Searchable Compression

Thomas R. Fischer
School of Electrical Engineering and Computer Science
Washington State University
Pullman, WA 99163, USA
Email: fischer@eecs.wsu.edu

***Abstract*— We are interested in structuring data compression to support efficient searching of information in the compressed data domain. Specifically, by "searchable" we mean a layered coding such that certain information can be extracted from the compressed bit-stream without requiring complete decoding of all compressed information. The chain rule for entropy provides a natural formulation for describing such a property for the case of a single discrete information source. The distributed lossless coding of correlated information sources is characterized by the Slepian-Wolf admissible rate region. For the problem of distributed coding of correlated information sources subject to our "searchability" constraint, we characterize the admissible rate region and demonstrate that it shares a region boundary in common with a portion of the boundary of the Slepian-Wolf admissible rate region.**

I. Introduction

The goal of data compression is to efficiently encode a source for transmission or storage. Compression efficiency is typically measured with respect to the source entropy rate for lossless compression, or to the source rate-distortion function for source coding subject to a fidelity criterion. Modern applications, however, tend to require additional functionality beyond traditional coding efficiency. For example, the JPEG2000 standard [1] for image compression supports manipulation of data *in the compressed domain* to allow region of interest coding, or data transmission/decoding that is progressive in signal-to-noise ratio or image size.

One important goal of information technology development is scalability. In the present context scalability refers to methods and algorithms that scale well with bandwidth, encoding rate, computational power, memory, or data size. The vast amounts of data that are acquired, archived, accessed, or shared, motivate information processing methods that are scalable. One way in which compression methods can scale well with data size is by supporting access to and manipulation of data in the compressed domain.

In this paper the terminology "searchable" compression is used loosely to describe compression methods that support data search. We are interested in three main problems related to searchable compression. First is the problem of structuring compression methods to support data search in the compressed data domain without requiring a complete decompression of all of the data before search. Second is the problem of what to search for - for example, in multimedia this is the problem of semantic retrieval of information, e.g., [2]. Third, and perhaps most fundamental, is the problem of relating the notions of semantics and meaning of data to the concepts of entropy and mutual information. This paper focuses on the first of these problems.

II. Basic Formulation

Let $\mathbf{X} = (X_1, X_2)$ be a discrete-time source of finite entropy. The chain rule for entropy [3] states that the joint entropy can be expressed as

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1). \tag{1}$$

So, in principle, there is no loss of coding efficiency in separately compressing X_1, and then compressing X_2 conditioned on X_1. This provides a natural formulation for combining efficient compression with searchability. Specifically, interpret X_1 as the information in the source that is used for search. Let B_1 denote the compressed version of X_1, and B_2 the compressed version of X_2 conditioned on X_1. In the compressed domain, B_1, B_2 represent all of the source data. However, B_1 can be readily extracted from B_1, B_2, and used alone to access X_1, which in turn can be used for data search. For searchable compression we thus require the decoding property that X_1 can be decoded independently of the encoded information about X_2. Such application of the chain rule for entropy is straightforward and provides a theoretical basis for efficient coding that also supports other decoding properties, such as progressive decoding or bit-plane decoding.

III. Distributed Coding

The problem becomes more interesting when one considers the distributed coding of correlated sources, generally known as Slepian-Wolf coding [4]. Let X and Y be two discrete, correlated information sources, to be separately encoded and jointly decoded, as illustrated in Figure 1. The admissible-rate region for such a situation is shown in Figure 2 [4]. The corner points of the admissible rate region are of special interest, namely point a

$$R_X^a = H(X|Y), \quad R_Y^a = H(Y), \tag{2}$$

and point b

$$R_X^b = H(X), \quad R_Y^b = H(Y|X). \tag{3}$$

These points lie along the line $R_X + R_Y = H(X, Y)$, with $H(X, Y)$ the minimum total rate necessary to reliably transmit both X and Y.

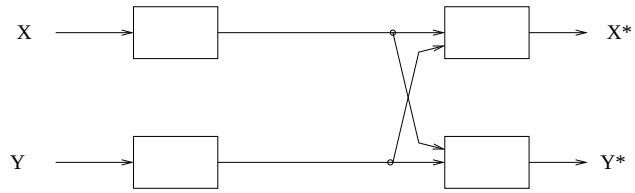

Fig. 1. Distributed coding of X and Y with joint decoding.

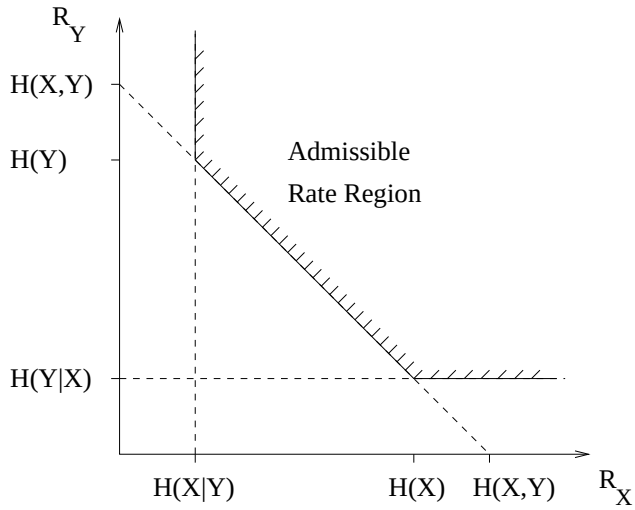

Fig. 2. Slepian-Wolf admissible rate region for distributed lossless coding of X and Y.

Now, slightly generalize the problem by allowing the two sources to be vector valued, so that $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (Y_1, Y_2)$. The corner points of the Slepian-Wolf admissible rate region are then given by point a

$$R_X^a = H(\mathbf{X}|\mathbf{Y}), \quad R_Y^a = H(\mathbf{Y}), \tag{4}$$

and point b

$$R_X^b = H(\mathbf{X}), \quad R_Y^b = H(\mathbf{Y}|\mathbf{X}). \tag{5}$$

Using the chain rule for entropy, corner point a can be rewritten as

$$\begin{aligned} R_X^a &= H(X_1|Y_1, Y_2) + H(X_2|Y_1, Y_2, X_1), \\ R_Y^a &= H(Y_1) + H(Y_2|Y_1). \end{aligned} \tag{6}$$

The admissible rate region is shown in Figure 3.

Let X_1 and Y_1 be interpreted as the portions of the data that are to be searchable. Then the structural property imposed on the decoder is that X_1 and Y_1 must be decodable without reference to the encoded information for X_2 or Y_2. The expression in (6) for R_Y^a satisfies this condition, but the expression for R_X^a does not, because of the conditioning on Y_2. Hence, the corners of the Slepian-Wolf admissible rate region in Figure 3 are generally not admissible with the additional searchability constraint.

Figure 4 describes the general distributed coding problem with the searchability constraint. A vector source generates the pair X_1, X_2. A different vector source generates the pair

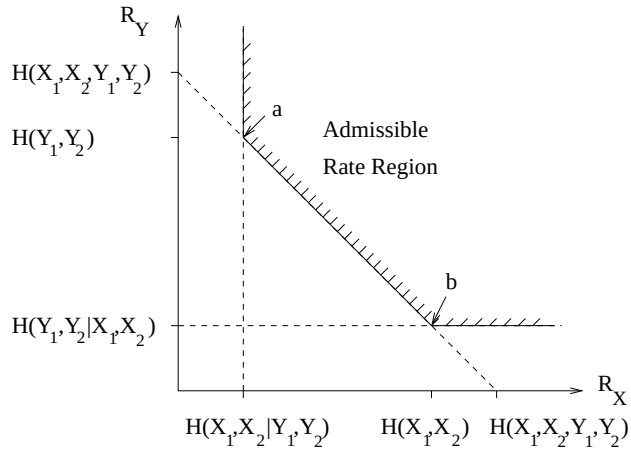

Fig. 3. Admissible rate region for the Slepian-Wolf coding of $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (Y_1, Y_2)$.

Y_1, Y_2. It is convenient to use the Slepian-Wolf distributed encoding model for encoding each pair. The decoder is structured to satisfy the searchability constraint. Hence it jointly has access only to the encoded versions of X_1 and Y_1 to decode X_1^* and Y_1^*. The decoder has access to all received information to decode X_2^* and Y_2^*.

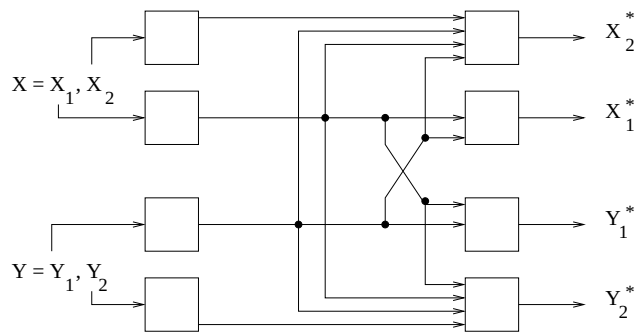

Fig. 4. Distributed source coding with searchability constraint.

Theorem The corner points of the admissible rate region for the distributed source coding problem described by Figure 4 are given by the points a' and b'

$$\begin{aligned} R_X^{a'} &= H(X_1|Y_1) + H(X_2|Y_1, Y_2, X_1), \\ R_Y^{a'} &= H(Y_1) + H(Y_2|Y_1, X_1), \end{aligned} \tag{7}$$

and point b

$$\begin{aligned} R_X^{b'} &= H(X_1) + H(X_2|X_1, Y_1), \\ R_Y^{b'} &= H(Y_1|X_1) + H(Y_2|X_1, Y_1, X_2). \end{aligned} \tag{8}$$

Proof The theorem can be proven using a random coding argument similar to that in [4].

Comments

1) Using the chain rule of entropy it follows that

$$\begin{aligned} H(X_1, X_2, Y_1, Y_2) \quad = \quad & H(Y_1) + H(X_1|Y_1) + \\ & H(Y_2|Y_1, X_1) + \\ & H(X_2|Y_1, X_1, Y_2) \end{aligned} \tag{9}$$

so it is evident that the points $(R_X^{a'}, R_Y^{a'})$ and $(R_X^{b'}, R_Y^{b'})$ are along the line $R_X + R_Y = H(X_1, X_2, Y_1, Y_2)$. Comparing (7) to (6) it is also evident that $R_Y^a \geq R_Y^{a'}$ and $R_X^a \leq R_X^{a'}$ so that the admissible rate region for the distributed source coding problem subject to the searchability constraint is a subset of the admissible rate region for the unconstrained problem. This is illustrated in Figure 5.

2) To establish the corner points a and b of the admissible rate region, Slepian and Wolf use a random coding argument. The line segment connecting the two points is then established by time sharing. A random coding argument can also be used to establish the corner points a' and b' of the admissible rate region subject to the searchability constraint. A time sharing argument can then be used to establish the points along the line segment connecting them. The codes used to establish the points a' and b' are, however, generally different than those used to establish the points a and b. Hence, the theorem establishes a different set of codes that, through time sharing, can cover the entire subset of the unconstrained distributed source coding problem lying along the line segment between the points a' and b'.

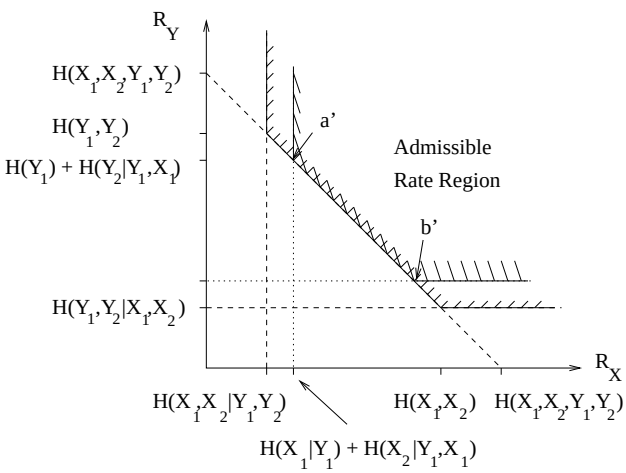

Fig. 5. Admissible rate region with searchability constraint.

IV. Conclusion

This paper has considered the problem of structuring data compression so that selected information can be separately decoded to support data search. Such layered coding can be thought of as a direct application of the chain rule of entropy. In the case of distributed source coding subject to the searchability constraint, the admissible rate region for the constrained problem is smaller than the Slepian-Wolf admissible rate region for the unconstrained problem, but shares a portion of the admissible rate region boundary for the unconstrained problem.

References

[1] ISO/IEC 15444-1. JPEG 2000 image coding system, 2000.

[2] D. Androutsous, G. Ling, A. N. Venetsanopoulos, Guest Editors, "Semantic retrieval of multimedia," *IEEE Signal Processing Magazine*, vol. 23, March 2006.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, 2nd Edition, 2007.

[4] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Th.*, vol. IT-19, pp. 471-480, July 1973.