On Lossless Universal Compression of Distributed Identical Sources

Ahmad Beirami and Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta GA 30332, USA
Email: {beirami, fekri}@ece.gatech.edu

***Abstract*—Slepian-Wolf theorem is a well-known framework that targets almost lossless compression of (two) data streams with symbol-by-symbol correlation between the outputs of (two) distributed sources. However, this paper considers a different scenario which does not fit in the Slepian-Wolf framework. We consider two identical but spatially separated sources. We wish to study the universal compression of a sequence of length n from one of the sources provided that the decoder has access to (i.e., memorized) a sequence of length m from the other source. Such a scenario occurs, for example, in the universal compression of data from multiple mirrors of the same server. In this setup, the correlation does not arise from symbol-by-symbol dependency of two outputs from the two sources. Instead, the sequences are correlated through the information that they contain about the unknown source parameter. We show that the finite-length nature of the compression problem at hand requires considering a notion of almost lossless source coding, where coding incurs an error probability $p_e(n)$ that vanishes with sequence length n. We obtain a lower bound on the average minimax redundancy of almost lossless codes as a function of the sequence length n and the permissible error probability p_e when the decoder has a memory of length m and the encoders do not communicate. Our results demonstrate that a strict performance loss is incurred when the two encoders do not communicate even when the decoder knows the unknown parameter vector (i.e., $m \to \infty$).**

I. Introduction

Many practical applications involve compression of data that are taken from multiple spatially separated sources. A key challenge in most of such applications is that the sources usually cannot communicate with each other. Theoretical results by Slepian and Wolf demonstrate that if the data streams from two sources have symbol-by-symbol correlation, the sequences can be compressed to their joint entropy even when the two encoders do not communicate [1]. In other words, as in Fig. 1, assume that sources S_1 and S_2 wish to transmit the sequences y^n and x^n, respectively, to a node R. As the length n of the sequences increases, the decoding of x^n at R with the help of y^n can be performed using a code with the average length that asymptotically approaches the conditional entropy, (i.e., $H(X^n|Y^n)$) with asymptotically zero error probability. If the decoder did not choose to use y^n in decoding, the encoder at S_2 would have to encode the sequence x^n irrespective to y^n with an average length that is lower bounded by $H(X^n)$. Note that the conditional entropy $H(X^n|Y^n)$ may be significantly smaller than the individual entropy $H(X^n)$. After recent development of practical Slepian-Wolf (SW) coding schemes by Pradhan and Ramchandran [2], SW coding has drawn a great deal of attention as a promising technique for sensor networks [3] and distributed video coding [4].

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1017234.

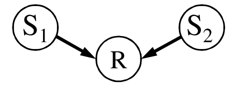

Fig. 1. The basic scenario for the compression of distributed sources.

The Slepian-Wolf theorem naturally suits applications where the (new) sequence x^n from S_2 (in Fig. 1) can be viewed as a noisy version of the (previously seen) sequence y^m that could possibly be exploited as side information to reduce the code length of x^n. Data gathering from sensors that measure the same phenomenon is one example. However, in many scenarios, the compression of distributed sources cannot be modeled by the SW framework. As an example, consider the universal compression of data from the mirrors of the same server, where the sources are exact copies of each other. Hence, it is plausible to assume that the sources (S_1 and S_2 in Fig. 1) follow the same statistical model. On the other hand, the source model might be unknown requiring universal compression [5]–[7]. The question is, assuming two identical sources S_1 and S_2 and having y^m from S_1 at the decoder, what is the achievable universal compression performance on x^n at S_2 provided that the encoders at S_1 and S_2 do not communicate.

We stress that the nature of this problems is fundamentally different from those addressed by the Slepian-Wolf (SW) theorem in [1]. Here, instead of symbol-by-symbol correlation between the sequences as in SW setup, the redundancy is due to the fact that when the source parameter is a priori unknown there is significant overhead in the universal compression of finite-length sequences [7]–[9]. Considering the example in Fig. 1 with two identical sources S_1 and S_2, y^m and x^n would be independent given that the source model is known. However, when the source parameter is unknown, y^m and x^n are *correlated* with each other through the information they contain about the unknown source parameter. The question is whether or not this correlation can be potentially leveraged by the encoder of S_2 and the decoder at R in the decoding of x^n using y^m in order to reduce the code length of x^n.

In this paper, we study the universal compression of distributed identical sources. By identical we mean that the sources (S_1 and S_2) share the same unknown source parameter. By distributed we mean that the sources are spatially separated and the encoders do not communicate with each other. This problem can also be viewed as universal compression with training data that is only available to the decoder. It is known that forming a statistical model from a training data set would improve the performance of universal compression [10], [11]. In [9], [12], we theoretically derived the *gain* that is obtained

978-1-4673-2579-0/12/$31.00 ©2012 IEEE

in the universal compression of the new sequence x^n from S_2 by memorizing (i.e., having access to) y^m from S_1 at *both* the decoder (at R) and the encoder (at S_2). This corresponds to the reduced case of our problem where the sources S_1 and S_2 are either co-located (a single source) or allowed to communicate. For the reduced problem case, in [11], [13], we further extended the setup to a network with a single source and derived bounds on the *network-wide gain* where a small fraction of the intermediate nodes in the network are capable of memorization. However, as we demonstrate in the present paper, the extension to the multiple spatially separated sources, where the training data is only available to the decoder, is non-trivial and raises a new set of challenges that we aim to address. The rest of this paper is organized as follows. In Sec. II, we briefly review the necessary background. In Sec. III, we describe the problem setup. In Sec. IV, we present our main results. In Sec. V, we provide discussion on the results. In Sec. VI, we present the technical analysis of the results. Finally Sec. VII concludes the paper.

II. Background Review

In this section, we review the necessary background, notations, and definitions followed by the formal problem setup. Following the notation in [12], let $\mathcal{A}$ be a finite alphabet. Let d be the number of the source parameters. Further, let $\theta = (\theta_1, ..., \theta_d)$ denote the d-dimensional parameter vector associated with the parametric source (that is a priori unknown). Let Θ^d denote the space of d-dimensional parameter vectors. We denote μ_θ as the probability measure that is defined by the parameter vector θ. Let $\mathcal{P}^d$ denote the *family* of sources that are described with the d-dimensional unknown parameter vector $\theta \in \Theta^d$. We use the notation $x^n = (x_1, ..., x_n) \in \mathcal{A}^n$ to present a sequence of length n from the alphabet $\mathcal{A}$. We further denote X^n as a random sequence of length n (that follows the probability distribution μ_θ). Let $H_n(\theta)$ be the source entropy given θ, i.e., $H_n(\theta) = \mathbf{E} \log \left(\frac{1}{\mu_\theta(X^n)} \right)$.[1]

Let $c_n : \mathcal{A}^n \to \{0,1\}^*$ be an injective mapping from the set $\mathcal{A}^n$ of the sequences of length n over $\mathcal{A}$ to the set $\{0,1\}^*$ of binary sequences. Next, we present the notions of strictly lossless and almost lossless source codes, which will be needed for the study of UC-DIS.

Definition 1 *The code $c_n(\cdot) : \mathcal{A}^n \to \{0,1\}^*$ is called strictly lossless (also called zero-error) if there exists a reverse mapping $d_n(\cdot) : \{0,1\}^* \to \mathcal{A}^n$ such that*

$$\forall x^n \in \mathcal{A}^n : \quad d_n(c_n(x^n)) = x^n.$$

All of the practical data compression schemes are examples of strictly lossless codes, namely, the arithmetic coding, Huffman, Lempel-Ziv, and Context-Tree-Weighting algorithms.

On the other hand, due to the distributed nature of the sources, we are concerned with the slightly weaker notion of almost lossless source coding in this paper.

Definition 2 *The code $\hat{c}_n^{p_e}(\cdot) : \mathcal{A}^n \to \{0,1\}^*$ is called almost lossless with permissible error probability $p_e(n) = o(1)$, if there exists a reverse mapping $\hat{d}_n^{p_e}(\cdot) : \{0,1\}^* \to \mathcal{A}^n$ such that*

$$\mathbf{E}\{\mathbf{1}_e(X^n)\} \leq p_e(n),$$

where $\mathbf{1}_e(x^n)$ denotes the error indicator function, i.e,

$$\mathbf{1}_e(x^n) = \begin{cases} 1 & \hat{d}_n^{p_e}(\hat{c}_n^{p_e}(x^n)) \neq x^n, \\ 0 & \textit{otherwise}. \end{cases}$$

The almost lossless codes allow a non-zero error probability $p_e(n)$ for any n while they are *almost surely* asymptotically error free. Note that strictly lossless codes correspond to $p_e(n) = 0$. The proofs of Shannon [14] for the existence of entropy achieving source codes are based on almost lossless random codes. Further, the proof of the SW theorem [1] also uses almost lossless codes. Further, all of the practical implementations of SW source coding are based on almost lossless codes (cf. [2], [3]). We stress that the nature of the almost lossless source coding is different from that incurred by the lossy source coding (i.e., the rate-distortion theory). In the rate-distortion theory, a code is designed to asymptotically achieve a given distortion level as the length of the sequence grows to infinity. Therefore, since the almost lossless coding asymptotically achieves a zero-distortion, in fact, it coincides with the special case of zero-distortion in the rate-distortion curve.

III. Problem Setup

We present the problem setup in the most basic scenario, shown in Fig. 1, consisting of two identical sources located in nodes S_1 and S_2, and the destination node R. We let the information sources at S_1 and S_2 be parametric with an identical d-dimensional parameter vector that is unknown a priori to the encoder and the decoder. Let y^m and x^n denote two sequences with lengths m and n, respectively, that are generated by the unknown information source model. In the sequel, we describe the communication scenario for universal compression of distributed identical sources. We assume that S_1 has transmitted the sequence y^m to R. Next, at some later time, S_2 wishes to send x^n to R. We further assume that R is a memory unit and is capable of memorizing the sequence y^m. We investigate the achievable saving in the compression of x^n in the S_2-R link when R has memorized the sequence y^m. Note that S_2 does not have access to the sequence y^m. If the node R did not have a memory unit, S_2 would have to apply an end-to-end universal compression to x^n. However, the side information provided by y^m at R about the source parameter can potentially result in a reduction in the amount of bits required to be transmitted in the S_2-R link. Throughout the paper, we refer to this problem setup as Universal Compression of Distributed Identical Sources (UC-DIS).

In the study of coding strategies for UC-DIS, we compare the following cases for the compression of x^n at S_2. Note that we assume that y^m is already universally compressed at S_1 and transmitted and decoded at R.

- UComp (Universal compression), which only applies end-to-end lossless universal compression to x^n at S_2 without regard to y^m.
- UCompM (Universal compression with memorization at both the encoder and the decoder), which assumes that

[1]Throughout this paper, all expectations are taken with respect to the probability measure μ_θ, and $\log(\cdot)$ denotes the logarithm in base 2.

the encoder (at S_2) and the decoder (at R) have access to a common memory (i.e., sequence y^m), which is utilized in the lossless compression of x^n at S_2.

- DUCompM (Distributed universal compression with memorization at the decoder), which assumes that decoder (at R) has memorized (i.e., has access to) y^m while the encoder (at S_2) only knows the length m of the side information but does not know the exact sequence y^m. The encoder then applies an almost lossless code to x^n that is decoded at R with permissible error probability p_e using y^m.

Note that UComp does not benefit from the memorization and is the conventional scheme. Further, UCompM is introduced as the benchmark for the purpose of evaluating the performance of DUCompM and is not practically useful since it requires the sequence y^m from S_1 to be available at the encoder of S_2.

Let $l_n(x^n)$ denote the strictly lossless length of the codeword associated with the sequence x^n. Further, let L_n denote the space of strictly lossless universal length functions on a sequence of length n. Denote $R_n(l_n, \theta)$ as the expected redundancy of such strictly lossless codes on a sequence of length n for the parameter vector θ, i.e., $R_n(l_n, \theta) = \mathbf{E}l_n(X^n) - H_n(\theta)$. Further, denote $\bar{R}_{\text{UComp}}(n)$ as the average minimax redundancy as given by

$$\bar{R}_{\text{UComp}}(n) \triangleq \min_{l_n \in L_n} \sup_{\theta \in \Theta^d} R_n(l_n, \theta). \quad (1)$$

In UCompM, let $l_{n|m}$ be the strictly lossless universal length function with a memory sequence of length m. Denote $L_{n|m}$ as the space of such strictly lossless universal length functions. Let $R_n(l_{n|m}, \theta)$ be the expected redundancy of encoding a sequence of length n form the source μ_θ using the length function $l_{n|m}$. Further, let $\bar{R}_{\text{UCompM}}(n, m)$ denote the corresponding average minimax redundancy, i.e.,

$$\bar{R}_{\text{UCompM}}(n, m) \triangleq \min_{l_{n|m} \in L_{n|m}} \sup_{\theta \in \Theta^d} R_n(l_{n|m}, \theta). \quad (2)$$

In DUCompM, let $\hat{l}^{p_e}_{n|m}$ denote the almost lossless universal length function with a memorized sequence of length m that is only available to the decoder, where the permissible error probability on decoding x^n is p_e. Further, denote $\hat{L}^{p_e}_{n|m}$ as the space of such universal length functions. Denote $R_n(\hat{l}^{p_e}_{n|m}, \theta)$ as the expected redundancy of encoding a sequence x^n of length n using the length function $\hat{l}^{p_e}_{n|m}$. Denote $\bar{R}^{p_e}_{\text{DUCompM}}(n, m)$ as the expected minimax redundancy as given by

$$\bar{R}^{p_e}_{\text{DUCompM}}(n, m) \triangleq \min_{\hat{l}^{p_e}_{n|m} \in \hat{L}^{p_e}_{n|m}} \sup_{\theta \in \Theta^d} R_n(\hat{l}^{p_e}_{n|m}, \theta). \quad (3)$$

Note that we denote $\bar{R}_{\text{DUCompM}}(n, m) \triangleq \bar{R}^0_{\text{DUCompM}}(n, m)$ as the expected minimax redundancy of *strictly lossless* DUCompM coding strategy.

IV. Performance Evaluation of UC-DIS: Results on the Average Minimax Redundancy

In this section, we provide results on the average minimax redundancy of the different coding strategies introduced in the previous section for the UC-DIS problem. Discussion on the implications of the results and the proof sketches are deferred to Sec. V and Sec. VI, respectively.

In the case of strictly lossless UComp, Clarke and Barron derived the expected minimax redundancy $\bar{R}_{\text{UComp}}(n)$ for memoryless sources [15], which was later generalized by Atteson for Markov sources, as the following [16]:

Theorem 1 *The average minimax redundancy of strictly lossless UComp coding strategy is given by*

$$\bar{R}_{\textit{UComp}}(n) = \frac{d}{2}\log\left(\frac{n}{2\pi e}\right) + \log\int |\mathcal{I}_n(\theta)|^{\frac{1}{2}} d\theta + O\left(\frac{1}{n}\right),$$

where $\mathcal{I}_n(\theta)$ is the Fisher information matrix.

In the case of strictly lossless UCompM (i.e., when the two encoders can communicate), we obtain the average minimax redundancy in the following theorem.

Theorem 2 *The average minimax redundancy of strictly lossless UCompM coding strategy is given by*

$$\bar{R}_{\textit{UCompM}}(n, m) = \frac{d}{2}\log\left(1 + \frac{n}{m}\right) + O\left(\frac{1}{m} + \frac{1}{n}\right).$$

In the next proposition, we confine ourselves to strictly lossless codes in the DUCompM strategy.

Proposition 3 *The average minimax redundancy of strictly lossless DUCompM coding strategy is equal to that of UComp coding strategy. That is $\bar{R}_{\textit{DUCompM}}(n, m) = \bar{R}_{\textit{UComp}}(n)$.*

Finally, in the case of almost lossless DUCompM, our main result is given in the following theorem.

Theorem 4 *The average minimax redundancy of almost lossless DUCompM coding strategy is upper bounded by*

$$\bar{R}^{p_e}_{\textit{DUCompM}}(n, m) \le \bar{R}_{\textit{UCompM}}(n, m) + \mathcal{F}(d, p_e) + O\left(\frac{1}{m} + \frac{1}{n}\right),$$

where $\mathcal{F}(d, p_e)$ is the penalty term due to the encoders not communicating, which is given by

$$\mathcal{F}(d, p_e) = \frac{d}{2}\log\left(1 + \frac{2}{d\log e}\log\frac{1}{p_e}\right). \quad (4)$$

V. Discussion on the Results

In this section, we provide some discussion on the significance of the results for different UC-DIS coding strategies. Figures 2 and 3 demonstrate the redundancy rate for the three coding strategies, namely, UComp, UCompM, and DUCompM for memoryless sources and first-order Markov sources with alphabet size $k = 256$, respectively. In the case of UComp, Theorem 1 defines the achievable average minimax redundancy for the compression of a sequence of length n encoded without regard to the previously seen sequence y^m.

According to Theorem 2, if the encoder and the decoder have access to a common memory y^m, i.e., UCompM coding strategy, the average minimax redundancy could be much smaller than that of UComp depending on how large m is. In particular, when $m \to \infty$ we have $\lim_{m\to\infty} \bar{R}_{\text{UCompM}}(n, m) = 0$.[2] This corresponds to the case where the parameter vector

[2] In this paper, we ignored the integer constraint on the length functions, which results in a negligible $O(1)$ redundancy analyzed in [17], [18].

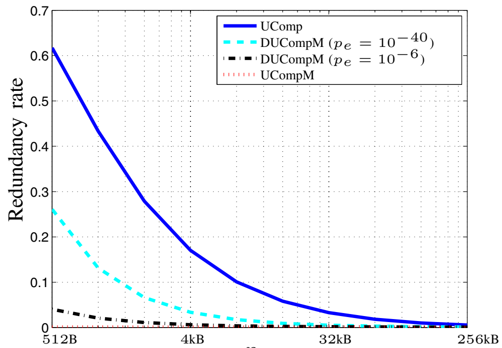

Fig. 2. The redundancy rate for the three coding strategies of interest for the UC-DIS problem. Memory size is $m = 32$kB and the source is memoryless with alphabet size $k = 256$.

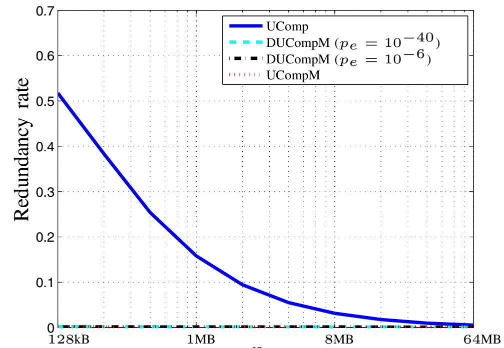

Fig. 3. The redundancy rate for the three coding strategies of interest for the UC-DIS problem. Memory size is $m = 16$MB and the source is first-order Markov with alphabet size $k = 256$.

is known to both the encoder and the decoder, and thus, the redundancy is zero similar to a perfect Shannon code. Hence, the fundamental limits are those of known source parameters and universality no longer imposes a compression overhead. This is also demonstrated in Figs. 2 and 3, where m has been chosen to be sufficiently large.

Proposition 3 demonstrates that if strictly lossless DUCompM coding strategy (i.e., $p_e = 0$) is to be used for the compression of x^n from S_2, the memorization of y^m from S_1 only at the decoder does not provide any compression benefit, assuming that the two encoders at S_1 and S_2 do not communicate. In other words, the best that S_2 can do is to simply apply a traditional universal compression on x^n.

Finally, according to Theorem 4, unlike the asymptotic behavior of the Slepian-Wolf problem, the distributed nature in this problem incurs an extra redundancy on the compression. As can be seen in Fig. 2, the overhead can be significant in the compression of memoryless sources. For example, when $n = 512$B, $m = 32$kB, and $p_e = 10^{-6}$, the redundancy rate is around 0.05, as compared with the almost zero redundancy rate of UCompM. On the other hand, as demonstrated in Fig. 3, when d is relatively larger, for medium length sequences even with extremely small error probability, DUCompM performs fairly close to UCompM. Further, DUCompM by far outperforms UComp in the compression of short to medium length sequences with reasonable permissible error probability, justifying usefulness of DUCompM in practice. If $\log \frac{1}{p_e} \ll d$, the penalty term can be simplified to be approximately equal to $\mathcal{F}(d, p_e) \approx \log \frac{1}{p_e}$ for the practical ranges of p_e.

VI. Technical Analysis

A. Sketch of the Proof of Theorem 2

We prove that the RHS is both an upper bound and a lower bound for $\bar{R}_{\text{UCompM}}(n, m)$. The upper bound is obtained using the KT-estimator [19] along with a proper Shannon code [14] and the proof follows the analysis of the redundancy of the KT-estimator. In the next lemma, we obtain the lower bound.

Lemma 1 *The average minimax redundancy of UCompM is lower-bounded by*

$$\bar{R}_{UCompM}(n, m) \geq \frac{d}{2} \log\left(1 + \frac{n}{m}\right) + O\left(\frac{1}{m} + \frac{1}{n}\right).$$

Proof: It can be shown that the minimax redundancy is equal to the capacity of the channel between the unknown parameter vector θ and the sequence x^n given the sequence y^m (cf. [8] and the references therein). Thus,

$$\begin{aligned} \bar{R}_{\text{UCompM}}(n, m) &= \sup_{\omega(\theta)} I(X^n; \theta | Y^m) \\ &= \sup_{\omega(\theta)} \{I(X^n, Y^m; \theta) - I(Y^m; \theta)\} \\ &\geq \{I(X^n, Y^m; \theta) - I(Y^m; \theta)\}|_{\theta \propto \omega_{\text{J}}(\theta)} \\ &= \bar{R}_{\text{UComp}}(n + m) - \bar{R}_{\text{UComp}}(m), \end{aligned} \tag{5}$$

where $\omega_{\text{J}}(\theta) \triangleq \frac{|\mathcal{I}(\theta)|^{\frac{1}{2}}}{\int |\mathcal{I}(\beta)|^{\frac{1}{2}} d\beta}$ denotes the Jeffreys' prior, and $\bar{R}_{\text{UComp}}(\cdot)$ is given in Theorem 1. Further simplification of (5) leads to the desired result in Lemma 1. ■

B. Sketch of the Proof of Proposition 3

Since the source is assumed to be from the family $\mathcal{P}^d$ of d-dimensional parametric sources, in particular, it is also an *ergodic* source. Thus, any pair (x^n, y^m) occurs with non-zero probability and the support set of (x^n, y^m) is equal to $\mathcal{A}^n \times \mathcal{A}^m$. Therefore, Proposition 3 trivially follows from the known results on strictly lossless compression (cf. [20] and the references therein).

C. Sketch of the Proof of Theorem 4

We provide a constructive optimal coding strategy at the encoder and obtain its achievable average minimax redundancy, which provides with an upper bound on the average minimax redundancy of the almost lossless DUCompM coding strategy.

Let $\hat{\theta}(x^n)$ (or $\hat{\theta}(y^m)$) denote the Maximum Likelihood (ML) estimate for the unknown source parameter given that the sequence x^n (or y^m) is observed, i.e., $\hat{\theta}(x^n) \triangleq \arg\max_\lambda \mu_\lambda(x^n)$. Further, let $\hat{\theta}_X \triangleq \hat{\theta}(x^n)$ and $\hat{\theta}_Y \triangleq \hat{\theta}(y^m)$. As discussed earlier $\mu_\theta(x^n)$ is the probability distribution induced by the parameter vector θ on the sequence x^n. It is straightforward to derive the pmf of the ML-estimate $p(\hat{\theta}_X|\theta)$ from $\mu_\theta(x^n)$ by summing over all the sequences that correspond to the same ML-estimate. Note that $\hat{\theta}_X$ follows a discrete distribution only taking values on a finite set of $(n+1)^d$ points in the space Θ^d. For any $\lambda, \theta \in \Theta^d$, let $D_n(\mu_\lambda || \mu_\theta)$ be the KL-divergence, i.e., $D_n(\mu_\lambda || \mu_\theta) \triangleq \mathbf{E} \log\left(\frac{\mu_\theta(X^n)}{\mu_\lambda(X^n)}\right)$.

It can be shown that expectations with respect to $p(\hat{\theta}_X|\theta)$ can be performed using a continuous RV $\tilde{\theta}_X$ (with uniformly vanishing error) whose distribution conditioned on θ is given by

$$p(\tilde{\theta}_X|\theta) = |\mathcal{I}_n(\tilde{\theta}_X)|^{\frac{1}{2}} \left(\frac{n}{2\pi}\right)^{\frac{d}{2}} \exp(-D_n(\mu_{\tilde{\theta}_X}||\mu_\theta)), \quad (6)$$

where n has to be large enough so that Stirling's approximation can be applied. Further, it is straightforward to show that this distribution can be approximated using a Gaussian distribution with mean θ and inverse covariance matrix $n\mathcal{I}_n(\theta)$.

Next, we will obtain an approximation for the distribution of $\hat{\theta}_X$ conditioned on $\hat{\theta}_Y$.

Lemma 2 *Let $\hat{\theta}_X$ and $\hat{\theta}_Y$ denote the ML-estimate parameter given observed sequences x^n and y^m, respectively. Further, let $p(\tilde{\theta}_X|\hat{\theta}_Y)$ follow a Gaussian distribution with mean $\hat{\theta}_Y$ and inverse covariance matrix $\frac{nm}{n+m}\mathcal{I}_m(\hat{\theta}_Y)$. Then, all expectations with respect to $p(\hat{\theta}_X|\hat{\theta}_Y)$ can be performed using $p(\tilde{\theta}_X|\hat{\theta}_Y)$ with uniformly vanishing error.*

Now, we are equipped to define $S_n(y^m, p_e)$ as the set with smallest Lebesgue volume such that

$$\int_{\tilde{\theta}_X \in \mathcal{S}_n(y^m, p_e)} p(\tilde{\theta}_X|\hat{\theta}_Y) d\tilde{\theta}_X \geq 1 - p_e. \quad (7)$$

The following lemma shows as to how $S_n(y^m, p_e)$ is determined.

Lemma 3 *Let $\hat{\theta}_Y$ denote the ML-estimate for the unknown parameter vector given sequence y^m is observed. Then, $\mathcal{S}_n(y^m, \epsilon)$ is given by*

$$\mathcal{S}_n(y^m, p_e) = \left\{\phi : r(\phi - \hat{\theta}_Y)'\mathcal{I}_m(\hat{\theta}_Y)(\phi - \hat{\theta}_Y) \leq \delta_d(p_e)\right\},$$

where $r = \frac{nm}{n+m}$, $\delta_d(p_e)$ satisfies $\Gamma\left(\frac{d}{2}, \delta_d(p_e)\right) = p_e\Gamma\left(\frac{d}{2}\right)$.[3]

The next lemma determines the probability measure of the set $\mathcal{S}_n(y^m, p_e)$ under Jeffreys' prior.

Lemma 4 *Assume that the parameter vector θ follows Jeffreys' prior. Then, the probability measure $P_\mathcal{S}(p_e)$ of the set $\mathcal{S}_n(y^m, p_e)$ is given by*

$$P_\mathcal{S}(p_e) = \int_{\theta \in S_n(y^m, p_e)} \omega_J(\theta) d\theta = \frac{C_d}{\int |\mathcal{I}(\beta)|^{\frac{1}{2}} d\beta} \left(\frac{2\delta_d(p_e)}{r \log e}\right)^{\frac{d}{2}},$$

where $r = \frac{nm}{n+m}$ and $C_d = \frac{\Gamma\left(\frac{1}{2}\right)^d}{\Gamma\left(\frac{d}{2}+1\right)}$.

Next, consider the following coding scheme. Let the space be partitioned into ellipsoids of the form $S_n(y^m, p_e)$. Then, each sequence is encoded within its respective ellipsoid without regard to the rest of the parameter space. The decoder chooses the decoding ellipsoid using the ML estimate $\hat{\theta}_Y$ and the permissible decoding error probability p_e . The probability measure covered by each ellipsoid is $P_\mathcal{S}(p_e)$ is independent of $\hat{\theta}_Y$, and provides with $-\log P_\mathcal{S}(p_e)$ reduction in the redundancy. Further, simplification of $P_\mathcal{S}(p_e)$ and the fact that $\delta_d(p_e) \approx \frac{d}{2}\log e + \log\frac{1}{p_e}$ will lead to the desired result.

[3] $\Gamma(s, x) \triangleq \int_0^x t^{s-1}e^{-t}dt$ denotes the incomplete Gamma function.

VII. Conclusion

In this paper, we introduced and studied the problem of Universal Compression of Distributed Identical Sources (UC-DIS), which is a more favorable framework as compared to the Slepian-Wolf (SW) framework in several applications, such as the compression of data from mirrors of a data server. In UC-DIS, the correlation among outputs of the sources is due to the finite-length universal compression constraint, departing from the nature of the correlation in the SW framework. For UC-DIS, involving two identical sources, we introduced DUCompM coding strategy (compression using the side information at the decoder when the two encoders do not communicate) and obtained an upper bound on its average minimax redundancy. We demonstrated that for finite-length sequences with reasonable permissible error probability, DUCompM coding strategy by far outperforms traditional universal compression, and hence, justifying the usefulness of DUCompM coding strategy in practice.

References

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[2] S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Trans. Info. Theory*, vol. 49, no. 3, pp. 626 – 643, Mar 2003.

[3] M. Sartipi and F. Fekri, "Distributed source coding using short to moderate length rate-compatible LDPC codes: the entire Slepian-Wolf rate region," *IEEE Trans. Commun.*, vol. 56, no. 3, pp. 400–411, 2008.

[4] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, 2005.

[5] M. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Info. Theory*, vol. 41, no. 3, pp. 643 –652, 1995.

[6] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Info. Theory*, vol. 30, no. 4, pp. 629 – 636, Jul 1984.

[7] A. Beirami and F. Fekri, "Results on the redundancy of universal compression for finite-length sequences," in *2011 IEEE International Symp. on Info. Theory (ISIT '2011)*, July 2011, pp. 1604–1608.

[8] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Info. Theory*, vol. 41, no. 3, pp. 714 –722, May 1995.

[9] A. Beirami and F. Fekri, "Memory-assisted universal source coding," in *2012 Data Compression Conference (DCC '2012)*, April 2012, p. 392.

[10] G. Korodi, J. Rissanen, and I. Tabus, "Lossless data compression using optimal tree machines," in *2005 Data Compression Conference (DCC '2005)*, March 2005, pp. 348 – 357.

[11] M. Sardari, A. Beirami, and F. Fekri, "Memory-assisted universal compression of network flows," in *IEEE INFOCOM 2012*, March 2012, pp. 91–99.

[12] A. Beirami, M. Sardari, and F. Fekri, "Results on the fundamental gain of memory-assisted universal source coding," in *2012 IEEE International Symposium on Information Theory (ISIT '2012)*, July 2012.

[13] M. Sardari, A. Beirami, and F. Fekri, "On the network-wide gain of memory-assisted source coding," in *2011 IEEE Information Theory Workshop (ITW' 2011)*, October 2011, pp. 476–480.

[14] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Jul, Oct 1948.

[15] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Info. Theory*, vol. 36, no. 3, pp. 453 –471, 1990.

[16] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Info. Theory*, vol. 45, no. 6, pp. 2104 –2109, 1999.

[17] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Info. Theory*, vol. 50, no. 11, pp. 2686–2707, 2004.

[18] W. Szpankowski, "Asymptotic average redundancy of Huffman (and other) block codes ," *IEEE Trans. Info. Theory*, vol. 46, no. 7, pp. 2434–2443, 2000.

[19] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Info. Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[20] N. Alon and A. Orlitsky, "Source coding and graph entropies," *IEEE Trans. Info. Theory*, vol. 42, no. 5, pp. 1329 –1339, September 1996.