Coding for Computing*

Alon Orlitsky[†] James R. Roche[‡]

Abstract

A sender communicates with a receiver who wishes to reliably evaluate a function of their combined data. We show that if only the sender can transmit, the number of bits required is a conditional entropy of a naturally defined graph. We also determine the number of bits needed when the communicators exchange two messages.

1 Introduction

Let f be a function of two random variables X and Y. A sender $P_{\mathcal{X}}$ knows X, a receiver $P_{\mathcal{Y}}$ knows Y, and both want $P_{\mathcal{Y}}$ to reliably determine $f(X,Y)$. How many bits must $P_{\mathcal{X}}$ transmit?

Embedding this communication-complexity scenario (Yao [22]) in the standard information-theoretic setting (Shannon [17]), we assume that (1) $f(X,Y)$ must be determined for a *block* of many independent (X,Y)-instances, (2) $P_{\mathcal{X}}$ transmits after observing the whole block of X-instances, (3) a vanishing block error probability is allowed, and (4) the problem's *rate* $L_f(X|Y)$ is the number of bits transmitted for the block, normalized by the number of instances.

Two simple bounds are easily established. Two naive bounds are easily established. Clearly $L_f(X|Y) \geq H(f(X,Y)|Y)$, the number of bits required when $P_{\mathcal{X}}$ knows Y in advance, and by a simple application of the Slepian-Wolf Theorem [19], $L_f(X|Y) \leq \min\{H(g(X)|Y) : g(X)$ and Y determine $f(X,Y)\}$. Both bounds are tight in special cases, but not in general.

Drawing on rate-distortion results, we show that for every X, Y, and f,

$$L_f(X|Y) = H_G(X|Y). \tag{1}$$

The graph G, defined by Witsenhausen [20] and used in [14, 6], is the *characteristic graph* of X, Y, and f. $H_G(X|Y)$ is the *conditional G-entropy* of X given Y. It extends $H_G(X)$, the *G-entropy* of X, or *graph entropy* of (G,X), defined by Körner [10]. Graph entropy has recently been used to derive an alternative characterization of perfect graphs [5], lower bounds on perfect hashing [7, 11, 12], lower bounds for Boolean formula size [13, 16], and algorithms for sorting [8]. For an excellent review of graph entropy and its applications, see [18].

---

*Submitted, IEEE Transactions on Information Theory. This version was printed on March 22, 1996.

[†]AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974.

[‡]Center for Communications Research, Thanet Road, Princeton, NJ 08540

The lower bound ($\geq$) in (1) is proven via an analogy between $H_G(X|Y)$ and rate-distortion results of Wyner and Ziv [21] and their extension in Csiszár and Körner [4]. The upper bound ($\leq$) strengthens these rate-distortion results, showing that in certain applications the same rate suffices to achieve small block- and not just bit-error probability. The proof uses *robust* typicality, a more restrictive form of the asymptotic equi-partition property.

We also consider the more general scenario in which the communicators can exchange two messages. $P_{\mathcal{Y}}$ sends a message based on the block of Y's, and $P_{\mathcal{X}}$ responds with a message based on $P_{\mathcal{Y}}$'s message and the block of X's. Again, $P_{\mathcal{Y}}$ must accurately evaluate all $f(X, Y)$'s. $P_{\mathcal{X}}$'s and $P_{\mathcal{Y}}$'s transmission *rates* are the number of bits they transmit, normalized by the block length.

We determine the region of possible rate pairs for all X, Y, and f. The *inner* bound is derived by generalizing the one-way achievability results. To prove the (matching) *outer* bound, we extend results of Kaspi and Berger [9] to a larger class of distortion measures.

The description of the two-way rate-region is somewhat involved, hence postponed to Section 5. In Section 4 we prove Characterization (1). Section 3 uses two naive bounds and three examples to motivate the results. To begin with, the next section formally defines $L_f(X|Y)$ and $H_G(X|Y)$.

2 Definitions

Characterization (1) relates $L_f(X|Y)$ to $H_G(X|Y)$, where G is the problem's characteristic graph. Each of the following subsections defines one of these quantities.

2.1 $L_f(X|Y)$

$\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ are finite sets, and $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. $\{(X_i, Y_i)\}_{i=1}^{\infty}$ are independent instances of a random pair (X, Y) ranging over $\mathcal{X} \times \mathcal{Y}$.

A *protocol* consists of an *encoding function* $\varphi : \mathcal{X}^n \to \{0,1\}^k$ and a *decoding function* $\psi : \{0,1\}^k \times \mathcal{Y}^n \to \mathcal{Z}^n$ where $n \in \mathbb{N}$ is the *block length* and $k \in \mathbb{N}$ is the number of bits transmitted. The protocol's *rate* is k/n and its *block error probability* is

$$p(\psi(\varphi(\mathbf{X}), \mathbf{Y}) \neq f(\mathbf{X}, \mathbf{Y}))$$

where $\mathbf{X} \stackrel{\text{def}}{=} X_1, \ldots, X_n$, $\mathbf{Y} \stackrel{\text{def}}{=} Y_1, \ldots, Y_n$, and $f(\mathbf{X}, \mathbf{Y}) \stackrel{\text{def}}{=} f(X_1, Y_1), \ldots, f(X_n, Y_n)$.

A rate r is *achievable* if for every $\epsilon > 0$ and $n \in \mathbb{N}$ there is a protocol with rate[1] $\leq r$, block length $\geq n$, and block-error probability $< \epsilon$. The *rate* $L_f(X|Y)$ of (X, Y) and f is the infimum of the set of achievable rates.

Intuitively, the sender $P_{\mathcal{X}}$ knows the X_i's, the receiver $P_{\mathcal{Y}}$ knows the Y_i's, and both want $P_{\mathcal{Y}}$ to reliably evaluate the $f(X_i, Y_i)$'s. $P_{\mathcal{X}}$ considers the n-instance *block* $\mathbf{X}$ and uses k bits to transmit $\varphi(\mathbf{X})$ at an average rate of k/n bits per copy. $P_{\mathcal{Y}}$ decides that $f(\mathbf{X}, \mathbf{Y})$ is $\psi(\varphi(\mathbf{X}), \mathbf{Y})$. He is correct for the block if the two vectors agree in every coordinate, and errs otherwise. $L_f(X|Y)$ is

[1] Inequality allows for irrational rates.

the minimum number of bits that $P_{\mathcal{X}}$ must transmit per instance for $P_{\mathcal{Y}}$'s block-error probability to be arbitrarily small.

2.2 $H_G(X|Y)$

A set of vertices of a graph G is *independent* if no two are connected to each other. $\Gamma(G)$ is the collection of independent sets of G.

Let X be a random variable and let a graph G be defined over the support set of X. Körner [10] defined the *G-entropy* of X (or *graph entropy* of (G, X)) to be

$$H_G(X) \stackrel{\text{def}}{=} \min_{X \in W \in \Gamma(G)} I(W;X).$$

Elaboration is in order. X induces a probability distribution over G's vertices. For every vertex x we select a transition probability distribution $p(w|x)$ ranging over the independent sets containing x: $p(w|x) \geq 0$ and $\sum_{w \ni x} p(w|x) = 1$. This specifies a joint distribution of X and a random variable W ranging over the independent sets and always containing X. The G-entropy of X is the smallest possible mutual information between X and W. By the Data-Processing Inequality, this minimization can be restricted to W's ranging over maximal independent sets of G. Note also that $0 \leq I(W;X) \leq H(X)$ for all W, hence $0 \leq H_G(X) \leq H(X)$ for all G and X.

Example 1 For an empty G, the set of all vertices is independent and always contains X, hence $H_G(X) = 0$. For a complete G, the only independent sets are singletons, hence $W = \{X\}$ yielding $H_G(X) = I(W;X) = H(X)$. More interestingly, let X be uniformly distributed over $\{1,2,3\}$ and let G consist of the single edge $1-3$. $\Gamma(G)$ contains only two maximal independent sets: $\{1,2\}$ and $\{2,3\}$. By convexity of mutual information, $I(W;X)$ is minimized when $p(\{1,2\}|2) = p(\{2,3\}|2) = \frac{1}{2}$. Therefore, $H_G(X) = H(W) - H(W|X) = 1 - \frac{1}{3} = \frac{2}{3}$. □

We extend the definition of graph entropy to include conditioning. Let (X, Y) be a random pair and let a graph G be defined over the support set of X. The *G-entropy of X given Y* is

$$H_G(X|Y) \stackrel{\text{def}}{=} \min_{\substack{W-X-Y \\ X \in W \in \Gamma(G)}} I(W;X|Y), \tag{2}$$

where $W-X-Y$ indicates that W, X, Y is a *Markov chain*: $p(w|x,y) = p(w|x)$ for every w, x, y. As with unconditional graph entropy, the minimization can be restricted to W's ranging over maximal independent sets of G. Also, $0 \leq I(W;X|Y) \leq H(X|Y)$ for all W, hence $0 \leq H_G(X|Y) \leq H(X|Y)$ for all G and (X, Y).

Example 2 For an empty G, we can again take $W = \mathcal{X}$, hence $H_G(X|Y) = 0$. For a complete G, we must again have $W = \{X\}$, yielding $H_G(X|Y) = I(W;X|Y) = H(X|Y)$. We will later use the following example: (X, Y) is uniformly distributed over $\{(x,y) : x, y \in \{1,2,3\}, x \neq y\}$ and G consists of the single edge $1-3$. $\Gamma(G)$ contains only two maximal independent sets: $\{1,2\}$ and $\{2,3\}$. By convexity, $I(W;X|Y)$ is minimized when $p(\{1,2\}|2) = p(\{2,3\}|2) = \frac{1}{2}$. Therefore, $H_G(X|Y) = H(W|Y) - H(W|XY) = \frac{1}{3} + \frac{2}{3} \cdot h(\frac{1}{4}) - \frac{1}{3} = \frac{2}{3} \cdot h(\frac{1}{4})$. □

2.3 G

The *characteristic graph* G of X, Y, and f was defined by Witsenhausen [20]. Its vertex set is the support set of X, and distinct vertices x, x' are connected if there is a y such that $p(x,y), p(x',y) > 0$ and $f(x,y) \neq f(x',y)$.

3 Motivation

We motivate the characterization $L_f(X|Y) = H_G(X|Y)$ in (1) via three examples and two naive bounds that are tight in some special cases, but not in general.

3.1 Naive bounds

Lemma 1 (Naive lower bound) For all (X,Y) and f,

$$L_f(X|Y) \geq H(f(X,Y)|Y) \ .$$

Proof: $P_{\mathcal{X}}$ would need that many bits per instance even if he knew $\mathbf{Y}$ in advance. $\square$

$g(X)$, where g is a function defined over $\mathcal{X}$, and Y *determine* $f(X,Y)$ if there is a function h such that

$$p[h(g(X),Y) \neq f(X,Y)] = 0.$$

For example, if $X, Y \in \mathbb{N}$ and $f(X,Y) = X + Y \bmod 4$, then $g(X) \stackrel{\text{def}}{=} X \bmod 4$ and Y determine $f(X,Y)$. If, in addition, $p((X+Y) \equiv 1 \bmod 2) = 0$, then $g(X) \stackrel{\text{def}}{=} \lfloor (X \bmod 4)/2 \rfloor$ and Y determine $f(X,Y)$ as well.

Lemma 2 (Naive upper bound) For all (X,Y) and f:

$$L_f(X|Y) \leq \min\{H(g(X)|Y) : g(X) \text{ and } Y \text{ determine } f\}.$$

Proof: By the Slepian-Wolf theorem [19], $P_{\mathcal{X}}$ can transmit that many bits per instance to reliably convey $g(X_1), \ldots, g(X_n)$ to $P_{\mathcal{Y}}$. $P_{\mathcal{Y}}$ can then determine $f(\mathbf{X}, \mathbf{Y})$. $\square$

3.2 Examples

We evaluate the two bounds via three, progressively more illuminating, examples. In the first, the naive bounds coincide, hence the graph-entropy characterization in (1) provides no additional information. In the second, the naive bounds diverge and (1) shows that the upper bound is tight. In the third, (1) agrees with neither bound, showing that optimal coding can save over the naive upper-bound approach. While the examples are presented in the single-instance description, the solutions assume the usual block interpretation.

Example 3 A satellite has ten k-bit parameters to report to base. Base knows eight of the parameters (satellite does not know which ones). How many bits must the satellite transmit for base to know all ten parameters? the $10k$ bits it would need if base knew nothing at all, or the $2k$ bits it would need if it knew which two parameters base was missing?

Formally, $X^1,\ldots,X^{10}$ are random k-bit strings representing the parameters, and I', I'' ($1 \le I' < I'' \le 10$) are the random indices of the parameters not known to base. All variables are distributed uniformly and independently. $X = (X^1,\ldots,X^{10})$ and $Y = (Y^1,\ldots,Y^{10})$ where $Y^i = X^i$ for the eight indices $i \notin \{I', I''\}$ and $Y^i = *$ for the two indices $i \in \{I_1, I_2\}$. Finally, $f(X,Y) = (X_{I'}, X_{I''})$.

It is easy to verify that given Y, there is a 1-1 correspondence between X and $f(X,Y)$. Hence $H(X|Y) = H(f(X,Y)|Y) = 2k$, and the above bounds imply

$$L_f(X|Y) = 2k \ .$$

For a simple illustration, imagine that base is missing just one parameter. Satellite transmits k bits representing the bit-wise parity of all 10 parameters. Base can deduce X, and hence $f(X,Y)$. In this atypical case, the minimal rate can be achieved without error using a single copy of (X,Y). When more parameters are missing, multiple copies and a nonzero (albeit vanishing) error probability may be unavoidable. $\square$

Example 4 As in Example 3, satellite has ten k-bit parameters to report to base and base needs to learn only two parameters (only base knows which ones). This time, however, base does not know the values of the other eight. How many bits must satellite transmit now?

As before, $H(f(X,Y)|Y) = 2k$. But now if $g(X)$ and Y determine $f(X,Y)$, then $H(g(X)|Y) = H(X|Y) = 10k$. Hence the bounds yield:

$$2k \le L_f(X|Y) \le 10k \ .$$

Yet a simple calculation shows that

$$H_G(X|Y) = 10k,$$

hence (1) implies that the upper bound is tight and the lower bound is not.

Note that when, as here, Y is independent of X, the same result can be obtained without using (1). However, we do not know of more direct proofs when X and Y are correlated. $\square$

Example 5 A top hat contains three cards, labeled 1, 2, and 3. Each of $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ selects one card (without replacement). What is the minimum number of bits that $P_{\mathcal{X}}$ must transmit for $P_{\mathcal{Y}}$ to determine who has the higher-valued card?

For $x, y \in \{1,2,3\}$,

$$p(x,y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{6} & \text{if } x \ne y, \\ 0 & \text{if } x = y, \end{cases} \qquad \text{and} \qquad f(x,y) = \begin{cases} 1 & \text{if } x > y, \\ 0 & \text{if } x < y. \end{cases}$$

It is easy to verify that $H(f(X,Y)|Y) = \frac{1}{3}$ and that $g(x) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } x = 1 \\ 1 & \text{if } x \in \{2,3\} \end{cases}$ and y determine $f(x,y)$. $H(g(X)|Y) = 2/3$ bits and this is the lowest entropy for all g's in the naive upper bound.

Hence the bounds yield

$$\frac{1}{3} \leq L_f(X|Y) \leq \frac{2}{3}.$$

The characteristic graph G consists of the vertices $\{1, 2, 3\}$ and the edge 1—3, hence Example 2 showed that

$$H_G(X|Y) = \frac{2}{3}h(\frac{1}{4}) \approx .541.$$

Therefore (1) implies that neither of the naive bounds is tight and that $\approx .125$ bits per instance can be saved over the naive encoding. □

3.3 Special cases

In special cases one or both naive bounds are tight.

1. When $p(x, y) > 0$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the upper bound is tight.

   Write $x \not\leftrightarrow x'$ if x is not connected to x' in G, namely, if $f(x, y) = f(x', y)$ for all y such that $p(x, y), p(x', y) > 0$. In general $\not\leftrightarrow$ can be arbitrarily complicated, but when, as we assume here, all (X, Y) values are possible, $\not\leftrightarrow$ is an equivalence relation. Let $[x]$ be x's equivalence class. It is easy to verify that $[X]$ minimizes $H(g(X)|Y)$ among all g's such that $g(X)$ and Y determine $f(X, Y)$. The naive upper bound is therefore $L_f(X|Y) \leq H([X]|Y)$.

   In this special case, the upper bound is tight. However, except when $[X]$ and Y are independent, the only way we know of proving that fact (i.e., establishing a matching lower bound) is via the general result. G is a complete partite graph connecting all $\mathcal{X}$ pairs belonging to different equivalence classes. A simple extension of Example 2 shows that for all such graphs $H_G(X|Y) = H([X]|Y)$. Hence,

   $$L_f(X|Y) = H([X]|Y).$$

   In the further restricted case where all (X, Y) values are possible and for every x, x' there is a y such that $f(x, y) \neq f(x, y')$, we have $[X] = \{X\}$, hence

   $$L_f(X|Y) = H(X|Y).$$

   This was the case in Example 4. We provide two additional examples. The first describes a typical scenario where the amount of information required by $P_{\mathcal{Y}}$ depends on what he already knows. While the one-way results are disappointing, we will revisit this scenario in Section 5 and show that interesting things happen when two messages are allowed. The second example relates the current result to a well-studied communication-complexity problem.

   Example 6 Consider satellite and base once more. Satellite knows a random variable X uniformly distributed over $\{1, \ldots, m\}$ and base knows an independent Bernoulli-p random variable Y ($Y \in \{0,1\}$, $p(Y = 1) = p$). If $Y = 1$ ("alert") base wants to know X. If $Y = 0$

("relax") — it doesn't. Though not necessarily useful for the proof, it may be worthwhile to note that base computes

$$f(x,y) = x \cdot y.$$

Since all values of (X,Y) are possible, and every two $\mathcal{X}$ values are distinguishable, satellite must transmit the maximum number of bits even when p is small:

$$L_f(X|Y) = H(X|Y) = \log m.$$

This of course is disappointing. When p is small, $P_{\mathcal{Y}}$ needs to learn X only in a small fraction of the instances. Yet $P_{\mathcal{X}}$ must transmit almost all the bits.

We will return to this example in Section where we show that in two-message communication coding can save over simple schemes. $\square$

Example 7 Consider the *equality function*

$$f(x,y) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise,} \end{cases}$$

where all $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ values are possible. Then

$$L_f(X|Y) = H(X|Y). \tag{3}$$

Rabin and Yao [15] showed that $O\Big(\log\log|\mathcal{X}| + \log(\frac{1}{\epsilon})\Big)$ bits suffice for $P_{\mathcal{Y}}$ to evaluate a single instance of the equality function with error probability $\leq \epsilon$. For fixed ϵ this number can smaller than $H(X|Y)$. Equation (3) shows that when only arbitrarily small error probability is allowed, exactly $H(X|Y)$ bits per instance are required even for multiple instances. $\square$

2. When $f(X,Y) = X$, both bounds are tight.

   The Slepian-Wolf Theorem says that

$$L_f(X|Y) = H(X|Y).$$

   This equality (and the implied tightness of the bounds) also follows from the fact that the naive bounds coincide. For a still more complicated proof, we employ (1). Distinct x and x' are connected in the characteristic graph G if $p(x,y), p(x',y) > 0$ for some $y \in \mathcal{Y}$. Hence, for every y, the set $\{x : p(x,y) > 0\}$ of possible x's is a clique in G. Since the intersection of a clique and an independent set is a singleton, Y and an independent set containing X determine X. Hence

$$H_G(X|Y) = \min_{\substack{W-X-Y\\ X\in W\in\Gamma(G)}} I(W;X|Y) = H(X|Y) - \max_{\substack{W-X-Y\\ X\in W\in\Gamma(G)}} H(X|W,Y) = H(X|Y).$$

4 Proofs

Recall that

$$H_G(X|Y) \stackrel{\text{def}}{=} \min_{\substack{W-X-Y \\ X\in W\in\Gamma(G)}} I(W;X|Y).$$

We prove

Theorem 1 For every X, Y, and f,

$$L_f(X|Y) = H_G(X|Y). \qquad \square$$

The next subsection establishes the lower bound ($\geq$). The upper bound is proved in Subsection 4.2.

4.1 Lower bound

To obtain the lower bound we relax the error demands. Instead of insisting on small block-error probability, we require only that $Z_i = f(X_i, Y_i)$ for all but a small fraction of the coordinates. Applying rate-distortion results, we show that $H_G(X|Y)$ bits per copy are needed even under this weaker error requirement.

We need the following rate-distortion definitions and results. For more details see Cover and Thomas [3]. Let $d : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}$ be a *distortion measure* over two finite alphabets. The *(normalized) distortion* between two sequences $\mathbf{x} = x_1,\ldots,x_n \in \mathcal{X}^n$ and $\mathbf{z} = z_1,\ldots,z_n \in \mathcal{Z}^n$ is

$$d(\mathbf{x},\mathbf{z}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} d(x_i, z_i).$$

The *expected distortion* between two random variables X and Z is

$$d(X,Z) \stackrel{\text{def}}{=} \sum_{x,z} p(x,z)d(x,z).$$

The *(normalized) expected distortion* between two random sequences $\mathbf{X} = X_1,\ldots,X_n \in \mathcal{X}^n$ and $\mathbf{Z} = Z_1,\ldots,Z_n \in \mathcal{Z}^n$ is therefore

$$d(\mathbf{X},\mathbf{Z}) = \sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x},\mathbf{z})d(\mathbf{x},\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} d(X_i, Z_i).$$

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be independent copies of a random pair $(X,Y) \in \mathcal{X}\times\mathcal{Y}$. $P_{\mathcal{X}}$ knows $\mathbf{X} = X_1,\ldots,X_n$ while $P_{\mathcal{Y}}$ knows $\mathbf{Y} = Y_1,\ldots,Y_n$. How many bits must $P_{\mathcal{X}}$ transmit for $P_{\mathcal{Y}}$ to determine $\mathbf{X}$ within distortion D?

Specifically, as in the introduction, an (n,k) *protocol* for $n,k \in \mathbb{N}$ consists of an *encoding function* $\varphi : \mathcal{X}^n \to \{0,1\}^k$ and a *decoding function* $\psi : \{0,1\}^k \times \mathcal{Y}^n \to \mathcal{Z}^n$. The protocol's *rate* is k/n and its *distortion* is $d(\mathbf{X}, \psi(\varphi(\mathbf{X}),\mathbf{Y}))$.

A *rate-distortion pair* (R,D) is *achievable* if for every $\epsilon > 0$ there is a protocol with rate $\leq R$ and distortion $< D+\epsilon$. The *rate-distortion function* of (X,Y) and d maps every distortion D to $R(D)$ — the infimum of the rates R such that (R,D) is an achievable pair.

In a seminal paper, Wyner and Ziv [21] showed that for every (X,Y) and d,

$$R(D) = \min_{\substack{V-X-Y \\ \exists g \ d(X,g(V,Y))\le D}} I(V;X|Y),$$

where the function g is defined over the support set of (V,Y).

We need a lesser known extension of this result, allowing the distortion $d : \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\to\mathbb{R}$ to depend on Y as well as on X and Z. All other definitions are modified accordingly. For example, the protocol's distortion is $d(\mathbf{X},\mathbf{Y},\psi(\varphi(\mathbf{X}),\mathbf{Y}))$. Csiszár and Körner [4] showed that for every (X,Y) and d,

$$R(D) = \min_{\substack{V-X-Y \\ \exists g \ d(X,Y,g(V,Y))\le D}} I(V;X|Y). \tag{4}$$

To express our problem in rate distortion terms, let

$$d(x,y,z) = \begin{cases} 0 & \text{if } z = f(x,y), \\ 1 & \text{otherwise,} \end{cases}$$

so that for every random variable $Z\in\mathcal{Z}$,

$$d(X,Y,Z) = p(Z\neq f(X,Y)).$$

For a sequence of variables,

$$d(\mathbf{X},\mathbf{Y},\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} d(X_i,Y_i,Z_i) = \frac{1}{n}\sum_{i=1}^{n} p(\mathcal{Z}_i\neq f(X_i,Y_i))$$

is the expected fraction of indices i for which $Z_i \neq f(X_i,Y_i)$. Viewing $\mathbf{Z}$ as an estimate of $f(\mathbf{X},\mathbf{Y})$, the distortion $d(\mathbf{X},\mathbf{Y},\mathbf{Z})$ is the *bit-error* probability.

We are especially interested in $R(0)$, the smallest rate ensuring that the bit error probability $d(\mathbf{X},\mathbf{Y},\mathbf{Z})$ diminishes to 0. In essence, the next lemma shows that this requirement is weaker than that of vanishing block-error probability.

Lemma 3 For every X,Y, and f,

$$L_f(X|Y)\ge R(0).$$

Proof: For every $i\in\{1,\ldots,n\}$,

$$p(\mathbf{Z}\neq f(\mathbf{X},\mathbf{Y})) \ge p(\mathcal{Z}_i\neq f(X_i,Y_i)).$$

Therefore,

$$p(\mathbf{Z}\neq f(\mathbf{X},\mathbf{Y})) \ge \frac{1}{n}\sum_{i=1}^{n} p(\mathcal{Z}_i\neq f(X_i,Y_i)) = d(\mathbf{X},\mathbf{Y},\mathbf{Z}).$$

It follows that a protocol's distortion is smaller than its block-error probability, hence every rate achievable with vanishing block error is also achievable with zero distortion. $\square$

Theorem 2 For every (X,Y) and f,

$$R(0) = H_G(X|Y).$$

Proof: In view of (4) and the definition of $H_G(X|Y)$, we need to prove that

$$\min_{\substack{V-X-Y \\ \exists g\ Ed(X,Y,g(V,Y))\leq 0}} I(V;X|Y) = \min_{\substack{W-X-Y \\ X\in W\in\Gamma(G)}} I(W;X|Y).$$

$\leq$: We show that if $X \in W \in \Gamma(G)$ then there is a (partial) function g over $\Gamma(G)\times\mathcal{Y}$ such that $f(x,y) = g(w,y)$ whenever $p(w,x,y) > 0$, hence $Ed(X,Y,g(W,Y)) = 0$.

Let $w \in \Gamma(G)$ and $y \in \mathcal{Y}$. If $p(x,y) = 0$ for all $x \in w$, leave g undefined. Otherwise, let $g(w,y) \stackrel{\text{def}}{=} f(x',y)$ for any (say the first) $x' \in w$ such that $p(x',y) > 0$. By definition of w, $f(x,y) = f(x',y) = g(w,y)$ for all $x \in w$ such that $p(x,y) > 0$, hence (as $X \in W$) whenever $p(w,x,y) > 0$.

$\geq$: Suppose that $V - X - Y$ and that there exists g such that $Ed(X,Y,g(V,Y)) \leq 0$. We define W such that $X \in W \in \Gamma(G)$ and $I(W;X|Y) \leq I(V;X|Y)$.

Let $p(v,x,y)$ be the probability distribution underlying (V,X,Y). Set

$$w(v) = \{x : p(v,x) > 0\}$$

to be the set of X values that could result in v, and define the Markov chain $W - V - XY$ by

$$p(w|v,x,y) = \begin{cases} 1 & \text{if } w = w(v) \\ 0 & \text{otherwise.} \end{cases}$$

The following claims, in order, show that $X \in W \in \Gamma(G)$, that $W - X - Y$, and that $I(W;X|Y) \leq I(V;X|Y)$.

Claim 1 $X \in W \in \Gamma(G)$.

Proof: $X \in W$: $p(w,x) > 0$ implies a v such that $w = w(v)$ and $p(v,x) > 0$, by definition of $w(v)$, $x \in w(v) = w$.

$W \in \Gamma(G)$: We want to show that if $p(w) > 0$ then $f(x',y) = f(x'',y)$ for all $x', x'' \in w$ such that $p(x',y), p(x'',y) > 0$. By definition, $p(w) > 0$ implies a v such that $w = w(v)$. If $x \in w$ then $p(v,x) > 0$, hence, by Markovity, $p(x,y) > 0$ implies $p(v,x,y) > 0$. To achieve $d(X,Y,g(V,Y)) \leq 0$, we must have $g(v,y) = f(x,y)$. If follows that if $x', x'' \in w$ and $p(x',y), p(x'',y) > 0$ then $f(x',y) = g(v,y) = f(x'',y)$.

Claim 2 $W - V - XY$ and $V - X - Y$ imply $W - X - Y$.

Proof: $W - V - XY$ alone implies $W - V - X$ as

$$H(W|VX) \leq H(W|V) = H(W|VXY) \leq H(W|VX).$$

Therefore,

$$p(w|x) = \sum_v p(v|x)p(w|vx)$$

$$
\begin{aligned}
&= \sum_v p(v|x)p(w|v) \\
&= \sum_v p(v|xy)p(w|vxy) \\
&= p(w|xy).
\end{aligned}
$$

Claim 3 $W - V - XY$ implies $I(W;X|Y) \leq I(V;X|Y)$.
Proof: First observe that $W - VY - X$ as

$$H(W|VXY) \leq H(W|VY) \leq H(W|V) = H(W|VXY).$$

Hence,

$$
\begin{aligned}
I(W;X|Y) &= H(X|Y) - H(X|W,Y) \\
&\leq H(X|Y) - H(X|W,V,Y) \\
&= H(X|Y) - H(X|V,Y) \\
&= I(V;X|Y),
\end{aligned}
$$

completing the proof of the theorem. □

Lemma 3 and Theorem 2 prove the lower bound in Theorem 1.

4.2 Upper bound

In the previous subsection we saw that $R(0) = H_G(X|Y)$. Namely, $P_{\mathcal{X}}$ must transmit about $H_G(X|Y)$ bits per copy to achieve vanishing *bit* error probability. We now show that the same rate suffices to achieve vanishing *block* error probability as well.

The first part of the proof of Theorem 2 constructs a partial function $g : \Gamma(G) \times \mathcal{Y} \to \mathcal{Z}$ such that $f(x,y) = g(w,y)$ whenever $x \in w$ and $p(x,y) > 0$; namely, for all $w \in \Gamma(G)$,

$$p[f(X,Y) = g(w,Y)|X \in w] = 1.$$

The protocol communicates to $P_{\mathcal{Y}}$ a sequence $\mathbf{W} = W_1,\ldots,W_n$ such that $X_i \in W_i$ for all i. $P_{\mathcal{Y}}$ can therefore evaluate $f(X_i,Y_i)$ for every $i \in \{1,\ldots,n\}$.

The proof uses various properties of *robust typicality*, defined and analyzed in Appendix B. For a random variable X distributed over a support set $\mathcal{X}$ and for $n \in \mathbb{N}$ and $\delta > 0$, we define a δ-robustly-typical set $T_{X,\delta} \subseteq \mathcal{X}^n$. Among the properties of robust typicality is that if $\mathbf{x} \in T_{X,\delta}$ then $p(x) = 0$ implies that for all i, $x_i \neq x$.

Let $p(w|x)$ achieve $H_G(X|Y)$ and let (W,X,Y) be distributed according to the implied joint distribution $p(w,x,y) = p(x,y) \cdot p(w|x)$. By definition, $x \notin w$ implies that $p(w,x) = 0$. Hence, if $(\mathbf{w},\mathbf{X}) \in T_{(W,X),\delta}$ for some $\delta > 0$, then $X_i \in w_i$ for all i.

Lemma 4 If $(\mathbf{w},\mathbf{X}) \in T_{(W,X),\delta}$ for some $\delta > 0$, then for all i,

$$p[f(X_i,Y_i) = g(w_i,Y_i)] = 1.$$

Hence, if $P_{\mathcal{Y}}$ knows *any* $\mathbf{w}$ robustly typical with $\mathbf{X}$, he can compute $f(X_i,Y_i)$ for all i. □

The protocol uses roughly $I(W;X|Y) = H_G(X|Y)$ bits to convey to $P_{\mathcal{Y}}$ a sequence $\mathbf{w}$ that is robustly typical with $\mathbf{X}$. It strengthens similar (strong-typicality) results of Wyner and Ziv [21] and Csiszár and Körner [4]. The proof follows the outlines in Berger [2] and Cover and Thomas [3]. It also addresses subtleties arising in the detailed proof.

For every rate $r > H_G(X|Y)$ we exhibit a collection of deterministic protocols. Each protocol's rate is $\leq r$, and the collection's average block-error probability is exponentially small. Standard arguments imply a single deterministic protocol with the same rate and block error probability. Before formally presenting the protocol, we need a number of definitions.

Let

$$0 < \delta < r - H_G(X|Y).$$

Set

$$s \stackrel{\text{def}}{=} \left\lceil 2^{(I(W;X)+\delta/2)n} \right\rceil \qquad \text{and} \qquad t \stackrel{\text{def}}{=} \left\lceil 2^{(I(W;X)-I(W;Y)+\delta)n} \right\rceil.$$

For $j \in \{1,\ldots,s\}$ let $\mathbf{W}^j \stackrel{\text{def}}{=} W_1^j,\ldots,W_n^j$, where each W_i^j is distributed according to $p(w)$, independently of $\mathbf{X}$, $\mathbf{Y}$, and all other W_i^j's. Let Φ be a random mapping from $\{1,\ldots,s\}$ to $\{1,\ldots,t\}$. Note that the $\mathbf{W}^j$'s are independent of the values of X and Y known to $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$. They are chosen ahead of time as part of the protocol.

We show that with high probability there is a $\mathbf{W}^j$ that is robustly jointly typical with $\mathbf{X}$ and that if $P_{\mathcal{X}}$ transmits $\Phi(j)$, $P_{\mathcal{Y}}$ can determine j and therefore $\mathbf{W}^j$. It will follow from Lemma 4 that $P_{\mathcal{Y}}$ can determine $f(\mathbf{X},\mathbf{Y})$ with no errors.

The actual proof requires additional work. Pick any

$$0 < \delta_1 < \delta_2 < \delta_3 < \delta/4,$$

and with

$$H \stackrel{\text{def}}{=} \max\{H(X,Y), 2H(X), 2H(Y)\}$$

define the robust typicality sets

$$T_1 \stackrel{\text{def}}{=} T_{X,\delta_1/H}, \qquad T_2 \stackrel{\text{def}}{=} T_{(W,X),\delta_2/H}, \qquad T_3 \stackrel{\text{def}}{=} T_{(W,Y),\delta_3/H},$$

and error bounds

$$\epsilon_1(n) \stackrel{\text{def}}{=} 2|S_X| e^{-(\delta_1/H)^2 \mu_X n/3} \qquad \text{and} \qquad \epsilon_3(n) \stackrel{\text{def}}{=} 2|S_{W,Y}| e^{-\frac{(\delta_3-\delta_1)^2}{H(H+\delta_1)}\cdot \mu_{W,Y} n/3},$$

where $S_X \stackrel{\text{def}}{=} \{x \in \mathcal{X} : p(x) > 0\}$ is the *support set* of X and $\mu_X = \min_{x\in S_X} p(x)$ is the smallest nonzero X-probability (similarly for $S_{W,Y}$ and $\mu_{W,Y}$). The following result is crucial in the proof.

Lemma 5 For every n,

1. $p[\mathbf{X} \notin T_1] \leq \epsilon_1(n)$.

2. For every $j \in \{1,\ldots,s\}$ and all $\mathbf{x} \in T_1$, $\quad p[(\mathbf{W}^j,\mathbf{x}) \in T_2] \geq 2^{-(I(W;X)+\delta_2)n}$.

3. For all $(\mathbf{w},\mathbf{x}) \in T_2$, $\quad p[(\mathbf{w},\mathbf{Y}) \notin T_3 | \mathbf{X} = \mathbf{x}] \leq \epsilon_3(n)$.

4. For every $j \in \{1,\ldots,s\}$, $p\big[(\mathbf{W}^j, \mathbf{Y}) \in T_3\big] \le 2^{-(I(W;Y)-\delta_3)n}$.

Proof: The four parts follow, in order, from Lemmas 17 and 25 and Corollaries 3 and 4. □

We now apply these assertions to the protocol.

Formally, define

$$\mathcal{J} \stackrel{\text{def}}{=} \{j : (\mathbf{W}^j, \mathbf{X}) \in T_2\}$$

to be the random index set of the $\mathbf{W}^j$'s that are robustly jointly typical with $\mathbf{X}$. If $\mathcal{J} \neq \emptyset$, the empty set, let J be the first[2] element of $\mathcal{J}$. If $\mathcal{J} = \emptyset$, J is undefined. Note that $\mathcal{J}$ and J are functions of $\mathbf{W}^1,\ldots,\mathbf{W}^s$ and $\mathbf{X}$, and that if $\mathcal{J} \neq \emptyset$ then $\mathbf{W}^J$ is robustly jointly typical with $\mathbf{X}$.

Finally, if $\mathcal{J} \neq \emptyset$, let

$$\mathcal{K} \stackrel{\text{def}}{=} \{k : (\mathbf{W}^k, \mathbf{Y}) \in T_3 \text{ and } \Phi(k) = \Phi(J)\}$$

be the random index set of $\mathbf{W}^k$'s that are robustly jointly typical with $\mathbf{Y}$ and are mapped by Φ to the same element as J. If $|\mathcal{K}| = 1$, let K be its unique element. Otherwise, K is undefined. When defined, K is the only k in $\Phi^{-1}\Phi(J)$ such that $\mathbf{W}^k$ is robustly jointly typical with $\mathbf{Y}$.

Protocol:

$P_{\mathcal{X}}$: If $\mathcal{J}$ is empty, transmits an error message. Otherwise, transmits $\Phi(J)$.

$P_{\mathcal{Y}}$: If $|\mathcal{K}| \neq 1$, declares an error. Otherwise, proceeds to determine $g(W_i^K, Y_i)$ for all i.

Error Analysis:

If $K = J$ then $(\mathbf{W}^K, X) \in T_2$. By Lemma 4, $f(X_i, Y_i) = g(W_i^K, Y_i)$ for every $i \in \{1,\ldots,n\}$, hence $P_{\mathcal{Y}}$ is correct on the whole block. An error can occur only if $\mathcal{J}$ is empty, if $|\mathcal{K}| \neq 1$, or if $K \neq J$; namely, if one or more of the following error events occurs:

E_1 : $\mathcal{J} = \emptyset$

E_2 : $\mathcal{J} \neq \emptyset$ and $(\mathbf{W}^J, \mathbf{Y}) \notin T_3$

E_3 : $\mathcal{J} \neq \emptyset$ and there is $k \neq J$ such that $(\mathbf{W}^k, \mathbf{Y}) \in T_3$ and $\Phi(k) = \Phi(J)$.

Lemmas 6 – 8 show that each of these error events occurs with exponentially small probability.

Corollary 1 $p[E_1 \cup E_2 \cup E_3] \le e^{-2^{\delta n/4}} + \epsilon_1(n) + \epsilon_3(n) + 2^{-\delta n/4}$. □

It follows that for every rate $r > H_G(X|Y)$ there is a deterministic protocol with rate $\le r$ and exponentially small block-error probability. The upper bound of Theorem 1 follows.

Lemma 6 $p(E_1) \le e^{-2^{\delta n/4}} + \epsilon_1(n)$.

Proof: By definition,

$$p(E_1) = p(\mathcal{J} = \emptyset) = p\Big[(\mathbf{W}^j, \mathbf{X}) \notin T_2 \text{ for all } j \in \{1,\ldots,s\}\Big].$$

[2] Any other choice, e.g., a random element of $\mathcal{J}$, will do.

By Part (2) of Lemma 5, for every $\mathbf{x} \in T_1$ and $j \in \{1,\ldots,s\}$,

$$p\Big[(\mathbf{W}^j, \mathbf{x}) \in T_2\Big] \geq 2^{-(I(W;X)+\delta_2)n}.$$

Hence, for every $\mathbf{x} \in T_1$,

$$p(E_1|\mathbf{X} = \mathbf{x}) \leq \Big(1 - 2^{-(I(W;X)+\delta_2)n}\Big)^s \leq e^{-s\cdot 2^{-(I(W;X)+\delta_2)n}} \leq e^{-2^{(\delta/2-\delta_2)n}} < e^{-2^{\delta n/4}},$$

and therefore,

$$p(E_1|\mathbf{X} \in T_1) \leq e^{-2^{\delta n/4}}.$$

Part (1) of Lemma 5 implies that

$$\begin{aligned} p(E_1) &= p(E_1 \text{ and } \mathbf{X} \in T_1) + p(E_1 \text{ and } \mathbf{X} \notin T_1) \\ &\leq p(E_1|\mathbf{X} \in T_1) + p(\mathbf{X} \notin T_1) \\ &\leq e^{-2^{\delta n/4}} + \epsilon_1(n). \end{aligned}$$

□

Lemma 7 $p(E_2) \leq \epsilon_3(n)$.
Proof: By definition of E_2 and standard calculations,

$$\begin{aligned} p(E_2) &= p\Big(\mathcal{J} \neq \emptyset, (\mathbf{W}^J, \mathbf{Y}) \notin T_3\Big) \\ &= \sum_{(\mathbf{w},\mathbf{x})\in T_2} p\Big(\mathcal{J} \neq \emptyset, (\mathbf{W}^J, \mathbf{Y}) \notin T_3, \mathbf{W}^J = \mathbf{w}, \mathbf{X} = \mathbf{x}\Big) \\ &= \sum_{(\mathbf{w},\mathbf{x})\in T_2} p\Big(\mathcal{J} \neq \emptyset, \mathbf{W}^J = \mathbf{w}, \mathbf{X} = \mathbf{x}\Big) \cdot p\Big((\mathbf{w}, \mathbf{Y}) \notin T_3 | \mathcal{J} \neq \emptyset, \mathbf{W}^J = \mathbf{w}, \mathbf{X} = \mathbf{x}\Big). \end{aligned}$$

If $W - X - Y$ is a Markov chain, so is $g(W, X) - X - Y$ for any function g. By construction, $(\mathbf{W}^1, \ldots, \mathbf{W}^s) - \mathbf{X} - \mathbf{Y}$ is a Markov chain, and both $\mathcal{J}$ and $\mathbf{W}^J$ depend only on $(\mathbf{W}^1, \ldots, \mathbf{W}^s)$ and X, hence $(\mathcal{J}, \mathbf{W}^J) - \mathbf{X} - \mathbf{Y}$ is a Markov chain. Incorporating Part (3) of Lemma 5, we obtain

$$\begin{aligned} p(E_2) &= \sum_{(\mathbf{w},\mathbf{x})\in T_2} p\Big(\mathcal{J} \neq \emptyset, \mathbf{W}^J = \mathbf{w}, \mathbf{X} = \mathbf{x}\Big) \cdot p((\mathbf{w}, \mathbf{Y}) \notin T_3 | \mathbf{X} = \mathbf{x}) \\ &\leq \sum_{(\mathbf{w},\mathbf{x})\in T_2} p\Big(\mathcal{J} \neq \emptyset, \mathbf{W}^J = \mathbf{w}, \mathbf{X} = \mathbf{x}\Big) \cdot \epsilon_3(n) \\ &\leq \epsilon_3(n). \end{aligned}$$

□

Lemma 8 $P(E_3) \leq 2^{-\delta n/4}$.
Proof: By definition,

$$\begin{aligned} p(E_3) &= p\Big[\mathcal{J} \neq \emptyset,\ \exists k \neq J \text{ s.t. } (\mathbf{W}^k, \mathbf{Y}) \in T_3 \text{ and } \Phi(k) = \Phi(J)\Big] \\ &\leq \sum_{j=1}^{s} \sum_{\substack{k=1 \\ k\neq j}}^{s} p\Big(\mathcal{J} \neq \emptyset, J = j, (\mathbf{W}^k, \mathbf{Y}) \in T_3\Big) p\Big(\Phi(k) = \Phi(j) | \mathcal{J} \neq \emptyset, J = j, (\mathbf{W}^k, \mathbf{Y}) \in T_3\Big). \end{aligned}$$

For $k \neq j$,

$$p\Big(\Phi(k) = \Phi(j) | \mathcal{J} \neq \emptyset, J = j, (\mathbf{W}^k, \mathbf{Y}) \in T_3\Big) = p(\Phi(k) = \Phi(j)) = \frac{1}{t}.$$

Reversing order of summation to overcome dependency of the events $J = j$ and $(\mathbf{W}^k, \mathbf{Y}) \in T_3$, and using Part (4) of Lemma 5,

$$\begin{aligned}\sum_{j=1}^{s}\sum_{\substack{k=1\\k\neq j}}^{s} p\Big(\mathcal{J}\neq\emptyset, J=j, (\mathbf{W}^k,\mathbf{Y})\in T_3\Big) &= \sum_{k=1}^{s}\sum_{\substack{j=1\\j\neq k}}^{s} p\Big(\mathcal{J}\neq\emptyset, J=j, (\mathbf{W}^k,\mathbf{Y})\in T_3\Big)\\ &\leq \sum_{k=1}^{s} p\Big(\mathcal{J}\neq\emptyset, (\mathbf{W}^k,\mathbf{Y})\in T_3\Big)\\ &\leq s\cdot p\Big((\mathbf{W}^1,\mathbf{Y})\in T_3\Big)\\ &\leq s\cdot 2^{-(I(W;Y)-\delta_3)n}.\end{aligned}$$

Therefore,

$$p(E_3) \leq \frac{s}{t}\cdot 2^{-(I(W;Y)-\delta_3)n} \leq 2^{-(\delta-\delta/2-\delta_3)n} < 2^{-\delta n/4}.$$

where the second inequality ignores the "ceiling" in the definition of s — an easily fixable flaw. □

5 Two-way communication

We consider the number of bits that must be transmitted when the communicators can exchange two messages. The next subsection defines the problem and the region of achievable rates. Subsection 5.2 describes the results. Subsection 5.3 shows that even in natural scenarios coding can reduce transmissions over simple-minded communication schemes. Subsection 5.4 gives the proof of the main theorem on two-message communication.

5.1 Definitions

As before, $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ are finite sets, and $f : \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$. $\{(X_i, Y_i)\}_{i=1}^{\infty}$ are independent instances of a random pair (X, Y) ranging over $\mathcal{X}\times\mathcal{Y}$.

A *two-message protocol* consists of a *Y-encoding function* $\varphi : \mathcal{Y}^n \to \{0,1\}^l$, an *$X$-encoding function* $\xi : \{0,1\}^l \times \mathcal{X}^n \to \{0,1\}^k$ and a *decoding function* $\psi : \{0,1\}^l \times \{0,1\}^k \times \mathcal{Y}^n \to \mathcal{Z}^n$. Here, $n \in \mathbb{N}$ is the *block length*, $k \in \mathbb{N}$ is the number of bits transmitted by $P_{\mathcal{X}}$ and $l \in \mathbb{N}$ is the number of bits transmitted by $P_{\mathcal{Y}}$. The protocol's *x-rate* is k/n and its *y-rate* is l/n. The *block error probability* is

$$p\Big[\psi(\varphi(\mathbf{Y}), \xi(\varphi(\mathbf{Y}), \mathbf{X}), \mathbf{Y}) \neq f(\mathbf{X}, \mathbf{Y})\Big],$$

where $\mathbf{X} \stackrel{\text{def}}{=} X_1,\ldots,X_n$, $\mathbf{Y} \stackrel{\text{def}}{=} Y_1,\ldots,Y_n$, and $f(\mathbf{X},\mathbf{Y}) \stackrel{\text{def}}{=} f(X_1,Y_1),\ldots,f(X_n,Y_n)$.

A rate pair (r_x, r_y) is *achievable* if for every $\epsilon > 0$ and $n \in \mathbb{N}$ there is a protocol with x-rate $\leq r_x$, y-rate $\leq r_y$, block length $\geq n$, and block-error probability $< \epsilon$. The two-message *rate region* $R_f^2(X|Y)$ of (X, Y) and f is the closure of the set of achievable rate pairs. The *two-message communication complexity* is

$$L_f^2(X|Y) \stackrel{\text{def}}{=} \min\{r_x + r_y : (r_x, r_y) \in R_f^2(X|Y)\}.$$

Intuitively, $P_{\mathcal{X}}$ knows the X_i's, $P_{\mathcal{Y}}$ knows the Y_i's, and both want $P_{\mathcal{Y}}$ to reliably evaluate the $f(X_i, Y_i)$'s. $P_{\mathcal{Y}}$ considers the n-instance *block* $\mathbf{Y}$ and uses l bits to transmit $\varphi(\mathbf{Y})$ at an average rate of l/n bits per instance. $P_{\mathcal{X}}$ then considers the n-instance *block* $\mathbf{X}$ and uses k bits to transmit $\xi(\varphi(\mathbf{Y}), \mathbf{X})$ at an average rate of k/n bits per instance. $P_{\mathcal{Y}}$ decides that $f(\mathbf{X}, \mathbf{Y})$ is $\psi(\varphi(\mathbf{Y}), \xi(\varphi(\mathbf{Y}), \mathbf{X}), \mathbf{Y})$. He is correct for the block if the two vectors agree in every coordinate, and errs otherwise. $R_f^2(X|Y)$ is the set of rate pairs that can be transmitted while permitting $P_{\mathcal{Y}}$'s block-error probability to be arbitrarily small.

The problem's communication complexity $L_f^2(X|Y)$ is the total number of bits both communicators must transmit. While a more complicated definition may seem necessary, one where $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$'s messages are of variable lengths, it can be shown that the current definition is equivalent.

5.2 Results

Two random variables U and V defined over finite alphabets are *admissible* if:

1. $U - Y - X$,
2. $V - UX - Y$,
3. U, V, and Y determine $f(X, Y)$.

Theorem 3 For every (X, Y) and f,

$$R_f^2(X|Y) = \Big\{(r_x, r_y) \ : \ r_x \geq I(V; X|UY) \text{ and } r_y \geq I(U; Y|X) \text{ for some admissible } U \text{ and } V\Big\}. \ \square$$

The theorem is illustrated in the next subsection and proved in Subsection 5.4.

5.3 Example

We use Theorem 3 to show that even for natural problems, coding can save transmission over simple-minded communication.

Example 8 Recall the scenario in Example 6. Satellite knows a random variable X uniformly distributed over $\{1, \ldots, m\}$ and base knows an independent Bernoulli-p random variable Y. If $Y = 1$ ("alert") base wants to know X. If $Y = 0$ ("relax") — it doesn't.

We showed that, since all values of (X, Y) are possible, one-way communication requires the maximum number of bits even for small p:

$$L_f(X|Y) = H(X|Y) = \log m.$$

When two messages are allowed, the communicators can save bits when $Y = 0$. $P_{\mathcal{Y}}$ transmits Y using $h(p)$ bits per instance, and $P_{\mathcal{X}}$ describes X only if $Y = 1$. Hence,

$$L_f^2(X|Y) \leq h(p) + p \log m.$$

For small p this is significantly smaller than $L_f(X|Y)$.

However, Theorem 3 shows that transmission can be further reduced. Define the Markov chain $X - Y - U$ where $U \in \{z, e\}$ by

$$p(U = e|Y = 1) = 1 \qquad \text{and} \qquad p(U = e|Y = 0) = \alpha,$$

namely, if $Y = 1$ then $U = e$, and if $Y = 0$ then $U = e$ with probability α. Then $U = z$ implies that Y is *zero*, while $U = e$ implies that Y is *either* 0 or 1. Let

$$q \stackrel{\text{def}}{=} p(U = e) = p + \overline{p} \cdot \alpha = 1 - \overline{p} \cdot \overline{\alpha},$$

where $\overline{p} = 1 - p$ and $\overline{\alpha} = 1 - \alpha$.

The Markovity $U - Y - X$ and the independence of X and Y imply that U and X are independent. Therefore,

$$I(U; Y|X) = H(U) - H(U|Y) = h(q) - \overline{p}h(\alpha).$$

Next define $Y - UX - V$, where $V \in \{1, \ldots, m\}$, via

$$p(v|U = e,\ x) = \begin{cases} 1 & \text{if } v = x, \\ 0 & \text{if } v \neq x \end{cases} \qquad \text{and } p(v|U = z,\ x) = \frac{1}{m}.$$

Then

$$I(V; X|UY) = H(V|UY) - H(V|UX) = H(V) - p(U = z)H(V|X, U = z) = q \log m.$$

Therefore

$$R_f^2(X|Y) \supseteq \{(r_x, r_y) : r_x \geq q \log m,\ r_y \geq h(q) - \overline{p}h(\overline{q}/\overline{p}),\ q \in [p, 1]\}.$$

With more work it can be shown that this is exactly the achievable region.

Next, consider the communication complexity. For every $q \in [p, 1]$,

$$L_f^2(X|Y) \leq h(q) - \overline{p}h(\overline{q}/\overline{p}) + q \log m = -q \log q - \overline{p} \log \overline{p} + (q - p)\log(q - p) + q \log m.$$

Differentiation with respect to q shows that $L_f^2(X|Y)$ is minimized when

$$q = \min\{\frac{p \cdot m}{m - 1}, 1\}.$$

For simplicity, assume that $p \leq 1 - \frac{1}{m}$. Then

$$L_f^2(X|Y) = h(p) + p \log(m - 1).$$

It is instructive to consider $m = 2$, equivalently, the scenario where $X \sim B(1/2)$ and $Y \sim B(p)$ are independent Bernoulli variables and $P_{\mathcal{Y}}$ wants to determine $X \wedge Y$. For conciseness, we assume that $p \leq \frac{1}{2}$.

In one-way communication, $P_{\mathcal{X}}$ must transmit 1 bit (per instance). In simple-minded two-message communication, $P_{\mathcal{Y}}$ transmits $h(p)$ bits and $P_{\mathcal{X}}$ responds with p bits for a total of $h(p) + p$ bits. In optimal two-message communication, $P_{\mathcal{Y}}$ transmits $h(p) - 2p$ bits and $P_{\mathcal{X}}$ responds using $2p$ bits for a total of $h(p)$ bits. $\square$

5.4 Proof

It remains to prove Theorem 3. We describe only the outer bound (converse) proof. The inner-bound (achievability) proof, omitted to save space, combines random-coding arguments similar to ones given by Kaspi and Berger [9] with robust typicality techniques like those used to prove one-way results in Section 4 of this paper. This combination proves vanishing block-error (not just bit-error) probability.

The remainder of this subsection outlines the outer-bound (converse) proof. As with one-way communication, we show that the outer bound holds even under the relaxed requirement of small bit-error (not just block-error) probability.

The proof consists of three main parts:

1. As in [9], we obtain an outer bound on the three-dimensional region $\mathcal{S}_3$, defined essentially as the set of all triples (r_x, r_y, D) that have associated protocols with x-rate $\leq r_x + \epsilon$, y-rate $\leq r_y + \epsilon$, and *bit*-error rate $\leq D + \epsilon$ for all $\epsilon > 0$. It will be shown that $\mathcal{S}_3 \subset \mathcal{S}_3^*$, where $\mathcal{S}_3^* = (\mathcal{S}_3')^{cl}$ is the closure of the set $\mathcal{S}_3'$, with $\mathcal{S}_3'$ defined in terms of mutual informations involving X, Y, and three auxiliary variables T, U, and V. The corresponding alphabets $\mathcal{T}$, $\mathcal{U}$, and $\mathcal{V}$ may be taken to have cardinalities bounded above by polynomials in $|\mathcal{X}|$ and $|\mathcal{Y}|$.

2. Manipulating Markov chains, we show that $\mathcal{S}_3'$ and its closure, $\mathcal{S}_3^*$, can be expressed in terms of X, Y, U, and V alone.

3. Using the continuity of the mutual information function and the uniform bound on $|\mathcal{U}|$ and $|\mathcal{V}|$, we argue that $\mathcal{S}_3'$ is closed and that the two-dimensional set $\mathcal{S}_3^* \cap \{D = 0\}$, or $(\mathcal{S}_3')^{cl} \cap \{D = 0\}$, is the same as $\mathcal{S}_3' \cap \{D = 0\}$. It follows that $\mathcal{S}_2$, the two-dimensional slice through $\mathcal{S}_3$ on which $D = 0$ (which is equivalent to the two-dimensional region $R_f^2(X|Y)$) is contained in the two-dimensional region $\mathcal{S}_2'$ given in Theorem 3.

We begin with some definitions. For $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, let

$$\rho(x,y,z) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if} \quad z = f(x,y), \\ 1 & \text{if} \quad z \neq f(x,y), \end{cases}$$

and for $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{y} \in \mathcal{Y}^n$, $\mathbf{z} \in \mathcal{Z}^n$, let

$$\rho_n \stackrel{\text{def}}{=} \rho_n(\mathbf{x},\mathbf{y},\mathbf{z}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, y_i, z_i).$$

Define $\mathcal{S}_3 \subset \mathbb{R}^3$ to be the closure of the set of all (r_x, r_y, D) such that for every $\epsilon > 0$ and $n \in \mathbb{N}$ there is a protocol with x-rate $\leq r_x$, y-rate $\leq r_y$, and

$$\begin{aligned} E(\rho_n) &= E\left[\frac{1}{n}\sum_{i=1}^{n} \rho(X_i, Y_i, Z_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} p\{Z_i \neq f(X_i, Y_i)\} \\ &\leq D + \epsilon \ . \end{aligned}$$

Let

$$\mathcal{S}_2 \stackrel{\text{def}}{=} \mathcal{S}_3 \cap \{D = 0\}.$$

Note that $\mathcal{S}_2$ is equivalent to the region $R_f^2(X|Y)$:

$$\mathcal{S}_2 = \{(r_x, r_y, 0) : (r_x, r_y) \in R_f^2(X|Y)\}\ .$$

Next we define the regions $\mathcal{S}_3'$ and $\mathcal{S}_3^*$ that provide an outer bound on $\mathcal{S}_3$. For $D \geq 0$ let $\mathcal{P}_3(D)$ be the set of all triples of r.v.'s (T, U, V) jointly distributed with X, Y and ranging over finite sets $\mathcal{T}, \mathcal{U}$, and $\mathcal{V}$, respectively, satisfying the following properties:

a) $U - Y - X$,

b) $V - UX - Y$,

c) $X - UY - T$,

d) $X - UVY - T$,

e) $f(X,Y) - UVY - T$,

f) $\exists F : \mathcal{T} \times \mathcal{U} \times \mathcal{V} \times \mathcal{Y} \to \mathcal{Z}$ such that $E[\rho(X, Y, F(T, U, V, Y))] \leq D$.

Then

$$\mathcal{S}_3' \stackrel{\text{def}}{=} \{(r_x, r_y, D) : r_x \geq I(V; X|TUY), r_y \geq I(U; Y|X) \ \text{ for some } \ (T, U, V) \in \mathcal{P}_3(D)\}$$

and

$$\mathcal{S}_3^* \stackrel{\text{def}}{=} (\mathcal{S}_3')^{cl}$$

is the closure of $\mathcal{S}_3'$.

As in [9], $\mathcal{S}_3^*$ is convex and the cardinalities $|\mathcal{T}|$, $|\mathcal{U}|$, $|\mathcal{V}|$ may be taken to be bounded above by polynomial functions of $|\mathcal{X}|$ and $|\mathcal{Y}|$. We now prove the first of the three steps establishing the outer bound.

Theorem 4

$$\mathcal{S}_3 \subset \mathcal{S}_3^*.$$

Proof: The proof is similar that of Theorem 4.1 in [9]. We provide only an outline. $\mathcal{S}_3^*$ is closed by definition. Hence it suffices to show that if $\mathbf{s} \stackrel{\text{def}}{=} (r_x, r_y, D) \in \mathcal{S}_3$, then for any $\epsilon > 0$,

$$\mathbf{s}_\epsilon \stackrel{\text{def}}{=} (r_x + \epsilon + \delta(\epsilon), r_y + \epsilon + \delta(\epsilon), D + \epsilon + \delta(\epsilon))$$

is in the interior of $\mathcal{S}_3^*$, where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Since $\mathbf{s} \in \mathcal{S}_3$, we know that for sufficiently large n there exists a code $\varphi\xi\psi$ that satisfies the conditions for the definition of $\mathcal{S}_3$.

We lower-bound the two rates for the code as in [9]. First, for $i = 1, 2, \ldots, n$ define the auxiliary r.v.'s

$$\begin{aligned} T_i &\stackrel{\text{def}}{=} (\mathbf{Y}_1^{i-1}, \mathbf{Y}_{i+1}^n), \\ U_i &\stackrel{\text{def}}{=} (\varphi(\mathbf{Y}), \mathbf{X}_1^{i-1}, \mathbf{Y}_{i+1}^n), \\ V_i &\stackrel{\text{def}}{=} \xi(\varphi(\mathbf{Y}), \mathbf{X}). \end{aligned}$$

For convenience we sometimes omit the arguments of the encoding functions φ and ξ (and of the decoding function ψ). We also often omit the index and limits of summation when a sum is over all i from 1 to n.

Proceeding as in [9], we find that for any given $\epsilon > 0$ and for any code satisfying the conditions that define the region $\mathcal{S}_3$,

$$\begin{aligned} n(r_x + \epsilon) &\geq \sum I(X_i; V_i | T_i U_i Y_i), \\ n(r_y + \epsilon) &\geq \sum I(Y_i; U_i | X_i). \end{aligned}$$

Let $\Delta_i \stackrel{\text{def}}{=} E[\rho(X_i, Y_i, Z_i)]$ be the probability of bit error that the given code achieves for the i^{th} character of the blocks, where Z_i is the i^{th} component of $\mathbf{Z} = \psi(\varphi, \xi, \mathbf{Y})$.

By definition, (T_i, U_i, V_i, Y_i) includes the coding functions φ and ξ, as well as the complete vector $\mathbf{Y}$. Therefore there exists a function F_i such that $Z_i = F_i(T_i, U_i, V_i, Y_i)$, namely $F_i(T_i, U_i, V_i, Y_i)$ is the i^{th} component of $\psi(\varphi, \xi, \mathbf{Y})$. As we shall see shortly, we also have the five following Markov conditions on T_i, U_i, V_i.

Condition 1: $U_i - Y_i - X_i$

Condition 2: $V_i - U_i X_i - Y_i$

Condition 3: $X_i - U_i Y_i - T_i$

Condition 4: $X_i - U_i V_i Y_i - T_i$

Condition 5: $f(X_i, Y_i) - U_i V_i Y_i - T_i$.

Taking these five conditions on faith for the moment, we see from the definition of $\mathcal{P}_3(D)$ that $(T_i, U_i, V_i) \in \mathcal{P}_3(\Delta_i)$. By a simple convexity lemma (proved as in [9]), if we let $\Delta = n^{-1} \sum \Delta_i$, then there exist $(T, U, V) \in \mathcal{P}_3(\Delta)$ such that

$$\begin{aligned} I(X; V | TUY) &= n^{-1} \sum I(X_i; V_i | T_i U_i Y_i) \ , \\ I(Y; U | X) &= n^{-1} \sum I(Y_i; U_i | X_i) \ . \end{aligned}$$

Combining previous results we obtain

$$\begin{aligned} r_x + \epsilon &\geq I(X;V|TUY), \\ r_y + \epsilon &\geq I(Y;U|X), \\ D + \epsilon &\geq \Delta\,. \end{aligned}$$

Thus $\mathbf{s}_\epsilon$ is in the interior of $\mathcal{S}_3^*$, as we wished to show. Furthermore, as in [9], we may without loss of generality impose uniform bounds on the cardinalities of the alphabets $\mathcal{T}$, $\mathcal{U}$, and $\mathcal{V}$. $\square$

Now we return to the five Markov conditions in Theorem 4 and outline their proofs. Conditions 1 and 2 are established as in [9], and Condition 3 is straightforward, so we concentrate on the last two conditions.

Lemma 9 (Condition 4) $X_i - U_iV_iY_i - T_i$.

Proof: This Markov constraint follows from the stronger condition

$$X_i^n - U_iV_iY_i - T_i$$

which, in turn, follows from the definitions of T_i, U_i, and V_i, and from the next lemma after letting

$$\begin{aligned} A_1 &= \mathbf{X}_1^{i-1}, \quad A_2 = \mathbf{X}_i^n, \\ B_1 &= \mathbf{Y}_1^{i-1}, \quad B_2 = \mathbf{Y}_i^n, \quad g = \xi, \quad h = \varphi. \end{aligned}$$

$\square$

Lemma 10 Let A_1, A_2, B_1, B_2 be r.v.'s with (A_1, A_2) independent of (B_1, B_2). Let h be a deterministic function $h(B_1, B_2)$ and let g be a deterministic function $g(h, A_1, A_2)$. Then

$$A_2 - (h, g, A_1, B_2) - B_1.$$

Proof: Straightforward. $\square$

Before establishing the fifth and final Markov condition, we prove two claims and two lemmas. Abbreviate

$$f_i \stackrel{\text{def}}{=} f(X_i, Y_i).$$

Claim 4 $f_iX_i - Y_i - U_iT_i$.

Proof: Follows readily from the definitions of T_i and U_i. $\square$

Lemma 11 (Lemma A.1 of [9]) Let A_i and B_i, $i = 1, 2$, be r.v.'s taking values on respective finite sets $\mathcal{A}_i$ and $\mathcal{B}_i$, with (A_1, B_1) independent of (A_2, B_2). Let $g = g(B_1, B_2)$ and $f = f(A_1, A_2, g)$ be deterministic functions. Then

$$f - (A_1, B_2, g) - B_1\,.$$

Proof: See [9]. $\square$

Claim 5 $\quad T_iY_i - f_iX_iU_i - V_i$

Proof: Follows almost immediately from the fact that $f_iT_iY_i - X_iU_i - V_i$, which can in turn be readily proved after using the previous lemma to show that $\xi - \mathbf{X}_1^i\mathbf{Y}_{i+1}^n\varphi - \mathbf{Y}_1^i$. $\square$

Lemma 12 Let A, B, T, U, V be r.v.'s defined over finite alphabets with $B - A - TU$ and $AT - BU - V$. Then

(a) $BV - AU - T$,

(b) $B - AUV - T$.

Proof: (a) For every realization a, b, t, u, v of A, B, T, U, V, we have

$$\begin{aligned} p(t|abuv) &= \frac{p(abtuv)}{p(abuv)} = \frac{p(ab)p(tu|a)p(v|bu)}{\sum\limits_u p(ab)p(tu|a)p(v|bu)} \\ &= \frac{p(tu|a)}{\sum\limits_u p(tu|a)} = \frac{p(tu|a)p(a)}{\sum\limits_u p(tu|a)p(a)} \\ &= \frac{p(atu)}{\sum\limits_u p(atu)} = \frac{p(atu)}{p(at)} = p(t|au) \ . \end{aligned}$$

Therefore

$$p(t|abuv) = p(t|au)$$

for every a, b, t, u, v, so $BV - AU - T$.

(b) $H(T|AUV) \geq H(T|ABUV) = H(T|AU) \geq H(T|AUV)$,
where the equality follows from part (a) and where the two inequalities follow from the fact that additional conditioning reduces entropy. Therefore

$$H(T|AUV) = H(T|ABUV) \ ,$$

so $B - AUV - T$. $\square$

Lemma 13 (Condition 5) $\quad f_i - U_iV_iY_i - T_i$.

Proof: Follows from the previous two claims and part (b) of the last lemma with $A = Y_i$ and $B = (f_i, X_i)$. $\square$

This completes the proof of Theorem 4. Next we show that the auxiliary variable T can be eliminated from the definition of the region $\mathcal{S}_3'$ and its closure, $\mathcal{S}_3^*$.

For $D \geq 0$, let $\tilde{\mathcal{P}}_3(D)$ be the set of all pairs of r.v.'s (U, V) jointly distributed with X, Y and ranging over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively, satisfying the following properties:

a) $U - Y - X$,

b) $V - UX - Y$,

c) $\exists F : \mathcal{U} \times \mathcal{V} \times \mathcal{Y} \to \mathcal{Z}$ such that

$$E[\rho(X, Y, F(U, V, Y))] \leq D.$$

Let

$$\tilde{\mathcal{S}}_3' \stackrel{\text{def}}{=} \left\{ (r_x, r_y, D) : r_x \geq I(V; X|UY), r_y \geq I(U; Y|X) \text{ for some } UV \in \tilde{\mathcal{P}}_3(D) \right\},$$

and let

$$\tilde{\mathcal{S}}_3^* \stackrel{\text{def}}{=} (\tilde{\mathcal{S}}_3')^{cl}$$

be its closure.

Theorem 5

$$\mathcal{S}_3 \subseteq \tilde{\mathcal{S}}_3^*.$$

In fact, $\tilde{\mathcal{S}}_3' = \mathcal{S}_3'$ and $\tilde{\mathcal{S}}_3^* = \mathcal{S}_3^*$.

Proof: Follows from the next two lemmas, showing that T can be eliminated in Theorem 4. $\square$

Lemma 14 If $X - UY - T$ and $X - UVY - T$, then

$$I(X; V|UTY) = I(X; V|UY).$$

Proof:

$$I(X; V|UTY) = H(X|UYT) - H(X|UVYT).$$

By the two Markov hypotheses, we have

$$H(X|UYT) = H(X|UY)$$

and

$$H(X|UVYT) = H(X|UVY),$$

so

$$I(X; V|UTY) = H(X|UY) - H(X|UVY) = I(X; V|UY). \qquad \square$$

Lemma 15 Given r.v.'s T, U, V, X, and Y defined over finite alphabets with a particular joint distribution, suppose that $f(X, Y) - UVY - T$ and that there exists a decoding function $F(T, U, V, Y)$ such that

$$p\{F(T, U, V, Y) \neq f(X, Y)\} \leq D.$$

Then there exists a decoding function $\tilde{F}(U, V, Y)$ such that

$$p\{\tilde{F}(U, V, Y) \neq f(X, Y)\} \leq D.$$

Proof: Define the r.v.'s $A = f(X, Y)$, $B = (U, V, Y)$, and $C = T$ over the corresponding finite alphabets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ with the joint distribution induced by T, U, V, X, Y, and f. Given any realization (b, c) of the random pair (B, C), there is some discrete distribution on the value of

A. The best possible decoding rule, F^*, is one that for each (b,c) sets $F^*(b,c)$ equal to the most probable value of A. (If there is more than one such value of A, it may be chosen arbitrarily.)

In particular, any F^* chosen as above must achieve

$$p\{F^*(B,C) \neq A\} \leq p\{F(B,C) \neq A\} \leq D.$$

But by the Markov condition $A - B - C$, we may choose F^* to be a function of B alone without altering $p\{F^*(B,C) \neq A\}$. Thus we may take

$$\tilde{F}(U,V,Y) = F^*(B) = F^*(U,V,Y). \qquad \square$$

The previous two lemmas and Theorem 4 establish Theorem 5. We are finally ready to prove the outer bound (converse) of Theorem 3. Recall that

$$\mathcal{S}_2 \stackrel{\text{def}}{=} \mathcal{S}_3 \cap \{D = 0\} = \{(r_x, r_y, 0) : (r_x, r_y) \in R_f^2(X|Y)\}.$$

Thus the outer bound of Theorem 3 will be established by the following theorem.

Theorem 6

$$\mathcal{S}_2 \subseteq \tilde{\mathcal{S}}_3' \cap \{D = 0\}.$$

Proof: It follows from Theorem 5 that if $\mathbf{s} \in \mathcal{S}_2$, then

$$\mathbf{s} \in \tilde{\mathcal{S}}_3^* \cap \{D = 0\} = (\mathcal{S}_3')^{cl} \cap \{D = 0\}.$$

To prove Theorem 6, we essentially need to show that $\tilde{\mathcal{S}}_3'$ already contains all its points of closure for which $D = 0$. The following claim establishes this result and thus, together with the inner bound on $R_f^2(X|Y)$, proves Theorem 3, our main theorem on two-way communication. $\square$

Claim 6 $\tilde{\mathcal{S}}_3'$ is closed.

Proof: We show that for every infinite sequence $\{(r_x, r_y, D)_k\}_{k\geq 1}$ from $\tilde{\mathcal{S}}_3'$ that has a limit (r_x^*, r_y^*, D^*), this limiting triple is also contained in $\tilde{\mathcal{S}}_3'$. We use the fact that we can place uniform upper bounds on the cardinalities of the alphabets $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{Z}$ in terms of $|\mathcal{X}|$ and $|\mathcal{Y}|$. Thus the number of possible decoding functions $F(U,V,Y)$ is bounded, and the space of admissible probability mass functions (p.m.f.'s) $p(x,y,u,v) = p(x,y)p(u|y)p(v|u,x)$ (the form allowed by the Markov constraints in the definition of $\tilde{\mathcal{P}}_3(D)$) is closed and bounded and has bounded dimension.

Now given any infinite sequence of triples $\{(r_x, r_y, D)_k\}$ from $\tilde{\mathcal{S}}_3'$, and any corresponding sequence $\{F_k(U,V,Y),\ p_k(u|y),\ p_k(v|u,x)\}$ achieving the given $\{(r_x, r_y, D)_k\}$, there must be an infinite subsequence having a common decoding function $F^*(U,V,Y)$. Within this subsequence, there must be an accumulation point $(F^*(U,V,Y), p^*(u|y), p^*(v|u,x))$. Since the space of p.m.f.'s is closed and since mutual information is a continuous function, it follows from the definition of $\tilde{\mathcal{S}}_3'$ that we can define U and V by $p^*(x,y,u,v) = p(x,y)p^*(u,y)p^*(v|u,x)$ and F by $F^*(U,V,Y)$ to achieve the point (r_x^*, r_y^*, D^*). The claim follows. $\square$

Acknowledgements

We are indebted to Toby Berger, János Körner, and especially Aaron Wyner, for elucidating discussions, literature references, and help in smoothing out the rough edges of some proofs.

Appendix: Robust typicality

Let X be a random variable distributed over a finite set $\mathcal{X}$ according to a probability distribution p, and let $\mathbf{X} = X_1,\ldots,X_n$ be a sequence of independent random copies of X. For large n, we expect roughly a $p(x)$-fraction of the copies to attain the value $x \in \mathcal{X}$. This fundamental property underlies many an information-theory result.

Formally, the *empirical frequency* of $x \in \mathcal{X}$ in a sequence $\mathbf{x} = x_1,\ldots,x_n \in \mathcal{X}^n$ is

$$\nu_{\mathbf{x}}(x) \stackrel{\text{def}}{=} \frac{|\{i : x_i = x\}|}{n}.$$

The sequence $\mathbf{x}$ is *δ-strongly typical* for $\delta > 0$ if for all $x \in \mathcal{X}$,

$$|\nu_{\mathbf{x}}(x) - p(x)| \le \delta \qquad \text{(or } \le \delta/|\mathcal{X}| \text{ in some definitions).}$$

Strong typicality is studied in detail by Berger [1], Csiszár and Körner [4], Cover and Thomas [3], and others.

We need a slightly modified form of typicality. The sequence $\mathbf{x}$ is *δ-robustly typical* (*δ-r.t.*) for $\delta > 0$ if for all $x \in \mathcal{X}$,

$$|\nu_{\mathbf{x}}(x) - p(x)| \le \delta \cdot p(x).$$

Suppressing sequence-length, n, for brevity, we let $T_{X,\delta}$, or T_δ for short, denote the set of δ-r.t. $\mathbf{x}$'s. In most cases, the random variable is implied by context, and we use the shorthand notation, for example, $\mathbf{x} \in T_\delta$ instead of $\mathbf{x} \in T_{X,\delta}$.

Pairs and, more generally, collections of random variables are special kinds of random variables, hence the definition of robust typicality applies to them too. For example, let (X, Y) be a random pair distributed over a finite set $\mathcal{X} \times \mathcal{Y}$ according to a probability distribution p and, using informal notation, let $(\mathbf{X}, \mathbf{Y}) \stackrel{\text{def}}{=} (X_1, Y_1),\ldots,(X_n, Y_n)$ be a sequence of independent copies of (X, Y). According to the definitions above, the empirical frequency of $(x, y) \in \mathcal{X} \times \mathcal{Y}$ in a sequence pair $(\mathbf{x}, \mathbf{y}) = (x_1, y_1),\ldots,(x_n, y_n)$ is $\nu_{\mathbf{x},\mathbf{y}}(x, y) = |\{i : (x_i, y_i) = (x, y)\}|/n$, the sequence $(\mathbf{x}, \mathbf{y})$ is δ-r.t. if $|\nu_{\mathbf{x},\mathbf{y}}(x, y) - p(x, y)| \le \delta \cdot p(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and the set of all δ-r.t. $(\mathbf{x}, \mathbf{y})$'s is denoted by $T_{(X,Y),\delta}$, usually abbreviated T_δ.

Robust typicality has certain advantages over its strong counterpart. The bounds have more natural expressions (e.g., Lemma 18), many proofs are simplified (cf. Lemma 21), and a single definition applies to arbitrary collections of random variables (as above). For our purposes, the crucial property distinguishing robust from strong typicality[3] is also the one that lends it its name:

[3] And from *weak typicality,* which is less relevant here.

If $\mathbf{x}$ is r.t. then $p(x) = 0$ implies $\nu_{\mathbf{x}}(x) = 0$. It is this property that allows us to achieve near-zero *block* error probability. In the protocol discussed in Subsection 4.2, $P_{\mathcal{Y}}$ attempts to find a $\mathbf{w} \in \Gamma(G)^n$ that is δ-r.t. with $\mathbf{x}$. If $x \notin w$ then $p(w|x) = 0$, and robust typicality guarantees that if $P_{\mathcal{Y}}$ finds $\mathbf{w}$ as above then $x \in w_i$ for *all* i and therefore $P_{\mathcal{Y}}$ is correct on the whole block.

In the rest of the appendix we prove robust-typicality equivalents of standard strong-typicality results (e.g., Cover and Thomas [3]). Many of the results hold under similar qualifications. To abbreviate the exposition, we describe all qualifiers here. All results hold for every random variable X and every $n \in \mathbb{N}$. Whenever δ appears (mainly Lemmas 16—21 and Corollary 2), the result holds for every $0 < \delta \le 1$; and, if used,

$$\epsilon_\delta(n) \stackrel{\text{def}}{=} 2|S_X| e^{-\delta^2 \mu_X n/3}$$

where $S_X \stackrel{\text{def}}{=} \{x \in \mathcal{X} : p(x) > 0\}$ is the *support set* of X and $\mu_X = \min_{x \in S_X} p(x)$ is its smallest nonzero probability. Whenever δ_1 and δ_2 appear (mainly following Lemma 22), the result holds for every $0 < \delta_1 < \delta_2 \le 1$, and, if used,

$$\epsilon_{\delta_1,\delta_2}(n) \stackrel{\text{def}}{=} 2|S_{X,Y}| e^{-\frac{(\delta_2-\delta_1)^2}{1+\delta_1} \cdot \mu_{X,Y} n/3}.$$

Note that $\epsilon_\delta(n)$ and $\epsilon_{\delta_1,\delta_2}(n)$ diminish to zero exponentially with n.

We first consider a single random variable X. We prove that with high probability a random sequence is typical, that every typical sequence has probability of about $2^{-nH(X)}$, and that there are roughly $2^{nH(X)}$ typical sequences. We use the following basic tool.

Lemma 16 (Chernoff Bound) For every $x \in \mathcal{X}$,

$$p[\nu_{\mathbf{X}}(x) \le (1-\delta)p(x)] \le e^{-\delta^2 p(x) n/2}$$

and

$$p[\nu_{\mathbf{X}}(x) \ge (1+\delta)p(x)] \le e^{-\delta^2 p(x) n/3}. \qquad \square$$

Off we go.

Lemma 17 $p(\mathbf{X} \in T_\delta) \ge 1 - \epsilon_\delta(n)$.

Proof: Using the Chernoff Bound,

$$p(\mathbf{X} \notin T_\delta) \le \sum_{x \in S_X} p[|\nu_{\mathbf{X}}(x) - p(x)| \ge \delta \cdot p(x)] \le 2|S_X| \cdot e^{-\delta^2 \mu_X n/3} = \epsilon_\delta(n). \qquad \square$$

Lemma 18 For every $\mathbf{x} \in T_\delta$,

$$2^{-(1+\delta)H(X)n} \le p(\mathbf{x}) \le 2^{-(1-\delta)H(X)n}.$$

Proof: For every $\mathbf{x} \in \mathcal{X}^n$,

$$p(\mathbf{x}) = 2^{-\left(\sum_{x \in \mathcal{X}} \nu_{\mathbf{x}}(x) \log \frac{1}{p(x)}\right) n} = 2^{-\left(H(X) + \sum_{x \in \mathcal{X}} (\nu_{\mathbf{x}}(x) - p(x)) \log \frac{1}{p(x)}\right) n}.$$

For every $\mathbf{x} \in T_\delta$,

$$\left|\sum_{x\in\mathcal{X}} (\nu_{\mathbf{x}}(x) - p(x)) \log \frac{1}{p(x)}\right| \leq \delta \cdot \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} = \delta H(X). \qquad \square$$

Lemma 19 $(1 - \epsilon_\delta(n)) \cdot 2^{(1-\delta)H(X)n} \leq |T_\delta| \leq 2^{(1+\delta)H(X)n}$.

Proof: Using Lemma 18,

$$1 \geq \sum_{\mathbf{x}\in T_\delta} p(\mathbf{x}) \geq |T_\delta| \cdot 2^{-(1+\delta)H(X)n}.$$

Conversely, incorporating Lemma 17,

$$1 - \epsilon_\delta(n) \leq \sum_{\mathbf{x}\in T_\delta} p(\mathbf{x}) \leq |T_\delta| \cdot 2^{-(1-\delta)H(X)n}. \qquad \square$$

We now turn to pairs of random variables. The next corollary summarizes some implication of the general results proven in the previous lemmas. The ensuing results are specific to random pairs.

Corollary 2

1. For every $(\mathbf{x}, \mathbf{y}) \in T_\delta$, $\quad 2^{-(1+\delta)H(X,Y)n} \leq p(\mathbf{x}, \mathbf{y}) \leq 2^{-(1-\delta)H(X,Y)n}$.
2. $p((\mathbf{X}, \mathbf{Y}) \in T_\delta) \geq 1 - \epsilon_\delta(n)$.
3. $(1 - \epsilon_\delta(n)) \cdot 2^{(1-\delta)H(X,Y)n} \leq |T_\delta| \leq 2^{(1+\delta)H(X,Y)n}$. $\square$

Using the results in Lemma 18 and Part (1) of the corollary, we can bound $p(\mathbf{y}|\mathbf{x})$ for every $(\mathbf{x}, \mathbf{y}) \in T_\delta$. However, an adaptation of the proof of Lemma 18 yields a stronger bound:

Lemma 20 For all $(\mathbf{x}, \mathbf{y}) \in T_\delta$,

$$2^{-(1+\delta)H(Y|X)n} \leq p(\mathbf{y}|\mathbf{x}) \leq 2^{-(1-\delta)H(Y|X)n}. \qquad \square$$

Lemma 21 $T_{(X,Y),\delta} \subseteq T_{X,\delta} \times T_{Y,\delta}$.

Proof: We need to show that if $(\mathbf{x}, \mathbf{y}) \in T_\delta$ then (using shameless notation) $\mathbf{x}, \mathbf{y} \in T_\delta$. For every $x \in \mathcal{X}$,

$$|\nu_{\mathbf{x}}(x) - p(x)| = |\sum_{y\in\mathcal{Y}} \nu_{\mathbf{x},\mathbf{y}}(x, y) - p(x, y)| \leq \sum_{y\in\mathcal{Y}} \delta \cdot p(x, y) = \delta \cdot p(x),$$

therefore $\mathbf{x} \in T_\delta$. Similarly, $\mathbf{y} \in T_\delta$. $\square$

Recall that the ensuing lemmas hold for all $n \in \mathbb{N}$ and all $0 < \delta_1 < \delta_2 \leq 1$, and that

$$\epsilon_{\delta_1,\delta_2}(n) \stackrel{\text{def}}{=} 2|S_{X,Y}| e^{-\frac{(\delta_2-\delta_1)^2}{1+\delta_1} \cdot \mu_{X,Y} n/3}.$$

Part (2) of Corollary 2 says that with high probability $(\mathbf{X}, \mathbf{Y}) \in T_\delta$. The same holds when we condition on every typical value of $\mathbf{X}$:

Lemma 22 For every $\mathbf{x} \in T_{\delta_1}$,

$$p[(\mathbf{x}, \mathbf{Y}) \in T_{\delta_2} | \mathbf{X} = \mathbf{x}] \geq 1 - \epsilon_{\delta_1,\delta_2}(n).$$

Proof: Let $\mathbf{x} \in T_{\delta_1}$, namely, for every $x \in \mathcal{X}$,

$$(1 - \delta_1)p(x) \leq \nu_{\mathbf{x}}(x) \leq (1 + \delta_1)p(x). \tag{5}$$

We show that if $\mathbf{X} = \mathbf{x}$ then for every (x, y), $\nu_{\mathbf{x},\mathbf{Y}}(x, y)$ is close to $p(x, y)$ with high probability.

If $p(x, y) = 0$, then $\nu_{\mathbf{x},\mathbf{Y}}(x, y) = \nu_{\mathbf{X},\mathbf{Y}}(x, y) = 0 = p(x, y)$. If $\nu_{\mathbf{x}}(x) = 0$, then, by (5), $p(x, y) = p(x) = 0$ and again $\nu_{\mathbf{x},\mathbf{Y}}(x, y) = p(x, y)$. For $(x, y) \in S_{X,Y}$ such that $\nu_{\mathbf{X}}(x) \neq 0$ we apply the Chernoff bound. The expected value of $\nu_{\mathbf{x},\mathbf{Y}}(x, y)$ is $\nu_{\mathbf{x}}(x) \cdot p(y|x)$. Hence,

$$\begin{aligned}
p[\nu_{\mathbf{x},\mathbf{Y}}(x, y) \geq (1 + \delta_2)p(x, y)] &= p\left[\nu_{\mathbf{x},\mathbf{Y}}(x, y) \geq (1 + \delta_2)\frac{p(x)}{\nu_{\mathbf{x}}(x)} \cdot \nu_{\mathbf{x}}(x) \cdot p(y|x)\right] \\
&< e^{-\left((1+\delta_2)\frac{p(x)}{\nu_{\mathbf{x}}(x)}-1\right)^2 \nu_{\mathbf{x}}(x)p(y|x)n/3} \\
&= e^{-\left((1+\delta_2)-\frac{\nu_{\mathbf{x}}(x)}{p(x)}\right)^2 \frac{p(x)}{\nu_{\mathbf{x}}(x)}p(x,y)n/3} \\
&\leq e^{-\frac{(\delta_2-\delta_1)^2}{1+\delta_1}p(x,y)n/3}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
p[\nu_{\mathbf{x},\mathbf{Y}}(x, y) \leq (1 - \delta_2)p(x, y)] &= p\left[\nu_{\mathbf{x},\mathbf{Y}}(x, y) \leq (1 - \delta_2)\frac{p(x)}{\nu_{\mathbf{x}}(x)} \cdot \nu_{\mathbf{x}}(x) \cdot p(y|x)\right] \\
&< e^{-\left(1-(1-\delta_2)\frac{p(x)}{\nu_{\mathbf{x}}(x)}\right)^2 \nu_{\mathbf{x}}(x)p(y|x)n/2} \\
&= e^{-\left(\frac{\nu_{\mathbf{x}}(x)}{p(x)}-(1-\delta_2)\right)^2 \frac{p(x)}{\nu_{\mathbf{x}}(x)}p(x,y)n/2} \\
&\leq e^{-\frac{(\delta_2-\delta_1)^2}{1+\delta_1}p(x,y)n/2}.
\end{aligned}$$

The lemma follows from the union bound. □

If $(\mathbf{W}, \mathbf{X}) = (\mathbf{w}, \mathbf{x}) \in T_{\delta_1}$ the lemma implies that with high probability $(\mathbf{w}, \mathbf{x}, \mathbf{Y}) \in T_{\delta_2}$. If $W - X - Y$ is a Markov chain, fewer conditions suffice.

Lemma 23 (Markov Lemma, Berger [2]) Let $W - X - Y$ be a Markov chain. For every $(\mathbf{w}, \mathbf{x}) \in T_{\delta_1}$,

$$p[(\mathbf{w}, \mathbf{x}, \mathbf{Y}) \in T_{\delta_2} | \mathbf{X} = \mathbf{x}] \geq 1 - \epsilon_{\delta_1,\delta_2}(n)$$

Proof: By Markovity and Lemma 22, for all $(\mathbf{w}, \mathbf{x}) \in T_{\delta_1}$,

$$p[(\mathbf{w}, \mathbf{x}, \mathbf{Y}) \in T_{\delta_2} | \mathbf{X} = \mathbf{x}] = p[(\mathbf{w}, \mathbf{x}, \mathbf{Y}) \in T_{\delta_2} | (\mathbf{W}, \mathbf{X}) = (\mathbf{w}, \mathbf{x})] \geq 1 - \epsilon_{\delta_1,\delta_2}(n). \quad \square$$

An obvious extension of Lemma 21 implies:

Corollary 3 Let $W - X - Y$ be a Markov chain. For every $(\mathbf{w}, \mathbf{x}) \in T_{\delta_1}$,

$$p[(\mathbf{w}, \mathbf{Y}) \in T_{\delta_2} | \mathbf{X} = \mathbf{x}] \geq 1 - \epsilon_{\delta_1,\delta_2}(n). \quad \square$$

For $\mathbf{x} \in \mathcal{X}^n$ and $\delta > 0$ let

$$T_\delta(\mathbf{x}) \stackrel{\text{def}}{=} \{\mathbf{y} : (\mathbf{x}, \mathbf{y}) \in T_\delta\}$$

be the set of $\mathbf{y}$'s that are δ-r.t. with $\mathbf{x}$.

Lemma 24 In the following, the upper bound holds for every $\mathbf{x} \in \mathcal{X}^n$ and the lower bound holds for every $\mathbf{x} \in T_{\delta_1}$.

$$(1 - \epsilon_{\delta_1,\delta_2}(n))2^{(1-\delta_2)H(Y|X)n} \le |T_{\delta_2}(\mathbf{x})| \le 2^{(1+\delta_2)H(Y|X)n}.$$

Proof: Lemma 20 implies that for every $\mathbf{x} \in \mathcal{X}^n$,

$$1 \ge \sum_{\mathbf{y} \in T_{\delta_2}(\mathbf{x})} p(\mathbf{y}|\mathbf{x}) \ge |T_{\delta_2}(\mathbf{x})| \cdot 2^{-(1+\delta_2)H(Y|X)n}.$$

Conversely, for every $\mathbf{x} \in T_{\delta_1}$,

$$1 - \epsilon_{\delta_1,\delta_2}(n) \le p[(\mathbf{x}, \mathbf{Y}) \in T_{\delta_2} | \mathbf{X} = \mathbf{x}] = \sum_{\mathbf{y} \in T_{\delta_2}(\mathbf{x})} p(\mathbf{y}|\mathbf{x}) \le |T_{\delta_2}(\mathbf{x})| \cdot 2^{-(1-\delta_2)H(Y|X)n}. \qquad \square$$

So far, we calculated the probability that a robustly typical set corresponding to some random variable contains a random sequence generated by the same random variable. For example, Lemma 22 and Corollary 2 show that both $p\Big[(\mathbf{x}, \mathbf{Y}) \in T_{(X,Y),\delta_2} | \mathbf{X} = \mathbf{x}\Big]$ for $\mathbf{x} \in T_{(X,Y),\delta_2}$, and $p\Big((X, Y) \in T_{(X,Y),\delta}\Big)$ are close to one. The next two results concern the probability that a random sequence falls in a typical set corresponding to another random variable. First, we ignore the probability distribution of X completely.

Lemma 25 In the following, the upper bound holds for every $\mathbf{x} \in \mathcal{X}^n$ and the lower bound holds for every $\mathbf{x} \in T_{\delta_1}$.

$$(1 - \epsilon_{\delta_1,\delta_2}(n))2^{-(I(X;Y)+2\delta_2 H(Y))n} \le p\Big((\mathbf{x}, \mathbf{Y}) \in T_{(X,Y),\delta_2}\Big) \le 2^{-(I(X;Y)-2\delta_2 H(Y))n}.$$

Proof: Lemmas 18, 21 and 24 imply that

$$p((\mathbf{x}, \mathbf{Y}) \in T_{\delta_2}) = \sum_{\mathbf{y} \in T_{\delta_2}(\mathbf{x})} p(\mathbf{y}) \le 2^{(1+\delta_2)H(Y|X)n} \cdot 2^{-(1-\delta_2)H(Y)n} \le 2^{-(I(X;Y)-2\delta_2 H(Y))n}.$$

The lower bound is proven similarly. $\square$

Recall that (X, Y) is distributed over $\mathcal{X} \times \mathcal{Y}$ according to a probability distribution p. Let p_X and p_Y be the marginal distributions of X and Y and let (X', Y') be distributed over $\mathcal{X} \times \mathcal{Y}$ according to $p_X(x') \cdot p_Y(y')$. Namely, X' and Y' have the same distribution as X and Y but are independent of each other. Let $(\mathbf{X}', \mathbf{Y}') = (X'_1, Y'_1), \ldots, (X'_n, Y'_n)$ be independent copies of (X', Y'). The next corollary implies that while $(\mathbf{X}, \mathbf{Y}) \in T_{(X,Y),\delta}$ with probability of about one, $(\mathbf{X}', \mathbf{Y}') \in T_{(X,Y),\delta}$ with probability of only about $2^{-I(X;Y)n}$.

Corollary 4

$$(1-\epsilon_\delta(n))\cdot 2^{-(I(X;Y)+2\delta H(Y))n} \le p\Big((\mathbf{X}',\mathbf{Y}')\in T_{(X,Y),\delta}\Big) \le 2^{-(I(X;Y)-2\delta H(Y))n}.$$

Proof:

$$\begin{aligned}
p\Big((\mathbf{X}',\mathbf{Y}')\in T_{(X,Y),\delta}\Big) &= \sum_{\mathbf{x}'\in\mathcal{X}^n} p(\mathbf{X}'=\mathbf{x}')p\Big((\mathbf{x}',\mathbf{Y}')\in T_{(X,Y),\delta}|\mathbf{X}'=\mathbf{x}'\Big)\\
&= \sum_{\mathbf{x}'\in\mathcal{X}^n} p(\mathbf{X}'=\mathbf{x}')p\Big((\mathbf{x}',\mathbf{Y}')\in T_{(X,Y),\delta}\Big)\\
&\le 2^{-(I(X;Y)-2\delta_2 H(Y))n}.
\end{aligned}$$

Conversely, for every $0<\delta_1<\delta$,

$$\begin{aligned}
p\Big((\mathbf{X}',\mathbf{Y}')\in T_{(X,Y),\delta}\Big) &\ge p(\mathbf{X}'\in T_{\delta_1})\cdot p\Big((\mathbf{X}',\mathbf{Y}')\in T_{(X,Y),\delta}|\mathbf{X}'\in T_{\delta_1}\Big)\\
&\ge (1-\epsilon_{\delta_1}(n))\cdot 2^{-(I(X;Y)+2\delta H(Y))n}.
\end{aligned}$$

The Corollary follows as $\lim_{\delta_1\to\delta}\epsilon_{\delta_1}(n)=\epsilon_\delta(n)$. □

References

[1] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice Hall, Englewood Cliffs, NJ, 1971.

[2] T. Berger. Multiterminal source coding. In G. Longo, editor, *The Information Theory Approach to Communications*, CISM Courses and Lectures No. 229, pages 171–231. Springer Verlag, 1978.

[3] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[4] I. Csiszár and J. Körner. Broadcast channels with confidential messages. *IEEE Transactions on Information Theory*, 24(3):339–348, 1978.

[5] I. Csiszár, J. Körner, L. Lovász, K. Marton, and G. Simonyi. Entropy splitting for antiblocking pairs and perfect graphs. *Combinatorica*, 10(1):27–40, 1990.

[6] T. Feder, E. Kushilevitz, and M. Naor. Amortized communication complexity. In *Proceedings of the 32nd Annual Symposium on Foundations of Computer Science*, pages 239–248, 1991.

[7] M.L. Fredman and J. Komlós. On the size of separating systems and families of perfect hash functions. *SIAM Journal on Algebraic and Discrete Methods*, 5(1):61–68, 1984.

[8] J. Kahn and J.H. Kim. Entropy and sorting. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, pages 178–187, 1992.

[9] A. H. Kaspi and T. Berger. Rate-distortion for correlated sources. *IEEE Transactions on Information Theory*, 28(6):828–840, 1982.

[10] J. Körner. Coding of an information source having ambiguous alphabet and the entropy of graphs. *Proceedings of the 6th Prague Conference on Information Theory*, pages 411–425, 1973.

[11] J. Körner. Fredman-Komlós bounds and information theory. *SIAM. J. Alg. Disc. Meth.*, 7:560–570, 1986.

[12] J. Körner and K. Marton. New bounds for perfect hashing via information theory. *European Journal of Combinatorics*, 9:523–530, 1988.

[13] I. Newman, P. Ragde, and A. Wigderson. Perfect hashing, graph entropy, and circuit complexity. In *Proceedings of the 5th Annual Conference on Structure in Complexity Theory*, pages 91–100, 1991.

[14] A. Orlitsky. Worst-case interactive communication I: Two messages are almost optimal. *IEEE Transactions on Information Theory*, 36(5):1111–1126, September 1990.

[15] M.O. Rabin and A.C. Yao. Unpublished manuscript, 1979. See Yao [22].

[16] J. Radhakrishnan. Better bounds for threshold formulas. In *Proceedings of the 32nd Annual Symposium on Foundations of Computer Science*, pages 314–323, October 1991.

[17] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[18] G. Simonyi. Graph entropy: A survey. In L. Lovász, P. Seymour, and W. Cook, editors, *DIMACS Volume on Special Year on Combinatorial Optimisation.* Rutgers University, to appear.

[19] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19:471–480, 1973.

[20] H. Witsenhausen. The zero-error side information problem and chromatic numbers. *IEEE Transactions on Information Theory*, 22(5):592–593, September 1976.

[21] A.D. Wyner and J. Ziv. The rate distortion function for source coding with side information at the receiver. *IEEE Transactions on Information Theory*, 22(1):1–11, January 1976.

[22] A.C. Yao. Some complexity questions related to distributive computing. In *Proceedings of the 11th Annual ACM Symposium on Theory of Computing*, pages 209–213, 1979.