \tilde{R} + \zeta D + \right.$$

$$\left. \left. \left. \inf_V \sup_M \inf_W \left\{ \sum_x Q_X(x) \log \max_u \frac{M^\rho(u)}{W^\lambda(u) 2^{\zeta d(x,u)} Z^{1+\lambda}(u,x,V,\lambda)} \right\} \right] \right) \right]$$

$$\triangleq \inf_{Q_X} \sup_{\theta \geq 0} \left[D(Q_X \| P_X) + \theta[D(Q_X \| P_X) - E_{\text{ecl}}] + \right.$$

$$\sup_{0 \leq \lambda \leq 1} \left(-\lambda H_Q(X) + \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \left[\rho \tilde{R} + \zeta D + \right. \right.$$

$$\left. \left. \left. \inf_V \sup_M \inf_W \sum_x Q_X(x) \log S(x, M, W, V, \rho, \zeta, \lambda) \right] \right) \right]$$

$$\geq \sup_{\theta \geq 0} \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \sup_M \inf_W \inf_{Q_X} \left\{ \rho \tilde{R} + \zeta D - \theta E_{\text{ecl}} + \right.$$

$$(1 + \theta) D(Q_X \| P_X) - \lambda H_Q(X) +$$

$$\left. \sum_x Q_X(x) \log S(x, M, W, V, \rho, \zeta, \lambda) \right\}$$

$$= \sup_{\theta \geq 0} \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \sup_M \inf_W \left\{ \rho \tilde{R} + \zeta D - \theta E_{\text{ecl}} + \right.$$

$$\left. (1 + \theta + \lambda)] \inf_{Q_X} \sum_x Q_X(x) \log \frac{Q_X(x)}{[P_X^{1+\theta}(x) / S(x, M, W, V, \rho, \zeta, \lambda)]^{1/(1+\theta+\lambda)}} \right\}$$

$$= \sup_{\theta \geq 0} \sup_{0 \leq \lambda \leq 1} \inf_{\rho \geq 0} \inf_{\zeta \geq 0} \inf_V \sup_M \inf_W \left\{ \rho \tilde{R} + \zeta D - \theta E_0 - \right.$$

$$\left. (1 + \theta + \lambda) \log \left(\sum_x \left[\frac{P_X^{1+\theta}(x)}{S(x, M, W, V, \rho, \zeta, \lambda)} \right]^{1/(1+\theta+\lambda)} \right) \right\}. \tag{104}$$

References

[1] J. Chen, D.-K. He, and A. Jagmohan, "On the duality between Slepian–Wolf coding and channel coding under mismatched decoding," *IEEE Trans. Inform. Theory*, vol. 55, no. 9, pp. 4006–4018, September 2009.

[2] J. Chen, D.–k. He, A. Jagmohan, and L. A. Lastras–Montaño, "On the reliability function of variable–rate Slepian–Wolf coding," *Entropy*, 2017, 19, 389, doi:10.3390/e19080389

[3] 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.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981. Second Edition: Cambridge University Press, New York, 2011.

[5] C. Heegard and T. Berger, "Rate-distortion when side–information may be absent," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 6, pp. 727–734, November 1985.

[6] A. H. Kaspi, "Rate-distortion function when side–information may be present at the decoder," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 2031–2034, November 1994.

[7] N. Merhav, "Universal decoding for asynchronous Slepian-Wolf encoding," submitted for publication. Available on-line at: `https://arxiv.org/pdf/2007.12863.pdf`

[8] N. Merhav, "False–accept/false–reject trade–offs for ensembles of biometric authentication systems," *IEEE Trans. Inform. Theory*, vol. 65, no. 8, pp. 4997-5006, August 2019.

[9] 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.

[10] N. Weinberger and N. Merhav, "Optimum tradeoffs 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. (See, in particular, the original version at `https://www.ee.technion.ac.il/people/merhav/papers/p166.pdf`, pp. 7–8).