On Coding for Data Analytics: New Information Distances

En-hui Yang

***Abstract*—Distance plays a vital role in many applications of data analytics. In this paper, the concept of distance between any two data objects X and Y is addressed from the perspective of Shannon information theory. Consider a coding paradigm where X and Y are encoded into a sequence of coded bits specifying a codeword (or method) which would in turn convert Y into $\hat{X}$, and X into $\hat{Y}$ such that both the distortion between X and $\hat{X}$ and the distortion between Y and $\hat{Y}$ are less than or equal to a prescribed threshold D. Given a class $\mathcal{C}$ of coding schemes within the coding paradigm, the information distance $R_{\mathcal{C}}(X,Y,D)$ between X and Y at the distortion level D is defined as the smallest number of coded bits afforded by coding schemes from $\mathcal{C}$. For two important classes $\mathcal{C}$, $R_{\mathcal{C}}(X,Y,D)$ is shown to be indeed a pseudo distance in some sense; it is further characterized or bounded. When $\mathcal{C}$ is the class of so-called separately precoded broadcast codes, it is shown that for any stationary, totally ergodic sources X and Y, $R_{\mathcal{C}}(X,Y,D)$ is equal to the maximum of the Wyner-Ziv coding rate of X with Y as side information and the Wyner-Ziv coding rate of Y with X as side information. In the general case where $\mathcal{C}$ consists of all codes within the coding paradigm, upper and lower bounds to $R_{\mathcal{C}}(X,Y,D)$ are established, and are further shown to be tight when X and Y are jointly Gaussian. The distance $R_{\mathcal{C}}(X,Y,D)$ generalizes the notion of information distance defined within the framework of Kolmogorov complexity.**

I. Introduction

Distance plays an important role in many applications of data analytics. For instance, in image retrieval, organization, and management, one needs a proper similarity distance to measure the perceptual similarity between any two images X and Y. Once such a similarity distance is defined for any two images, it could be used to retrieve images in a database which are perceptually similar to a query image according to the similarity distance in image retrieval [1], and to organize images into different groups according to their mutual similarity in image management [2]. Likewise, in data clustering and bioinformatics, the notion of distance also plays a dominant role [3].

In the literature of image retrieval [1], a typical approach to determining a perceptual distance between two images is to first extract features from each image, then derive a signature of each image from its respectively extracted features, and finally determine the perceptual distance based on their respective signatures. Euclidean distance, Hausdorff distance, Kullback-Leibler divergence, etc. have all been used as a distance between signatures [1]. The variation in feature extraction, signature derivation, and distance between signatures leads to many different image perceptual distances. In general, however, as one moves from original images to features to signatures, the notion of distance becomes less intuitive and is increasingly disconnected from the original images.

To alleviate this issue, a different approach was taken recently in [4]. Instead of extracting each image into its signatures, the paper [4] first expanded each image X conceptually into a set $\phi(X)$ of images, which may contain images perceptually similar to X, and then defined the perceptual distance between X and Y as the smallest average distortion per pixel between any pair of images, one from $\phi(X)$ and the other from $\phi(Y)$. The resulting distance is dubbed SMID and denoted by $d_\phi(X,Y)$. It was demonstrated in [4] that when compared with other standard perceptual distances reported in the literature [5]-[12], SMID indeed shows better discriminating power on image similarity. An interesting property relevant to our discussion in this paper is that the optimization solution in SMID $d_\phi(X,Y)$ gives a method which converts X to $\hat{Y}$, and Y to $\hat{X}$ such that $d_\phi(X,Y)$ is equal to the average distortion per pixel between X and $\hat{X}$, i.e., $d(X,\hat{X})$, and between Y and $\hat{Y}$, i.e., $d(Y,\hat{Y})$. Nonetheless, the descriptive complexity of the conversion method is completely ignored in SMID $d_\phi(X,Y)$.

Based on descriptive complexity, particularly Kolmogorov complexity [13], [14], the notion of information distance was proposed in [15] for discrete data objects such as strings over a finite alphabet. Give any two finite strings x and y, their information distance $E_0(x,y)$ was defined in [15] to be the length of the shortest program which, when running on a universal computer (i.e., Turing machine), will convert x into y when x is the input, and convert y into x when y is the input. An inspiring property of $E_0(x,y)$ is its universality [15], which says in theory $E_0(x,y)$ captures all patterns and regularities that can be utilized computationally and are shared by x and y, and hence is the best cognitive distance one could hope for to certain extent. In [3] and references therein, this notion was successfully applied to bioinformatics, music clustering, and machine translation.

However, the information distance as defined in [15] has two major issues. First, since it is based on Kolmogorov complexity, it is uncomputable. Second, more importantly it is not applicable to continuous valued data such as images & videos. Therefore, it is desirable to develop a notion of distance which could combine the best of both worlds: the universality from the information distance as defined in [15], and the computability and applicability to both discrete and

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN203035-11, and by the Canada Research Chairs Program.

En-hui Yang is with the Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (Email: ehyang@uwaterloo.ca).

continuous valued data as in SMID $d_\phi(X,Y)$.

In this paper, we address the notion of distance between any two data objects X and Y (continuous or discrete) from the perspective of Shannon information theory. We bring distortion into the information distance $E_0(x,y)$, and descriptive complexity into SMID $d_\phi(X,Y)$. To this end, we formulate a new coding paradigm where X and Y are encoded into a sequence of coded bits specifying a codeword (or method) which would in turn convert Y into $\hat{X}$, and X into $\hat{Y}$ such that both the distortion between X and $\hat{X}$ and the distortion between Y and $\hat{Y}$ are less than or equal to a prescribed threshold D. To have universality to some extent, we consider a class $\mathcal{C}$ of coding schemes within the coding paradigm. Given $\mathcal{C}$, the information distance $R_{\mathcal{C}}(X,Y,D)$ between X and Y at the distortion level D is then defined as the smallest number of coded bits afforded by coding schemes from $\mathcal{C}$. We then characterize and analyze the information distance $R_{\mathcal{C}}(X,Y,D)$ for some classes $\mathcal{C}$.

The rest of the paper is organized as follows. In Section II, we formally formulate the new coding paradigm and define the information distance $R_{\mathcal{C}}(X,Y,D)$. In Section III, we analyze the distance property of $R_{\mathcal{C}}(X,Y,D)$ when $\mathcal{C}$ consists of all coding schemes allowed in the coding paradigm, and establish upper and lower bounds to $R_{\mathcal{C}}(X,Y,D)$, which are further shown to be tight when X and Y are jointly Gaussian. In Section IV, $R_{\mathcal{C}}(X,Y,D)$ is characterized in terms of the Wyner-Ziv coding rate of X with Y as side information and the Wyner-Ziv coding rate of Y with X as side information when $\mathcal{C}$ consists only of all so-called separately precoded broadcast codes within the coding paradigm; its distance property among different sources is also presented.

II. Formal Definitions: Codes and New Information Distances

Let $\mathbf{A}$ and $\hat{\mathbf{A}}$ be two abstract alphabets. They could be either continuous or discrete. The sets $\mathbf{A}$ and $\hat{\mathbf{A}}$ will serve as our source alphabet and reproduction alphabet, respectively. Let $\mathcal{A}$ be a σ-field of subsets of $\mathbf{A}$, and let $\hat{\mathcal{A}}$ be a σ-field of subsets of $\hat{\mathbf{A}}$. (Here we implicitly assume that any element of $\hat{\mathbf{A}}$ belongs to the σ-field $\hat{\mathcal{A}}$.) Let the measurable space

$$(\mathbf{A}^\infty, \mathcal{A}^\infty) = \prod_{k=1}^{\infty} (\mathbf{A}_k, \mathcal{A}_k)$$

be the infinite Cartesian product of exemplars $(\mathbf{A}_k, \mathcal{A}_k)$ of the measurable space $(\mathbf{A}, \mathcal{A})$. The measurable space $(\hat{\mathbf{A}}^\infty, \hat{\mathcal{A}}^\infty)$ is defined similarly. If $x = (x_i)$ is a finite or infinite sequence of symbols from $\mathbf{A}$ or $\hat{\mathbf{A}}$, let $x_m^n = (x_m, x_{m+1}, \cdots, x_n)$ and, for simplicity, write x_1^n as x^n. The same conventions apply to sequences of random variables taking their values in these sets as well. We denote the set of all n-tuples drawn from $\mathbf{A}$ ($\hat{\mathbf{A}}$) by $\mathbf{A}^n$ ($\hat{\mathbf{A}}^n$).

Without loss of generality, we assume that each of data objects X, Y, Z, etc is a sequence of symbols from $\mathbf{A}$, and its lossy version is a sequence of symbols of the same length from $\hat{\mathbf{A}}$. (Discussions and results below can be easily extended to data objects from different alphabets.) In most cases, we model each data object as a stationary source taking values in $\mathbf{A}$. For example, $X = \{X_i\}_{i=1}^\infty$ will be a stationary source with each X_i being a random variable taking values in $\mathbf{A}$, and its lossy version $\hat{X} = \{\hat{X}_i\}_{i=1}^\infty$ will be a sequence of random variables taking values in $\hat{\mathbf{A}}$. Let $d : \mathbf{A} \times \hat{\mathbf{A}} \to [0,\infty)$ be a measurable function. Let $\{d_n\}_{n=1}^\infty$ be the single-letter fidelity criterion generated by d, by which we mean that for each n, $d_n : \mathbf{A}^n \times \hat{\mathbf{A}}^n \to [0,\infty)$ is the map in which $d_n(x^n, y^n) = n^{-1}\sum_{i=1}^n d(x_i, y_i)$ for any $x^n \in \mathbf{A}^n$ and $y^n \in \hat{\mathbf{A}}^n$. The distortion between X^n and $\hat{X}^n$ is measured by $d(X^n, \hat{X}^n)$.

Graphically, our coding paradigm for data analytics is illustrated in Figure 1, where X^n and Y^n are encoded jointly into a sequence of coded bits at rate R in bits per symbol. The coded bits specify a codeword (or method) which would in turn convert X^n into $\hat{Y}^n$ at Decoder 1, and Y^n into $\hat{X}^n$ at Decoder 2 such that for all sufficiently large n, both $d(X^n, \hat{X}^n)$ and $d(Y^n, \hat{Y}^n)$ are less than or equal to a prescribed threshold D.

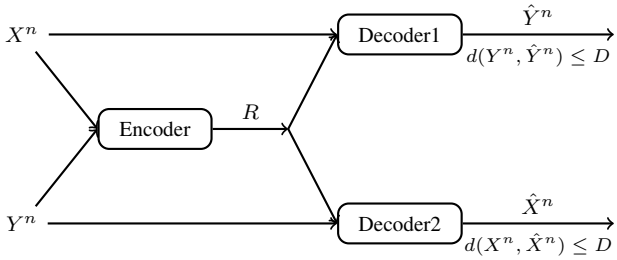

Fig. 1. Coding for data analytics.

For any $R > 0$ and n, let

$$\Omega(n,R) = \{1, 2, \cdots, \lfloor 2^{nR} \rfloor\}.$$

Formally, we have the following definition.

Definition 1: A block code C_n of order n and rate R consists of one encoding mapping

$$f : \mathbf{A}^n \times \mathbf{A}^n \to \Omega(n,R)$$

and two decoding mappings

$$g_1 : \Omega(n,R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n$$

and

$$g_2 : \Omega(n,R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n.$$

For any two data objects x^n and y^n, the encoder f encodes x^n and y^n into $f(x^n, y^n)$ of nR bits. On the decoding side, the encoded message $f(x^n, y^n)$ then converts x^n into $\hat{y}^n = g_1(f(x^n, y^n), x^n)$, and y^n into $\hat{x}^n = g_2(f(x^n, y^n), y^n)$.

Impose no other constraints on C_n, and let $\mathcal{C}$ consist of all possible block codes of order n for all n. We want to seek the best trade-off between R and the maximum distortion $\max\{d(X^n, \hat{X}^n), d(Y^n, \hat{Y}^n)\}$ attainable by $\mathcal{C}$ for any stationary sources $X = \{X_i\}_{i=1}^\infty$ and $Y = \{Y_i\}_{i=1}^\infty$ and sufficiently large n.

Definition 2: Given stationary sources $X = \{X_i\}_{i=1}^\infty$ and $Y = \{Y_i\}_{i=1}^\infty$, a rate distortion pair (R, D) is said to be achievable for (X, Y) if for any $\epsilon > 0$, there exists, for all

sufficiently large n, a block code $C_n = (f, g_1, g_2)$ of order n and rate $R + \epsilon$ such that

$$\Pr\{d(X^n, \hat{X}^n) > D + \epsilon\} \le \epsilon \tag{2.1}$$

and

$$\Pr\{d(Y^n, \hat{Y}^n) > D + \epsilon\} \le \epsilon \tag{2.2}$$

where $\hat{X}^n = g_2(f(X^n, Y^n), Y^n)$, and $\hat{Y}^n = g_1(f(X^n, Y^n), X^n)$.

Let $\mathcal{R}(X, Y)$ denote the set of all achievable (R, D) pairs for (X, Y). It can be verified that $\mathcal{R}(X, Y)$ is closed. Given $D \ge 0$, define the information distance between X and Y at the distortion level D as

$$R(X, Y, D) \stackrel{\Delta}{=} \min\{R : (R, D) \in \mathcal{R}(X, Y)\}. \tag{2.3}$$

One of our purposes in this paper is to characterize $R(X, Y, D)$, and analyze its relationship among different sources X, Y, Z, etc as a notion of distance.

Remark 1: The diagram shown in Figure 1 resembles the butterfly network in network coding [16]. As such, the coding diagram illustrated Figure 1 may be regarded as lossy network coding in the context of transmission. Also related are Wyner-Ziv coding [17] and coding with multiple decoders accessing different side information considered by Kaspi [18] and Heegard & Berger [19]. However, in addition to characterizing $R(X, Y, D)$, we are also interested in its relationship among different sources X, Y, Z, etc as a notion of distance.

Let us now impose some constraints on C_n. In particular, we split the encoding process into two steps. At Step 1, data objects x^n and y^n are separately precoded into nR_1 and nR_2 bits, respectively. At Step 2, the precoded bits are then jointly encoded into nR bits. The resulting type of code is called a separately precoded broadcast code. Formally, we have the following definition.

Definition 3: A separately precoded broadcast code C_n of order n and rate R with precoded rates R_1 and R_2 consists of two separate precoding mappings

$$f_1 : \mathbf{A}^n \to \Omega(n, R_1)$$

$$f_2 : \mathbf{A}^n \to \Omega(n, R_2)$$

a joint encoding mapping

$$f : \Omega(n, R_1) \times \Omega(n, R_2) \to \Omega(n, R)$$

and two decoding mappings

$$g_1 = (g_{11}, g_{12}) : \Omega(n, R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n \times \Omega(n, R_2)$$

and

$$g_2 = (g_{21}, g_{22}) : \Omega(n, R) \times \mathbf{A}^n \to \hat{\mathbf{A}}^n \times \Omega(n, R_1).$$

Figure 2 illustrates the encoding and decoding processes of a separately precoded broadcast code $C_n = (f_1, f_2, f, g_1, g_2)$ of order n and rate R with precoded rates R_1 and R_2. X^n and Y^n are first separately precoded into $f_1(X^n)$ of nR_1 bits and $f_2(Y^n)$ of nR_2 bits, and then jointly encoded into $f(f_1(X^n), f_2(Y^n))$ of nR bits. On the decoder side, the jointly encoded message $f(f_1(X^n), f_2(Y^n))$ converts: (1) X^n via Decoder 1 into an estimate $\hat{Y}^n = g_{11}(f(f_1(X^n), f_2(Y^n)), X^n)$ of Y^n and an estimate $\hat{f}_2(Y^n) = g_{12}(f(f_1(X^n), f_2(Y^n)), X^n)$ of $f_2(Y^n)$; and (2) Y^n via Decoder 2 into an estimate $\hat{X}^n = g_{21}(f(f_1(X^n), f_2(Y^n)), Y^n)$ of X^n and an estimate $\hat{f}_1(X^n) = g_{22}(f(f_1(X^n), f_2(Y^n)), Y^n)$ of $f_1(X^n)$.

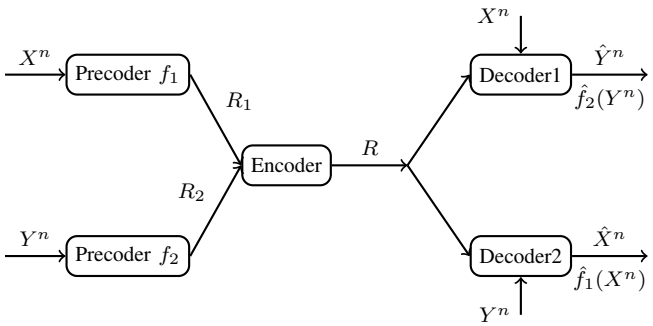

Fig. 2. Illustration of a separately precoded broadcast code.

Definition 4: Let $\mathcal{C}_{sb}$ consist of all separately precoded broadcast codes. Given stationary sources $X = \{X_i\}_{i=1}^{\infty}$ and $Y = \{Y_i\}_{i=1}^{\infty}$, a rate distortion pair (R, D) is said to be $\mathcal{C}_{sb}$-achievable for (X, Y) if for any $\epsilon > 0$, there exist a finite set $\mathbf{B} \subseteq \hat{\mathbf{A}}$ and, for all sufficiently large n, a separately precoded block code $C_n = (f_1, f_2, f, g_1, g_2)$ of order n and rate $R + \epsilon$ with precoded rates $R_1 + \epsilon$ and $R_2 + \epsilon$ such that

$$\Pr\{d(X^n, \hat{X}^n) > D + \epsilon\} \le \epsilon \tag{2.4}$$

$$\Pr\{d(Y^n, \hat{Y}^n) > D + \epsilon\} \le \epsilon \tag{2.5}$$

$$\Pr\{f_1(X^n) \ne \hat{f}_1(X^n)\} \le \epsilon \tag{2.6}$$

and

$$\Pr\{f_2(Y^n) \ne \hat{f}_2(Y^n)\} \le \epsilon \tag{2.7}$$

where $(\hat{X}^n, \hat{f}_1(X^n)) = g_2(f(f_1(X^n), f_2(Y^n)), Y^n)$, $(\hat{Y}^n, \hat{f}_2(Y^n)) = g_1(f(f_1(X^n), f_2(Y^n)), X^n)$, and both $\hat{X}^n$ and $\hat{Y}^n$ take values in $\mathbf{B}^n$.

Let $\mathcal{R}_{sb}(X, Y)$ denote the set of all $\mathcal{C}_{sb}$-achievable (R, D) pairs for (X, Y). It can be verified that $\mathcal{R}_{sb}(X, Y)$ is closed. Given $D \ge 0$, define the information distance between X and Y at the distortion level D with respect to $\mathcal{C}_{sb}$ as

$$R_{sb}(X, Y, D) \stackrel{\Delta}{=} \min\{R : (R, D) \in \mathcal{R}_{sb}(X, Y)\}. \tag{2.8}$$

As in the case of $R(X, Y, D)$, we also aim to characterize $R_{sb}(X, Y, D)$, and analyze its relationship among different sources X, Y, Z, etc as a notion of distance.

III. $R(X, Y, D)$: Distance Property and Bounds

Unless otherwise specified, in this section X, Y, Z, etc denote arbitrary stationary sources. We begin with the distance property of $R(X, Y, D)$ among different sources for a fixed $D \ge 0$.

A. Finite Alphabets

Suppose that both $\mathbf{A}$ and $\hat{\mathbf{A}}$ are finite, and

$$\max_{x\in\mathbf{A}}\min_{\hat{x}\in\hat{\mathbf{A}}} d(x,\hat{x}) = 0. \tag{3.1}$$

For any $D \geq 0$, define

$$H(D) \stackrel{\Delta}{=} \max H(U|\hat{U}) \tag{3.2}$$

where the maximum is taken over all random variables U and $\hat{U}$ taking values in $\mathbf{A}$ and $\hat{\mathbf{A}}$, respectively, such that $\mathbf{E}[d(U,\hat{U})] \leq D$, and $H(U|\hat{U})$ denotes the conditional entropy of U given $\hat{U}$. (All information quantities in this paper are expressed in bits, and the function $\log$ is to base 2.) It is easy to see that in the case where $\mathbf{A} = \hat{\mathbf{A}}$ and d is the Hamming distance measure on $\mathbf{A}$,

$$H(D) = h(D) + D\log(|\mathbf{A}| - 1) \tag{3.3}$$

for any $0 \leq D \leq 1/2$, where $h(D) = -D\log D - (1-D)\log(1-D)$, and $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$ if $\mathcal{S}$ is a finite set.

Theorem 1: Fix $D \geq 0$. Let

$$R^*(X,Y,D) = \begin{cases} R(X,Y,D) & \text{if } X = Y \\ R(X,Y,D) + H(D) & \text{otherwise.} \end{cases} \tag{3.4}$$

Then $R^*(X,Y,D)$ is a pseudo distance over the set of all stationary sources, i.e., satisfying the following three properties:

(1) $R^*(X,Y,D) = 0$ if $X = Y$.
(2) $R^*(X,Y,D) = R^*(Y,X,D)$.
(3) $R^*(X,Z,D) \leq R^*(X,Y,D) + R^*(Y,Z,D)$.

Proof: Properties (1) and (2) above follow immediately from the definition of $R(X,Y,D)$ along with Definitions 1 and 2. To prove Property (3), i.e., the triangle inequality, we first show that allowing random encoding mappings in the definition of block codes C_n will not enlarge the set of all achievable (R,D) pairs for any (X,Y). Given any X, Y, and Z, we then show that $(R(X,Y,D)+R(Y,Z,D)+H(D),D)$ is achievable by block codes with random encoding mappings for (X,Z). Thus,

$$R(X,Z,D) \leq R(X,Y,D) + R(Y,Z,D) + H(D)$$

from which Property (3) follows. The detailed proof in these two steps along with proofs of other resutls can be found in the full paper [20]. ∎

B. Abstract Alphabets

In this subsection, both $\mathbf{A}$ and $\hat{\mathbf{A}}$ are abstract. As usual, however, we assume that the distortion measure d and stationary sources $X = \{X_i\}_{i=1}^{\infty}$ satisfy the following condition:

$$\mathbf{E}[d(X_1,\hat{x})] < \infty \tag{3.5}$$

for some $\hat{x} \in \hat{\mathbf{A}}$. Then we have the following result.

Theorem 2: Let (X,Y) be a stationary, ergodic pair with each of X and Y satisfying (3.5). Let $R_{X|Y}(D)$ ($R_{Y|X}(D)$, resp.) denote the conditional rate distortion function of X (Y, resp.) given Y (X, resp.). Then the following holds:

$$\max\{R_{X|Y}(D), R_{Y|X}(D)\} \leq R(X,Y,D) \tag{3.6}$$

$$\leq \inf\{\max\{I(Y_1;U|X_1), I(X_1;U|Y_1)\} + I(X_1;\hat{X}_1|Y_1U) + I(Y_1;\hat{Y}_1|X_1U) : U, \hat{X}_1, \hat{Y}_1\} \tag{3.7}$$

where the infimum is taken over all random variables U, $\hat{X}_1$, and $\hat{Y}_1$ such that $\mathbf{E}[d(X_1,\hat{X}_1)] \leq D$ and $\mathbf{E}[d(Y_1,\hat{Y}_1)] \leq D$, and I denotes the mutual information and conditional mutual information, as the case may be.

Example 1: Let X_1 and Y_1 be two real-valued random variables with

$$X_1 = aY_1 + N_1$$

where $|a| \geq 1$ and N_1 is independent of Y_1. Let $d(x,\hat{x}) = |x-\hat{x}|^r$ with $r \geq 1$. Suppose that $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$ is independent and identically distributed (iid). In this case, it follows from Theorem 2 that

$$R(X,Y,D) = R_{X|Y}(D).$$

Example 2: Let X_1 and Y_1 be two binary random variables with

$$X_1 = Y_1 \oplus N_1$$

where $\oplus$ denotes the binary addition, and N_1 is independent of Y_1 with $\Pr\{Y_1 = 1\} = 1/2$ and $\Pr\{N_1 = 1\} = p < 1/2$. Let d be the Hamming distortion measure over $\{0,1\}$. Suppose that $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$ is iid. In this case, it follows from Theorem 2 that

$$R(X,Y,D) = R_{X|Y}(D).$$

Corollary 1: Suppose that X_1 and Y_1 are jointly Gaussian, and $(X,Y) = \{(X_i,Y_i)\}_{i=1}^{\infty}$ is iid. Then with $d(x,\hat{x}) = (x-\hat{x})^2$, both the lower bound (3.6) and upper bound (3.7) are tight, and

$$R(X,Y,D) = \begin{cases} \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D} & \text{if } 0 \leq D < (1-\rho^2)\sigma^2 \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

where ρ is the correlation between X_1 and Y_1, and σ^2 is the maximum of the variances of X_1 and Y_1.

Proof: Without loss of generality, assume that $\mathbf{E}X_1 = \mathbf{E}Y_1 = 0$, and $\sigma^2 = \sigma_X^2 \geq \sigma_Y^2$, where σ_X^2 and σ_Y^2 are the variances of X_1 and Y_1, respectively. Consider three cases: (1) $D \geq (1-\rho^2)\sigma^2$, (2) $D < (1-\rho^2)\sigma_Y^2$, and (3) $(1-\rho^2)\sigma_Y^2 \leq D < (1-\rho^2)\sigma^2$. In Case (1), X and Y can be estimated from each other. The resulting $\hat{X}$ and $\hat{Y}$ satisfy the distortion requirement, and no information needs to be sent from the encoder. Hence, $R(X,Y,D) = 0$.

In Case (2), let

$$a = \frac{(1-\rho^2)\sigma^2 - D}{(1-\rho^2)\sigma^2} \text{ and } b = \frac{(1-\rho^2)\sigma_Y^2 - D}{(1-\rho^2)\sigma_Y^2}.$$

Define

$$V = a(X_1 + N_1) \text{ and } W = b(Y_1 + N_2) \tag{3.9}$$

where N_1 and N_2 are zero mean Gaussian random variables with variances

$$\frac{D(1-\rho^2)\sigma^2}{(1-\rho^2)\sigma^2 - D} \text{ and } \frac{D(1-\rho^2)\sigma_Y^2}{(1-\rho^2)\sigma_Y^2 - D}$$

respectively. Furthermore, N_1 and N_2 are independent of each other and of both X_1 and Y_1. Let

$$\hat{X}_1 = V + (1-a)\rho\frac{\sigma}{\sigma_Y}Y_1 \text{ and } \hat{Y}_1 = W + (1-b)\rho\frac{\sigma_Y}{\sigma}X_1. \tag{3.10}$$

One can verify that

$$\mathbf{E}[X_1 - \hat{X}_1]^2 = D \text{ and } \mathbf{E}[Y_1 - \hat{Y}_1]^2 = D. \tag{3.11}$$

Let $U = (V, W)$. Plugging U, $\hat{X}_1$, and $\hat{Y}_1$ into the respective information quantities in (3.7), we have

$$I(Y_1; \hat{Y}_1 | X_1 U) = 0 \tag{3.12}$$

$$I(X_1; \hat{X}_1 | Y_1 U) = 0 \tag{3.13}$$

and

$$\begin{aligned}
& I(X_1; U|Y_1) \\
&= I(X_1; VW|Y_1) \\
&= I(X_1; V|Y_1) + I(X_1; W|Y_1 V) \\
&= I(X_1; V|Y_1) && (3.14) \\
&= H(V|Y_1) - H(V|Y_1 X_1) \\
&= H(V|Y_1) - H(V|X_1) && (3.15) \\
&= H(V - a\rho\frac{\sigma}{\sigma_Y}Y_1|Y_1) - H(aN_1) \\
&= H(V - a\rho\frac{\sigma}{\sigma_Y}Y_1) - H(aN_1) && (3.16) \\
&= \frac{1}{2}\log 2\pi e a^2 \frac{[(1-\rho^2)\sigma^2]^2}{(1-\rho^2)\sigma^2 - D} - H(aN_1) && (3.17) \\
&= \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D} && (3.18)
\end{aligned}$$

where (3.14) and (3.15) are due to (3.9) which implies the conditional independence of W and (X_1, V) given Y_1, and the conditional independence of V and Y_1 given X_1; and (3.16) follows from the fact that under the joint Gaussian assumption, $V - a\rho\frac{\sigma}{\sigma_Y}Y_1$ is independent of Y_1. In parallel with (3.18), we have

$$\begin{aligned}
& I(Y_1; U|X_1) \\
&= I(Y_1; VW|X_1) \\
&= I(Y_1; W|X_1) + I(Y_1; V|X_1 W) \\
&= I(Y_1; W|X_1) \\
&= H(W|X_1) - H(W|Y_1 X_1) \\
&= H(W|X_1) - H(W|Y_1) \\
&= H(W - b\rho\frac{\sigma_Y}{\sigma}X_1|X_1) - H(bN_2) \\
&= H(W - b\rho\frac{\sigma_Y}{\sigma}X_1) - H(bN_2) && (3.19) \\
&= \frac{1}{2}\log 2\pi e b^2 \frac{[(1-\rho^2)\sigma_Y^2]^2}{(1-\rho^2)\sigma_Y^2 - D} - H(bN_2) && (3.20) \\
&= \frac{1}{2}\log\frac{(1-\rho^2)\sigma_Y^2}{D}. && (3.21)
\end{aligned}$$

By combining the information quantities in (3.21), (3.18), (3.13), and (3.12) together, it follows from (3.7) that

$$R(X, Y, D) \le \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D}.$$

This, together with (3.6) and

$$R_{X|Y}(D) = \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D}$$

implies (3.8) in Case (2).

In Case (3), Y can be estimated directly from X. Specifically, let $\hat{Y}_1 = \rho\frac{\sigma_Y}{\sigma}X_1$. With V and $\hat{X}_1$ defined as in (3.9) and (3.10), respectively, we now let $U = V$. Plug U, $\hat{X}_1$, and $\hat{Y}_1$ into the respective information quantities in (3.7). Again, (3.8) follows from a similar argument to the above. This completes the proof of Corollary 1. ∎

Careful examination reveals that the equal sign $=$ in (3.16), (3.17), (3.19), and (3.20) can be replaced by $\le$ when X_1 and Y_1 are not necessarily jointly Gaussian. Therefore, the above argument also shows that the right side of (3.8) is actually an upper bound to $R(X, Y, D)$ for any real-valued sources X and Y satisfying (3.5) with $d(x, \hat{x}) = (x - \hat{x})^2$, which is stated as a corollary below.

Corollary 2: Let $X = \{X_i\}_{i=1}^{\infty}$ and $Y = \{Y\}_{i=1}^{\infty}$ be a real-valued, stationary, and ergodic source pair with each satisfying (3.5) with $d(x, \hat{x}) = (x - \hat{x})^2$. Then

$$R(X, Y, D) \le \begin{cases} \frac{1}{2}\log\frac{(1-\rho^2)\sigma^2}{D} & \text{if } 0 \le D < (1-\rho^2)\sigma^2 \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

where ρ is the correlation between X_1 and Y_1, and σ^2 is the maximum of the variances of X_1 and Y_1.

We conclude this section by pointing out that the single-letter characterization of $R(X, Y, D)$ remains open in general even when $(X, Y) = \{(X_i, Y_i)\}_{i=1}^{\infty}$ is iid.

IV. $R_{sb}(X, Y, D)$: Distance Property and Characterization

In this section, we analyze the distance property of $R_{sb}(X, Y, D)$ over the set of stationary sources X, Y, Z, etc, and characterize it in terms of Wyner-Ziv coding rates for stationary, ergodic source pairs (X, Y). Again, we begin with its distance property.

A. Finite Alphabets

Suppose that both $\mathbf{A}$ and $\hat{\mathbf{A}}$ are finite, and the condition (3.1) is met. In parallel with Theorem 1, we have the following result.

Theorem 3: Fix $D \ge 0$. Let

$$R_{sb}^*(X, Y, D) = \begin{cases} R_{sb}(X, Y, D) & \text{if } X = Y \\ R_{sb}(X, Y, D) + H(D) & \text{otherwise.} \end{cases} \tag{4.1}$$

Then $R_{sb}^*(X, Y, D)$ is a pseudo distance over the set of all stationary sources.

Remark 2: In view of the definitions of $R(X, Y, D)$ and $R_{sb}(X, Y, D)$, it follows that

$$R(X, Y, D) \le R_{sb}(X, Y, D). \tag{4.2}$$

However, the above inequality, together with Theorem 1, does not imply Theorem 3 directly.

An approach similar to the proof of Theorem 1 can be used to show Theorem 3. To prove the corresponding triangle inequality, we first show that allowing random joint encoding mappings f in the definition of separately precoded broadcast codes does not enlarge the set $\mathcal{R}_{sb}(X,Y)$ of all $\mathcal{C}_{sb}$-achievable pairs (R,D) for any (X,Y). Given any X, Y, and Z, we then show that $(R_{sb}(X,Y,D)+R_{sb}(Y,Z,D)+H(D),D)$ is achievable by separately precoded broadcast codes with random joint encoding mappings f for (X,Z).

B. Abstract Alphabets

Suppose now that both $\mathbf{A}$ and $\hat{\mathbf{A}}$ are abstract. As in [21], we make the following assumptions on $d:\mathbf{A}\times\hat{\mathbf{A}}\to[0,\infty)$ and stationary, totally ergodic sources $X=\{X_i\}_{i=1}^{\infty}$:

- A1 For any $\hat{x}\in\hat{\mathbf{A}}$, $\mathbf{E}d(X_1,\hat{x})<\infty$.
- A2 For any $\epsilon>0$ and any $\hat{X}_1$ with $\mathbf{E}d(X_1,\hat{X}_1)<\infty$, there exist a finite set $\mathbf{B}\subseteq\hat{\mathbf{A}}$ and a measurable mapping $q:\hat{\mathbf{A}}\to\mathbf{B}$ such that

$$\mathbf{E}d(X_1,q(\hat{X}_1))\le(1+\epsilon)\mathbf{E}d(X_1,\hat{X}_1). \tag{4.3}$$

For any stationary, totally ergodic source pair $X=\{X_i\}_{i=1}^{\infty}$ and $Y=\{Y_i\}_{i=1}^{\infty}$ with each satisfying Conditions A1 and A2, let $R_{X|Y}^{WZ}(D)$ ($R_{Y|X}^{WZ}(D)$, resp.) denote the Wyner-Ziv coding rate of X (Y, resp.) with Y (X, resp.) as side information available at the decoder. Then we have the following result.

Theorem 4: For any stationary, totally ergodic source pair $X=\{X_i\}_{i=1}^{\infty}$ and $Y=\{Y_i\}_{i=1}^{\infty}$ with each satisfying Conditions A1 and A2,

$$R_{sb}(X,Y,D)=\max\{R_{X|Y}^{WZ}(D),R_{Y|X}^{WZ}(D)\}.$$

It is instructive to compare $R_{sb}(X,Y,D)$ with $R(X,Y,D)$. When X_1 and Y_1 are jointly Gaussian, and $(X,Y)=\{(X_i,Y_i)\}_{i=1}^{\infty}$ is iid with $d(x,\hat{x})=(x-\hat{x})^2$, we have

$$R_{sb}(X,Y,D)=R(X,Y,D).$$

In general, however, it is expected that the inequality in (4.2) is strict, which is the case, for example, when (X,Y) is the source pair in Example 2.

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age", *ACM Computing Surveys*, Vol. 40, No. 2, Article 5, pp.1–60, April 2008.

[2] E.-H. Yang, J. Meng, and X. Yu, "Display, Visualization, and Management of Photos Based on Content Analytics," US Patent Application No. 14/790,650, July 2, 2015.

[3] F. Emmert-Streib and M. Dehmer, Eds., *Information Theory and Statistical Learning*. Springer Science+Business Media, LLC 2009.

[4] E.-H. Yang, X. Yu, and J. Meng, "Set mapping induced image perceptual similarity distance," *Proc. of the 2015 Information Theory and Applications Workshop*, San Diego, California, U.S.A., Feb. 1–Feb. 6, 2015.

[5] O. Chapelle, P. Haffner, and V. Vapnik, "Support vector machines for histogram-based image classification", *IEEE Trans. Neural Netw.*, Vol. 10, No. 5, pp.1055–1064, Sept. 1999.

[6] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification", *Proc. of the 2003 Int. Conf. on Image process.*, Vol. 3, pp. 513–516, Sept. 2003.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints" *Int. Journal of Computer Vision*, 60, 2, pp. 91–110, 2004.

[8] T. Lindeberg, "Scale invariant feature transform", *Scholarpedia*, 7 (5): 10491, 2012.

[9] A. Vedaldi, "An implementation of SIFT detector and descriptor", http://www.robots.ox.ac.uk/ vedaldi/code/sift.html.

[10] L. Kang, C. Hsu, H. Chen, C. Lu, C. Lin, and S. Pei, "Feature-based sparse representation for image similarity assessment" *IEEE Trans. Multimedia*, Vol. 13, No. 5, pp. 1019–1030, October 2011.

[11] H. Ling, K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, No. 5, pp. 840–853, May 2007.

[12] J. Huang, S. Kumar, M. Mitra, W.J. Zhu, and R. Zabih, "Spatial color indexing and applications", *Int. Journal of Computer Vision*, 35(3), pp. 245–268, 1999.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory (second edition)*. Hoboken, NJ: Wiley, 2006.

[14] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Trans.*, Vol. 1, pp. 1–7, 1965.

[15] C. H. Bennet, *et al*, "Information distance," *IEEE Trans. Inf. Theory*, Vol. 44, No. 4, pp. 1407 – 1423, Jul. 1998.

[16] R. W. Yeung, *Information Theory and Network Coding*. Springer Science+Business Media, LLC 2008.

[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. Kaspi, Rate-distortion function when side-information may be present at the decoder, *IEEE Trans. Inf. Theory*, Vol. 40, No. 6, pp. 2031 – 2034, Nov. 1994.

[19] C. Heegard and T. Berger, Rate distortion when side information may be absent, *IEEE Trans. Inf. Theory*, Vol. 31, No. 6, pp. 727 – 734, Nov. 1985.

[20] E.-H. Yang, "Coding for data analytics: New information distances," *in preparation*.

[21] A. D. Wyner, "The rate distortion function for source coding with side information at the decoder-II: General sources," *Infor. Contr.*, Vol. 38, pp. 60 – 80, 1978.