Fifty-second Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 1 - 3, 2014

A Novel Correlation Model for Universal Compression of Parametric Sources

Ahmad Beirami[1]
Department of Electrical and Computer Engineering
Duke University, Durham, NC USA
Email: ahmad.beirami@duke.edu

Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA USA
Email: fekri@ece.gatech.edu

***Abstract*—In this paper, we consider k parametric sources with unknown source parameter vectors. In this setup, we propose a novel correlation model where the degree of correlation of each parameter vector is governed by a single variable. We derive the properties of the parameter vectors. In particular, we derive bounds on the correlation between the parameter vectors and show show that this will include independence all the way to convergence in mean square sense. Then, we set up the minimax and maximin games in universal compression and characterize the compression risk under the proposed correlation model when side information from one other source is available at both the encoder and the decoder.**

***Index Terms*—Universal Compression; Distributed Source Coding; Parametric Sources; Side Information.**

I. Introduction

The premise of data compression broadly relies on the correlation in the data. For instance, data that are gathered from multiple sensors measuring the same phenomenon (e.g., temperature) are clearly correlated. As another example, when chunks of the same file/content are acquired by a client in a content-centric network, the data chunks are correlated because they are originated from the same file/content. There are several works in the literature that consider the universal compression problem [1]–[11] and more recently in the context of of one-to-one codes without prefix constraint [12]–[16].

In this paper, we study *universal compression of parametric sources with correlated parameter vectors*. We assume k parametric sources with unknown parameter vectors $\theta^{(1)}, \ldots, \theta^{(k)}$, respectively. Each of the parameter vectors is assumed to live in a d-dimensional space Λ such that $\Lambda \subset \mathbb{R}^d$. We further assume that $\theta^{(1)}, \ldots, \theta^{(k)}$ are correlated according to the correlation model in Fig. 1. In this setup, we assume that nature pick ϕ according to some prior q. Then, let $Z^{t_1}, \ldots, Z^{t_k}$ be independent samples of length $t_1, \ldots, t_k$ from a parametric source with parameter vector ϕ, respectively. Finally, $\theta^{(1)}, \ldots, \theta^{(k)}$ are samples from a posteriori distribution of ϕ from observation $Z^{t_1}, \ldots, Z^{t_k}$, respectively. We establish some key properties about this correlation model as a function of $t_1, \ldots, t_k$

Next, we consider an application of such correlation model where X^n is a sample of length n from the parametric source with parameter $\theta^{(1)}$ and Y^m is a sample of length m from the parametric source with source parameter vector $\theta^{(2)}$. We would like to characterize the average minimax and maximin redundancy in the compression of Y^n when the side information sequence X^n is available to the encoder and/or the decoder. The problem in which the side information is only available at the decoder coincides with Wyner-Ziv problem [17]. The special case of such problem where Y^m and X^n were samples from the same parametric source was studied in [18]. This corresponds to the reduced case of our problem where the correlation between the source parameter vectors is in the form of exact equality, i.e., $\theta^{(2)} = \theta^{(1)}$. In the present paper, the extension to the spatially separated sources with correlated parameter vectors.

The rest of this paper is organized as follows. In Section II, we provide the notations and definitions. In Section III, we present our correlation model for sources with correlated parameter vectors. In Section IV, we present the formal problem setup. In Section V, we provide the coding strategies using two correlated parameter vectors for the universal compression of sources with correlated parameter vectors. Sections VI gives the main results on the average redundancy of strictly lossless codes. Finally, Section VII concludes the paper.

II. Notations and Definitions

Let $\mathcal{X}$ be a finite alphabet with alphabet size $|\mathcal{X}|$. We define a parametric source by using a d-dimensional parameter vector $\theta = (\theta_1, ..., \theta_d) \in \Lambda$, where d denotes the number of the source parameters and $\Lambda \subset \mathbb{R}^d$ is the space of d-dimensional parameter vectors of interest. Denote μ_θ as the probability measure defined by a parameter vector θ on sequences of length n from the source.

Let $x^n = (x_1, ..., x_n) \in \mathcal{X}^n$ be a sequence of length n from the alphabet $\mathcal{X}$. Further, let X^n be a random sequence of length n that follows probability distribution function μ_θ. Let $H_n(\theta)$ be the source entropy given the source parameter

[1] This research was carried out when A. Beirami was affiliated with Georgia Institute of Technology.

978-1-4799-8009-3/14/$31.00 ©2014 IEEE

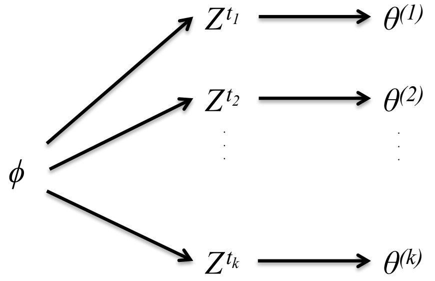

Fig. 1: The correlation model for source coding with correlated parameter vectors.

vector θ, i.e.,

$$H_n(\theta) \triangleq \mathbf{E}\log\left(\frac{1}{\mu_\theta(X^n)}\right) = \sum_{x^n} \mu_\theta(x^n)\log\left(\frac{1}{\mu_\theta(x^n)}\right).^1 \tag{1}$$

Note that $\log(\cdot)$ denotes the logarithm in base 2 in this paper.

Let $\mathcal{I}(\theta)$ be the Fisher information matrix,[2] i.e.,

$$\mathcal{I}(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n\log e}\mathbf{E}\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\left(\frac{1}{\mu_\theta(X^n)}\right)\right\}. \tag{2}$$

Roughly speaking, Fisher information quantifies the average amount of information that each symbol in a sample sequence x^n from the source conveys about the source parameter vector θ. Let the least favorable Jeffreys' prior on the space of the parameter vectors be denoted by

$$p_J(\theta) \triangleq \frac{|\mathcal{I}(\theta)|^{\frac{1}{2}}}{\int_{\lambda\in\Lambda} |\mathcal{I}(\lambda)|^{\frac{1}{2}} d\lambda}. \tag{3}$$

Jeffreys' prior is optimal in the sense that the minimum average redundancy is asymptotically achieved by an optimal code when the source parameter vector is assumed to follow Jeffreys' prior [4]. This prior distribution is particularly interesting because it also corresponds to the least favorable prior for the compression performance of the best coding scheme, i.e., it is the capacity achieving distribution.

We need some regularity conditions to hold for the parametric model so that our results can be derived.

P1. The parametric model is smooth, i.e., twice differentiable with respect to θ in the interior of Λ so that the Fisher information matrix can be defined. Further, the limit in (2) exists.

P2. The determinant of fisher information matrix is finite for all θ in the interior of Λ and the normalization constant in the denominator of (3) is finite.

P3. The parametric model has a minimal d-dimensional representation, i.e., $\mathcal{I}(\theta)$ is full-rank. Hence, $\mathcal{I}^{-1}(\theta)$ exists.

P4. We require that the central limit theorem holds for the maximum likelihood estimator $\hat{\theta}(x^n)$ of each θ in the interior of Λ so that $(\hat{\theta}(X^n) - \theta)\sqrt{n}$ converges to a normal distribution with zero mean and covariance matrix $\mathcal{I}^{-1}(\theta)$.

III. Correlation Model

In this section, we present our model for the nature of the correlation between the parameter vectors $\theta^{(1)}, \ldots, \theta^{(k)}$. In this model, as we shall see, the correlation between the sources is controlled using the parameter $t_1, \ldots, t_k$. We assume that the unknown (and unobserved) parameter vector ϕ follows a prior distribution q, i.e., $p_\phi(\phi) = q(\phi)$. Let Z^{t_i} be a random sequence of length t_i that follows μ_ϕ. We further assume that given Z^{t_i}, the parameter vectors ϕ and $\theta^{(i)}$ are independent and identically distributed, i.e., $p_{\theta^{(i)}|z^{t_i}}(\cdot) = p_{\phi|z^{t_i}}(\cdot)$.

Lemma 1: The conditional distribution of $\theta^{(i)}$ given ϕ, i.e., $p^{t_i}_{\theta^{(i)}|\phi}(\cdot)$, is given by

$$p^{t_i}_{\theta^{(i)}|\phi}(\theta^{(i)}|\phi) = q(\theta^{(i)})f(t_i, \phi, \theta^{(i)}), \tag{4}$$

where $f(t_i, \phi, \theta^{(i)})$ is defined as

$$f(t_i, \phi, \theta^{(i)}) \triangleq \sum_{z^{t_i}\in\mathcal{X}^{t_i}} \left(\frac{\mu_\phi(z^{t_i})\mu_{\theta^{(i)}}(z^{t_i})}{\int_\Lambda \mu_\lambda(z^{t_i})q(\lambda)d\lambda}\right). \tag{5}$$

Proof:

$$p^{t_i}_{\theta^{(i)}|\phi}(\theta^{(i)}|\phi) = \sum_{z^{t_i}\in\mathcal{X}^{t_i}} p(\theta^{(i)}, z^{t_i}|\phi) \tag{6}$$

$$= \sum_{z^{t_i}\in\mathcal{X}^{t_i}} p_{\phi|z^{t_i}}(\theta^{(i)}|z^{t_i})\mu_\phi(z^{t_i}) \tag{7}$$

$$= q(\theta^{(i)}) \sum_{z^{t_i}\in\mathcal{X}^{t_i}} \left(\frac{\mu_\phi(z^{t_i})\mu_{\theta^{(i)}}(z^{t_i})}{\int_\Lambda \mu_\lambda(z^{t_i})q(\lambda)d\lambda}\right), \tag{8}$$

[1] Throughout this paper expectations are taken over functions of the random sequence X^n with respect to the (unknown) probability measure $\mu_{\theta^{(2)}}$ unless otherwise stated.

[2] We assume that the parametric sources of interest are smooth so that the Fisher information matrix exists, and is positive definite, and all its elements are continuous.

where (22) follows from the fact that $\theta^{(i)}$ and ϕ are independent and identically distributed given Z^{t_i}, and (8) follows from the Bayes rule. ■

Lemma 2:

$$\int_\Lambda f(t_i,\phi,\theta^{(i)})q(\phi)d\phi = 1. \tag{9}$$

Proof:

$$\int_\Lambda f(t_i,\phi,\theta^{(i)})q(\phi)d\phi$$

$$=\sum_{z^{t_i}}\left(\frac{\mu_{\theta^{(i)}}(z^{t_i})\int_\Lambda \mu_\phi(z^{t_i})q(\phi)d\phi}{\int_\Lambda \mu_\lambda(z^{t_i})q(\lambda)d\lambda}\right)$$

$$=\sum_{z^{t_i}}\mu_{\theta^{(i)}}(z^{t_i}) \tag{10}$$

$$=1, \tag{11}$$

where (10) is obtained since the two integrals in the numerator and denominator cancel. ■

Next, we find the marginal distribution of $\theta^{(i)}$, i.e., $p^{t_i}_{\theta^{(i)}}(\theta^{(i)})$.

Lemma 3: $p^{t_i}_{\theta^{(i)}}(\theta^{(i)}) = q(\theta^{(i)})$.

Proof:

$$p^{t_i}_{\theta^{(i)}}(\theta^{(i)}) = \int_\Lambda p^{t_i}_{\theta^{(i)}|\phi}(\theta^{(i)}|\phi)q(\phi)d\phi$$

$$= q(\theta^{(i)})\int_\Lambda f(t_i,\phi,\theta^{(i)})q(\phi)d\phi \tag{12}$$

$$= q(\theta^{(i)}), \tag{13}$$

where (12) follows from Lemma 2. ■

Lemma 4: If $t_i = 0$, then $\theta^{(i)}$ is independent of ϕ.

Proof: By definition of $f(\cdot)$, and the fact that $\mu_\lambda(z^0) = 1$, we have

$$f(0,\phi,\theta^{(i)}) = \left(\frac{1}{\int_\Lambda q(\lambda)d\lambda}\right) = 1. \tag{14}$$

Hence, using Lemma 1, we have

$$p^{0}_{\theta^{(i)}|\phi}(\theta^{(i)}|\phi) = q(\theta^{(i)})f(0,\phi,\theta^{(i)}) = q(\theta^{(i)}), \tag{15}$$

which proves the claim. ■

Lemma 5: $\theta^{(i)}$ converges in mean square to ϕ as $t_i \to \infty$, that is

$$\lim_{t_i\to\infty}\mathbf{E}||\theta^{(i)} - \phi||^2 = 0. \tag{16}$$

Proof: Let $\hat{\theta}(Z^{t_i})$ be the maximum likelihood estimate of ϕ from the observation Z^{t_i}. By definition, $\hat{\theta}(Z^{t_i}$ also serves as the maximum likelihood estimate of $\theta^{(i)}$. Then,

$$\mathbf{E}||\theta^{(i)} - \phi||^2 \leq \mathbf{E}||\theta^{(i)} - \hat{\theta}(Z^{t_i})||^2 + \mathbf{E}||\phi - \hat{\theta}(Z^{t_i})||^2 \tag{17}$$

$$= 2\mathbf{E}||\phi - \hat{\theta}(Z^{t_i})||^2 \tag{18}$$

$$= \frac{2}{t_i}|\mathcal{I}(\phi)|^{-1}, \tag{19}$$

and hence,

$$\lim_{t_i\to\infty}\mathbf{E}||\theta^{(i)} - \phi||^2 \leq \lim_{t_i\to\infty}\frac{|\mathcal{I}(\phi)|^{-1}}{t_i} = 0, \tag{20}$$

which completes the proof. ■

Next, we will derive the joint probability distribution of $\theta^{(i)}$ and $\theta^{(j)}$ for $i \neq j$.

Lemma 6: For all $i,j \in \{1,\ldots,k\}$ such that $i \neq j$, we have $p^{t_i,t_j}_{\theta^{(i)},\theta^{(j)}}(\theta^{(i)},\theta^{(j)})$ is given by

$$p^{t_i,t_j}_{\theta^{(i)},\theta^{(j)}}(\theta^{(i)},\theta^{(j)}) = q(\theta^{(i)})q(\theta^{(j)}) \times \int_\Lambda f(t_i,\phi,\theta^{(i)})f(t_j,\phi,\theta^{(j)})q(\phi)d\phi. \tag{21}$$

Proof:

$$p^{t_i,t_j}_{\theta^{(i)},\theta^{(j)}}(\theta^{(i)},\theta^{(j)}) = \int_\Lambda p^{t_i,t_j}(\theta^{(i)},\theta^{(j)}|\phi)q(\phi)d\phi$$

$$= \int_\Lambda p^{t_i}(\theta^{(i)}|\phi)p^{t_j}(\theta^{(j)}|\phi)q(\phi)d\phi \tag{22}$$

$$= q(\theta^{(i)})q(\theta^{(j)}) \times \int_\Lambda f(t_i,\phi,\theta^{(i)})f(t_j,\phi,\theta^{(j)})q(\phi)d\phi, \tag{23}$$

where (22) follows from the independence of $\theta^{(i)}$ and $\theta^{(j)}$ given ϕ, and (23) follows from Lemma 1 and the definition of $g(\cdot)$. ■

Please note that the strategy above can be extended to derive the joint distribution of all of the source parameter vectors.

Lemma 7:

$$p^{t_1,t_2,\ldots,t_k}_{\theta^{(1)},\theta^{(2)},\ldots\theta^{(k)}}(\theta^{(1)},\theta^{(2)},\ldots\theta^{(k)}) =$$

$$\prod_{i=1}^{k}\left(q(\theta^{(i)})\right)\int_\Lambda\prod_{i=1}^{k}\left(f(t_i,\phi,\theta^{(i)})\right)q(\phi)d\phi. \tag{24}$$

Lemma 8: If $t_i = 0$, then $\theta^{(i)}$ is independent of all $\theta^{(j)}$ ($i \neq j$), i.e.,

$$p^{0,t_j}_{\theta^{(i)},\theta^{(j)}}(\theta^{(i)},\theta^{(j)}) = q(\theta^{(i)})q(\theta^{(j)}).$$

Proof:

$$p^{0,t_j}_{\theta^{(i)},\theta^{(j)}}(\theta^{(i)},\theta^{(j)}) = q(\theta^{(i)})q(\theta^{(j)}) \times \int_\Lambda f(0,\phi,\theta^{(i)})f(t_j,\phi,\theta^{(j)})q(\phi)d\phi. \tag{25}$$

$$= q(\theta^{(i)})q(\theta^{(j)})\int_\Lambda f(t_j,\phi,\theta^{(j)})q(\phi)d\phi \tag{26}$$

$$= q(\theta^{(i)})q(\theta^{(j)}), \tag{27}$$

where (25) follows from Lemma 6 and (26) follows from (14) and (27) follows from Lemma 2. ■

According to Lemma 8, we can make $\theta^{(i)}$ independent of the rest of the parameter vectors by setting $t_i = 0$.

Lemma 9: The parameter vector $\theta^{(j)}$ converges in mean square to $\theta^{(i)}$ as $t_i, t_j \to \infty$, i.e.,

$$\lim_{t_i,t_j \to \infty} ||\theta^{(i)} - \theta^{(j)}|| = 0. \tag{28}$$

Proof: By triangle inequality, we have

$$||\theta^{(i)} - \theta^{(j)}|| \leq ||\theta^{(i)} - \phi|| + ||\theta^{(j)} - \phi|| \tag{29}$$

and the rest follows from Lemma 5. ■

According to Lemma 9, when $t \to \infty$, we have $\theta^{(2)} \to \theta^{(1)}$ in probability, which reduces to the universal compression of identical sources studied in [18].

Remark: The degree of correlation between the two parameter vectors $\theta^{(i)}$ and $\theta^{(j)}$ is determined via the parameters t_i and t_j. This degree of correlation varies from independence of the two parameter vectors at $t_i = 0$ or $t_j = 0$ all the way to the vectors being equal (convergence in mean square) when $t_i, t_j \to \infty$. Please also note that the covariance matrix of the parameter vectors $\theta^{(i)}$ and $\theta^{(j)}$ for sufficiently large t_i and t_j converges to $\frac{2}{t^\star}\mathcal{I}^{-1}(\phi)$, where

$$\frac{1}{t^\star} = \frac{1}{t_i} + \frac{1}{t_j}. \tag{30}$$

IV. Problem Setup

In this section, we present the basic setup of the problem. As shown in Fig. 2, the setup is comprised of two sources with parameter vectors $\theta^{(1)}$ and $\theta^{(2)}$. Let X^n and Y^m denote two random sequences (samples) of lengths m and n, respectively, that are generated by parameter vectors $\theta^{(1)}$ and $\theta^{(2)}$, respectively.We consider four coding strategies (according to the orientation of the switches s_e and s_d in Fig. 2) for the compression of x^n from $\theta^{(1)}$ provided that the sequence y^m from $\theta^{(2)}$ is available to the encoder/decoder or not.[3]

- Ucomp (Universal compression without side information), where the switches s_e and s_d in Fig. 2 are both *open*. In this case, the encoder input space is given by $\mathcal{C} = \mathcal{X}^n \times \mathbb{Z}^*$. We let $C = (x^n, m)$ denote the input to the encoder. The decoder input space is denoted by $\mathcal{D} = \{0,1\}^* \times \mathbb{Z}^*$ and we let $D = (c(C), m)$ denote the input to the decoder.
- UcompE (Universal compression with encoder side information), where the switch s_e in Fig. 2 is *closed* but the switch s_d is *open*. In this case, the encoder input space is given by $\mathcal{C}^E = \mathcal{X}^n \times \mathcal{X}^m$. We let $C^E = (x^n, y^m)$ denote the input to the encoder. The decoder input space is denoted by $\mathcal{D}^E = \{0,1\}^* \times \mathbb{Z}^*$ and we let $D^E = (c(C^E), m)$ denote the input to the decoder.
- UcompD (Universal compression with decoder side information), where the switch s_e in Fig. 2 is *open* but the switch s_d is *closed*. In this case, the encoder input space is given by $\mathcal{C}^D = \mathcal{X}^n \times \mathbb{Z}^*$. We let $C^D = (x^n, m)$ denote the input to the encoder. The decoder input space is denoted by $\mathcal{D}^D = \{0,1\}^* \times \mathcal{X}^m$ and we let $D^D = (c(C^D), y^m)$ denote the input to the decoder.
- UcompED (Universal compression with encoder-decoder side information), where the switches s_e and s_d in Fig. 2 are both *closed*. In this case, the encoder input space is given by $\mathcal{C}^{ED} = \mathcal{X}^n \times \mathcal{X}^m$. We let $C^{ED} = (x^n, y^m)$ denote the input to the encoder. The decoder input space is denoted by $\mathcal{D}^{ED} = \{0,1\}^* \times \mathcal{X}^m$ and we let $D^{ED} = (c(C^{ED}), y^m)$ denote the input to the decoder.

Please note that, from the viewpoint of applications, the interesting coding strategy in this study is UcompD, where the side information sequence from S_1 is available at the decoder while it is not known to the encoder. The Ucomp coding is the benchmark for the achievable universal compression when no side information is present. Further, UcompED is the benchmark in the evaluation of UcompD but it may not be always practically feasible since it requires the sequence y^m from S_1 to be available at the encoder of S_2, i.e., coordination between the two encoders, which may not be always possible. Finally, UcompE is presented for the sake of completeness and as expected, UcompE provides no significant improvement over Ucomp.

In this paper, we focus on the family of fixed-to-variable length codes that map an n-vector to a variable-length binary sequence [19]. We only consider codes that are uniquely decodable, i.e., satisfy Kraft inequality.

Definition 1: The code $c : \mathcal{C} \to \{0,1\}^*$ is called strictly lossless (also called zero-error) if there exists a reverse mapping $d : \mathcal{D} \to \mathcal{X}^n$ such that

$$\forall x^n \in \mathcal{X}^n : \quad d(D) = x^n.$$

Most of the practical data compression schemes are examples of strictly lossless codes, namely, the arithmetic coding [20], Huffman [21], Lempel-Ziv [5], [6], and context-tree-weighting (CTW) algorithm [8]. Please note that based on the orientation of the switches in Fig. 2 and the input and output spaces, it is straightforward to extend the definition of strictly lossless codes to UcompE, UcompD, and UcompED.

V. Minimax and Maximin Redundancy

Let $l : \mathcal{C} \to \mathbb{R}$ denote the universal (strictly lossless) length function for Ucomp coding.[4] Denote L as the space of almost lossless universal length functions. Denote $R(l, \theta)$

[3] In this paper, we assume that m and n are a priori known to both the encoder and the decoder.

[4] Note that we have ignored the integer constraint on the length functions in this paper, which will result in a negligible redundancy smaller than 1 bit and is exactly analyzed in [19], [22].

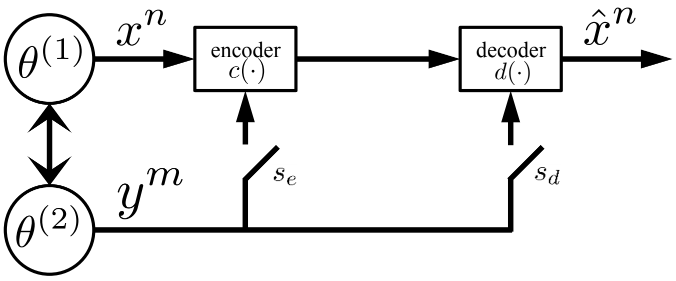

Fig. 2: The compression model for universal source coding with two correlated parameter vectors.

as the expected redundancy of the code with length function $l(\cdot)$, defined as

$$R(l, \theta) = \mathbf{E}l(C) - H_n(\theta). \tag{31}$$

Define $\bar{R}$ as the minimax redundancy of Ucomp coding, i.e.,

$$\bar{R} = \min_{l} \max_{\theta \in \Lambda} R(l, \theta). \tag{32}$$

Let $\underline{R}$ denote the maximin redundancy of Ucomp coding, i.e.,

$$\underline{R} = \max_{w} \min_{l} \int_{\theta \in \Lambda} R(l, \theta) w(\theta) d\theta. \tag{33}$$

It is straightforward to extend the definitions of the length function and the minimax and maximin redundancy to UcompE, UcompD, and UcompED coding strategies that are denoted by $\bar{R}_E$, $\bar{R}_D$, and $\bar{R}_{ED}$, respectively. In light of Gallager's result in [23], it can be deduced that the minimax and maximin risks in the abovementioned problems are equal, i.e.,

$$\begin{cases} \bar{R} = \underline{R} \\ \bar{R}_D = \underline{R}_D \\ \bar{R}_E = \underline{R}_E \\ \bar{R}_{ED} = \underline{R}_{ED} \end{cases}. \tag{34}$$

The following intuitive inequalities demonstrate that the redundancy decreases with the availability of the side information.

Lemma 10: The following set of inequalities hold:

$$\begin{cases} \bar{R}_{ED} \leq \bar{R}_E \leq \bar{R} \\ \bar{R}_{ED} \leq \bar{R}_D \leq \bar{R} \end{cases}. \tag{35}$$

Proof: Let $\check{l} \in L$ denote the optimal code for Ucomp coding strategy. Then, it is straightforward to see that $\check{l} \in L_E$ (i.e., $\check{l}$ is a code with encoder side information) since the encoder can choose not to use the side information sequence y^m in the coding. Likewise, if $\check{l}_E \in L_E$ is the optimal code with encoder side information. We have $\check{l}_E \in L_{ED}$ as it is a candidate code for the encoder-decoder side information case when the coding system is only a function of the side information sequence length at the decoder and not the side information sequence itself. This completes the proof of the first set of inequalities. The proof for the second set of the inequalities is similar and is omitted for brevity. ■

Before we delve into the main results of this paper, we present another result that will be useful in characterizing the redundancy in later sections.

Lemma 11: If $t_1 = 0$ or $t_2 = 0$, we have

$$\bar{R}_{ED} = \bar{R}_D = \bar{R}_E = \bar{R}.$$

Proof: It suffices to show that $\bar{R}_{ED} = \bar{R}$. Then, by Lemma 10, the rest follows. As pointed out in Lemma 8, when $t_1 = 0$ or $t_2 = 0$, then $\theta^{(1)}$ and $\theta^{(2)}$ are *independent*. Hence, X^n and Y^m are also independent and the result follows. ■

According to Lemma 11, there is no benefit provided by the side information when the two parameter vectors of the sources S_1 and S_2 are independent. This is not surprising as when $\theta^{(1)}$ and $\theta^{(2)}$ are independent, then X^n (produced by S_1) and Y^m (produced by S_2) are also independent. Thus, the knowledge of y^m does not affect the distribution of x^n. Hence, y^m cannot be used toward the reduction of the codeword length for x^n.

VI. Performance of Strictly Lossless Codes

In this section, we present our main results on the minimax redundancy for strictly lossless codes. As previously discussed, we only consider the case where $m = \omega(n)$, i.e., when the size of the side information sequence is sufficiently large. In other words, our focus is not on the transient period where the memory is populated with data traffic. Instead, we would like to analyze how much performance improvement is obtained when a sufficiently large side information sequence is used in the compression of a new sequence.

In the case of Ucomp, the side information sequence is not utilized at the encoder/decoder for the compression of x^n, and hence, the minimum number of bits required to represent x^n is $H(X^n)$. Thus, it can be shown that

Theorem 1 ([2], [4]): The minimax redundancy for strictly lossless Ucomp coding is

$$\bar{R} = \frac{d}{2}\log\left(\frac{n}{2\pi e}\right) + \log\int_{\Lambda} |\mathcal{I}(\lambda)|^{\frac{1}{2}} d\lambda + o(1).^{5}$$

Next, we confine ourselves to UcompE strategy and establish that the side information provided by y^m only at the encoder does not provide any benefit on the strictly lossless universal compression of the sequence x^n.

Theorem 2: The minimax redundancy for strictly lossless UcompE coding is

$$\bar{R}_E = \bar{R}.$$

Proof: In the case of UcompE coding, since the side information sequence y^m is not available to the decoder, then the minimum number of average bits required at the decoder to describe the random sequence X^n is indeed $H(X^n)$. On the other hand, it is straightforward to see that $H(X^n) = H_n(\theta^{(2)}) + I(X^n;\theta^{(2)})$. Further, it is clear that

$$I(X^n;\theta^{(2)}) = \bar{R}. \tag{36}$$

by the redundancy-capacity theorem (cf. [7]). ∎

Considering the UcompD strategy, we establish a result that the side information provided by y^m at the decoder does not provide any performance improvement in the strictly lossless universal compression of the sequence x^n.

Theorem 3: The minimax redundancy for strictly lossless UcompD coding is

$$\bar{R}_D = \bar{R}.$$

Proof: Since the two sources $\mu_{\theta^{(1)}}$ and $\mu_{\theta^{(2)}}$ are assumed to be from the d-dimensional parametric sources, in particular, they are also *ergodic*. In other words, any pair $(x^n, y^m) \in \mathcal{X}^n \times \mathcal{X}^m$ occurs with non-zero probability and the support set of (x^n, y^m) is equal to the entire $\mathcal{X}^n \times \mathcal{X}^m$. Therefore, the knowledge of the side information sequence y^m at the decoder does not rule out any possibilities for x^n at the decoder, and hence, the probability distribution of x^n remains unchanged (equal to the prior distribution) after y^m has been observed. Proposition 3 is then completed by using the known results of strictly lossless compression (cf. [24] and the references therein). ∎

Finally, we present our main result on the strictly lossless UcompED coding. In this case, since a side information sequence y^m is known to both the encoder and the decoder, the achievable codeword length for representing x^n is given by $H(X^n|Y^m)$. Hence, the redundancy can be shown to be obtained from the following theorem.

Theorem 4: For strictly lossless UcompED coding, if $\min\{t_1, t_2, m\} = O(1)$, then[6]

$$\bar{R}_{ED} = \bar{R} - O(1),$$

and if $\min\{t_1, t_2, m\} = \omega(1)$, then[7]

$$\bar{R}_{ED} = \hat{R}(n, m, t) + o(1),$$

where $\hat{R}(n, m, t)$ is defined as

$$\hat{R}(n, m, t) = \frac{d}{2}\log\left(1 + \frac{n}{m^\star}\right), \tag{37}$$

and $m^\star$ is given by the following.

$$\frac{1}{m^\star} = \frac{1}{m} + \frac{2}{t_1} + \frac{2}{t_2}. \tag{38}$$

Sketch of the proof: First, note that the minimax and maximin redundancies are equal, i.e, $\bar{R}_{ED} = \underline{R}_{ED}$. Also, it is not difficult to see that Jeffreys' prior is still capacity achieving in this case. Hence, we can focus on redundancy of the best code for Jeffreys' prior.

If $\min\{t_1, t_2, m\} = O(1)$, then the estimate $\hat{\theta}(Y^m)$, which is the maximum likelihood estimator of $\theta^{(1)}$ from the observation of Y^m, is going to have a variance that is bounded away from zero. Hence, the impact of Y^m would be to reshape the prior but since it will still be bounded away from zero on the space Λ, the redundancy will still be of the form $\frac{d}{2}\log n + \Theta(1)$ [2]. Hence, since Jeffreys' prior maximizes the redundancy, the reduction is going to be $O(1)$

Now, if $\min\{t_1, t_2, m\} = \omega(1)$, due to the property P4 of the source parameter vectors, central limit theorem holds and $\hat{\theta}(Y^m)$ will be distributed around $\theta^{(1)}$ with a variance $m^\star$. This will become similar to the case where a sequence $z^{m^\star}$ from a parametric source with source parameter vector $\theta^{(1)}$ is available to the encoder and is studied in [18]. The result then follows from Theorem 2 of [18]. ∎

VII. Conclusion

In this paper, we introduced a novel correlation model for the problem of universal compression of parametric sources with correlated parameter vectors. We formally defined a correlation model, which departs from the nature of the correlation in the Slepian-Wolf framework or the CEO problems. Involving two source parameter vectors, we investigated the minimax and maximin redundancy of lossless compression for four different coding strategies (based on whether or not the side information was available to the encoder and/or the decoder). These strategies are: 1. Universal compression without side information, 2.Universal compression with encoder side information, 3. Universal compression with decoder side information, and 4. Universal compression with encoder-decoder side information. We proved that the interesting case is only when the side information is available at both the encoder and the decoder.

Future work will investigate the case where several side information sequences from sources with correlated parameter vectors are available. Also, future work will investigate the notion of almost lossless coding (cf. [18]), where decoder side information can offer performance improvement.

[5] $f(n) = o(g(n))$ if and only if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

[6] $f(n) = O(g(n))$ if and only if $\lim_{n\to\infty} \sup \frac{f(n)}{g(n)} < \infty$.

[7] $f(n) = \omega(g(n))$ if and only if $g(n) = o(f(n))$.

References

[1] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783 – 795, Nov. 1973.

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

[3] ——, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40 –47, Jan. 1996.

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

[5] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, May 1977.

[6] ——, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, Sept. 1978.

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

[8] F. Willems, Y. Shtarkov, and T. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.

[9] M. Effros, K. Visweswariah, S. Kulkarni, and S. Verdu, "Universal lossless source coding with the Burrows Wheeler transform ," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1061–1081, May 2002.

[10] D. Baron and Y. Bresler, "An O(N) semipredictive universal encoder via the BWT," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 928–937, May 2004.

[11] A. Beirami and F. Fekri, "Results on the redundancy of universal compression for finite-length sequences," in *2011 IEEE International Symposium on Information Theory (ISIT '11)*, Jul. 2011, pp. 1604–1608.

[12] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1670–1672, Sept. 1994.

[13] W. Szpankowski, "A one-to-one code and its anti-redundancy," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4762–4766, Oct. 2008.

[14] I. Kontoyiannis and S. Verdu, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.

[15] S. Leung-Yan-Cheong and T. Cover, "Some equivalences between shannon entropy and kolmogorov complexity," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 331–338, May 1978.

[16] W. Szpankowski and S. Verdu, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4017–4025, Jul. 2011.

[17] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[18] A. Beirami and F. Fekri, "On lossless universal compression of distributed identical sources," in *2012 IEEE International Symposium on Information Theory (ISIT '12)*, Jul. 2012, pp. 561–565.

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

[20] G. G. Langdon Jr., "An Introduction to Arithmetic Coding," *IBM J. Res. Develop.*, vol. 28, no. 2, pp. 135–149, Mar. 1984.

[21] D. A. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proceedings of the I.R.E.*, pp. 1098–1102, Sept. 1952.

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

[23] R. G. Gallager, "Source coding with side information and universal coding," *unpublished*.

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