Quadratic Similarity Queries on Compressed Data

Amir Ingber, Thomas Courtade and Tsachy Weissman
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
Email: {ingber, courtade, tsachy}@stanford.edu

Abstract

The problem of performing similarity queries on compressed data is considered. We study the fundamental tradeoff between compression rate, sequence length, and reliability of queries performed on compressed data. For a Gaussian source and quadratic similarity criterion, we show that queries can be answered reliably if and only if the compression rate exceeds a given threshold – the *identification rate* – which we explicitly characterize. When compression is performed at a rate greater than the identification rate, responses to queries on the compressed data can be made exponentially reliable. We give a complete characterization of this exponent, which is analogous to the error and excess-distortion exponents in channel and source coding, respectively.

For a general source, we prove that the identification rate is at most that of a Gaussian source with the same variance. Therefore, as with classical compression, the Gaussian source requires the largest compression rate. Moreover, a scheme is described that attains this maximal rate for any source distribution.

I. Introduction

For a database consisting of many long sequences, it is natural to perform queries of the form: *which sequences in the database are similar to a given sequence* $\mathbf{y}$? In this paper, we study the problem of compressing this database so that queries about the original data can be answered reliably given only the compressed version. This goal stands in contrast to the traditional compression paradigm, where data is compressed so that it can be reconstructed – either exactly or approximately – from its compressed form.

Specifically, for each sequence $\mathbf{x}$ in the database we only keep a short *signature*, denoted $T(\mathbf{x})$, where $T(\cdot)$ is a signature assignment function. Queries are performed using only $\mathbf{y}$ and $T(\mathbf{x})$ as input, rather than the original (uncompressed) sequence $\mathbf{x}$. This setting is illustrated in Fig. 1.

As alluded to above, we generally do not require that the original data be reproducible from the signatures. Therefore the set of signatures is not meant to replace the database itself. Nevertheless, there are many instances where such compression is desirable. For example, the set of signatures can be thought of as a cached version of the original database (possibly hosted at many locations due to its relatively small size). By performing queries only on the cached (i.e., compressed) database,

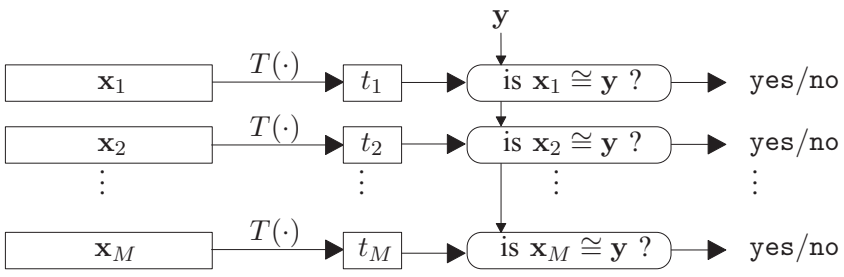

Fig. 1. Answering a query from compressed data

1068-0314/13 $26.00 © 2013 IEEE
DOI 10.1109/DCC.2013.52

query latency can be reduced and the computational burden on the server hosting the uncompressed database can be lessened.

In many scenarios (e.g,. querying a criminal forensic database), query responses which are false negatives are not acceptable. A false negative occurs if a query performed on $T(\mathbf{x})$ and $\mathbf{y}$ indicates that $\mathbf{x}$ and $\mathbf{y}$ are not similar, but they are in truth. Therefore, we impose the restriction in our model that false negatives are not permitted. With this in mind, we regard the query responses from the compressed data as either "`no`" or "`maybe`". Since minimizing the probability that a query returns `maybe` is equivalent to minimizing the probability of returning a false positive[1], any good compression scheme will have a corresponding query function which returns `maybe` with small probability. We note briefly that a false positive does not cause an error *per se*. Rather, it only introduces a computational burden due to the need for further verification.

In our setting we assume that the query and database sequences are independent from one another, and all entries are drawn i.i.d. according to a given distribution. We note that a similar problem has been considered by Ahlswede et al. [1], where the focus was only on discrete sources and false negatives are permitted. The results in [1] are parameterized by an auxiliary random variable with unbounded alphabet cardinality. This renders the various quantities incomputable, and therefore the results contained therein are of limited practical interest. In addition, [1] does not provide a lower bound on (i.e., a converse for) the identification rate.

Related ideas in the literature include Bloom Filters [2], which are efficient data structures enabling queries without false negatives. The Bloom Filter only applies for exact matches (where here we are interested in similarity queries) so it is not applicable to our problem. Nevertheless, as surveyed in [3], Bloom filters demonstrate the potential of answering queries from compressed data.

Another related notion is that of Local Sensitivity Hashing (LSH), which is a framework for data structures and algorithms for finding similar items in a given set (see [4] for a survey). LSH trades off accuracy with computational complexity and space, and false negatives are allowed. Two fundamental points are different in our approach. First, we study the information-theoretic aspect of the problem, i.e. concentrate on space only (compression rate) and ignore computational complexity in an attempt to understand the amount of information *relevant to querying* that can be stored in the short signatures. Second, we do not allow false negatives. As discussed above, false negatives are inherent for LSH.

This paper is organized as follows. In the next section we formally define the problem and the quantities we study (i.e., the identification rate and the identification exponent). In Section III we state and discuss our main results. Section IV sketches the proofs of these results, and Section V delivers concluding remarks.

II. Problem Formulation

Boldface notation $\mathbf{x}$ denotes vectors of elements $[x_1, ...x_n]$. Capital letters denote random variables (e.g. X, Y), and $\mathbf{X}, \mathbf{Y}$ denote random vectors. Throughout the paper $\log(\cdot)$ denotes the base-2 logarithm, while $\ln(\cdot)$ denotes the usual natural logarithm. We focus on the basic notion of quadratic similarity (sometimes called mean square error, or MSE). To this end, for any length-n real sequences $\mathbf{x}$ and $\mathbf{y}$ define

$$d(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 = \frac{1}{n} \|\mathbf{x} - \mathbf{y}\|^2, \tag{1}$$

where $\|\cdot\|$ denotes the standard Euclidean norm. We say that $\mathbf{x}$ and $\mathbf{y}$ are *D-similar* when $d(\mathbf{x}, \mathbf{y}) \leq D$, or simply *similar* when D is clear from the context.

[1]Complementary to false negatives, a false positive occurs if a query performed on $T(\mathbf{x})$ and $\mathbf{y}$ indicates that $\mathbf{x}$ and $\mathbf{y}$ are similar (i.e., returns `maybe`), but they are not in truth.

A rate-R identification system (T, g) consists of a signature assignment[2] $T : \mathbb{R}^n \to [1 : 2^{nR}]$, and a query function $g : [1 : 2^{nR}] \times \mathbb{R}^n \to \{\texttt{no}, \texttt{maybe}\}$. A system is said to be *D-admissible*, if for any $\mathbf{x}, \mathbf{y}$ satisfying $d(\mathbf{x}, \mathbf{y}) \leq D$, we have

$$g(T(\mathbf{x}), \mathbf{y}) = \texttt{maybe}. \tag{2}$$

We wish to minimize the probability $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$, subject to the scheme being admissible. Note that this is equivalent to minimizing the probability $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe} | d(\mathbf{X}, \mathbf{Y}) > D\}$ because $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe} | d(\mathbf{X}, \mathbf{Y}) \leq D\} = 1$. In other words, we desire to minimize the probability of false positive.

Generally, let $\mathbf{X}, \mathbf{Y}$ be random vectors representing the sequence from the database and the query sequence, respectively. Unless specified otherwise[3], we assume that $\mathbf{X}$ and $\mathbf{Y}$ are independent, with entries drawn independently from the same distribution P_X. In other words, $(\mathbf{X}, \mathbf{Y}) \sim \prod_{i=1}^{n} P_X(x_i) P_X(y_i)$. In the first part of the paper we discuss Gaussian sources, and in the second part we discuss the general case.

For a given similarity threshold D, we study the tradeoff between compression rate and reliability. Formally:

Definition 1: For a given source distribution P_X and similarity threshold D, a rate R is said to be *D-achievable* if there exists a sequence of rate-R admissible schemes $(T^{(n)}, g^{(n)})$, s.t.

$$\lim_{n \to \infty} \Pr \left\{ g^{(n)} \left(T^{(n)}(\mathbf{X}), \mathbf{Y} \right) = \texttt{maybe} \right\} = 0. \tag{3}$$

Definition 2: For a similarity threshold D, the *identification rate* $R_{\mathrm{ID}}(D)$ is the infimum of D-achievable rates. That is,

$$R_{\mathrm{ID}}(D) \triangleq \inf\{R : R \text{ is } D\text{-achievable}\}. \tag{4}$$

The above definitions are in the same spirit of the rate distortion function (the rate above which a vanishing probability for excess distortion is achievable), and also in the spirit of the channel capacity (the rate below which a vanishing probability of error can be obtained). See, for example, Gallager [5].[4]

Having defined $R_{\mathrm{ID}}(D)$, the rate at which the probability of `maybe` vanishes is also of significant interest. We expect the vanishing rate to be exponential as in the traditional source coding setting, motivating the following definition:

Definition 3: Fix $R \geq R_{\mathrm{ID}}(D)$. The *identification exponent* is defined as

$$\mathbf{E}_{\mathrm{ID}}(R) \triangleq \limsup_{n \to \infty} -\frac{1}{n} \log \inf \Pr \left\{ g^{(n)} \left(T^{(n)}(\mathbf{X}), \mathbf{Y} \right) = \texttt{maybe} \right\}, \tag{5}$$

where the infimum is over all D-admissible schemes $g^{(n)}, T^{(n)}$ at rate R and length n.

A few remarks are in order. First, the quantity $\mathbf{E}_{\mathrm{ID}}(R)$ implicitly depends on the parameters of the problem (i.e., D and P_X). Second, the equivalent quantity in source coding is the excess distortion exponent, first studied by Marton [7] for discrete sources and by Ihara and Kubo [8] for the Gaussian source.

III. Main Results: Statements and Discussion

This section contains our main results. All proofs are delayed until Section IV.

[2] $[1 : k]$ denotes the set $\{1, 2, ..., k\}$.

[3] Later on in the paper we also state the results for different distributions for $\mathbf{X}$ and $\mathbf{Y}$.

[4] See, for example, Cover and Thomas [6] for the alternative approach based on average distortion rather than excess distortion probability

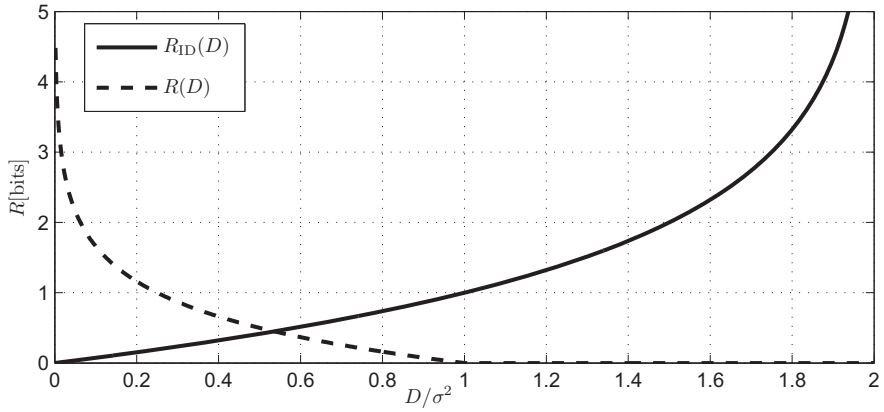

Fig. 2. The rate distortion function $R(D)$ and the identification rate $R_{\mathrm{ID}}(D)$ of a Gaussian source with variance σ^2.

A. The Identification Rate for Gaussian Sources

Our first result characterizes the identification rate for the Gaussian case:

Theorem 1: When P_X is given by the Gaussian distribution $N(0, \sigma^2)$,

$$R_{\mathrm{ID}}(D) = \begin{cases} \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right) & \text{for } D < 2\sigma^2 \\ \infty & \text{for } D \geq 2\sigma^2. \end{cases} \tag{6}$$

The fact that $R_{\mathrm{ID}}(D) = \infty$ for $D \geq 2\sigma^2$ is to be expected since the MSE between independent sequences $\mathbf{X}, \mathbf{Y}$ is $2\sigma^2$. Hence, $\mathbf{X}$ is D-similar to $\mathbf{Y}$ with non-negligible probability when $D \geq 2\sigma^2$.

We remark that (6) is reminiscent of the Gaussian rate distortion function $R(D) = \left[\frac{1}{2}\log\frac{\sigma^2}{D}\right]^+$ (see e.g. [6]). However, upon closer inspection, we observe that $R(D)$ is monotonically decreasing in D, while $R_{\mathrm{ID}}(D)$ is monotone *increasing*. In other words, as the similarity criterion is relaxed, more information is required about the compressed sequence in order to answer queries reliably. The identification rate $R_{\mathrm{ID}}(D)$ and rate distortion function $R(D)$ for a Gaussian source are plotted in Fig. 2.

B. Identification Exponent for Gaussian Sources

Our second result gives a complete characterization of the identification exponent $\mathbf{E}_{\mathrm{ID}}(R)$ for a Gaussian source:

Theorem 2: For a Gaussian source with variance σ^2 and a rate $R > R_{\mathrm{ID}}(D)$,

$$\mathbf{E}_{\mathrm{ID}}(R) = \min_{\rho\in(0,1]} \frac{2}{\ln 2}\mathbf{E}_Z(\rho) - \log\sin\min\left[\frac{\pi}{2}, \left(\arcsin(2^{-R}) + \arccos\frac{\rho - \frac{D}{2\sigma^2}}{\rho}\right)\right] \tag{7}$$

where

$$\mathbf{E}_Z(\rho) \triangleq \frac{\rho}{2} - \frac{1}{2} - \frac{1}{2}\ln\rho. \tag{8}$$

Since (7) only involves a minimization in the single variable ρ over $(0, 1]$, the function $\mathbf{E}_{\mathrm{ID}}(R)$ is easily computed. Furthermore, the following properties are readily observed:

- As expected, $\mathbf{E}_{\mathrm{ID}}(R_{\mathrm{ID}}(D)) = 0$. This can be easily seen since $2^{-R_{\mathrm{ID}}(D)} = 1 - \frac{D}{2\sigma^2}$, so (7) is minimized at $\rho = 1$.
- $\mathbf{E}_{\mathrm{ID}}(R)$ is strictly positive for $R > R_{\mathrm{ID}}(D)$. Therefore, the direct part of Theorem 1 is implied by Theorem 2. However, the converse part of Theorem 1 is *not* implied by Theorem 2.

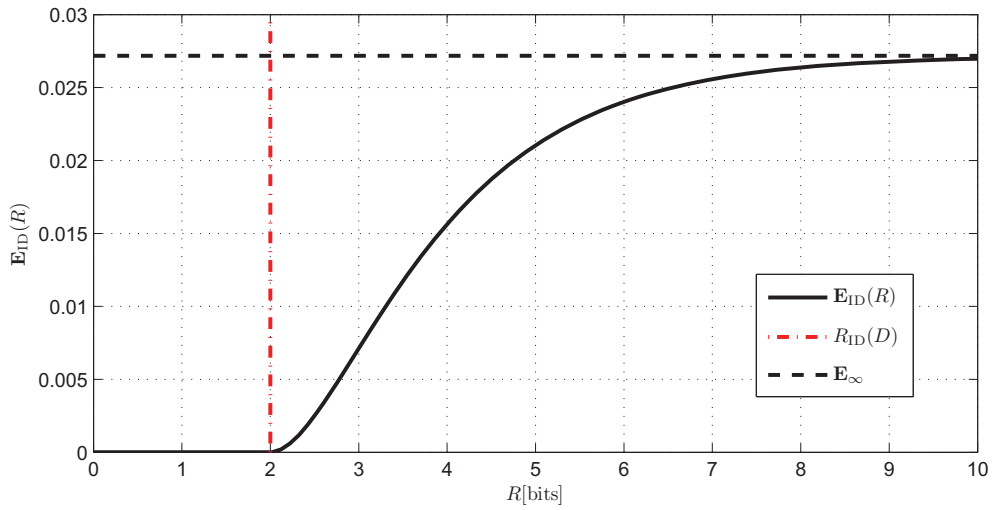

Fig. 3. $\mathbf{E}_{\mathrm{ID}}(R)$ for $R_{\mathrm{ID}}(D) = 2$ bits/sym.

- $\mathbf{E}_{\mathrm{ID}}(R)$ as $R \to \infty$ has an interesting value, which we denote $\mathbf{E}_{\infty}$. It can be calculated independently as the exponential decaying speed of the event $\Pr\{\frac{1}{n}\|\mathbf{X} - \mathbf{Y}\| \leq D\}$.

The exponent $\mathbf{E}_{\mathrm{ID}}(R)$ for the case of $R_{\mathrm{ID}}(D) = 2$bit is shown in Fig. 3.

C. Gaussian Source with Different Variance for $\mathbf{X}$ and $\mathbf{Y}$

The results for the identification rate and exponent, given in Theorems 1 and 2 respectively, can be extended to the setting where $\mathbf{X}$ and $\mathbf{Y}$ have different variances. We state these results below (without proof).

Theorem 3: Let X and Y be independent Gaussian with variances σ_X^2 and σ_Y^2 respectively. Then the identification rate at which $\mathbf{X}$ can be compressed is given by

$$R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2) = \begin{cases} \log \frac{2\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2-D} & \text{for } D < \sigma_X^2 + \sigma_Y^2 \\ \infty & \text{for } D \geq \sigma_X^2 + \sigma_Y^2. \end{cases} \tag{9}$$

Before proceeding to our discussion of $\mathbf{E}_{\mathrm{ID}}(R)$, we make several observations:

- $R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2) = R_{\mathrm{ID}}(D, \sigma_X^2)$ when either $\sigma_Y^2 = \sigma_X^2$ or $\sigma_Y^2 = \sigma_X^2(1 - \frac{D}{\sigma_X^2})^2$.
- $R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2)$ is symmetric w.r.t. σ_X^2 and σ_Y^2.
- For fixed σ_Y^2, the function $R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2)$ is maximized when $\sigma_X^2 = \sigma_Y^2 + D$.
- Similar to the case of equal variance, $R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2) = \infty$ when $D \geq \sigma_X^2 + \sigma_Y^2$ since independent $\mathbf{X}, \mathbf{Y}$ have an expected MSE of $\sigma_X^2 + \sigma_Y^2$.

The extension of Theorem 2 reads as follows:

Theorem 4: Let X and Y be Gaussian with variances σ_X^2 and σ_Y^2, respectively. For any fixed rate $R > R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2)$,

$$\begin{aligned} \mathbf{E}_{\mathrm{ID}}(R) = \min_{\rho_X, \rho_Y > 0} & \frac{1}{\ln 2} \left[\mathbf{E}_Z(\rho_X) + \mathbf{E}_Z(\rho_Y)\right] \\ & - \log \sin \min \left[\frac{\pi}{2}, \left(\arcsin(2^{-R}) + \arccos \frac{\rho_X\sigma_X^2 + \rho_Y\sigma_Y^2 - D}{2\sigma_X\sigma_Y\sqrt{\rho_X\rho_Y}}\right)\right]. \end{aligned} \tag{10}$$

D. General sources: an upper bound on the identification rate

Until now we have only considered a Gaussian source. For general sources we have the following upper bound on the identification rate.

Theorem 5: Suppose that both $\mathbf{X}$ and $\mathbf{Y}$ are drawn according to a general distribution P_X with finite second moment. Then $R_{\mathrm{ID}}(D) \leq \inf_{P_{\hat{X}|X}} I(X;\hat{X})$, where the infimum is taken over all test channels $P_{\hat{X}|X}$ satisfying

$$\sqrt{\mathbb{E}_{P_X \times P_{\hat{X}}}(X-\hat{X})^2} \geq \sqrt{\mathbb{E}_{P_{X,\hat{X}}}(X-\hat{X})^2} + \sqrt{D}. \tag{11}$$

In the statement of Theorem 5, $I(\cdot;\cdot)$ denotes mutual information, $\mathbb{E}_{P_{X,\hat{X}}}$ denotes expectation with respect to $P_{X,\hat{X}}$ (i.e., the joint distribution of X and $\hat{X}$), and $\mathbb{E}_{P_X \times P_{\hat{X}}}$ denotes expectation with respect to $P_X \times P_{\hat{X}}$ (i.e., the product of marginal distributions P_X and $P_{\hat{X}}$).

Remark 1: Theorem 5 readily extends to more general similarity metrics. Since the focus of this paper is on quadratic similarity, we omit the details in favor of a simple presentation.

For a general source distribution P_X, we lack a matching lower bound on $R_{\mathrm{ID}}(D)$. However, such a converse was proved in the Gaussian setting (Theorem 1). The key ingredient in the proof of Theorem 1 is the isoperimetric inequality on the surface of a hypersphere – the set on which the probability of a high dimensional Gaussian random vector concentrates (see the following section for details). In general, precise isoperimetric inequalities are unknown and therefore establishing a general converse appears to be extremely difficult.

The following theorem is a corollary of the Theorem 5, and states that among all sources with the same variance, the Gaussian source is most difficult to compress. This is analogous to the setting of classical lossy compression, where the Gaussian distribution has the same extremal property.

Theorem 6: If X is a random variable with finite variance σ^2, then $R_{ID}(D) \leq \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right)$. In particular, a Gaussian source X requires the largest identification rate for a given variance.

In the spirit of Theorem 6, we now describe a single[5] identification system (g^*, T^*) of rate arbitrarily close to $\log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right)$ which permits reliable queries for any source with variance σ^2. To this end, assume $n = 2^\ell$ and let $\mathbb{X} = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)]$ be a matrix, where each column is an independent copy of $\mathbf{X} \sim \prod_{i=1}^n P_X(i)$ (without loss of generality, let $\mathbb{E}X = 0$). Now, define $\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \ldots, \tilde{\mathbf{X}}(n)$ via the matrix multiplication:

$$[\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \ldots, \tilde{\mathbf{X}}(n)] = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)] \times H_\ell, \tag{12}$$

where H_ℓ is the Hadamard matrix, which can be defined recursively as follows, starting with $H_0 = 1$:

$$H_m = \frac{1}{\sqrt{2}} \begin{pmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{pmatrix}. \tag{13}$$

Define $\mathbb{Y} = [\mathbf{Y}(1), \mathbf{Y}(2), \ldots, \mathbf{Y}(n)]$ and $\tilde{\mathbf{Y}}(1), \tilde{\mathbf{Y}}(2), \ldots, \tilde{\mathbf{Y}}(n)$ in an identical manner.

Fix $R > -\log\left(1 - \frac{D}{2\sigma^2}\right)$, and let (g, T) be a rate-R identification system for the Gaussian source $N(0, \sigma^2)$. Define $T^*(\mathbb{X})$ to be the concatenation of the signatures $T(\tilde{\mathbf{X}}(1)), \ldots, T(\tilde{\mathbf{X}}(n))$. Next, define g^* so that $g^*(T^*(\mathbb{X}), \mathbb{Y}) = \texttt{maybe}$ iff $g(T(\tilde{\mathbf{X}}(i)), \tilde{\mathbf{Y}}(i)) = \texttt{maybe}$ for some $i = 1, \ldots, n$.

We claim that (g^*, T^*) is D-admissible and that $\lim_{n\to\infty} \Pr\{g^*(T^*(\mathbb{X}), \mathbb{Y}) = \texttt{maybe}\} = 0$. To see that (g^*, T^*) is D-admissible, note that H_ℓ is an isometry, and therefore $\frac{1}{n^2}\sum_{i=1}^n \|\mathbf{X}(i) - \mathbf{Y}(i)\| =$

[5]Note that in order to prove Theorem 6 we may choose different schemes for different sources. Therefore the existence of this scheme accomplishes more than simply proving Theorem 6. In the following section, we provide an elementary proof of Theorem 6, as a direct corollary of Theorem 5.

$\frac{1}{n^2}\sum_{i=1}^{n}\|\tilde{\mathbf{X}}(i)-\tilde{\mathbf{Y}}(i)\|$. As a consequence, D-admissibility follows since $d(\mathbb{X},\mathbb{Y})\leq D$ only if $d(\tilde{\mathbf{X}}(i),\tilde{\mathbf{Y}}(i))\leq D$ for some i.

Crucially, observe that $\tilde{\mathbf{X}}(i)$ (resp. $\tilde{\mathbf{Y}}(i)$) is a random vector with entries drawn i.i.d. according to some distribution $\tilde{P}_i$. Furthermore, $\tilde{P}_i\to N(0,\sigma^2)$ in distribution (by the central limit theorem) uniformly[6] in i. Therefore, it is relatively easy to show that $\Pr\left\{g\left(T(\tilde{\mathbf{X}}(i)),\tilde{\mathbf{Y}}(i)\right)=\texttt{maybe}\right\}$ is exponentially small in n for each i (we omit the details here due to space constraints). By the union bound, $\lim_{n\to\infty}\Pr\left\{g^*\left(T^*(\mathbb{X}),\mathbb{Y}\right)=\texttt{maybe}\right\}=0$ as desired.

Remark 2: Note that $\mathbf{X}$ and $\mathbf{Y}$ need not have the same distribution for the above scheme to work.

IV. Proofs

Here we provide proof outlines for Theorems 1, 2, 5 and 6. The proofs of Theorems 3 and 4 are technical variations of Theorems 1 and 2 and are omitted due to space limitation.

A. Identification Rate for a Gaussian Source

Proof of Theorem 1: The proof relies on the fact that, at high dimensions, a Gaussian random vector concentrates near a thin hyper-spherical shell of radius $r_0\triangleq\sqrt{n\sigma^2}$, denoted S_{r_0}. Formally, define the *typical spherical shell* $S_{\text{typ}}\triangleq\{\mathbf{x}\in\mathbb{R}^n:\ |\frac{1}{n}\|\mathbf{x}\|^2-\sigma^2|<\eta\}$ for some fixed η.

Direct Part: Following a result on spherical shell covering by Dumer [9], it can be shown that for any $0<D_0<\sigma^2$, there exists a code $\mathcal{C}$ that D_0-covers the spherical shell of radius r_0 with rate being essentially $R_0=\frac{1}{2}\log\frac{\sigma^2}{D_0}$, i.e. the Gaussian rate distortion function for distortion D_0. Moreover, the covering property guarantees that each quantization cell (the set of points on the shell that are mapped to a single reconstruction point) is contained in a spherical cap of radius r_0 and half-angle θ_0 given by

$$\theta_0\triangleq\arcsin(\sqrt{D_0/\sigma^2})<\frac{\pi}{2}. \tag{14}$$

The mapping T is defined as follows: if $\mathbf{x}$ is on a the typical spherical shell S_{typ}, then $T(\mathbf{x})$ is the index to any point $\mathbf{u}$ from the covering code $\mathcal{C}$ that D_0-covers $\frac{r_0}{\|\mathbf{x}\|}\mathbf{x}$ (the projection of $\mathbf{x}$ on the spherical shell S_{r_0}). If $\mathbf{x}\notin S_{\text{typ}}$, then $T(X)=\varepsilon$, denoting erasure. The overall rate of the scheme is essentially R_0, and is not affected by the introduction of the erasure signature ε.

The decision function $g(t,\mathbf{y})$ is defined as follows:

$$g(t,\mathbf{y})=\begin{cases}\texttt{maybe}, & \text{If } t=\varepsilon \text{ or if } \arccos\frac{\mathbf{y}\cdot\mathbf{u}(t)}{\|\mathbf{y}\|\|\mathbf{u}(t)\|}\leq\theta_0+\theta_1;\\ \texttt{no}, & \text{otherwise,}\end{cases} \tag{15}$$

where $\theta_1\triangleq\arccos\frac{2(\sigma^2-\eta)-D}{2(\sigma^2-\eta)}$.

It follows from the triangle inequality for geodesic distance that whenever $d(\mathbf{x},\mathbf{u})\leq D_0$ (for $\mathbf{x}\in S_{\text{typ}}$) and $d(\mathbf{x},\mathbf{y})\leq D$, then the condition above involving $\arccos\frac{\mathbf{y}\cdot\mathbf{u}}{\|\mathbf{y}\|\|\mathbf{u}\|}$ holds, and therefore the proposed scheme is admissible.

Because of the spherical symmetry of the pdf of $\mathbf{Y}$, it follows that the probability for `maybe` is upper bounded by the fraction of the spherical surface area occupied by a cap with half angle $\theta_0+\theta_1$, denoted $\Omega(\theta_0+\theta_1)$. By [10, Corrolary 3.2], it follows that[7] $\Omega(\theta)\doteq\sin(\theta)^n$ when $\theta<\pi/2$. The condition $\theta_0+\theta_1<\pi/2$ translates to $R>R^*(D,\eta)$, where $R^*(D,\eta)\to R_{\text{ID}}(D)$ when $\eta\to 0$, leading to the direct part of the proof.

Remark 3: The alert reader will observe that the direct part also follows from the direct part of Theorem 2, and also from Theorem 6. However, we have chosen to include an explicit proof here to introduce the notations and ideas crucial for proving the converse part and Theorem 2.

[6]The fact that convergence takes place uniformly follows from the observation that $\tilde{P}_i=\tilde{P}_j$ for $i,j>1$.

[7]$a\doteq b$ denotes $\lim_{n\to\infty}\frac{1}{n}\log\frac{a}{b}=0$.

Converse Part: Given any rate-R scheme (T, g), we consider the behavior of the mapping T on the typical sphere S_{typ}. Standard arguments show that it suffices to consider schemes with quantization cells (on the typical sphere) of equal surface area. Since there are 2^{nR} signatures, each cell occupies $\Omega(\theta_0)$ of the entire surface area of the shell, where $\theta_0 \cong \arcsin(2^{-R})$.

Next, we wish to give a lower bound on the surface area of the set of typical $\mathbf{y}$ sequences that are D-similar to any $\mathbf{x}$ in a quantization cell (sometimes called a D-expansion of the cell), since it is proportional to the probability of this set (which, in turn, lower bounds the probability for `maybe`). The isoperimetric inequality on the surface of the sphere (known as Levy's lemma, see e.g. [11, Theorem 1.1]) gives exactly this. It says that for a given surface area α, the set (on the spherical shell) with area α that minimizes the surface area of its D-expansion is a spherical cap. From here the analysis follows the same way as in the direct part, since the quantization cells there are indeed spherical caps, and since for $\theta > \pi/2$, $\Omega(\theta) \to 1$ when n grows. ■

B. Identification Exponent for a Gaussian Source

Proof of Theorem 2: Characterizing the optimal exponent requires a slightly more sophisticated scheme than only quantizing the typical sphere. The required extension is to partition the entire space (up to a large radius $r_{\max}$) to thin spherical shells, and quantize each shell separately using a code as in the achievability part of the identification rate proof. If $\|\mathbf{X}\| > r_{\max}$, an erasure ε is declared. By a judicious selection of $r_{\max}$ and the thickness of the shells, we have :

- The radius $\|\mathbf{x}\|$ is revealed (arbitrarily accurately) through the signature,
- The specification of which shell $\mathbf{x}$ was in has essentially no effect on the rate,
- The probability for an erasure ε vanishes super-exponentially and does not affect the exponent.

We denote $Z_X \triangleq \frac{1}{\sigma^2}\|\mathbf{X}\|^2$ and $Z_Y \triangleq \frac{1}{\sigma^2}\|\mathbf{Y}\|^2$. Then the probability for `maybe` (assuming $\|\mathbf{x}\|, \|by\| < r_{\max}$) can be written as

$$\Pr\{\texttt{maybe}\} = \int_0^{r_{\max}^2/\sigma^2}\int_0^{r_{\max}^2/\sigma^2} \Pr\{\texttt{maybe}|Z_X = z_X, Z_Y = z_Y\} f_Z(z_X) f_Z(z_Y) dz_X dz_Y \quad (16)$$

$$\doteq \max_{0\le\rho_X,\rho_Y} \Pr\{\texttt{maybe}|Z_X = n\cdot\rho_X, Z_Y = n\cdot\rho_Y\} f_Z(n\cdot\rho_X) f_Z(n\cdot\rho_Y) \quad (17)$$

where $f_Z(z)$ is the pdf of a chi-square random variable with n degrees of freedom.

By the definition of $f_Z(\cdot)$ and the Stirling approximation it can be shown that for any $\rho > 0$,

$$f_Z(n\rho) \doteq \exp\left[-n\mathbf{E}_Z(\rho)\right], \quad (18)$$

with $\mathbf{E}_Z(\cdot)$ defined in (8). By arguments similar to those in the proof of the identification rate, the conditional probability for `maybe` is upper bounded by $\Omega\left(\theta_0 + \theta_1(\rho_X, \rho_Y)\right)$, where $\theta_1(\rho_X, \rho_Y) \triangleq \arccos\frac{\rho_x+\rho_Y-\frac{D}{\sigma^2}}{2\sqrt{\rho_X\cdot\rho_Y}}$. Recalling that $\Omega(\theta) \doteq \sin^n\theta$ for $\theta < \pi/2$ gives the exponential form as in (10) with $\sigma_X = \sigma_Y = \sigma$, and symmetry arguments lead to the simplified (7). Note that the minimum w.r.t. ρ is always attained because $\mathbf{E}_Z(\rho) \to \infty$ for both $\rho \to 0$ and $\rho \to \infty$.

For the converse part, take the value ρ^* that minimizes (7) and construct a thin spherical shell S^* with radius $\sqrt{n\rho^*\sigma^2}$ and small, fixed, nonzero thickness. The converse then follows the same steps of the converse for the identification rate, where the typical shell S_{typ} is replaced by the shell S^*. ■

C. General Sources and the Extremal Property of the Gaussian

Proof of Theorem 5: We can assume that X is a discrete random variable with finite support. The extension to continuous distributions (with finite variance) follows by the usual quantization

arguments and continuity of $\|\cdot\|$. Fix $\epsilon > 0$ and a conditional pmf $P_{\hat{X}|X}(\hat{x}|x)$. Let $\mathcal{T}_\epsilon^{(n)}$ denote the usual ϵ-typical set (cf. [12, Chapter 1]).

Signature assignment. Randomly and independently generate 2^{nR} sequences $\hat{\mathbf{x}}(t), t \in [1 : 2^{nR}]$, each according to $\prod_{i=1}^n P_{\hat{X}}(\hat{x}_i)$. Given a sequence $\mathbf{x}$, find an index t such that $(\mathbf{x}, \hat{\mathbf{x}}(t)) \in \mathcal{T}_\epsilon^{(n)}$ and put $T(\mathbf{x}) = t$. If there is more than one such index, break ties arbitrarily. If there is no such index, put $T(\mathbf{x}) = \varepsilon$. Again, the rate R is negligibly affected by the addition of the additional signature ε.

Definition of the query function. For a signature $t \in [1 : 2^{nR}] \cup \{\varepsilon\}$ and a sequence $\mathbf{y}$, define

$$g(t, \mathbf{y}) = \begin{cases} \texttt{maybe} & \text{if} \begin{cases} t = \varepsilon, \text{ or} \\ \frac{1}{\sqrt{n}}\|\mathbf{y} - \hat{\mathbf{x}}(t)\| \le \sqrt{(1+\epsilon)\mathbb{E}_{P_{X,\hat{X}}}(X - \hat{X})^2} + \sqrt{D} \text{ and } t \neq \varepsilon \end{cases} \\ \texttt{no} & \text{otherwise.} \end{cases}$$

Scheme analysis. We first check to ensure that $g(\cdot, \cdot)$ does not produce any false negatives. It suffices to restrict our attention to the case where $t \in [1 : 2^{nR}]$. By definition of the signature assignment, if $T(\mathbf{x}) = t \in [1 : 2^{nR}]$, then $(\mathbf{x}, \hat{\mathbf{x}}(t)) \in \mathcal{T}_\epsilon^{(n)}$, and therefore the typical average lemma [12] implies

$$\frac{1}{n}\|\mathbf{x} - \hat{\mathbf{x}}(t)\|^2 \le (1+\epsilon)\mathbb{E}_{P_{X,\hat{X}}}(X - \hat{X})^2. \tag{19}$$

Thus, for any $\mathbf{y} \in \mathcal{X}^n$, the triangle inequality implies that $\frac{1}{n}\|\mathbf{x} - \mathbf{y}\|^2 \le D$ only if

$$\frac{1}{\sqrt{n}}\|\mathbf{y} - \hat{\mathbf{x}}(t)\| \le \sqrt{(1+\epsilon)\mathbb{E}_{P_{X,\hat{X}}}(X - \hat{X})^2} + \sqrt{D}. \tag{20}$$

Hence, our choice of $g(\cdot, \cdot)$ does not produce any false negatives, as desired.

Next, we check to ensure that $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$ is small. To this end, consider the events

$$\mathcal{E}_1 = \{T(\mathbf{X}) = \varepsilon\}, \qquad \mathcal{E}_2 = \left\{\frac{1}{\sqrt{n}}\|\mathbf{Y} - \hat{\mathbf{x}}(T(\mathbf{X}))\| \le \sqrt{(1+\epsilon)\mathbb{E}_{P_{X,\hat{X}}}(X - \hat{X})^2} + \sqrt{D}\right\},$$

and observe that $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\} \le \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2)$.

Here, we adopt the usual convention by letting $\delta(\epsilon)$ denote a positive quantity satisfying $\lim_{\epsilon \to 0} \delta(\epsilon) = 0$. Standard arguments yield

$$\Pr(\mathcal{E}_1) \le \exp\left(-2^{n(R - I(X;\hat{X}) - \delta(\epsilon))}\right), \tag{21}$$

which is exponentially small in n if $R > I(X; \hat{X}) + \delta(\epsilon)$.

Next, we recall the sequence $\mathbf{Y}$ is independent of $\mathbf{X}$, and therefore is also independent of $\hat{\mathbf{x}}(T(\mathbf{X}))$. An application of Hoeffding's inequality implies

$$\Pr\left(\frac{1}{\sqrt{n}}\|\mathbf{Y} - \hat{\mathbf{x}}(T(\mathbf{X}))\| \le \sqrt{\mathbb{E}_{P_X \times P_{\hat{X}}}(X - \hat{X})^2} - \epsilon\right) \le \exp(-n\delta(\epsilon)). \tag{22}$$

From this, we can conclude that $\Pr(\mathcal{E}_2)$ is exponentially small in n if

$$\sqrt{(1+\epsilon)\mathbb{E}_{P_{X,\hat{X}}}(X - \hat{X})^2} + \sqrt{D} \le \sqrt{\mathbb{E}_{P_X \times P_{\hat{X}}}(X - \hat{X})^2} - \epsilon. \tag{23}$$

■

Proof of Theorem 6: Assume without loss of generality that $\mathbb{E}[X] = 0$. Consider the test channel $P_{\hat{X}|X}$ defined by $\hat{X} = \rho X + Z$, where $Z \sim N(0, \sigma_Z^2)$ is independent of X and ρ, σ_Z^2 are given by

$$\rho = \frac{(4\sigma^2 - D)}{(2\sigma^2)} \qquad\qquad \sigma_Z^2 = \frac{(4\sigma^2 - D)(2\sigma^2 - D)^2}{4\sigma^2 D}.$$

With $P_{\hat{X}|X}$ defined in this way, the following identities are readily verified

$$\sqrt{\mathbb{E}_{P_X \times P_{\hat{X}}}(X-\hat{X})^2} = \sqrt{\sigma^2(1+\rho^2)+\sigma_Z^2} = \frac{2\sigma^2}{\sqrt{D}}$$
$$\sqrt{\mathbb{E}_{P_{X,\hat{X}}}(X-\hat{X})^2} = \sqrt{\sigma^2(1-\rho)^2+\sigma_Z^2} = \frac{2\sigma^2-D}{\sqrt{D}}.$$

Therefore, we have $\sqrt{\mathbb{E}_{P_X \times P_{\hat{X}}}(X-\hat{X})^2} \geq \sqrt{\mathbb{E}_{P_{X,\hat{X}}}(X-\hat{X})^2} + \sqrt{D}$, as desired. Since $\hat{X}$ has density and the Gaussian distribution maximizes differential entropy for a given variance (cf. [6]), we have the inequality $h(\hat{X}) \leq \frac{1}{2}\log\left(2\pi e(\rho^2\sigma^2+\sigma_Z^2)\right)$. It follows that

$$I(X;\hat{X}) \leq \frac{1}{2}\log\left(\frac{\rho^2\sigma^2+\sigma_Z^2}{\sigma_Z^2}\right) = \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right).$$

An application of Theorem 5 completes the proof. ■

V. Concluding Remarks

We studied the problem of answering similarity queries from compressed data from an information-theoretic perspective. We focused on the setting where the similarity criterion is the (normalized) quadratic distance. For the case of i.i.d. Gaussian data, we gave an explicit characterization of the minimal compression rate which permits reliable queries (i.e., the identification rate). Furthermore, we characterize the exponential rate at which the probability for false positives vanishes.

For general sources, we derive an upper bound on the identification rate, and prove that it is at most that of the Gaussian source of the same variance. Finally, we presented a single scheme that compresses *any* source at the Gaussian identification rate, while permitting reliable queries.

Acknowledgement

The authors would like to thank Golan Yona for stimulating discussions that motivated this work.

References

[1] R. Ahlswede, E.-h. Yang, and Z. Zhang, "Identification via compressed data," *Information Theory, IEEE Transactions on*, vol. 43, no. 1, pp. 48 –70, Jan 1997.

[2] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.

[3] A. Z. Broder and M. Mitzenmacher, "Survey: Network applications of Bloom filters: A survey," *Internet Mathematics*, vol. 1, no. 4, pp. 485–509, 2003.

[4] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, no. 1, pp. 117–122, 2008.

[5] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & sons, 1991.

[7] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. on Information Theory*, vol. 20, no. 2, pp. 197 – 199, mar 1974.

[8] S. Ihara and M. Kubo, "Error exponent of coding for memoryless Gaussian sources with a fidelity criterion," *IEICE Transactions*, vol. 83-A, no. 10, pp. 1891–1897, 2000.

[9] I. Dumer, "Covering spheres with spheres." *Discrete & Computational Geometry*, vol. 38, no. 4, pp. 665–679, 2007.

[10] K. Böröczky Jr. and G. Wintsche, "Covering the sphere by equal spherical balls," in *Discrete and Computational Geometry: The Goodman-Pollack Festschrift*. Springer, 2003, pp. 237–253.

[11] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, ser. Ergebnisse der Mathematik Und Ihrer Grenzgebiete. Springer, 2011.

[12] A. Gamal and Y. Kim, *Network Information Theory*. Cambridge University Press, 2011.