A CODING THEOREM FOR MULTITERMINAL ESTIMATION

Amin Zia, James P. Reilly, Timothy R. Field, Shahram Shirani
Department of ECE, McMaster University, Hamilton, Ontario, Canada.

1. ABSTRACT

In this paper a coding theorem for multiterminal estimation is presented. The theorem is a generalization of the distributed coding theorem first proved by Slepian and Wolf [7], where the goal is to estimate the joint probability distribution of correlated sources, rather than to reconstruct them at the receiver. For this, it is shown first that the joint-type of the received sequences is a sufficient statistic for estimation. Then, it is proved that for sufficiently large sequences, only a sum-rate lower bounded by the mutual information of the correlated sources is *"sufficient"* to perfectly reconstruct the sufficient statistic at the receiver. Simulation results for the special case of estimation with side information at the receiver is provided.

2. INTRODUCTION

Suppose two spatially separated "non-cooperative" sensors measure a random phenomenon characterized by a bivariate binary probability distribution (PD) $p(X, Y;\ t)$ where $t \in \Re$ is scalar parameter by which the PD is uniquely identified. The sensors transmit the sequence of measurements via generally limited capacity communication channels to a common receiver. Given the set of possibly compressed sequences, it is desirable for the receiver to estimate the parameter, i.e. identify the PD. It was Berger [1] that introduced this interesting problem also referred to as "multiterminal estimation". Later through a series of papers by Zhang and Berger [2], Han and Amari [3] [4] the multiterminal estimation problem was studied and necessary and sufficient conditions for designing efficient estimators under *zero-rate* as well as *positive-rate* conditions were derived. The analytical solutions provided in this series of papers are mathematically intractable and too complicated to implement; they can only be applied to elementary scenarios [3]. The main questions in this category of solution have been "Given constrained communication resources, what degree of accuracy is attainable and, whether estimation is efficient, and to what extent?"

In contrast we approach the problem by answering the question "what rate (or sum of rates) is *"sufficient"* to have perfect estimation?" The answer is an effort to provide a practical solution to the general multiterminal estimation problem. Our approach is closer to the distributed source coding theorem first proved by Slepian-Wolf [7]. In fact, we extend the distributed source coding theorem to the case in which the accuracy of parameter estimation is the main concern, rather than the perfect reconstruction of the sources. For this purpose, in Section (3) we first show that in order to have optimal estimation it is sufficient to preserve the information residing in the *joint-type* of the correlated sequences. Then in Section (4) the rate region for noiseless transmission of such information as well as a coding scheme for achieving the rates are presented. It is proved that the necessary sum-rate is $I(X, Y)$, i.e. the mutual information between the sources. In Section (5) simulation results for the special case of estimation with side information at the receiver are presented. Concluding remarks close the paper.

3. MULTITERMINAL ESTIMATION

Suppose a phenomenon is characterized by a *discrete* bivariate joint probability distribution $p(X, Y;\ t)$, where $t \in \Re$ is a scalar parameter by which the PD is uniquely identified. The phenomenon is measured by two "non-cooperative" spatially separated sensors S_x and S_y. The two sensors can be considered as monitoring two correlated sources generating a pair of correlated sequences $x^n = (x_1, ..., x_n)$ and $y^n = (y_1, ..., y_n)$ of n *i.i.d.* samples respectively drawn from $p(X, Y;\ t)$. It is assumed that the random variables X and Y choose their values from a discrete set of alphabets $x, y \in= \{0, 1\}$. The sequences are encoded into corresponding message sequences separately by encoder functions f and g, respectively, and transmitted to a common receiver via communication channels which are generally limited in capacity. The messages are chosen from sets $M_x = \{1, 2, ..., 2^{nR_x}\}$ and $M_y = \{1, 2, ..., 2^{nR_y}\}$ respectively. Here R_x and R_y are called the *code rates* defined as $R_x = \lim_{n\to\infty} \frac{\log |M_x|}{n}$ and $R_y = \lim_{n\to\infty} \frac{\log |M_y|}{n}$ where $|M_x|$ and $|M_y|$ are the cardinalities of the sets M_x and M_y, respectively. Since the communication resources are usually limited, the design of the codebooks and the messages might involve compression. By appropriately incorporating the compression such that the rate of the codes is "no less" than the available channel capacity, the decoder receives the noiseless encoded messages. At the receiver, it is desired to estimate the parameter of the underlying phenomenon using the received sequences and a decoder/estimator function h, i.e. $\hat{t} = h(x^n, y^n, m_{x^n}, m_{y^n})$.

A. Repeated Observations and Sufficient Statistics

The bivariate binary PD $p(X, Y;\ t) = \begin{bmatrix} p_{00}(t) & p_{10}(t) \\ p_{01}(t) & p_{11}(t) \end{bmatrix}$ is assumed to be a member

of the K-dimensional *exponential family* defined generally as [3]:

$$p(X,Y;\theta) = \exp\left[C + \sum_{k=1}^{K} \theta^k F_k - \psi(\theta)\right], \quad (1)$$

where $\theta(t) \in \Theta$ is the vector of *canonical parameters*, and the functions $F_k(X,Y)$ are called *sufficient statistics*, and $\psi(\theta(t))$ is the normalization function. Notice that the canonical parameters are generally functions of the scalar parameter t. Therefore throughout the paper, estimation of the canonical parameters implies estimation of the parameter t implicitly. (The PD with scalar parameter t is a member of the curved exponential family [3], and therefore for estimating t, it is sufficient to estimate the corresponding canonical parameters $\theta(t)$. To avoid confusion we do not show this relationship explicitly.) The representation of the PD is minimal and the θ-parameters are *"minimal canonical"* parameters. The F_k functions are minimal sufficient statistics when the parameters and functions are linearly independent [8].

For the bivariate binary PD the functions describing the minimal sufficient statistics are $C(X,Y) = 0$, $F_1(X,Y) = \delta_1(X)$, $F_2(X,Y) = \delta_1(Y)$, and $F_3(X,Y) = \delta_{11}(X,Y)$, where $\delta_a(X) = 1$ iff $X = a$ is the *Kronecker delta* function and $\delta_{11}(X,Y)$ is defined similarly. Also the canonical parameters known as θ-*coordinates* are defined as $\theta^1 \triangleq \log\frac{p_{10}}{p_{00}}$, $\theta^2 \triangleq \log\frac{p_{01}}{p_{00}}$, and $\theta^3 \triangleq \log\frac{p_{00}p_{11}}{p_{01}p_{10}}$. Also $\psi(\theta) = -\log p_{00}$.

Note that the functions F_1, F_2, and F_3 are related to the probability of occurrences of $X = 1$, $Y = 1$ and $X, Y = 1, 1$, respectively. The relative frequency of occurrence of "1" in the sequences x^n and y^n as well as the relative frequency of joint occurrences of "$(1,1)$" in (x^n, y^n) are sufficient statistics for estimation of the PD parameters.

Definitions-Joint and Marginal Types: [6] Given n *i.i.d.* samples $(x^n, y^n) = (x_1, y_1),\ ...,\ (x_n, y_n)$ taken from the binary PD $p(X,Y;\theta)$, the *joint-type* $\tilde{p}_{ab}(x,y)$ is defined as the relative frequency of the occurrence of $(x,y) = (a,b)$ in n observations:

$$\tilde{p}_{ab}(x,y) = \frac{1}{n}\sum_{s=1}^{n} \delta_{ab}(x_s, y_s), \quad (2)$$

where $\delta_{ab}(x_s, y_s) = 1$ iff $(x_s, y_s) = (a,b)$; $\forall\ a, b \in 0, 1$ is the Kronecker delta function. The *marginal-types* $\tilde{p}_a(x)$ and $\tilde{p}_b(y)$ are defined similarly as the relative occurrence of a and b in the samples, respectively.

The parameter t can be estimated using the ML estimation equation obtained by the equating the partial derivatives of the likelihood function defined for the PS in Eq. (1) to zero. The *Cramer-Rao* lower-bound (CRLB) defined by the inverse of the *Fisher information* (FI) matrix (FIM), respectively [8].

Theorem 1: Given n *i.i.d.* samples $(x^n, y^n) = (x_1, y_1),\ ...,\ (x_n, y_n)$ taken from a binary PD $p(X,Y,\theta)$. (a) The marginal types $\tilde{p}_1(x)$ and $\tilde{p}_1(y)$ and the joint-type $\tilde{p}_{11}(x,y)$ are minimal sufficient statistics. (b) They achieve the variance of error *Cramer-Rao lower bound*(CRLB) for maximum likelihood (ML) estimation of θ.

Proof: The proof is straightforward by forming the log-likelihood of the available samples using the exponential representation of the binary PD $p(X,Y;\theta)$, and the definition of the joint as well as marginal-types. The functions describing the sufficient statistics are then solved by maximum likelihood procedures. For a detailed proof, refer to [8]. □

Therefore in order to have efficient estimation, it is sufficient to preserve the marginal-types as well as the joint-type $\tilde{p}_{11}(x,y)$. By carefully examining these types, it becomes clear that any compression/transmission strategy that preserves *the number of ones* in each sequence as well as *the relative occurrence of ones in both sequences* will be sufficient for this purpose.

4. Multiterminal Coding Theorem

In this section we provide *"sufficient"* conditions and a coding scheme to preserve the marginal types as well as the joint-type at the receiver in multiterminal estimation scenarios. Suppose we generate a pair of sequences (PoS) (x^n, y^n) and two sequences x^n and y^n consisting of n i.i.d samples from probability distribution $p(X,Y)$ and its marginal distributions $p(X)$ and $p(Y)$, respectively. We then pair up the sequences x^n and y^n referring to as paired-up pair of sequences (PPofS) $(x^n, y^n)_p$. We appreciate the importance of noticing the different notations used for the PoS ((x^n, y^n))and PPoS ($(x^n, y^n)_p$). Define:

Definition- The Set of Correlation-Typical Sequences: The set A of *correlation-typical* PPoS $(x^n, y^n)_p$ with respect to the the marginal distribution $p(X)$ and $p(Y)$ and the joint PD $p(X,Y)$ is the set of PPoS $n-$sequences with empirical mutual entropy $\epsilon-$close to the true mutual entropy, i.e.,

$$A = \{(x^n, y^n)_p \in \mathcal{X}^n \times \mathcal{Y}^n : \quad (3)$$

$$\left|\frac{1}{n}\log\frac{p(x^n, y^n)}{p(x^n)p(y^n)} - I(X,Y)\right| < \epsilon\}. \quad (4)$$

Definition- The Set of Correlation-Typical "Pair of Sequences": The set A^I of PPoS $(x^n, y^n)_p \in A$ which are also jointly-typical, i.e. $(x^n, y^n)_p \in A^{XY}$ with respect to the joint distribution $p(X,Y)$. The A^{XY} is the corresponding joint-typical set that contains all the joint-typical pairs of sequences (x^n, y^n) with respect to PD $p(X,Y)$ (see [5]).

Lemma: (*Correlational Asymptotic Equipartition Property, CAEP*)
For sufficiently large n and for arbitrarily chosen $\epsilon > 0$:

1) $\Pr(A) > 1 - \epsilon$,
2) $\Pr((x^n, y^n)_p \in A) \leq 1 - \epsilon$,
3) $\Pr((x^n, y^n)_p) \in A^I \leq 2^{-n(I(X,Y)+\epsilon)}$,
4) $|A| \leq 2^{n(H(X)+H(Y)+\epsilon)}$,
5) $|A^I| \leq 2^{n(H(X,Y)+\epsilon)}$,
6) $|A| \leq 2^{n(I(X,Y)+\epsilon)}|A^I|$
7) There are at most $2^{n(I(X,Y)+\epsilon)}$ sets of $|A^I|$.

Proof: To save the space we explain briefly the outline of the proofs. For more detail please refer to

[8]. (1) The proof is immediate from the definition of A_I. (2) The proof is similar to the proof of the Asymptotic Equipartition Property (AEP, [5], page 52) and is based on the *weak law of large numbers*. (3) The proof is similar to the second part of Lemma by considering the fact that the the probability of a joint-typical sequence is close to $2^{nH(X,Y)}$ (c.f. [5], page 196). (4) The proof is immediate from the definition of A_I and the fact that typical sets of sequences x^n and y^n have close to $2^{nH(X)}$ and $2^{nH(Y)}$ members, respectively ([5], page 52). (5) The result is immediate from (4). (6) The result is immediate by combining the results of previous parts. □

According to this Lemma, the set of PPoS A is partitioned into $2^{nI(X,Y)}$ sets of correlation-typical sets $A_i^I \;\; i = 1, ..., 2^{nI(X,Y)}$. Members of each of these sets (on average) are paired pair of sequences with the same amount of correlation information no matter which sequences x^n and y^n are paired-up. Therefore, in order to extract the correlation information of the source, there needs to be at least one pair of sequences from each of these sets that when paired-up carries the correlation information. This observation leads to the following multiterminal coding theorem. Before that we introduce the *"distributed version of random-bin coding scheme"* and the corresponding definition of *probability of error* in the multiterminal estimation framework.

A. Distributed Random Bin Scheme, Achievability

Suppose we partition the space of $\mathcal{X}^n$ into 2^{nR_x} and the space of $\mathcal{Y}^n$ into 2^{nR_y} bins.

Random code generation: The source S_x represented by the random variable X assigns every $x^n \in \mathcal{X}^n$ to one of 2^{nR_x} bins according to a uniform distribution on $\{1, 2, ..., 2^{nR_x}\}$. Independently, the source S_y represented by random variable Y randomly assigns the sequences $y^n \in \mathcal{Y}^n$ to the 2^{nR_y} bins $\{1, 2, ..., 2^{nR_y}\}$. We call the sequences x^n and y^n as the *codewords* and the set of bins for x^n and y^n as the *codebooks* M_x and M_y, respectively. The assignment functions f and g are revealed to both the encoder and decoder. Note that the generation of the sequences are according to their marginal distributions rather than the joint distribution.

Encoding and Decoding: Sender S_x sends the index m_{x^n} of the bin to which X belongs to the decoder. Similarly for S_y. Given the received index pair (m_{x^n}, m_{y^n}), declare $(\hat{x}^n, \hat{y}^n) = (x^n, y^n)$ if there is one and only one pair of sequences (x^n, y^n) such that $f(x^n) = m_{x^n}$, $g(y^n) = m_{y^n}$ and $(x^n, y^n) \in A$.

Probability of error P_e. In conventional decoding (e.g. Slepian-Wolf decoding [5]) an error is declared if (x^n, y^n) is not in the joint-typical set (A^{XY}) or if there is another *jointly typical* sequence in the same bin. Instead, since we are interested in *just the correlation information* between the of sequences x^n and y^n we relax the probability of error and declare an error if the PPoS $(x^n, y^n)_p$ is not in A or if there is another PPoS in the same bin. By this we relax the constraint to receive and decode exactly the same sequences at the decoder, and keep our interest in constraining the decoded pair sequences $(\hat{x}^n, \hat{y}^n)$ to maintain their correlational information at the receiver.

Definition-Achievability: The rate R is called *achievable* if there exists at least one pair of encoders (f, g) and a decoder h with probability converging to 1 by which one can construct sequences of codes that provide transmission of the correlational-type information of the sequences (x^n, y^n) to the receiver with probability of error converging to 0 as n becomes sufficiently large. It is important to notice that in the multiterminal distribution coding scheme case, the convergence of P_e to zero has a different meaning than for the conventional coding schemes- given a PoS at the transmitters, decoding a PPoS in a different correlation-typical set converges to zero.

Theorem 2: The following rates are achievable:

$$R_x \geq 0, \tag{5}$$

$$R_y \geq 0, \tag{6}$$

$$R_x + R_y \geq I(X, Y). \tag{7}$$

The achievability proof: The proof is based on the random bin argument similar to the proof of Slepian-Wolf Theorem given in ([5], page 412). The proof is based on the particular definitions of achievability and the correlation-typical set. The main idea is to relax the conditions corresponding to detecting errors in the random bin argument, to the sole case where more than a PPoS are detected in one of the bins. See [8] for more discussion. □

B. Coding Scheme

Transmission of marginal-types The upper bound on the number of types with denominator n (here the *marginal-type set*) is polynomial in n (e.g. $(n+1)^2$ [5]). Therefore a codebook containing all possible types with denominator n can be indexed by not more than $\log(n+1)^2$ bits. This gives a rate $R = \frac{2\log(n+1)}{n}$ which for a sufficiently large n vanishes asymptotically. Note that in this stage the codebooks are not random anymore and consist of all the possible marginal-types rather than the sequences themselves. Therefore, after this stage, and with almost zero rate, the marginal types are accessible to both transmitters and the receiver (estimator).

Transmission of correlation information It is assumed that each one of the sources have access to their corresponding marginal types. *Codebook design:* Two random codebooks M_x and M_y are generated by two sources S_x and S_y using the marginal distributions $\tilde{p}_x$ and $\tilde{p}_y$, respectively. The length of the two codebooks are $|M_x| = 2^{nR_x}$ and $|M_y| = 2^{nR_y}$, respectively, such that $R_X + R_Y \geq I(X, Y)$. *Encoding:* The encoding functions f and g assign to each sequence x^n and y^n their corresponding messages m_{x^n} and m_{y^n}, respectively. The messages are selected as the indices (addresses) of the closest codewords from the corresponding codebooks M_x and M_y, respectively. *Decoding:* Given the received index pair (m_{x^n}, m_{y^n}),

the decoder retrieves the corresponding codewords from the codebooks M_x and M_y at the receiver.

Joint-type Retrieval: The joint-type $\tilde{p}_{11}(x,y)$, of the source can be retrieved by using the correlation information retrieved. This, in addition to the marginal types $\tilde{p}_1(x)$ and $\tilde{p}_1(y)$, provide the necessary sufficient statistics for estimation (cf. *Theorem (1)*). *Correlation computation:* Since "on average" and for sufficiently large n, the ratio of $\frac{p_{11}}{p_{x1}p_{y1}}$ of the correlation-typical PoS is almost equal to those of the source. Noticing that the marginal information is assumed to be conveyed and therefore known at the estimator, the joint-type $\tilde{p}_{11}(x,y)$, of the source can be retrieved by using the correlation information retrieved.

Estimation: The estimation is performed by available sufficient statistics at the receiver, i.e., $\tilde{p}_x$, $\tilde{p}_y$, and $\tilde{p}_{11}(x,y)$.

5. Simulation Results: Estimation with Side Information at the Decoder

For this purpose the estimation of a bivariate binary PD with side information at the receiver is considered. The particular scenario presented here is a corner point of the rate region; i.e. $R_y = I(X,Y)$, $R_x = 0$, which is when complete knowledge of the correlation information transmitted by the source Y is available at the decoder. For this purpose the marginal types $\tilde{p}_x$ and $\tilde{p}_y$ of the measured sequences x^n and y^n are computed at the transmitters X and Y, respectively, and conveyed to the receiver with almost zero rates. Then the random codebook M_y of size $2^{nI(X,Y)}$ is generated at the transmitter S_y according to the marginal type $\tilde{p}_y$. This codebook is transmitted to the receiver before the estimation has begun. The codebook M_x is generated at the receiver according to the marginal type $\tilde{p}_x$. The size of the M_x is chosen equal to the size of the type class of the marginal distribution, i.e. $2^{nH(X)}$. For performing estimation, the index (address) of the closest random codeword to the measured sequence y^n is chosen from M_y and transmitted to the receiver where the corresponding codeword $\hat{y}^n$ is retrieved from the same codebook. The corresponding sequence $\hat{x}^n$ is chosen from M_x such that it is closest to the reconstructed correlated sequence $\hat{y}^n$.

The performance; i.e., variance of error (trace of the error covariance matrix), for 50 Monte Carlo runs of distributed estimation of the bivariate binary symmetric source $p_{00} = p_{11} = 0.4$ and $p_{10} = p_{01} = 0.1$ is compared with the ideal case where the estimation is performed locally at the transmitters with perfect knowledge of the pair of measured sequences (x^n, y^n). The ideal case provides the Cramer-Rao lower bound (CRLB) for ML estimation. For this scenario, $H(X) = H(Y) = 1$ and $I(X,Y) = 0.2781$.

As can be seen from Figure 1, as the length n of sequences become larger, the CRLB as well as the error variance of the proposed distributed estimation scheme decay exponentially. Additionally, the error variance for distributed estimation asymptotically decays towards the CRLB. Simulations with larger n were not feasible, due to the exponentially increasing size of the codebooks. Obviously, in order to be able to achieve the rate limits to have efficient estimation (in the sense of achieving the CRLB), implementation of more sophisticated coding schemes is necessary. This is the subject of future investigation.

6. Conclusions

In this paper, a distributed coding theorem for parameter estimation in multiterminal estimation is presented. It is shown that when communication is performed for the purpose of parameter estimation, the sum-rate of codes for correlated sources can be as low as the mutual information of the correlated sources. It is shown that this amount of information is sufficient to preserve the information conveyed in the *joint-type*, which in turn is a sufficient statistic for parameter estimation. This result is in contrast with the limits of conventional communication systems that aim to reconstruct the sources perfectly, i.e. $H(X,Y)$ (cf. [7]). In order to achieve the theoretical rate limits for obtaining efficient estimators (in the sense of achieving the CRLB), it is required to design and implement more sophisticated distributed source coding schemes.

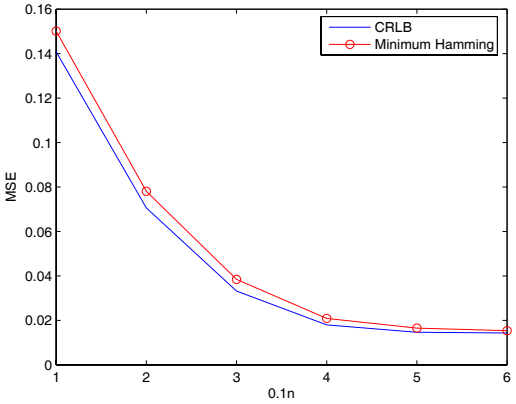

Fig. 1. Estimation Error Variance for Uncompressed and Compressed Sequences (50 Monte Carlo).

References

[1] T. Berger, "Decentralized estimation and decision theory," *presented at the IEEE 7th Spring Workshop on Information Theory*, Mt. Kisco, NY, Sept. 1979.

[2] Z. Zhang, and T. Berger, "Estimation via Compressed Information," *IEEE T-IT, Vol. 34, No. 2, pp. 198-211, March 1988.*

[3] T. S. Han and S. Amari, Parameter estimation with multiterminal data compression, *IEEE Trans. Inform. Theory, vol. 41, pp. 18021833, Nov. 1995.*

[4] T. S. Han and S. Amari, "Statistical Inference Under Multiterminal Data Compression," *IEEE T-IT, Vol. 44, No. 6, pp. 2300-2324, Oct. 1998*

[5] T. M.Cover and J.A. Thomas, *Elements of Information Theory.*

[6] I. Csiszar, "Method of Types," *IEEE T-IT, Vol. 44, No. 6, Oct 1998, pp. 2505-2523.*

[7] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory, vol. 19, pp. 471-480, July 1973.*

[8] A. Zia, Multiterminal estimation, an information theoretic approach, *Technical Report, Department of ECE, McMaster University June 2005.*