

# Multiple Description Coding of Discrete Ergodic Sources

Shirin Jalali and Tsachy Weissman

**Abstract**—We investigate the problem of Multiple Description (MD) coding of discrete ergodic processes. We introduce the notion of MD stationary coding, and characterize its relationship to the conventional block MD coding. In stationary coding, in addition to the two rate constraints normally considered in the MD problem, we consider another rate constraint which reflects the conditional entropy of the process generated by the third decoder given the reconstructions of the two other decoders. The relationship that we establish between stationary and block MD coding enables us to devise a universal algorithm for MD coding of discrete ergodic sources, based on simulated annealing ideas that were recently proven useful for the standard rate distortion problem.

## I. INTRODUCTION

Consider a packet network where a signal is to be described to several receivers. In a basic setup, the source is coded by a lossy encoder, and several copies of the packet containing the source description is sent over the network to make sure that each receiver gets at least one copy. Receiving more than one copy of these packets is not advantageous, because all the packets contain similar information. In contrast to this setup, one can think of a more reasonable scenario where the packets flooded into the network are not exactly the same; They are designed such that receiving each one of them is sufficient for recovering the source, but receiving more packets improves the quality of the reconstructed signal. The described scenario is referred to as *multiple description*.

The information-theoretic statement of the MD problem, and early results on the MD problem can be found in [1]–[4]. Even for the seemingly simple case where there are only two receivers, and the source is i.i.d., the characterization of the achievable rate-distortion region is not known in general. For this case, there are two well-known inner bounds due to El Gamal-Cover [5] and Zhang-Berger [6]. There is also a combined region, introduced in [7], which includes both regions, but recently shown to be no better than the Zhang-Berger region [8]. In any case, full characterization of the achievable region is not yet known.

Since even for i.i.d. sources, the single-letter characterization of the achievable rate-distortion region is not known in general, there are few works done on the MD of non-i.i.d. sources. The rate-distortion region of Gaussian processes is derived in [10], and is shown to be achievable using a scheme based on transform lattice quantization. In

[9], a multi-letter characterization of the achievable weighted rate-distortion region of discrete stationary ergodic sources is derived.

In this paper, we consider the MD of discrete ergodic processes where the distribution of the source is not known to the encoder and decoder. We introduce a universal algorithm which can asymptotically achieve any point in the achievable rate-distortion region. In order to get this result, we start by defining two notions of MD coding, namely, (i) conventional block coding, and (ii) *stationary* coding. In the normal block-coding MD, there are two rates but three reconstruction processes. In the stationary coding setup, there are three rates and three reconstruction processes. The additional rate corresponds to the *conditional entropy rate* of the the ergodic process reconstructed by the privileged decoder, which receives two descriptions of the source, given the two other ergodic reconstruction processes. We show that these two setups are closely related and, in fact, characterize each other. The beneficial point of the new definition is that it enables us to devise a universal MD algorithm. The introduced algorithm takes advantage of simulated annealing which was used recently in [15] to design an asymptotically optimal universal algorithm for lossy compression of discrete ergodic sources.

The outline of this paper is as follows: In Section II some preliminary notation, and definitions are presented. Section III studies a simple example, which, as made clear later, is closely related to the MD problem. Section IV formally defines the MD problems, and the two notions of block MD coding and stationary MD coding, and shows the relationship between the two. Based on these results, a universal MD algorithm is described in Section V, and in Section VI some simulation results demonstrating the performance of the proposed algorithm on simulated data are presented. Finally, Section VII discusses some future research directions.

## II. NOTATION

Let  $\mathbf{X} = \{X_i; \forall i \in \mathbb{N}^+\}$  be a stochastic process defined on a probability space  $(\mathbf{X}, \Sigma, \mu)$ , where  $\mu$  is a probability measure defined on  $\Sigma$ , the  $\sigma$ -algebra generated by the cylinder sets  $\mathcal{C}$ . For a process  $\mathbf{X}$ , let  $\mathcal{X}$  denote the alphabet set of  $X_i$ , which is assumed to be finite in this paper. The shift operator  $T : \mathcal{X}^\infty \rightarrow \mathcal{X}^\infty$  is defined by

$$(T\mathbf{x})_n = x_{n+1}, \quad \mathbf{x} \in \mathcal{X}^\infty, n \geq 1.$$

Moreover, for a stationary process  $\mathbf{X}$ , let  $\bar{H}(\mathbf{X})$  denote its entropy rate defined as  $\bar{H}(\mathbf{X}) = \lim_{n \rightarrow \infty} H(X_{n+1} | X^n)$ .

Let  $\mathcal{X}$  and  $\hat{\mathcal{X}}$  denote the source and reconstruction alphabets respectively. For  $y^n \in \mathcal{Y}^n$ , define the matrix  $\mathbf{m}(y^n)$  to

S. Jalali is a postdoctoral fellow at the Center for the Mathematics of Information, California Institute of Technology, Pasadena, CA 91125, USA [shirin@caltech.edu](mailto:shirin@caltech.edu)

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA [tsachy@stanford.edu](mailto:tsachy@stanford.edu)

be the  $|\mathcal{Y}| \times |\mathcal{Y}|^k$  matrix representing the  $(k+1)^{\text{th}}$  order empirical distribution of  $y^n$ , i.e., its  $(\beta, \mathbf{b})^{\text{th}}$  element is defined as

$$m_{\beta, \mathbf{b}}(y^n) = \frac{1}{n} |\{1 \leq i \leq n : y_{i-k}^{i-1} = \mathbf{b}, y_i = \beta\}|, \quad (1)$$

where  $\mathbf{b} \in \mathcal{Y}^k$ , and  $\beta \in \mathcal{Y}$ . In (1) and throughout we assume a cyclic convention whereby  $y_i \triangleq y_{n+i}$  for  $i \leq 0$ . Let  $H_k(y^n)$  denote the conditional empirical entropy of order  $k$  induced by  $y^n$ , i.e.

$$H_k(y^n) = H(Y_{k+1}|Y^k), \quad (2)$$

where  $Y^{k+1}$  on the right hand side of (2) is distributed according to

$$P(Y^{k+1} = [\mathbf{b}, \beta]) = m_{\beta, \mathbf{b}}(y^n). \quad (3)$$

The conditional empirical entropy in (2) can be expressed as a function of  $\mathbf{m}(y^n)$  as follows

$$H_k(y^n) = \frac{1}{n} \sum_{\mathbf{b}} \mathcal{H}(\mathbf{m}_{\cdot, \mathbf{b}}(y^n)) \mathbf{1}^T \mathbf{m}_{\cdot, \mathbf{b}}(y^n), \quad (4)$$

where  $\mathbf{1}$  and  $\mathbf{m}_{\cdot, \mathbf{b}}(y^n)$  denote the all-ones column vector of length  $|\mathcal{Y}|$ , and the column in  $\mathbf{m}(y^n)$  corresponding to  $\mathbf{b}$  respectively. For a vector  $\mathbf{v} = (v_1, \dots, v_\ell)^T$  with non-negative components, we let  $\mathcal{H}(\mathbf{v})$  denote the entropy of the random variable whose probability mass function (pmf) is proportional to  $\mathbf{v}$ . Formally,

$$\mathcal{H}(\mathbf{v}) = \begin{cases} \sum_{i=1}^{\ell} \frac{v_i}{\|\mathbf{v}\|_1} \log \frac{\|\mathbf{v}\|_1}{v_i} & \text{if } \mathbf{v} \neq (0, \dots, 0)^T \\ 0 & \text{if } \mathbf{v} = (0, \dots, 0)^T. \end{cases} \quad (5)$$

Let  $\mathbf{m}(w^n|y^n, z^n)$  denote the conditional  $k^{\text{th}}$  order empirical distribution of  $w^n$  given  $y^n$  and  $z^n$ , whose  $(\beta, \mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2)^{\text{th}}$  element is defined as

$$m_{\beta, \mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2} = \frac{1}{n} \left| \left\{ i : w_i = \beta, w_{i-k}^{i-1} = \mathbf{b}_0, y_{i-k_1}^{i+k_1} = \mathbf{b}_1, z_{i-k_1}^{i+k_1} = \mathbf{b}_2 \right\} \right|, \quad (6)$$

where  $\beta \in \mathcal{W}$ ,  $\mathbf{b}_0 \in \mathcal{W}^k$ ,  $\mathbf{b}_1 \in \mathcal{Y}^{2k_1+1}$ , and  $\mathbf{b}_2 \in \mathcal{Z}^{2k_1+1}$ . Now define the conditional empirical entropy of  $w^n$  given  $y^n$  and  $z^n$ ,  $H_{k, k_1}(y^n|w^n, z^n)$ , in terms of  $\mathbf{m}(w^n|y^n, z^n)$  as

$$H_{k, k_1}(w^n|y^n, z^n) = \sum_{\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2} \mathbf{1}^T \mathbf{m}_{\cdot, \mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2} \mathcal{H}(\mathbf{m}_{\cdot, \mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2}). \quad (7)$$

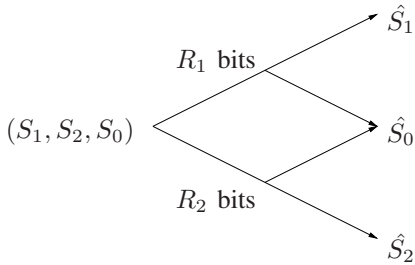


Fig. 1. Example setup

### III. SIMPLE EXAMPLE

Before formally defining the MD problem, consider the setup shown in Fig. 1. This example is meant to provide some insight into the MD problem. Also, the results of this section will be used in the proof of Theorem 2 in Appendix A. Here  $S_1 \in \mathcal{S}_1$ ,  $S_2 \in \mathcal{S}_2$  and  $S_0 \in \mathcal{S}_0$  denote three correlated discrete-valued random variables, and  $(S_1, S_2, S_0) \sim P(s_1, s_2, s_0)$ . The Encoder's goal is to send  $R_1$  bits to Decoder 1, and  $R_2$  bits to Decoder 2 such that Decoder 1 and 2 are able to reconstruct  $S_1$  and  $S_2$  respectively. Moreover, the transmitted bits are required to be such that receiving both of them enables Decoder 0 to reconstruct  $S_0$ . In all three cases, the probability of error is assumed to be zero. Let  $M_1 \in \{1, \dots, 2^{R_1}\}$ , and  $M_2 \in \{1, \dots, 2^{R_2}\}$  denote the messages sent to the decoders 1 and 2 respectively. The question is to find the set of achievable rates  $(R_1, R_2)$ . The following theorem states some necessary conditions for  $(R_1, R_2)$  to be achievable. It is very similar to Theorem 2 of [5], and the two theorems are in fact easily seen to prove each other. The version we give here is most suited for our later needs.

*Theorem 1:* For any achievable rate  $(R_1, R_2)$  for the setup shown in Fig. 1,

$$R_1 \geq H(S_1)$$

$$R_2 \geq H(S_2)$$

$$R_1 + R_2 \geq H(S_1) + H(S_2) + H(S_0|S_1, S_2). \quad (8)$$

*Proof:*  $R_1 \geq H(M_1)$  and  $R_2 \geq H(M_2)$  follow from Shannon's lossless coding Theorem. It is also clear that we should have

$$\begin{aligned} R_1 + R_2 &\geq H(S_1, S_2, S_0) \\ &= H(S_1, S_2) + H(S_0|S_1, S_2). \end{aligned} \quad (9)$$

But, perhaps somewhat counterintuitively, (9) is just an outer bound, and is not enough.  $R_1 + R_2$  in fact satisfies the tighter condition stated in (8), as can be seen via the following chain of inequalities:

$$\begin{aligned} R_1 + R_2 &\geq H(M_1) + H(M_2), \\ &= H(M_1, S_1) + H(M_2, S_2), \\ &= H(S_1) + H(M_1|S_1) + H(S_2) + H(M_2|S_2), \\ &\geq H(S_1) + H(S_2) + H(M_1|S_1, S_2) + \\ &\quad H(M_2|S_1, S_2), \\ &\geq H(S_1) + H(S_2) + H(M_1, M_2|S_1, S_2), \\ &\geq H(S_1) + H(S_2) + H(M_1, M_2, S_0|S_1, S_2), \\ &\geq H(S_1) + H(S_2) + H(S_0|S_1, S_2). \end{aligned} \quad (10)$$

### IV. MULTIPLE DESCRIPTION PROBLEM

Consider the basic setup of MD problem shown in Fig. 2. In this figure,  $X^n$  is generated by a stationary ergodic source  $\mathbf{X}$ .

**Remark:** In order to see the connection between the example described in Section III, and the MD problem, note

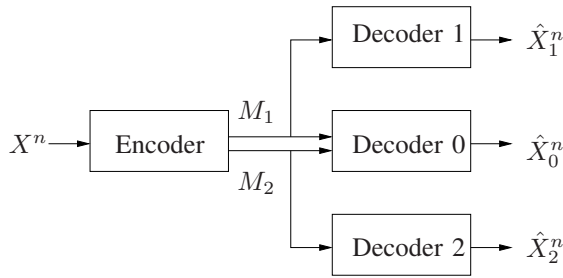


Fig. 2. MD coding setup

that letting  $S_i = \hat{X}_i^n$ ,  $i \in \{1, 2\}$ , and  $S_0 = \hat{X}_0^n$ , the MD problem can be described as the problem of describing  $(S_1, S_2, S_0)$  to the respected receivers error-free. In other words, for each code design, we have a problem equivalent to the one described in Section III.

#### A. Block coding:

MD coding problem can be described in terms of encoding mapping  $f$ , and decoding mappings  $(g_1, g_2, g_0)$  as follows

- 1)  $f : \mathcal{X}^n \rightarrow [1 : 2^{nR_1}] \times [1 : 2^{nR_2}]$ ,
- 2)  $g_i : [1 : 2^{nR_i}] \rightarrow \hat{\mathcal{X}}^n$ , for  $i = 1, 2$ ,
- 3)  $g_0 : [1 : 2^{nR_1}] \times [1 : 2^{nR_2}] \rightarrow \hat{\mathcal{X}}^n$ ,
- 4)  $(M_1, M_2) = f(X^n)$ ,
- 5)  $\hat{X}_i^n = g_i(M_i)$ , for  $i = 1, 2$ ,
- 6)  $\hat{X}_0^n = g_0(M_1, M_2)$ .

$(R_1, R_2, D_1, D_2, D_0)$  is said to be achievable for this setup, if there exists a sequence of codes  $(f^{(n)}, g_1^{(n)}, g_2^{(n)}, g_0^{(n)})$  such that

$$\limsup_n \mathbb{E} d_n(X^n, \hat{X}_i^n) \leq D_i, \text{ for } i = 1, 2,$$

$$\limsup_n \mathbb{E} d_n(X^n, \hat{X}_0^n) \leq D_0.$$

Let  $\mathcal{R}^B$  be the set of all  $(R_1, R_2, D_1, D_2, D_0)$  that are achievable by block MD coding of source  $\mathbf{X}$ .

#### B. Stationary coding:

Define  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0)$  to be achievable by stationary coding of source  $\mathbf{X}$ , if for any  $\epsilon > 0$ , there exist processes  $\hat{\mathbf{X}}_1^{(\epsilon)}$ ,  $\hat{\mathbf{X}}_2^{(\epsilon)}$  and  $\hat{\mathbf{X}}_0^{(\epsilon)}$  such that  $(\mathbf{X}, \hat{\mathbf{X}}_1^{(\epsilon)}, \hat{\mathbf{X}}_2^{(\epsilon)}, \hat{\mathbf{X}}_0^{(\epsilon)})$  are jointly stationary ergodic processes, and

$$\bar{H}(\hat{\mathbf{X}}_1^{(\epsilon)}) \leq R_{11} + \epsilon \quad (11)$$

$$\bar{H}(\hat{\mathbf{X}}_2^{(\epsilon)}) \leq R_{22} + \epsilon \quad (12)$$

$$\bar{H}(\hat{\mathbf{X}}_0^{(\epsilon)} | \hat{\mathbf{X}}_1^{(\epsilon)}, \hat{\mathbf{X}}_2^{(\epsilon)}) \leq R_0 + \epsilon \quad (13)$$

$$\mathbb{E} d(X_0, \hat{X}_{1,0}^{(\epsilon)}) \leq D_1 + \epsilon \quad (14)$$

$$\mathbb{E} d(X_0, \hat{X}_{2,0}^{(\epsilon)}) \leq D_2 + \epsilon \quad (15)$$

$$\mathbb{E} d(X_0, \hat{X}_{0,0}^{(\epsilon)}) \leq D_0 + \epsilon. \quad (16)$$

Let  $\mathcal{R}^P$  denote the set of all  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0)$  that are achievable by stationary MD coding of source  $\mathbf{X}$ . The following theorem characterizes  $\mathcal{R}^B$  in terms of  $\mathcal{R}^P$ .

**Theorem 2:** Let  $\mathbf{X}$  be a stationary ergodic source. For any  $(R_1, R_2, D_1, D_2, D_0) \in \mathcal{R}^B$ , there exists  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0) \in \mathcal{R}^P$  such that

$$R_{11} \leq R_1 \quad (17)$$

$$R_{22} \leq R_2 \quad (18)$$

$$R_{11} + R_{22} + R_0 \leq R_1 + R_2 \quad (19)$$

On the other hand, if  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0) \in \mathcal{R}^P$ , any point  $(R_1, R_2, D_1, D_2, D_0)$  satisfying (17)-(19) belongs to  $\mathcal{R}^B$ .

*Proof:* Refer to Appendix A for an outline of the proof. ■

**Remark:** The theorem implies that  $\mathcal{R}^B$  can be characterized as the set of  $(R_1, R_2, D_1, D_2, D_0)$  such that

$$\bar{H}(\hat{\mathbf{X}}_1) \leq R_1$$

$$\bar{H}(\hat{\mathbf{X}}_2) \leq R_2$$

$$\bar{H}(\hat{\mathbf{X}}_1) + \bar{H}(\hat{\mathbf{X}}_2) + \bar{H}(\hat{\mathbf{X}}_0 | \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2) \leq R_1 + R_2,$$

for some jointly stationary ergodic processes  $(\mathbf{X}, \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \hat{\mathbf{X}}_0)$  which satisfy (14)-(16).

### V. UNIVERSAL MULTIPLE DESCRIPTION CODING

Equipped with the characterization of the achievable region established in the previous section, we now turn to our construction of a universal scheme for this problem. Consider the following MD algorithm for the setup shown in Fig. 2. Let

$$(\hat{x}_1^n, \hat{x}_2^n, \hat{x}_0^n) \triangleq \arg \min_{(y^n, z^n, w^n)} [\gamma_1 H_k(y^n) + \gamma_2 H_k(z^n) + \gamma_0 H_{k,k_1}(w^n | y^n, z^n) + \alpha_1 d_n(x^n, y^n) + \alpha_2 d_n(x^n, z^n) + \alpha_0 d_n(x^n, w^n)], \quad (20)$$

Assume that  $\gamma_i \geq 0$  and  $\alpha_i \geq 0$ , for  $i \in \{0, 1, 2\}$ , are given Lagrangian coefficients. Also, assume that  $k_1 \leq k = o(\log n)$  such that  $k_1 \rightarrow \infty$  as  $n \rightarrow \infty$ .

**Theorem 3:** Let  $\mathbf{X}$  be a stationary ergodic process, and  $(\hat{X}_1^n, \hat{X}_2^n, \hat{X}_0^n)$  denote the output of the above algorithm to input sequence  $X^n$ . Then,

$$\limsup_n \left[ \gamma_1 H_k(\hat{X}_1^n) + \gamma_2 H_k(\hat{X}_2^n) + \gamma_0 H_{k,k_1}(\hat{X}_0^n | \hat{X}_1^n, \hat{X}_2^n) + \alpha_1 d_n(X^n, \hat{X}_1^n) + \alpha_2 d_n(X^n, \hat{X}_2^n) + \alpha_0 d_n(X^n, \hat{X}_0^n) \right] = \min [\gamma_1 R_{11} + \gamma_2 R_{22} + \gamma_0 R_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_0 D_0] \quad (21)$$

almost surely, where the minimization is over all  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0) \in \mathcal{R}^P$ .

The proof of Theorem 3 is presented in Appendix B.

After finding  $(\hat{x}_1^n, \hat{x}_2^n, \hat{x}_0^n)$ ,  $\hat{x}_1^n$  and  $\hat{x}_2^n$  will be described to Decoders 1 and 2 respectively using one of the well-known universal lossless compression algorithms, e.g., Lempel Ziv algorithm. Then Encoder forms a description of  $\hat{x}_0^n$  conditioned on knowing  $\hat{x}_1^n$  and  $\hat{x}_2^n$  using conditional Lempel

Ziv algorithm or some other universal algorithm for lossless coding with side information [11]. A portion  $0 \leq \theta \leq 1$  of these bits will be included in the message  $M_1$  and the rest in message  $M_2$ .

For finding an approximate solution of (20) instead of doing the required exhaustive search directly, as done in [15], one can employ simulated annealing [14]. To do this, we assign a cost to each  $(y^n, z^n, w^n) \in \hat{\mathcal{X}}^n \times \hat{\mathcal{X}}^n \times \hat{\mathcal{X}}^n$  as follows

$$\begin{aligned} \mathcal{E}(y^n, z^n, w^n) := & \\ & \gamma_1 H_k(y^n) + \gamma_2 H_k(z^n) + \gamma_0 H_{k, k_1}(w^n | y^n, z^n) \\ & + \alpha_1 d_n(x^n, y^n) + \alpha_2 d_n(x^n, z^n) + \alpha_0 d_n(x^n, w^n), \end{aligned}$$

and then define the Boltzmann probability distribution at temperature  $T = 1/\beta$  as

$$p_\beta(y^n, z^n, w^n) := \frac{1}{Z} e^{-\beta \mathcal{E}(y^n, z^n, w^n)}, \quad (22)$$

where  $Z$  is a normalizing constant. Sampling from this distribution at a very low temperature yields  $(\hat{X}_1^n, \hat{X}_2^n, \hat{X}_0^n)$  with energy close to the minimum possible energy, i.e.,

$$\mathcal{E}(\hat{X}_1^n, \hat{X}_2^n, \hat{X}_0^n) \approx \min_{(y^n, z^n, w^n)} \mathcal{E}(y^n, z^n, w^n). \quad (23)$$

Since sampling from (22) at low temperatures is almost as hard as doing the exhaustive search, we turn to simulated annealing (SA) which is a known method for solving discrete optimization problems. The SA procedure works as follows: it first defines Boltzmann distribution over the optimization space, and then tries to sample from the defined distribution while gradually decreasing the temperature from some high  $T$  to zero according to a properly chosen *annealing schedule*.

Given  $\mathcal{E}(y^n, z^n, w^n)$ , similarly as in [15], the number of computations required for calculating  $\mathcal{E}(y^{i-1} a y_{i+1}^n, z^{i-1} b z_{i+1}^n, w^{i-1} c w_{i+1}^n)$ , when only one of the following is true:  $a \neq y_i$ ,  $b \neq z_i$ , or  $c \neq w_i$ , for some  $i \in \{1, \dots, n\}$  and  $a, b, c \in \hat{\mathcal{X}}$ , is linear in  $k$  and  $k_1$ , and is independent of  $n$ . Therefore, this energy function lends itself to a heat bath type algorithm as simply and naturally as the one in the original setting of [15] did.

Now consider Algorithm 1 which is based on the Gibbs sampling method for sampling from  $p_\beta$ , and let  $(\hat{X}_{1,r}^n, \hat{X}_{2,r}^n, \hat{X}_{0,r}^n)$  denote its random outcome for the input sequence  $X^n$  after  $r$  iterations<sup>1</sup>, when taking  $k_1 = k_{1,n}$ ,  $k = k_n$  and  $\beta = \{\beta_t\}_t$  to be deterministic sequences satisfying  $k_{1,n} = o(\log n)$ ,  $k_n = o(\log n)$  such that  $k, k_1 \rightarrow \infty$  as  $n \rightarrow \infty$ , and  $\beta_t = \frac{1}{T_0^{(n)}} \log(\lfloor \frac{t}{n} \rfloor + 1)$ , for some

<sup>1</sup>Here and throughout it is implicit that the randomness used in the algorithms is independent of the source, and the randomization variables used at each drawing are independent of each other.

$T_0^{(n)} > n \max(\Delta_1, \Delta_2, \Delta_0)$ , where

$$\begin{aligned} \Delta_1 = \max & |\mathcal{E}(y^{i-1} a y_{i+1}^n, z^n, w^n) - \mathcal{E}(y^{i-1} b y_{i+1}^n, z^n, w^n)|, \\ & i \in \{1, \dots, n\} \\ & y^{i-1} \in \hat{\mathcal{X}}^{i-1}, y_{i+1}^n \in \hat{\mathcal{X}}^{n-i}, \\ & a, b \in \hat{\mathcal{X}}, \\ & z^n \in \hat{\mathcal{X}}^n, w^n \in \hat{\mathcal{X}}^n, \end{aligned} \quad (24)$$

$$\begin{aligned} \Delta_2 = \max & |\mathcal{E}(y^n, z^{i-1} a z_{i+1}^n, w^n) - \mathcal{E}(y^n, z^{i-1} b z_{i+1}^n, w^n)|, \\ & i \in \{1, \dots, n\} \\ & z^{i-1} \in \hat{\mathcal{X}}^{i-1}, z_{i+1}^n \in \hat{\mathcal{X}}^{n-i}, \\ & a, b \in \hat{\mathcal{X}}, \\ & y^n \in \hat{\mathcal{X}}^n, w^n \in \hat{\mathcal{X}}^n, \end{aligned} \quad (25)$$

$$\begin{aligned} \Delta_0 = \max & |\mathcal{E}(y^n, z^n, w^{i-1} a w_{i+1}^n) - \mathcal{E}(y^n, z^n, w^{i-1} b w_{i+1}^n)|, \\ & i \in \{1, \dots, n\} \\ & w^{i-1} \in \hat{\mathcal{X}}^{i-1}, w_{i+1}^n \in \hat{\mathcal{X}}^{n-i}, \\ & a, b \in \hat{\mathcal{X}}, \\ & y^n \in \hat{\mathcal{X}}^n, z^n \in \hat{\mathcal{X}}^n, \end{aligned} \quad (26)$$

As discussed before, the computational complexity of the algorithm at each iteration is independent of  $n$  and linear in  $k$  and  $k_1$ . Following exactly the same steps as in the proof of Theorem 2 in [15], we can prove the following theorem which established universal optimality of Algorithm 1.

*Theorem 4:* For any ergodic process  $\mathbf{X}$ ,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \lim_{r \rightarrow \infty} \mathcal{E}(\hat{X}_1^n, \hat{X}_2^n, \hat{X}_0^n) \\ & = \min [\gamma_1 R_{11} + \gamma_2 R_{22} + \gamma_0 R_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_0 D_0] \end{aligned} \quad (27)$$

almost surely, where the minimization is over all  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0) \in \mathcal{R}^P(\mathbf{X})$ .

---

#### Algorithm 1 Generating the reconstruction sequences

---

**Require:**  $x^n, k_1, k, \{\alpha_i\}_{i=0}^2, \{\beta_i\}_{i=0}^2, \{\beta_t\}_{t=1}^r, r$

**Ensure:** a reconstruction sequences  $(\hat{x}_1^n, \hat{x}_2^n, \hat{x}_0^n)$

- 1:  $y^n \leftarrow x^n$
  - 2:  $z^n \leftarrow x^n$
  - 3:  $w^n \leftarrow x^n$
  - 4: **for**  $t = 1$  to  $r$  **do**
  - 5:   Draw an integer  $i \in \{1, \dots, n\}$  uniformly at random
  - 6:   For each  $y \in \hat{\mathcal{X}}$  compute  $q_1(y) = p_{\beta_t}(Y_i = y | Y^{n \setminus i} = y^{n \setminus i}, Z^n = z^n, W^n = w^n)$
  - 7:   Update  $y^n$  by letting  $y_i = V_1$ , where  $V_1 \sim q_1$
  - 8:   For each  $z \in \hat{\mathcal{X}}$  compute  $q_2(z) = p_{\beta_t}(Z_i = z | Y^n = y^n, Z^{n \setminus i} = z^{n \setminus i}, W^n = w^n)$
  - 9:   Update  $z^n$  by letting  $z_i = V_2$ , where  $V_2 \sim q_2$
  - 10:   For each  $y \in \hat{\mathcal{X}}$  compute  $p_{\beta_t}(Y_i = y | Y^{n \setminus i} = y^{n \setminus i})$
  - 11:   Update  $w^n$  by letting  $w_i = V_0$ , where  $V_0 \sim q_0$
  - 12:   Update  $\mathbf{m}(y^n)$ ,  $\mathbf{m}(z^n)$  and  $\mathbf{m}(w^n | y^n, z^n)$
  - 13: **end for**
  - 14:  $\hat{x}_1^n \leftarrow y^n$
-



## VI. SIMULATION RESULTS

In this section, we present some results showing the actual implementation of the algorithm described in Section V. The simulated source here is a sym metric binary Markov source with transition probability  $p = 0.2$ . The considered block length is  $n = 10^4$ , and the context sizes are  $k = 5$  and  $k_1 = 1$ . The annealing schedule was chosen according to

$$T(t) = \frac{1}{2nt^{1/10}},$$

where  $t$  is the iteration number. The number of iterations,  $r$ , is equal to  $50n$ . The algorithm with the specified parameters, for  $\gamma_1 = \gamma_2 = \gamma_0 = \alpha_1 = \alpha_2 = a_0 = 1$ , achieves the following set of rates and distortions:

$$\begin{aligned} H_k(\hat{x}_1^n) &= 0.5503, \\ H_k(\hat{x}_2^n) &= 0.5586, \\ H_{k,k_1}(\hat{x}_0^n | \hat{x}_1^n, \hat{x}_2^n) &= 0.0038, \\ d_n(x^n, \hat{x}_1^n) &= 0.0505, \\ d_n(x^n, \hat{x}_2^n) &= 0.0483, \\ d_n(x^n, \hat{x}_0^n) &= 0.0036. \end{aligned}$$

Fig. 3 shows how the total cost is reducing in this case, as the number of iterations increases. One interesting thing to note here is that although the sequences  $\hat{x}_1^n$  and  $\hat{x}_2^n$  have almost the same distance from the original sequence  $x^n$ , they are far from being equal. In fact,  $d_n(\hat{x}_1^n, \hat{x}_2^n) = 0.0966$ , which, given  $d_n(x^n, \hat{x}_1^n) = 0.0505$  and  $d_n(x^n, \hat{x}_2^n) = 0.0483$ , suggests that they are almost maximally distant.

As another example, consider the case where  $n = 5 \times 10^4$  and  $\alpha_1 = \alpha_2 = 2$ . The rest of the parameters are left unchanged. The achieved point in this case is going to be

$$\begin{aligned} H_k(\hat{x}_1^n) &= 0.6091, \\ H_k(\hat{x}_2^n) &= 0.5951, \\ H_{k,k_1}(\hat{x}_0^n | \hat{x}_1^n, \hat{x}_2^n) &= 0, \\ d_n(x^n, \hat{x}_1^n) &= 0.0200, \\ d_n(x^n, \hat{x}_2^n) &= 0.0240, \\ d_n(x^n, \hat{x}_0^n) &= 0.0010. \end{aligned}$$

Here,  $H_{k,k_1}(\hat{x}_0^n | \hat{x}_1^n, \hat{x}_2^n) = 0$  implies that  $\hat{x}_{0,i}$  is a deterministic function of its context,  $(\hat{x}_{0,i-k_1}^{i-1}, \hat{x}_{1,i-k_1}^{i+k_1}, \hat{x}_{2,i-k_1}^{i+k_1})$ . This of course does not mean that no additional rate is required for describing  $\hat{x}_0^n$  when the decoder already knows  $\hat{x}_1^n$  and  $\hat{x}_2^n$ , because this deterministic mapping itself is not known to the decoder beforehand. Here again  $\hat{x}_1^n$  and  $\hat{x}_2^n$  are almost maximally distant because  $d_n(\hat{x}_1^n, \hat{x}_2^n) = 0.0436$ .

Note that the fundamental performance limits are unknown even for memoryless sources and, a fortiori, for the Markov source in our experiment. Thus the performance of our algorithm cannot be compared to the corresponding optimum performance. The results of the preceding section, however, imply that our algorithm attains that performance in the limit of many iterations and large block length. Thus, the performance attained by our algorithm, can alternatively be viewed as approximating the unknown optimum.

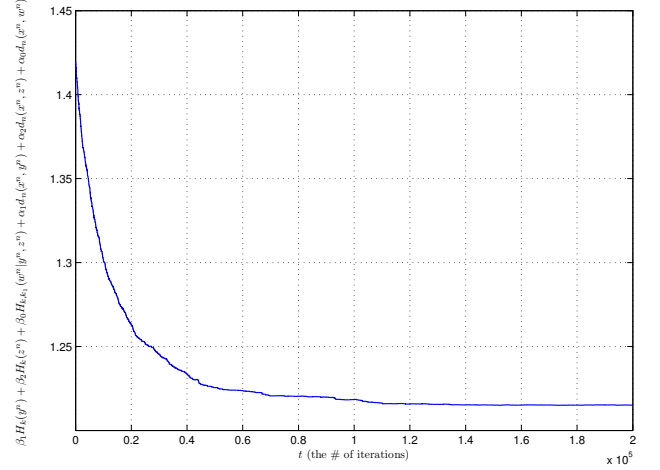


Fig. 3. Reduction in the cost. At the end of the process, the final achieved point is:  $(H_k(\hat{x}_1^n), H_k(\hat{x}_2^n), H_{k,k_1}(\hat{x}_0^n | \hat{x}_1^n, \hat{x}_2^n), d_n(x^n, \hat{x}_1^n), d_n(x^n, \hat{x}_2^n), d_n(x^n, \hat{x}_0^n)) = (0.5503, 0.5586, 0.0038, 0.0505, 0.0483, 0.0036)$

## VII. FUTURE DIRECTIONS

Simulated annealing was recently employed in [15] to design a universal lossy compression algorithm. In this paper, we proved that in fact the same tool can be applied to devise a universal MD algorithm. We started by defining the equivalent of MD problem for ergodic processes, and defined the idea of stationary MD coding which includes three rate constraints instead of two. Extensions of these results to additional distributed coding scenarios are under current investigation.

## ACKNOWLEDGMENT

We thank Jun Chen for suggesting the current proof of Theorem 1, in lieu of our original proof which was more complicated.

## APPENDIX A: OUTLINE OF THE PROOF OF THEOREM 2

*Outline of the proof of the first part:* Let  $(R_1, R_2, D_1, D_2, D_0) \in \mathcal{R}^B$ . We need to find  $(R_{11}, R_{22}, R_0)$  such that  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0) \in \mathcal{R}^P$ , and (17)–(19) are satisfied.

Let  $(f^{(n)}, g_1^{(n)}, g_2^{(n)}, g_0^{(n)})$  be a sequence of codes at rate  $(R_1, R_2)$  that achieves the point  $(R_1, R_2, D_1, D_2, D_0) \in \mathcal{R}^B$ . Note that for a given code,  $(\hat{X}_1^n, \hat{X}_2^n, \hat{X}_0^n)$  is a deterministic function of  $X^n$ . Using the same method used in [12], we can generate jointly stationary ergodic processes  $(\hat{X}_1^{(n)}, \hat{X}_2^{(n)}, \hat{X}_0^{(n)})$  by appropriately embedding these block codes into ergodic processes. Here the superscript  $(n)$  indicates the dependence of the constructed processes on  $n$ . In order to code an ergodic process into another ergodic process using a block code of length  $n$ , we need to cover an infinite length sequence by non-overlapping blocks of length  $n$  up to a set of negligible measure, and then replace each block by its reconstruction generated by the block code. The challenging

part is the partitioning which should preserve the ergodicity. This can be done using R-K Theorem [13] which states that:

**Theorem 5 (Rohlin-Kakutani Theorem):** Given the ergodic source  $\mathbf{X}$ , integers  $N, n \leq N$ , and  $\epsilon > 0$ , there exists an event  $F$  (called the *base*) such that

- 1)  $F, TF, \dots, T^{N-1}F$  are disjoint,
- 2)  $P\left(\bigcup_{i=0}^{N-1} T^i F\right) \geq 1 - \epsilon$ ,
- 3)  $P(\mathcal{S}(a^n)|F) = P(\mathcal{S}(a^n))$ , where  $\mathcal{S}(a^n) = \{\mathbf{x} : x^n = a^n\}$ .

Since the sequence of MD block codes were assumed to achieve the point  $(R_1, R_2, D_1, D_2, D_0)$ , the constructed process  $(\hat{\mathbf{X}}_1^{(n)}, \hat{\mathbf{X}}_2^{(n)}, \hat{\mathbf{X}}_0^{(n)})$  satisfies the distortion constraints given in (14)-(16) at  $(D_1 + \epsilon_n, D_2 + \epsilon_n, D_0 + \epsilon_n)$ , where  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ . Therefore,  $(\bar{H}_n(\hat{\mathbf{X}}_1^{(n)}), \bar{H}_n(\hat{\mathbf{X}}_2^{(n)}), \bar{H}_n(\hat{\mathbf{X}}_0^{(n)}|\hat{\mathbf{X}}_1^{(n)}, \hat{\mathbf{X}}_2^{(n)}), D_1 + \epsilon_n, D_2 + \epsilon_n, D_0 + \epsilon_n) \in \mathcal{R}^P$ . Let

$$R_{11}^{(n)} := \frac{1}{n} H(\hat{X}_1^n), \quad (\text{A-1})$$

$$R_{22}^{(n)} := \frac{1}{n} H(\hat{X}_2^n), \quad (\text{A-2})$$

$$R_0^{(n)} := \frac{1}{n} H(\hat{X}_0^n|\hat{X}_1^n, \hat{X}_2^n), \quad (\text{A-3})$$

where  $\hat{X}_i^n = g_i^{(n)}(M_i)$ , for  $i \in \{1, 2\}$  and  $\hat{X}_0^n = g_0^{(n)}(M_1, M_2)$ . Note that since the encoder knows  $(\hat{X}_1^n, \hat{X}_2^n, \hat{X}_0^n)$ , by Theorem 1,  $R_{11}^{(n)} \leq R_1$ ,  $R_{22}^{(n)} \leq R_2$ , and  $R_{11}^{(n)} + R_{22}^{(n)} + R_0^{(n)} \leq R_1 + R_2$ . The only remaining step is to find the relationship between  $(\bar{H}_n(\hat{\mathbf{X}}_1^{(n)}), \bar{H}_n(\hat{\mathbf{X}}_2^{(n)}), \bar{H}_n(\hat{\mathbf{X}}_0^{(n)}|\hat{\mathbf{X}}_1^{(n)}, \hat{\mathbf{X}}_2^{(n)}))$  and  $(R_{11}^{(n)}, R_{22}^{(n)}, R_0^{(n)})$ , which is not hard from the way the processes are constructed.

*Outline of the proof of the second part:* Let  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0) \in \mathcal{R}^P$ . This means that there exist processes  $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2$  and  $\hat{\mathbf{X}}_0$  jointly stationary and ergodic with  $\mathbf{X}$  which satisfy (11)-(16). Based on these processes, for block length  $n$ , we use the following block coding strategy: For coding sequence  $X^n$ , describe  $\hat{X}_1^n$  and  $\hat{X}_2^n$  losslessly to the decoders 1 and 2 using  $n(\bar{H}(\hat{\mathbf{X}}_1) + \epsilon_n)$  and  $n(\bar{H}(\hat{\mathbf{X}}_2) + \epsilon_n)$  bits respectively. Given  $\hat{X}_1^n$  and  $\hat{X}_2^n$ ,  $n(\bar{H}(\hat{\mathbf{X}}_0|\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2) + \epsilon_n)$  bits suffice to describe  $\hat{X}_0^n$  losslessly to Decoder 0. These bits can be divided into two parts: the first part will be included in the message  $M_1$ , and the rest in the message  $M_2$ . Decoders 1 and 2 just ignore these extra bits, but Decoder 0 combines them with the two other messages to reconstruct  $\hat{X}_0^n$ . Since  $R_1$  and  $R_2$  satisfy (17)-(19), it is possible to do this.

## APPENDIX B: PROOF OF THEOREM 3

For an ergodic source  $\mathbf{X}$ , let

$$\mu(\gamma, \alpha) := \min_{\mathcal{R}^P(\mathbf{X})} [\gamma_1 R_{11} + \gamma_2 R_{22} + \gamma_0 R_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_0 D_0]. \quad (\text{B-1})$$

No coding strategy can beat  $\mu(\gamma, \alpha)$  on a set of non-zero probability. Therefore, the left hand side of (21) is lower bounded by its right hand side. Therefore, we only need to prove the other direction. By definition, for any  $(R_{11}, R_{22}, R_0, D_1, D_2, D_0) \in \mathcal{R}^P(\mathbf{X})$ , there exist processes  $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2$  and  $\hat{\mathbf{X}}_0$  such that (11)-(16) are satisfied. On the other hand, if  $(\hat{X}_1^n, \hat{X}_2^n, \hat{X}_0^n)$  is generated by jointly ergodic processes  $(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \hat{\mathbf{X}}_0)$ , then for  $k = o(\log n)$  and  $k_1 = o(\log n)$  such that  $k, k_1 \rightarrow \infty$  as  $n \rightarrow \infty$ ,  $H_k(\hat{X}_i^n) \rightarrow \bar{H}(\hat{\mathbf{X}}_i)$ , for  $i \in \{1, 2\}$ , and moreover  $H_{k, k_1}(\hat{X}_0^n|\hat{X}_1^n, \hat{X}_2^n) \rightarrow \bar{H}(\hat{\mathbf{X}}_0|\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2)$ . This implies that

$$\begin{aligned} \limsup \min & [\gamma_1 H_k(\hat{X}_1^n) + \alpha_2 d_n(X^n, \hat{X}_2^n) + \\ & \gamma_2 H_k(\hat{X}_2^n) + \alpha_1 d_n(X^n, \hat{X}_1^n) + \\ & \gamma_0 H_{k, k_1}(\hat{X}_0^n|\hat{X}_1^n, \hat{X}_2^n) + \alpha_0 d_n(X^n, \hat{X}_0^n)] \end{aligned} \quad (\text{B-2})$$

is upper-bounded by  $\mu(\gamma, \alpha) + \epsilon_n$ , where  $\epsilon_n \rightarrow 0$ . Combining these two results in the desired conclusion.

## REFERENCES

- [1] H. Witsenhausen, "On source networks with minimal breakdown degradation," *Bell Syst. Tech. J.*, vol. 59, no. 6, pp. 1083-1087, July-Aug. 1980.
- [2] J. Wolf, A. Wyner, and J. Ziv, "Source coding for multiple descriptions," *Bell Syst. Tech. J.*, vol. 59, no. 8, pp. 1417-1426, Oct. 1980.
- [3] L. Ozarow, "On a source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, no. 10, pp. 1909-1921, Dec. 1980.
- [4] H.S. Witsenhausen and A.D. Wyner, "Source coding for multiple descriptions II: A binary source," *Bell Lab. Tech. Rep. TM-80-1217*, Dec. 1980.
- [5] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 851-857, Nov. 1982.
- [6] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. 33, no. 4, pp. 502-521, July 1987.
- [7] R. Venkataramani, G. Kramer, and V.K. Goyal, "Multiple description coding with many channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, pp. 2106-2114, Sept. 2003.
- [8] L. Zhao, P. Cuff and H. Permuter, "Consolidating Achievable Regions of Multiple Descriptions," submitted to *IEEE Inter. Symp. on Inf. Theory*, Seoul, Korea, 2009.
- [9] M. Fleming and M. Effros, "The rate distortion region for the multiple description problem," *Proc. IEEE Int. Symp. Information Theory Sorrento*, Italy, Jun. 2000, p. 208.
- [10] J. Chen, C. Tian, S. Diggavi, "Multiple description coding for stationary and ergodic sources," *Proc. Data Compression Conference (DCC)*, pp. 73-82, 2007.
- [11] H. Cai, S. R. Kulkarni and S. Verdú, "An Algorithm for Universal Lossless Compression With Side Information," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, Sept. 2006.
- [12] R. M. Gray, "Sliding-block source coding," *IEEE Trans. on Inform. Theory*, vol. 21, pp. 357-368, July 1975.
- [13] P. C. Shields, *The theory of Bernoulli shifts*, Univ. of Chicago press, Chicago, 1973.
- [14] P. Brémaud, *Markov chains, Gibbs fields, Monte Carlo simulation, and queues*, Springer, New York, 1998.
- [15] S. Jalali, T. Weissman, "Rate-Distortion via Markov Chain Monte Carlo," submitted to *IEEE Trans. on Info. Theory*. (available on arxiv at [http://arxiv.org/PS\\_cache/arxiv/pdf/0808/0808.4156v1.pdf](http://arxiv.org/PS_cache/arxiv/pdf/0808/0808.4156v1.pdf))