Correlated Source Coding for Fusion Storage and Selective Retrieval

Jayanth Nayak, Sharadh Ramaswamy and Kenneth Rose
Dept. of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106.
Email: {jayanth,rose}@ece.ucsb.edu, rsharadh@engineering.ucsb.edu

***Abstract*— Motivated by the sensor network setting, we consider lossless storage of correlated discrete memoryless sources. The underlying tradeoff is between exploitation of inter-source correlation for low rate storage and efficient (low rate) *selective* retrieval from the fusion storage. We define the problem of shared descriptions (SD) source coding and relate it to the storage and retrieval problem. We present an achievable rate region for the SD problem and use it to characterize the storage vs. retrieval tradeoff.**

I. Introduction

The data acquired by networks of sensors generally exhibit a high degree of correlation. There has been considerable work on utilizing this correlation to minimize the capacity of the link from each sensor to the fusion center. Issues in the storage of this data at the fusion center, however, have received less attention. We suspect that this problem may turn out to be critical in practice. The fusion center could store (many) received data streams separately, but this would clearly be wasteful due to the inter-stream correlation. Hence, joint coding for storage is naturally called for. However, an end user may eventually be interested in retrieving only a subset of the available streams at any given time. It would be highly undesirable for the decoder to retrieve all the streams in order to satisfy the user's request for reconstructing a small subset of them. For example, consider a network of cameras that cover a scene or an area from a variety of angles (and/or spectral bands, etc.). A user will eventually want to retrieve a small subset of these streams to review an event of interest, as it would be impossible for a human observer to view all of these simultaneously. It is clearly highly inefficient to retrieve all the data in order to reconstruct a small subset of the streams.

Motivated by this type of scenarios, we investigate the tradeoff between storage cost and retrieval cost in the compression of correlated streams of data. We focus on the scenario where the sources are discrete and memoryless and the reconstruction is required to be (asymptotically) lossless. The problem is formally defined in Section II. The storage problem is cast as a source coding problem, which we term the shared descriptions (SD) problem in Section III. Adopting earlier results by Han and Kobayashi [1] for general multiterminal source coding scenarios, we obtain the (non-single letter) achievable rate region for the SD problem. By relating the multiple encoder problem to the multiple descriptions problem, we also obtain a single letter characterization of a (partial) achievable rate region. The results for the SD problem lead to a characterization of the storage-retrieval cost tradeoff problem (Section IV). Finally, we observe that the general solution that emerges from the analysis involves an inordinately large computational complexity. This can be mitigated by the imposition of structural constraints on the encoder. We also study one such structural constraint in Section IV.

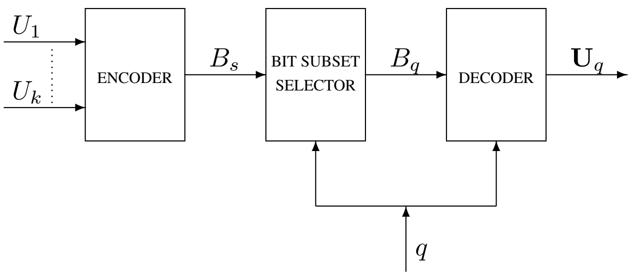

Fig. 1. Storage of correlated sources

II. Problem Definition

Let $\{(U_{t1},\ldots,U_{tK})\}_{t=1}^{T}$ be a sequence of random vectors defined over a discrete alphabet. The sequence is memoryless and represents the realizations of a random vector $(U_1,\ldots,U_K) \sim p_{U_1,\ldots,U_K}(u_1,\ldots,u_K), u_k \in \mathcal{U}_k$ at each instant t. Alice observes the sequence of random vectors and stores her observations as the set of bits B_s. At some later time, Bob queries Alice for a non-empty subset $q \subset \mathcal{K} \triangleq \{1,\ldots,K\}$ of the streams. Alice retrieves a subset $B_q \subset B_s$ of the stored bits to reconstruct the data streams indexed by q (See Fig. 1). We are interested in the tradeoff between the *average retrieval* cost and the *total storage* cost. We will assume infinite data streams, $T \to \infty$, to simplify the derivation of results, in the expectation that they will nevertheless provide significant insight into the tradeoff for finite T.

In the following, for any set s, $|s|$ denotes the cardinality of the set and 2^s denotes the power set of s, that is the set of all subsets of s. For a function r, $\|r\|$ denotes the cardinality of its range. For any set $q = \{k_1,\ldots,k_{|q|}\} \subset \mathcal{K}$, we shall denote $(U_k, k \in q)$ by $\mathbf{U}_q$ and the corresponding alphabet

This work was supported in part by the NSF under grant IIS-0329267, and by the University of California MICRO program, Applied Signal Technology, Inc., Dolby Laboratories, Inc., and Qualcomm, Inc.

$\mathcal{U}_{k_1} \times \cdots \times \mathcal{U}_{k_{|q|}}$ by $\mathcal{U}_q$. A sequence of n realizations of $\mathbf{U}_q$ will be denoted $\mathbf{U}_q^n$.

An ϵ-storage code $(f, g, h_q, q \subset \mathcal{K})$ of block length n for the source $(U_1, \ldots, U_K)$ is the following set of mappings:

1) the encoder: $f : \mathcal{U}_{\mathcal{K}}^n \rightarrow \{0,1\}^{M_s}$, where M_s is some positive integer,
2) the query to bit subset mapping : $g : 2^{\mathcal{K}} \rightarrow 2^{\{1,\ldots,M_s\}}$, and
3) the query decoders : For each $q \subset \mathcal{K}$, $h_q : \{0,1\}^{M_q} \rightarrow \mathcal{U}_q^n$, where $M_q = |g(q)|$,

that satisfies for each $q \subset \mathcal{K}$

$$\Pr\left[\mathbf{U}_q^n \neq h_q(f(U_{\mathcal{K}}^n)|_{g(q)})\right] < \epsilon,$$

where for a b bit word B and $a \subset \{1, \ldots, b\}$, $B|_a$ denotes the $|a|$ bit word formed by extracting from B the bits at the positions indicated by a.

A tuple of rates $(R_s, R_q, q \subset \mathcal{K})$ is termed achievable if for every $\epsilon > 0$, there exists an ϵ-storage code at some block length $n = n(\epsilon)$ that satisfies

$$M_s \leq n(R_s + \epsilon) \quad (1)$$

$$M_q \leq n(R_q + \epsilon), \forall q \subset \mathcal{K} \quad (2)$$

Suppose that the subset q is requested with probability or frequency $P(q)$. To avoid unnecessary complications, we shall assume that all sources have positive probability of being retrieved, i.e., for all $k \in \mathcal{K}$, there exists $q \ni k$ such that $P(q) > 0$. We are interested in the tradeoff between the average retrieval cost $\sum_{q \subset \mathcal{K}} P(q) R_q$ and the storage cost R_s for achievable rate tuples $(R_s, R_q, q \subset \mathcal{K})$. To better understand the tradeoff, it is instructive to consider the two extreme scenarios.

Case 1. Suppose $R_s = \sum_{q \subset \mathcal{K}} H(\mathbf{U}_q)$, where $\epsilon > 0$ and $H(X)$ denotes the Shannon entropy of the random variable X. This extreme case allows for enough storage rate to encode each possible subset of sources that may be requested by Bob, and store *separately* in the database. (Since the Shannon entropy $H(X)$ of random variable X signifies the minimum asymptotic per symbol bit rate that is necessary and sufficient for lossless compression of the corresponding i.i.d. sequence.) Here, for any given query subset q, only the corresponding encoded bits (at rate $R_q = H(\mathbf{U}_q)$) need to be retrieved. The average per letter retrieval cost in this scheme is

$$C(R_s) = \sum_{q \subset \mathcal{K}} P(q) H(\mathbf{U}_q).$$

Since reconstructing the streams $\mathbf{U}_q$ would require a bit rate of at least $H(\mathbf{U}_q)$, this is the smallest *retrieval* cost that can be attained by any scheme and with any total storage bit rate constraint.

Case 2. Suppose $R_s = H(\mathbf{U}_{\mathcal{K}})$. By jointly encoding all the streams we remove all redundancy in the database and obtain a single joint description the entire data that uses $H(\mathbf{U}_{\mathcal{K}})$ bits per instant. Note that this is the lowest achievable R_s for lossless compression of the observed data. To obtain a reconstruction, regardless of the query subset q, the entire description of $H(\mathbf{U}_{\mathcal{K}})$ bits would in general need to be retrieved. The retrieval cost that is incurred is therefore

$$C(R_s) = H(\mathbf{U}_{\mathcal{K}}).$$

Most interesting is, of course, the intermediate tradeoff between these extremes. In the following sections, we study more general coding strategies and characterize a region of achievable rate tuples.

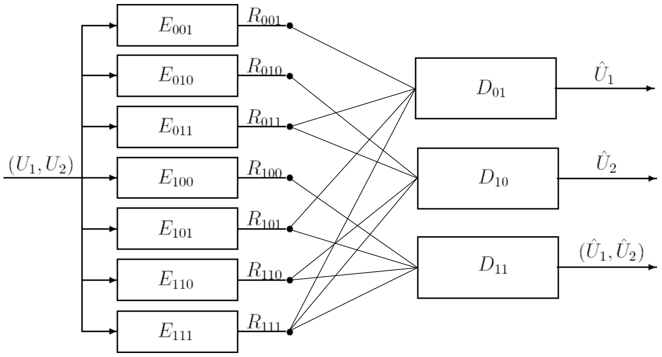

Fig. 2. General Storage Scheme for a Two-Source

III. The Shared Descriptions Problem

We transform the storage problem into a source coding problem with multiple encoders as follows. Since the query q can be one of $2^K - 1$ possibilities, there are $J \triangleq 2^K - 1$ decoders. For any code $(f, g, h_q, q \subset \mathcal{K})$, each of the M_s bits can be accessed by any subset of the decoders. Therefore each bit position could be indexed to identify the corresponding one of the subsets of $\{1, \ldots, J\}$. Alternatively, we can divide the total bit rate R_s into $2^D - 1$ components and can view the coding system as consisting of $2^J - 1$ encoders, one for each non-empty subset of the decoders. For $K = 2$ (a two-source), this system is depicted in Fig. 2. The binary subscript on the decoders indicate the sources that they need to reconstruct. Similarly, the binary subscript on the encoders indicate the decoders that they service.

The basic multiple encoder scenario, which we term the shared descriptions (SD) scenario consists of a set of encoders $\{E_i, i \in \Sigma\}$ $(|\Sigma| = I)$ and a set of decoders $\{D_j, j \in \Delta\}(|\Delta| = J)$. There are J discrete memoryless correlated sources $(X_j, j \in \Delta)$, each of which is associated with a distinct decoder. For every subset of decoders $i \subset \Delta$, there is some encoder E_i that transmits information at rate R_i to exactly those decoders. Each of the $I = 2^J - 1$ co-located encoders observes all the J sources. Let Σ_j denote the set of encoders that service a particular decoder j. We are interested in characterizing the set of rate tuples $(R_i, i \in \Sigma)$ that would allow all decoders to reconstruct their corresponding sources noiselessly in the Shannon sense. The SD problem with two decoders is represented in Fig. 3. This special case has been considered earlier by Gray and Wyner [4]

The asymptotically achievable rate region for the SD problem remains unchanged even if we constrain the encoders to

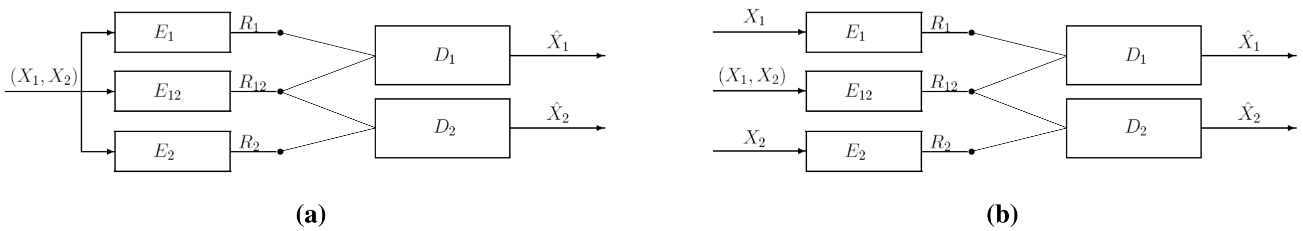

Fig. 3. (a) An SD scenario with two sources, (b) An equivalent multiterminal scenario with modified encoder inputs

be independent since all encoders observe the same source. This enables us to use known results from multiterminal source coding to obtain a non-single letter characterization of the rate region. In the example in Fig. 3, of the three encoders, $E_{\{1\}}$ and $E_{\{2\}}$ service a single decoder. We claim that each such encoder need only access that source that needs to be reconstructed at its corresponding decoder. This is a consequence of the following lemma for general multiterminal source coding systems. A multiterminal source coding system consists of a set of *independent* encoders $\{E_i, i \in \Sigma\}$ and a set of decoders $\{D_j, j \in \Delta\}$. Each encoder i observes a set of correlated sources. Each decoder j is connected to a subset of the encoders $\{E_i, i \in \Sigma_j \subset \Sigma\}$ and wishes to reconstruct some subset of the sources observed by the encoders corresponding to Σ_j. Let J, I and I_j denote the cardinalities of Δ, Σ and Σ_j respectively. Achievable rates for this network are defined in the usual manner (see [1]).

Lemma 1: Consider an arbitrary multiterminal source coding network. Suppose there exists an encoding terminal that observes a correlated pair of discrete memoryless sources with generic random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ and the sole decoder that it services needs to losslessly reconstruct only the X process. Then the region of achievable rates would not change were the encoder to only depend on the X process.

Proof: Let $(R_i, i \in \Sigma)$ be an achievable rate for this network. Let encoder 1 and decoder 1 be as in the statement of the lemma. We wish to show that for every $\epsilon > 0$, there exists encoders $(f_i, i \in \Sigma)$ at block length n that allow decoding at each decoder with probability of error less than ϵ, $\frac{1}{n}\log\|f_i\| \leq R_i + \epsilon$ and where f_1 is solely a function of the X sequence. Since $(R_i, i \in \Sigma)$ is an achievable rate, for every $\epsilon' > 0$, there exists a set of encoders $(f'_i, i \in \Sigma)$ at block length n' such that $\frac{1}{n'}\log\|f'_i\| \leq R_i + \epsilon'$ and the probability of decoding error at each decoder is less than ϵ' (f_1 is possibly a function of the Y sequence as well). Let $V \in \mathcal{V}$ be the random variable denoting the codewords from all encoders except f'_1 that decoder 1 has observed. The converse part of the Slepian-Wolf theorem [3] tells us that

$$n'(R_1 + \epsilon') \geq H(X^n|V) - \epsilon'[\log|\mathcal{X}^n| + \log|\mathcal{V}|] - 1. \quad (3)$$

On the other hand, from the direct half of the Slepian-Wolf theorem, for all $\epsilon'' > 0$ there exists a supercode that encodes m blocks of the X sequence of length n' each using a binning scheme such that the probability of reconstruction error is less than ϵ'' and the rate per n'-length block $\tilde{R}_1$ is less than $H(X^{n'}|V) + \epsilon''$. Choosing $\epsilon'' < \epsilon'$, we therefore have a new code with block length mn' where the rate for encoder 1 is

$$R'_1 \leq \frac{1}{n'}(H(X^{n'}|V) + \epsilon''), \quad (4)$$

and all the other rates remain unchanged. Note that the behavior at the other decoders also remains unchanged. The probability of error for the new code is therefore less than ϵ' at each decoder, while the first encoder only needs to observe the X sequence. Comparing the new rate with the old one, we have

$$\begin{aligned} R'_1 - R_1 &\leq \frac{1}{n'}(\epsilon'' + \epsilon'[\log|\mathcal{X}^{n'}| + \log|\mathcal{V}|] + 1) + \epsilon' \\ &\leq \frac{1}{n'} + \epsilon'[\log|\mathcal{X}| + \sum_{i>1} R_i + 1 + \frac{1}{n'}]. \quad (5) \end{aligned}$$

By choosing ϵ' sufficiently small and n' sufficiently large, we can make the right hand side of (5) less than the given ϵ, thus obtaining a code of the type we wanted to construct. This proves the lemma. ■

Remark 1: The phenomenon described in Lemma 1 had been observed earlier for certain special cases, most notably the Slepian-Wolf case [3] and the Gray-Wyner case [4], [5].

Lemma 1 enables us to modify the inputs of encoders that only service a single decoder to be that subset of the input data that needs to be reconstructed at the decoder. For the two source SD problem, the input to $E_{\{1\}}$ and $E_{\{2\}}$ can be assumed to be the X_1 and X_2 sequences respectively (see Fig. 3 (b)). We are now left with a coding problem that has been considered earlier in [1] (see also [2]): For each decoder D_j, let μ_j denote the index of the encoder that solely services D_j (and whose input the decoder needs to reconstruct). For each encoder E_i, let $\mathbf{X}_i$ denote the source that is input to E_i in the modified coding system. Let $\tilde{\mu} \triangleq \{\mu_j, j \in \Delta\}$. In the two-source case (Fig. 3 (b)), we have $\Sigma = \{\{1\}, \{2\}, \{1, 2\}\}$, $\tilde{\mu} = \{\{1\}, \{2\}\}$, and for example, $\Sigma_1 = \{\{1\}, \{1, 2\}\}$ and $\mu_1 = \{1\}$.

A non-single letter characterization of the entire rate region is now possible by applying the results of [1] (Theorem 5 in [2] provides an alternate, but equivalent characterization of the achievable rate region). For some positive integer n, define auxiliary finite random variables $V_i, i \in \Sigma$ such that

1) The random variables $V_i, i \in \Sigma$ are conditionally independent given $\mathbf{X}^n_\Delta$.
2) For each $i \in \tilde{\mu}$, the conditional distribution of V_i given $\mathbf{X}^n_\Delta$ depends only on $\mathbf{X}^n_i$.

For every such set of auxiliary random variables, define $\mathcal{R}^n_{HK,\text{Total}}(V_\Sigma)$, where the subscript stands for "Han-Kobayashi", as the set of all rate tuples $(R_i, i \in \Sigma)$ satisfying for each $j \in \Delta$ and each $S \subset \Sigma_j$

$$\sum_{i \in S} R_i \quad \geq \quad \frac{1}{n} I(V_S; \mathbf{X}^n_\Delta | V_{\Sigma_j \setminus S}) + \psi^n_j(S), \tag{6}$$

where

$$\psi^n_j(S) \triangleq \begin{cases} H(\mathbf{X}^n_{\mu_j} | V_{\Sigma_j}) & \text{if } S \ni \mu_j \\ 0 & \text{Otherwise.} \end{cases}$$

The entire achievable rate region

$$\mathcal{R}^*_{HK,\text{Total}} = \text{ closure of } \cup_{n=1}^{\infty} \cup_{V_\Sigma} \mathcal{R}^n_{HK,\text{Total}}(V_\Sigma). \tag{7}$$

Han and Kobayashi also present a single letter partial achievable rate region. However, their result assumes independent encoders, which is an unnecessary restriction for the SD problem. We present a more general rate region by relating the SD problem to the multiple descriptions (MD) problem [6], [7]. In the MD scenario, there are I encoders, all of which observe the same source and $2^I - 1$ decoders. For every nonempty subset of the encoders, there is a decoder that observes the output of exactly those encoders. The objective is to perform a lossy encoding of the common source such that the reconstruction quality improves as the number of encoder outputs observed at a decoder increases. Although the complete MD achievable region for general sources is unknown, some partial achievable rate regions have been characterized all of which rely on encoding schemes that transmit some set of auxiliary random variables to the decoders. The SD problem is in a dual of the MD problem in the following sense: in the MD problem I encoders communicate with $2^I - 1$ decoders, while in the SD problem, $2^J - 1$ encoders communicate with J decoders. However, we can employ the MD strategy of communicating auxiliary random variables correlated with the source to the decoder for the SD problem. In applying results for the MD problem to the SD problem, we impose the constraint that only the MD decoders corresponding to the SD decoders are active. Since only a subset of the MD decoders is active, we can enlarge the rate region given by the MD strategy by using binning schemes.

For the SD scenario, let $\Sigma_v = \cup_j 2^{\Sigma_j}$. Define $\mathcal{R}^*_{\text{ach}}$ to be the closure of the set of I tuples $(R_i, i \in \Sigma)$ such that there exist both an auxiliary rate tuple $(R'_i, i \in \Sigma)$ and a set of auxiliary finite random variables $(W_S, S \in \Sigma_v)$ satisfying

1) For all $S \subset \Sigma, S \neq \emptyset$,

$$\sum_{i \in S} R'_i \geq \phi_j(S) - H(W_{2^S \cap \Sigma_v} | X_\Delta) + \sum_{\substack{S' \subseteq S \\ S' \in \Sigma_v}} H(W_{S'} | W_{2^{S'} \setminus \{S'\}}), \tag{8}$$

where

$$\phi_j(S) \triangleq \begin{cases} 0 & \text{if } S \ni \Delta \\ -I(W_\emptyset; \mathbf{X}_\Delta) & \text{otherwise.} \end{cases}$$

2) For all $j, S \subset \Sigma_j, S \neq \emptyset$

$$\sum_{i \in S} R_i \geq \sum_{i \in S} R'_i - \sum_{\substack{S' \in \Sigma_j \\ S' \not\subseteq \Sigma_j \setminus S}} H(W_{S'} | W_{2^{S'} \setminus \{S'\}}) + H(W_{2^{\Sigma_j} \setminus 2^{(\Sigma_j \setminus S)}} | W_{2^{(\Sigma_j \setminus S)}}) + \psi_j(S), \tag{9}$$

where

$$\psi_j(S) \triangleq \begin{cases} H(\mathbf{X}_{\mu_j} | W_{2^{\Sigma_j}}) & \text{if } S \ni \mu_j \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 1: All rate tuples in $\mathcal{R}^*_{\text{ach}}$ are achievable.

Proof: (Sketch) Fix $\epsilon > 0$. Let $(R_i, i \in \Sigma) \in \mathcal{R}_{\text{ach}}$ and $(R'_i, i \in \Sigma)$ and $(W_S, S \in \Sigma_v)$ be the associated auxiliary rate tuple and auxiliary random variables. Define $R''_{\mu_j} = R_{\mu_j} - \psi_j(\{\mu_j\}) + \frac{\epsilon}{2}, \forall j \in \Delta$ and $R''_i = R_i + \epsilon$ otherwise. We now describe a code at block length n.

Codebook Design: For every $\mathbf{c}' = (c'_i, i \in \Sigma), 1 \leq c'_i \leq 2^{n(R'_i + \epsilon')}$ and every $S \in \Sigma_v$, associate a vector $\mathbf{W}^n_S(\mathbf{c}'_S)$. The components of $\mathbf{W}^n_S(\mathbf{c}'_S) = (W^{(1)}_S, \ldots, W^{(t)}_S, \ldots, W^{(n)}_S)$ at every instant $1 \leq t \leq n$ are drawn independently from the distribution $p_{W_S | W_{2^S \setminus \{S\}}}(\cdot | W^{(t)}_{2^S \setminus \{S\}}(\mathbf{c}'_S))$. Once the codebook has been constructed, for all $i \in \Sigma$ distribute $1 \leq c'_i \leq 2^{n(R'_i + \epsilon')}$ uniformly among $2^{nR''_i}$ bins indexed by $\mathbf{c} = (c_i, i \in \Sigma), 1 \leq c_i \leq 2^{nR''_i}$. Each bin contains $2^{n(R'_i - R''_i + \epsilon')}$ elements. Finally, for every $j \in \Delta$ assign the elements of $\mathcal{X}^n_j$ uniformly at random into one of $2^{n(\psi_j(\{\mu_j\}) + \frac{\epsilon}{2})}$ bins uniformly at random.

Encoding: An observation $\mathbf{X}^n_\Delta$ is encoded in two steps

1) Find the $\mathbf{c}'$ such that $(\mathbf{X}^n_\Delta, \mathbf{W}^n_\Sigma(\mathbf{c}'))$ belongs to the typical set (The typical set at block length n with respect to some probability distribution P is the set of all vectors of length n over the alphabet of P whose empirical distribution is close to P. See [8] for a formal definition and a discussion of properties). If there is no such $\mathbf{c}'$, set $\mathbf{c}' = (0, \ldots, 0)$. Encoder i outputs the bin index vector $\mathbf{c}$ corresponding to $\mathbf{c}$.
2) For every $j \in \Delta$, encoder E_{μ_j} transmits the bin index of $\mathbf{X}^n_{\mu_j}$.

Note that the total rate from encoder i is $R_i + \epsilon$.

Decoding: Using the bin indices $\mathbf{c}_{\Sigma_j}$ output after the first encoding step, decoder j finds the unique $\hat{\mathbf{c}}'_{\Sigma_j}$ from the bins corresponding to $\mathbf{c}_{\Sigma_j}$ such that $\mathbf{W}_{2^{\Sigma_j}}(\hat{\mathbf{c}}'_{\Sigma_j})$ is jointly typical. The decoder declares an error if there are no jointly typical sequences or more than one such sequence. If the first decoding step is successful, the decoder tries to find the $\mathbf{X}^n_{\mu_j}$ that is jointly typical with $\mathbf{W}_{2^{\Sigma_j}}(\hat{\mathbf{c}}'_{\Sigma_j})$. From the results of Slepian and Wolf [3] and the choice of the number of bins in the second encoding step, the second decoding step is successful with high probability conditioned on the success of the first step if the block length n is large enough. So to prove that the entire encoding scheme is successful, we need to show that the first decoding step succeeds with high probability.

The first decoding step fails if one of the following occurs:

a) $\mathbf{X}^n_\Sigma$ is not in the typical set. From the law of large numbers, this happens with small probability if n is large.

b) Given that $\mathbf{X}_\Sigma^n$ is in the typical set, the first encoding step fails, that is there is no $\mathbf{c}'$ such that $(\mathbf{X}_\Delta^n, \mathbf{W}_\Sigma^n(\mathbf{c}'))$ belongs to the typical set. Since the $(R_i', i \in \Sigma)$ satisfy equation (8), this event occurs with negligible probability at large n using the proof of Theorem 1 in [7]. Since all decoders receive the output of decoder E_Σ, unlike in [7], the information corresponding to $W_\emptyset$ can be included solely in the output of encoder E_Σ, which leads to an enlargement of the rate region for a given set of auxiliary random variables.

c) Given that the above two events have not occurred, for some $j \in \Delta$, there is more than one jointly typical $\hat{W}_{2^{\Sigma_j}}$ that belongs to the bin indexed by $\mathbf{c}_{\Sigma_j}$. Using Slepian-Wolf type arguments (see the proof of Theorem 1 in [1]), it can be shown that this happens with low probability if equation (9) is satisfied. A key component in applying the argument to the SD scenario is the fact that for a given j if two codeword index subvectors $\mathbf{c}'_{\Sigma_j,1}$ and $\mathbf{c}'_{\Sigma_j,2}$ do not match at the coordinates corresponding to some set $S \subset \Sigma_j$, the codewords for the two index subvectors corresponding to subsets of $\Sigma_j \setminus S$ will match while all other codewords will be conditionally independent.

We have therefore shown as required that for any given ϵ, if the block length is sufficiently large, then there is a code with rates $(R_i + \epsilon, i \in \Sigma)$ if $(R_i, i \in \Sigma) \in \mathcal{R}^*_{\text{ach}}$. ■

Since we employ a binning strategy in addition to the MD-like strategy, $\mathcal{R}^*_{\text{ach}}$ is certainly at least as large than the rate region obtained by using a pure MD-like strategy. If, on the other hand, we consider only those cases where the auxiliary random variables $W_S, |S| = 1$ are conditionally independent given X_Σ, while the rest of the auxiliary random variables are constants, we obtain the achievable region characterized by Han and Kobayashi for the multiterminal source coding problem. Therefore $\mathcal{R}^*_{\text{ach}}$ is potentially larger than the region predicted by the constituent strategies.

For the two source case, if all auxiliary random variables other than $W_{\{\{1,2\}\}}$ are constants, $R'_{\{1\}} = R'_{\{2\}} = 0$ and $R'_{\{1,2\}} = I(W_{\{\{1,2\}\}}; X_{1,2})$, both (8) and (9) are satisfied by $(R_{\{1\}}, R_{\{2\}}, R_{\{1,2\}}) = (H(X_1|W_{\{\{1,2\}\}}), H(X_2|W_{\{\{1,2\}\}}), I(W_{\{\{1,2\}\}}; X_{1,2}))$. The region of rates obtained by varying $W_{\{\{1,2\}\}}$ is known to be the entire rate region for this problem [4]. Hence, for the two source SD problem, $\mathcal{R}^*_{\text{ach}}$ is the entire achievable rate region.

IV. The storage vs. retrieval tradeoff

As described earlier, the storage and retrieval problem is equivalent to an SD problem where $\Delta = 2^{\mathcal{K}}$ and $X_j = (U_k, k \in j)$. Therefore, the achievable rate region for the SD scenario $\mathcal{R}^*_{HK,\text{Total}}$ contains all the information required for characterizing the rate region for the storage problem. From the way we defined the SD problem from the storage problem, it is apparent that the rate region of our interest $\mathcal{S}^*$ is a section of $\mathcal{R}^*_{HK,\text{Total}}$ where the sum of the rate tuples does not exceed R_s. Therefore the minimum retrieval cost at a given storage cost is given by

$$C(R_s) = \min_{(R_i, i \in \Sigma) \in \mathcal{R}^*_{HK,\text{Total}}} \{ \sum_{q \subset \mathcal{K}} P(q) \sum_{i \in \Sigma_q} R_i : \sum_{i \in \Sigma} R_i \le R_s \} \quad (10)$$

We can obtain a computable upper bound on $C(R_s)$ by replacing $\mathcal{R}^*_{HK,\text{Total}}$ by $\mathcal{R}^*_{\text{ach}}$ in equation (10).

Observe that for a given number of data streams K, the optimal number of decoders is exponential in K (specifically $2^K - 1$), and the number of encoders is doubly exponential in K (i.e., $2^{2^K - 1} - 1$). This means that the computational complexity of the optimal system becomes unbearably high even for moderately large K. It is of interest to incorporate some means for performance-complexity tradeoff. While the number of decoders cannot be modified without altering the problem setting altogether, we can impose a complexity constraint on the encoding side by restricting the number of encoders to L. $\mathcal{R}^*_{HK,\text{Total}}$ is sufficient to characterize the retrieval cost vs storage cost tradeoff for this case as well. The minimum retrieval cost at a given storage cost and for a given number of encoders is:

$$C(R_s, L) = \min_{(R_i, i \in \Sigma) \in \mathcal{R}^*_{HK,\text{Total}}} \{ \sum_{q \subset \mathcal{K}} \sum_{i \in \Sigma_q} P(q) R_i : \sum_{i \in \Sigma} R_i \le R_s \quad |\{i : R_i > 0\}| \le L \}.$$

In the two-source example, if $R_s = H(U_1, U_2)$ and $L = 2$, the minimum achievable retrieval cost is

$$C(H(U_1, U_2), 2) = C_{\min} + \min[P(\{1\})H(U_2|U_1), P(\{2\})H(U_1|U_2)],$$

where $C_{\min} \triangleq \sum_{q \subset \mathcal{K}} P(q) H(\mathbf{U}_q)$ is the minimum retrieval cost when there are no storage or complexity constraints.

References

[1] T. S. Han and K. Kobayashi, "A unified achievable rate region for a general class of multiterminal source coding systems," *IEEE Trans. on Information Theory*, vol. 26, no. 3, pp. 277–288, May 1980.

[2] I. Csiszár and J. Körner, "Towards a general theory of source networks,"*IEEE Trans. on Information Theory*, vol. 26, no. 2, pp. 155–165 Mar 1980.

[3] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Information Theory*, vol. 19, no. 4, pp. 471–80, Jul. 1973.

[4] R. M. Gray and A. D. Wyner, "Source coding for a simple network," *Bell Syst. Tech. J.*, vol. 53, no. 9, pp. 1681–1721, Nov. 1974.

[5] A. D. Wyner, "On source coding with side-information at the decoder," *IEEE Trans. on Information Theory*, vol. 21, no. 3, pp. 294–300, May 1975.

[6] A. A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. on Information Theory*, vol. 28, no. 6, pp. 851–857, Nov. 1982.

[7] R. Venkataramani, G. Kramer, V. K. Goyal, "Multiple descriptions coding with many channels," *IEEE Trans. on Information Theory*, vol. 49, no. 9, pp. 2106–2114, Sep. 2003.

[8] I. Csiszár and J. Körner, *Information theory: Coding theorems for discrete memoryless systems*. Academic Press, New York, 1982.