Estimating with Randomized Encoding the Joint Empirical Distribution in a Correlated Source

R. Ahlswede and Z. Zhang

1 Introduction

In order to put the present model and our results into the right perspectives we describe first key steps in multiuser source coding theory.

We are given a discrete memoryless double source (DMDS) with alphabets $\mathcal{X}$, $\mathcal{Y}$, and generic variables X, Y, i.e., a sequence of independent replicas (X_t, Y_t), $t = 1, 2, \dots$ of the pair of random variables (X, Y) taking values in the finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively.

I. Slepian and Wolf considered the problem of encoding the source output blocks $X^n \triangleq X_1 \dots X_n$ resp. $Y^n \triangleq Y_1 \dots Y_n$ by two separate encoders in such a way that a common decoder could reproduce both blocks with small probability of error. They proved that such an encoding is possible with rates (R_1, R_2) if and only if

$$R_1 \geq H(X|Y),\ R_2 \geq H(Y|X),\ R_1 + R_2 \geq H(X, Y). \tag{1.1}$$

II. It may happen, however, that what is actually required at the decoder is to answer a certain question concerning (X^n, Y^n). Such a question can of course be described by a function F of (X^n, Y^n). The authors of [5] are interested in those functions for which the number k_n of possible values of $F(X^n, Y^n)$ satisfies

$$\lim_{n \to \infty} \frac{1}{n} \log k_n = 0. \tag{1.2}$$

This means that the questions asked have only "a few" possible answers. For example, X_t and Y_t may be the results of two different quality control tests performed on the ith item of a lot. Then for certain purposes, e.g., for determining the price of the lot, one may be interested only in the frequencies of the various possible pairs (x, y) among the results, their order, i.e., the knowledge of the individual pairs (X_t, Y_t), being irrelevant. In this case $k_n \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}$, and (1.2) holds. A natural first question is whether or not it is always true in this case that, for large n, arbitrarily small encoding rates permit the decoder to determine $F(X^n, Y^n)$.

The authors of [5] also consider other choices of F and first obtain the following result. For every DMDS with

$$H(X|Y) > 0,\ H(Y|X) > 0$$

there exists a binary question (function F with only two possible values) such that in order to answer this question (determine $F(X^n, Y^n)$) one needs encoding rates as specified in (1.1).

R. Ahlswede et al. (Eds.): Information Transfer and Combinatorics, LNCS 4123, pp. 535–546, 2006.

As a matter of fact, almost all randomly selected functions F are of this kind. Since the reason for this unexpected phenomenon might be that randomly selected functions are very irregular, we next study more regular functions. A function F of special interest is the joint composition (joint type) of the two source blocks hinted at in the quality control example. In this respect our main result is that for determining the joint type of X^n and Y^n when Y^n is completely known at the decoder, X^n must be encoded with just as large a rate as if X^n were to be fully reproduced except for (exactly specified) singular cases. Actually, this analogous result is proved in [5] for a class of functions F which include, in addition to the joint type, the Hamming distance and — for alphabet size at least three — the parity of the Hamming distance.

As a consequence of these results one obtains that in the case of encoding both X^n and Y^n, the rates must satisfy

$$R_1 \geq H(X|Y), \ R_2 \geq H(Y|X), \tag{1.3}$$

in order that the joint type or the Hamming distance of X^n and Y^n can be determined by the decoder. In particular, it follows that for a DMDS with independent components (i.e., when X and Y are independent random variables (RV's)) nothing can be gained in rates, if instead of (X^n, Y^n) only the joint type or the Hamming distance of X^n and Y^n is to be determined by the decoder. For a DMDS with dependent components such a rate gain is possible, although it remains to be seen whether this always happens and to what extent. At present a complete solution to this problem is available only in the binary symmetric case. In fact, it readily follows from a result of Körner and Marton, that our necessary conditions (1.3) are also sufficient. Let us emphasize that their result concerns "componentwise" functions F

$$F(X^{n\prime}, Y^n) \triangleq \big(F_1(X_1, Y_1), F_1(X_2, Y_2), \dots, F_1(X_n, Y_n)\big), \tag{1.4}$$

where F_1 is defined on $\mathcal{X} \times \mathcal{Y}$.

In the binary symmetric case (i.e. $Pr\{X = Y = 0\} = Pr\{X = Y = 1\}$, $Pr\{X = 0, Y = 1\} = Pr\{X = 1, Y = 0\}$), they proved for the particular F with $f_1(x, y) \triangleq x + y \pmod 2$ that (R_1, R_2) is an achievable rate pair for determining $F(X^n, Y^n)$ if and only if (1.3) holds. Now observe that the types of X^n and of Y^n can be encoded with arbitrarily small rates and that those two types and the mod 2 sum $F(X^n, Y^n)$ determine the Hamming distance and also the joint type of X^n, Y^n.

Notice that the problem of F–codes is outside the usual framework of rate–distortion theory except for "componentwise" functions F, cf. (1.4). Still, a complete description of the achievable F rate region, e.g., for $F(x, y) \triangleq P_{x,y}$, may be as hard a problem as to determine the achievable rate region for reproducing X^n, Y^n within a prescribed distortion measure. We draw attention to the fact that for the latter problem it is also the projection of the achievable rate region to the R_1–axis which could be determined (Wyner–Ziv, [10]).

III. The authors of [9] consider a new model: identification via compressed data. To put it in perspective, let us first review the traditional problems in traditional rate–distortion theory for sources. Consider the diagram shown in Fig 1,

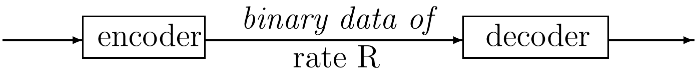

Fig. 1. Model for source coding

where $\{X_t\}_{t=1}^{\infty}$ is an independent and identically distributed (i.i.d.) source taking values in a finite alphabet $\mathcal{X}$. The encoder output is a binary sequence which appears at a rate of R bits per symbol. The decoder output is a sequence $\{\hat{X}_n\}_{n=1}^{\infty}$ which takes values in a finite reproduction alphabet $\mathcal{Y}$. In traditional source coding theory, the decoder is required to recover $\{X_t\}_{t=1}^{\infty}$ either completely or with some allowable distortion. That is, the output sequence $\{\hat{X}_t\}_{t=1}^{\infty}$ of the decoder must satisfy

$$\frac{1}{n}\sum_{i=1}^{n} E_{\rho}(X_t, \hat{X}_t) \leq d \tag{1.5}$$

for sufficiently large n, where $\mathbf{E}$ denotes the expected value,

$$\rho : \mathcal{X} \times \mathcal{Y} \rightarrow [0, +\infty)$$

is a distortion measure, and d is the allowable distortion between the source sequence and the reproduction sequence. The problem is then to determine the infimum of the rate R such that the system shown in Fig. 1 can operate in such a way that (1.5) is satisfied. It is known from rate distortion theory that the infimum is given by the rate distortion function of the source $\{X_t\}_1^{\infty}$.

Let us now consider the system shown in Fig. 2,

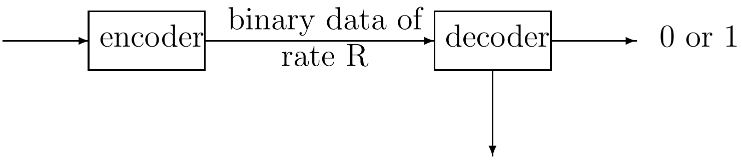

Fig. 2. Model for joint source coding and identification

where the sequence $\{Y_t\}_1^{\infty}$ is a sequence of i.i.d. random variables taking values from $\mathcal{Y}$. Knowing Y^n, the decoder is now required to be able to identify whether or not the source sequence X^n and the sequence Y^n have some prescribed relation F in such a way that two kinds of error probabilities, the probabilities for misacceptance (false identification) and the probabilities for misrejection, satisfy some prescribed conditions. In parallel with rate distortion theory, we consider in this paper the following relation F defined by:

$$n^{-1}\sum_{t=1}^{n}\rho(X_t,Y_t)\le d. \tag{1.6}$$

That is, the values X^n and Y^n are said to have relation F if (1.6) is satisfied. The problem we are interested in is to determine the infimum of the rate R such that the system shown in Fig. 2 can operate so that the error probability of misrejection, that is the decoder votes for 0 even though F holds, and the error probability of misacceptance, that is the decoder votes for 1 even though F does not hold, satisfy constraints on the error exponents α and β, say. So the goal of the decoder is to identify whether X^n is close to Y^n (in the sense of relation F) or not. The encoder is cooperative.

It must be remarked that in this model the minimum achievable rate is shown to always equal zero, if we only require that the two kinds of error probabilities go to zero as n goes to infinity. So the exponential decay of error probabilities makes the problem meaningful. The regions of pairs of exponents (α,β) are studied as functions of rate R and fidelity criterion d for general correlated sources. Complete characterizations are obtained, if X^n and Y^n are independent.

IV. Now we come to our new model of estimating the joint empirical distribution (joint type) not exactly like in [5], but within some accuracy only. This "computational" aspect was motivated by [9]. Furthermore the help of randomization was understood in [6] and [7].

We consider the following model. The encoder knows a word $x^n\in\mathcal{X}^n$ and the receiver knows a word $y^n\in\mathcal{Y}^n$. The encoder sends information of at most ℓ bits to the receiver, who uses these bits and his own observation $y^n\in\mathcal{Y}^n$ to estimate the joint type. The question is how accurate the estimate can be. It can be formalized as follows:

A randomized encoding is a pair $\mathcal{E}=\big\{\mathcal{M},Q(\cdot|\cdot)\big\}$, where

$$\mathcal{M}=\{1,2,\dots,M\}, M=2^k, \quad\text{and}\quad Q(\cdot|x^n)\in\mathcal{P}(\mathcal{M}), x^n\in\mathcal{X}^n. \tag{1.5}$$

Here and elsewhere $\mathcal{P}(\cdot)$ denotes the set of probability distributions (abbreviated as PD) or probability vectors of a set in brackets.

The decoder uses a decoding function

$$g:\mathcal{M}\times\mathcal{Y}^n\to\mathcal{P}(\mathcal{X}\times\mathcal{Y}). \tag{1.6}$$

Next we describe performance criteria for the code $\mathcal{C}=(\mathcal{E},g)$. For any two PD's $P=(P_1,\dots,P_s)$ and $Q=(Q_1,\dots,Q_s)$ define the norms

$$\|P-Q\|_1=\sum_{i=1}^{s}|P_i-Q_i|, \tag{1.7}$$

$$\|P-Q\|_2=\sqrt{\sum_{i=1}^{s}|P_i-Q_i|^2}, \tag{1.8}$$

and the "individual errors" based on them for the code $\mathcal{C} = (\mathcal{E}, g)$

$$\overline{e}_{\mathcal{C}}^{(i)}(x^n, y^n) = \sum_{j \in \mathcal{M}} Q(j|x^n) \|g(j, y^n) - P_{x^n y^n}\|_i \quad (i = 1, 2). \tag{1.9}$$

This leads to two notions of maximal errors of the code $(\mathcal{C}, g)$

$$\overline{e}_{\mathcal{C}}^{(i)} = \max_{x^n, y^n} \overline{e}_{\mathcal{C}}^{(i)}(x^n, y^n); \; (i = 1, 2). \tag{1.10}$$

Finally, we get the best possible maximal errors (for parameters n and M)

$$\overline{e}^{(i)}(n, M) = \min_{\mathcal{C}: |\mathcal{M}| = M} \overline{e}_{\mathcal{C}}^{(i)}. \tag{1.11}$$

We mention two other kinds of criteria for the measurement of the estimation error.

Let J be a RV with distribution $\Pr(J = j) = Q(j|x^n)$ and use the RV's

$$\varDelta_{x^n y^n}^{(i)}(J) = \|g(J, y^n) - P_{x^n y^n}\|_i \tag{1.12}$$

to define

$$e_{\mathcal{C}}^{(i)}(x^n, y^n, \delta) = \Pr\bigl(\varDelta_{x^n y^n}^{(i)}(J) > \delta\bigr); \; i = 1, 2; \tag{1.13}$$

and

$$e_{\mathcal{C}}^{(i)}(x^n, y^n, \varepsilon) = \min\bigl\{\delta : \Pr\bigl(\varDelta_{x^n y^n}^{(i)}(J) > \delta\bigr) < \varepsilon\bigr\}; \; i = 1, 2. \tag{1.14}$$

Actually, all these definitions lead to similar results and we start here with $e_{\mathcal{C}}^{(2)}(x^n, y^n, \varepsilon)$ for which we define

$$e_{\mathcal{C}}(\varepsilon) = \max_{x^n, y^n} e_{\mathcal{C}}^{(2)}(x^n, y^n, \varepsilon) \tag{1.15}$$

and

$$e(n, M, \varepsilon) = \min_{\mathcal{C}: |\mathcal{M}| = M} e_{\mathcal{C}}(\varepsilon). \tag{1.16}$$

An appropriate scaling

$$\alpha(D, \varepsilon) = \sup_{n, M: \frac{\log \log M}{\log n} < D} \frac{-\log e(n, M, \varepsilon)}{\log n} \tag{1.17}$$

leads to a striking result.

Theorem

$$\alpha(D, \varepsilon) = D \text{ for all } \varepsilon \in (0, 1). \tag{1.18}$$

2 Direct Coding Theorem

We use the following simple coding method. Label the members of $\binom{[n]}{\ell_n}$, the set of all ℓ_n–element subsets of $[n] = \{1, 2, \ldots, n\}$. The sender randomly selects one such subset and transmits its label and the components of x^n within this subset to the receiver. The receiver uses the joint type of y^n and x^n *within this subset* as the estimate of the joint type.

We now evaluate the performance of this method. First we count the number L of subsets where (x^n, y^n)'s local joint type is at least $\sqrt{\ell_n^{-1} \log_n^2}$ away from the true type.

For this we need the definitions

$$n(x, y) := P_{x^n y^n}(x, y)n, \tag{2.1}$$

$$\ell(x, y) := \text{local frequencies of } \ (x^n, y^n) \text{ in } \ell_n\text{–subset considered} \tag{2.2}$$

and

$$\overrightarrow{\ell} := \big(\ell(x, y)\big)_{(x,y)\in\mathcal{X}\times\mathcal{Y}}. \tag{2.3}$$

Clearly $\sum\limits_{x,y} \ell(x, y) = \ell_n$.

Now

$$L = \sum_{\overrightarrow{\ell}:\sum_{x,y}\left|\frac{\ell(x,y)}{\ell_n}-\frac{n(x,y)}{n}\right|^2>\frac{\log^2 n}{\ell_n}} \prod_{x,y} \binom{n(x,y)}{\ell(x,y)}$$

and

$$L \cdot \binom{n}{\ell_n}^{-1} \leq O(n^{ab-1}) \max_{\overrightarrow{\ell}:\sum_{x,y}\left|\frac{\ell(x,y)}{\ell_n}-\frac{n(x,y)}{n}\right|^2>\frac{\log^2 n}{\ell_n}} \frac{\sqrt{\ell_n}}{\prod\limits_{x,y}\sqrt{\ell(x,y)}}$$

$$\cdot \exp\left\{\sum_{x,y} n(x,y)h\left(\frac{\ell(x,y)}{n(x,y)}\right) - nh\left(\frac{\ell_n}{n}\right)\right\} \tag{2.4}$$

by Stirling's formula.

This can be bounded from above by using the following auxiliary result.

Lemma. *Let positive integers $n(x, y)$, $\ell(x, y)$, ℓ, n satisfy*

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}} n(x, y) = n, \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \ell(x, y) = \ell, \ell(x, y) \leq n(x, y) \ \ \textit{for} \ \ x \in \mathcal{X}, y \in \mathcal{Y}.$$

Then

$$\theta \triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} n(x,y)\left[h\left(\frac{\ell(x,y)}{n(x,y)}\right) - h\left(\frac{\ell}{n}\right)\right] \leq -\frac{n}{2ab\ell} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} n(x,y)\left(\frac{\ell(x,y)}{n(x,y)} - \frac{\ell}{n}\right)^2,$$

where $a = |\mathcal{X}|$, $b = |\mathcal{Y}|$.

Proof. For $\gamma(x,y) \triangleq \frac{\ell(x,y)}{n(x,y)} - \frac{\ell}{n}$ obviously $\sum\limits_{x,y} n(x,y)\gamma(x,y) = 0$. With $C \triangleq \frac{\ell}{n}$ we can now write

$$\theta = \sum_{x,y} n(x,y)\big[h\big(C + \gamma(x,y)\big) - h(C)\big].$$

By Lagrange's interpolation formula

$$h\big(C + \gamma(x,y)\big) - h(C) = h'(C)\gamma(x,y) + \frac{h''(\xi(x,y))}{2}\,\gamma^2(x,y),$$

where $\xi(x,y)$ is between C and $C + \gamma(x,y)$.

Thus

$$\begin{aligned}
\theta &= \sum_{x,y} n(x,y)h'(C)\gamma(x,y) + \sum_{x,y} n(x,y)\frac{h''(\xi(x,y))}{2}\,\gamma^2(x,y) \\
&= \sum_{x,y} n(x,y)\frac{h''(\xi(x,y))}{2}\,\gamma^2(x,y) \\
&\le \sum_{x,y:\gamma(x,y)\le 0} n(x,y)\frac{h''(\xi(x,y))}{2}\,\gamma^2(x,y) \\
&= \sum_{x,y:\gamma(x,y)\le 0} n(x,y)\left(-\frac{1}{2\xi(x,y)(1-\xi(x,y))}\right)\,\gamma^2(x,y) \\
&\le \sum_{x,y:\gamma(x,y)\le 0} n(x,y)\left(-\frac{1}{2C}\right)\,\gamma^2(x,y) \\
&= -\frac{n}{2\ell}\sum_{x,y:\gamma(x,y)\le 0} n(x,y)\gamma^2(x,y).
\end{aligned}$$

Clearly, the claimed inequality follows from the identity

$$\mu \triangleq \min_{\rho:\sum\limits_{x,y} n(x,y)\rho(x,y)=0} \frac{\sum\limits_{x,y:\rho(x,y)\le 0} n(x,y)\rho^2(x,y)}{\sum\limits_{x,y} n(x,y)\rho^2(x,y)} = \frac{1}{ab}, \tag{2.5}$$

which remains to be proved.

Obviously, the optimizing ρ has the properties:

1. $|\{(x,y) : \rho(x,y) > 0\}| = 1$
2. There exists a constant ν such that $\rho(x,b) \le 0$ implies $\rho(a,b) = \nu$.

These two properties imply

$$\mu = \frac{(ab-1)\left(\frac{a}{ab-1}\right)^2}{a^2 + (ab-1)\left(\frac{a}{ab-1}\right)^2} = \frac{1}{ab}.$$

□

We apply now the Lemma to upper bound the exponent in the exponential function and get

$$\begin{aligned}
L\cdot\binom{n}{\ell_n}^{-1} \le & O(n^{ab-1}) \max_{\vec{\ell}:\sum_{x,y}\left|\frac{\ell(x,y)}{\ell_n}-\frac{n(x,y)}{n}\right|^2>\frac{\log^2 n}{\ell_n}} \frac{\sqrt{\ell_n}}{\prod_{x,y}\sqrt{\ell(x,y)}} \\
& \cdot\exp\left\{-\frac{n}{2ab\ell_n}\sum_{x\in\mathcal{X},y\in\mathcal{Y}} n(x,y)\right\}\left(\frac{\ell(x,y)}{n(x,y)}-\frac{\ell_n}{n}\right)^2 \\
\le & \max_{\vec{\ell}:\sum_{x,y}\left|\frac{\ell(x,y)}{\ell_n}-\frac{n(x,y)}{n}\right|^2>\frac{\log^2 n}{\ell_n}} O(n^{ab}) \\
& \cdot\exp\left\{-\frac{\mu}{2}\frac{n}{\ell_n}\sum_{x,y} n(x,y)\right\}\left(\frac{\ell(x,y)}{n(x,y)}-\frac{\ell_n}{n}\right)^2 \\
= & \max_{\vec{\ell}:\sum_{x,y}\left|\frac{\ell(x,y)}{\ell_n}-\frac{n(x,y)}{n}\right|^2>\frac{\log^2 n}{\ell_n}} O(n^{ab}) \\
& \cdot\exp\left\{-\frac{\mu}{2}\sum_{x,y}\frac{n\cdot\ell_n}{n(x,y)}\right\}\left(\frac{\ell(x,y)}{\ell_n}-\frac{n(x,y)}{n}\right)^2 \\
\le & O(n^{ab})\exp\left\{-\frac{\mu}{2}\log^2 n\right\}\to 0 \text{ as } n\to\infty.
\end{aligned}$$

Now the number of bits needed for sending an element of $\binom{[n]}{\ell_n}$ is $\log\binom{n}{\ell_n}$ and for sending the ℓ_n bits is ℓ_n. This amounts to a total number of $\log\binom{n}{\ell_n}+\ell_n$ bits. The accuracy achieved is $\ell_n\log^2 n$.

Therefore we get

$$\frac{\log\delta}{\log n}=\frac{\log\ell_n}{\log n}+\frac{2\log\log n}{\log n}$$

and

$$\frac{\log\log M}{\log n}=\frac{\log(\ell_n\log n+\ell_n)}{\log n}=\frac{\log\ell_n}{\log n}+\frac{\log(\log n+1)}{\log n}.$$

If $\ell_n\gg\log n$, then

$$\frac{\log\delta}{\log n}\approx\frac{\log\log M}{\log n}$$

and the direct part is proved.

3 Converse of Coding Theorem (Proof in Binary Case, Exercise in General Case)

Let $\mathcal{C}=(\mathcal{E},g)$ be a (D,α,n) code and let

$$\mathcal{M}(x^n,y^n)=\left\{m\in\mathcal{M}:\|g(m,y^n)-P_{x^ny^n}\|_2^2\le\exp\left\{\alpha\log n+o(\log n)\right\}\right\}.$$

We have

$$Q\big(\mathcal{M}(x^n, y^n)|x^n\big) > 1 - \varepsilon.$$

Select now n^β codewords independently at random according to the PD $Q(\cdot|x^n)$. Abbreviate the random code $\big(X_1^n(x^n), \dots, X_{n^\beta}^n(x^n)\big)$ as $B(x^n)$ and use $\tilde{F}(\cdot|x^n)$ to denote the uniform distribution on $B(x^n)$.

$$\Pr\left(\tilde{F}\big(\mathcal{M}(x^n, y^n)|x^n\big) < \frac{1}{2} + \varepsilon\right) \approx \sum_{k > \left(\frac{1}{2} - \varepsilon\right) n^\beta} \binom{n^\beta}{k} \varepsilon^k (1-\beta)^{n^\beta - k}$$
$$\approx \exp\left\{-n^\beta D\left(\frac{1}{2} - \varepsilon \middle\| \varepsilon\right)\right\} \gtrsim \lambda n^\beta.$$

A y^n is called irregular with respect to x^n for a particular $B(x^n)$ or $\tilde{F}(\cdot|x^n)$ iff $\tilde{F}\big(\mathcal{M}(x^n, y^n|x^n)\big) < \frac{1}{2} + \varepsilon$.

The average number of irregular y^n is $2^{n-\lambda n^\beta}$. Therefore a choice of $B(x^n)$ exists such that the number of irregular y^n's is at most $2^{n-\lambda n^\beta}$.

According to this principle we make choices for every x^n. So we get a whole family $\big(\tilde{F}(\cdot|x^n)\big)_{x^n \in \mathcal{X}^n}$, where each member has at most $2^{n-\lambda n^\beta}$ irregular y^n's.

Now we use a constant weight error correcting code of cardinality $2^{\gamma n}$ and of minimum distance μn, where γ, μ are constants (independent of n).

Let x_1^n and x_2^n be two codewords of this code. We prove that for suitable β, $B(x_1^n) \neq B(x_2^n)$. Actually, we count the number of y^n's with

$$\left(\sum_{x,y} \big(n_{x_1^n y^n}(x,y) - n P_{x_2^n y^n}(x,y)\big)^2\right)^{\frac{1}{2}} \geq 2n^{\alpha + o(1)}.$$

For this define

$A = \big\{t \in [n] : x_{1t} = 1 \text{ and } x_{2t} = 0\big\}, B = \big\{t \in [n] : x_{1t} = 1 \text{ and } x_{2t} = 1\big\},$
$C = \big\{t \in [n] : x_{1t} = 0 \text{ and } x_{2t} = 1\big\},$ and $D = \big\{t \in [n] : x_{1t} = 0 \text{ and } x_{2t} = 0\big\}.$

This number of y^n's exceeds

$$2^{|B|+|D|} \sum_{|u-v| > 2n^{\frac{1}{2}\alpha + o(1)}} \binom{|A|}{u}\binom{|C|}{v} = 2^{|B|+|D|} \sum_{\ell > 2n^{\frac{1}{2}\alpha + o(1)}} \binom{|A| + |C|}{|A| - \ell} = 2^{n - \mu n^{\alpha + o(1)}}.$$

Now, if $B(x_1^n) = B(x_2^n)$, then those y^n must be irregular for at least one of x_1^n, x_2^n. Hence $2^{n-\mu n^\alpha + o(1)} \leq 2^{n-\lambda n^\beta}$ and thus $\alpha \geq \beta + o(1)$. Finally $M^{n^\beta} \geq 2^{rn}$ implies $M \geq 2^{rn^{1-\beta}} \geq 2^{n^{1-\alpha-o(1)}}$. The converse is proved in the binary case.

4 Other Problems

A. The existing work on statistical inference (hypothesis testing and estimation in [4] and [3]) under communication constraints uses a "one shot" side information. It seems important to introduce and analyze interactive models.

B. **Permutation invariant functions**
A function F, defined on $\mathcal{X}^n \times \mathcal{Y}^n$, is called permutation invariant iff for all permutations π of the set $\{1, 2, \ldots, n\}$

$$F(x^n, y^n) = F(\pi x^n, \pi y^n), \tag{4.1}$$

where $x^n = (x_1, x_2, \ldots, x_n)$

$$\pi x^n = (x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}) \tag{4.2}$$

and y^n, πy^n are defined analogously.

Permutation invariant functions are actually functions of the joint empirical distribution $P_{x^n y^n}$ of the sequences x^n and y^n, where for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$

$$P_{x^n y^n}(x, y) = |\{t : x_t = x, y_t = y\}| n^{-1}. \tag{4.3}$$

Examples of permutation invariant functions include, but are not limited to, sum–type functions f^n,

$$f^n(x^n, y^n) = \sum_{t=1}^{n} f(x_t, y_t), \tag{4.4}$$

such as the Hamming distance function. In identification problems, we can be interested in Boolean functions. When the problem is permutation invariant, we need to study permutation invariant Boolean functions. If we estimate the joint empirical distribution of x^n and y^n. Then $\big(P_{x^n y^n}(x, y)\big)_{x \in \mathcal{X}, y \in \mathcal{Y}}$ is a permutation invariant vector–valued function on $\mathcal{X}^n \times \mathcal{Y}^n$.

C. **Approximation of continuous permutation invariant functions**
Let F be a continuous function defined on the compact set $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Define

$$\hat{F}(x^n, y^n) = F(P_{x^n y^n}). \tag{4.5}$$

If the task of the receiver is not to estimate $P_{x^n y^n}$, but to compute $\hat{F}(x^n, y^n)$, what is then the trade–off between the computation accuracy and the "communication rate" D?

This problem is closely related to the joint empirical distribution estimation problem — actually, it generalizes it.

D. **Classification Problem**
Let $\{\mathcal{A}_0, \mathcal{A}_1\}$ be a partition of $\mathcal{X}^n \times \mathcal{Y}^n$ and let both sets in this partition be permutation invariant. If in the model treated in this paper the task of the receiver is to determine whether or not $(x^n, y^n) \in \mathcal{A}_0$, then this is a new "classification" problem.

In case we want to determine this exactly, then we have to transmit for "most" partitions almost all bits of x^n to the receiver. We introduce now a model, which allows a much lower rate.

Let $d_1(P, P') = \|P - P'\|_1$ be the L_1–distance of P and P' in $\mathcal{P}(\mathcal{X}^n \times \mathcal{Y}^n)$. For $\mathcal{A} \subset \mathcal{X}^n \times \mathcal{Y}^n$ and $\delta > 0$ let

$$\Gamma_\delta(\mathcal{A}) = \big\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : d_1(P_{x^n y^n}, P_{x'^n y'^n}) \le \delta \ \text{ for some } \ (x'^n, y'^n) \in \mathcal{A} \big\},$$

and for permutation invariant $\mathcal{A} \subset \mathcal{X}^n \times \mathcal{Y}^n$ let

$$N(\mathcal{A}) = \left|\left\{P_{x^n y^n} : (x^n, y^n) \in \mathcal{A}\right\}\right|.$$

Now, for $\varepsilon > 0$, find maximal $\delta_0, \delta_1 \geq 0$ such that

$$\frac{N\big(\varGamma_{\delta_0}(\mathcal{A}_0) \cap \mathcal{A}_1\big)}{N(\mathcal{A}_1)} \leq \varepsilon, \tag{4.6}$$

$$\frac{N\big(\varGamma_{\delta_1}(\mathcal{A}_1) \cap \mathcal{A}_0\big)}{N(\mathcal{A}_0)} \geq \varepsilon. \tag{4.7}$$

Finally, let

$$g : \mathcal{M} \times \mathcal{Y}^n \to \{0, 1\}$$

be a binary–valued function such that for all $(x^n, y^n) \in \mathcal{A}_0 \smallsetminus \varGamma_{\delta_1}(\mathcal{A}_1)$

$$Q\big(g(J, y^n) = 0 | x^n\big) \geq 1 - \varepsilon$$

and for all $(x^n, y^n) \in \mathcal{A}_1 \smallsetminus \varGamma_{\delta_0}(\mathcal{A}_0)$

$$Q\big(g(J, x^n) = 1 | x^n\big) \geq 1 - \varepsilon.$$

What is the minimum number of bits $\lceil \log M \rceil$ needed? This problem is also closely related to the joint empirical distribution estimation problem.

References

1. R. Ahlswede, Channel capacities for list codes, J. Appl. Probability, 10, 824–836, 1973.
2. R. Ahlswede, Coloring hypergraphs: A new approach to multi–user source coding, Part I, Journ. of Combinatorics, Information and System Sciences, Vol. 4, No. 1, 76–115, 1979; Part II, Journ. of Combinatorics, Information and System Sciences, Vol. 5, No. 3, 220–268, 1980.
3. R. Ahlswede and M. Burnashev, On minimax estimation in the presence of side information about remote data, Ann. of Stat., Vol. 18, No. 1, 141–171, 1990.
4. R. Ahlswede and I. Csiszár, Hypothesis testing under communication constraints, IEEE Trans. Inform. Theory, Vol. 32, No. 4, 533–543, 1986.
5. R. Ahlswede and I. Csiszár, To get a bit of information may be as hard as to get full information, IEEE Trans. Inform. Theory, Vol. 27, 398–408, 1981.
6. R. Ahlswede and G. Dueck, Identification via channels, IEEE Trans. Inform. Theory, Vol. 35, No. 1, 15–29, 1989.
7. R. Ahlswede and G. Dueck, Identification in the presence of feedback — a discovery of new capacity formulas, IEEE Trans. Inform. Theory, Vol. 35, No. 1, 30–39, 1989.
8. R. Ahlswede and J. Körner, Source coding with side information and a converse for degraded broadcast channels, IEEE Trans. Inf. Theory, Vol. 21, 629–637, 1975.
9. R. Ahlswede, E. Yang, and Z. Zhang, Identification via compressed data, IEEE Trans. Inform. Theory, Vol. 43, No. 1, 22–37, 1997.
10. R. Ahlswede and Z. Zhang, Worst case estimation of permutation invariant functions and identification via compressed data, Preprint 97–005, SFB 343 “Diskrete Strukturen in der Mathematik”, Universität Bielefeld.

11. T. Berger, Rate Distortion Theory, Englewood Cliffs, NJ, Prentice–Hall, 1971.
12. I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, New York, Academic, 1981.
13. D. Slepian and J.K. Wolf, Noiseless coding of correlated information sources, IEEE Trans. Inform. Theory, Vol. 19, 471–480, 1973.
14. A.D. Wyner and J. Ziv, The rate–distortion function for source coding with side information at the decoder, IEEE Trans. Inform. Theory, Vol. 22, 1–10, 1976.