Compression Schemes for Similarity Queries

Idoia Ochoa, Amir Ingber and Tsachy Weissman
Dept. of Electrical Engineering, Stanford University, Stanford CA 94305
Email: {iochoa, ingber, tsachy}@stanford.edu

Abstract

We consider compression of sequences in a database so that similarity queries can be performed efficiently in the compressed domain. The fundamental limits for this problem setting, which characterize the tradeoff between compression rate and reliability of the answers to the queries, have been characterized in past work. However, how to approach these limits in practice has remained largely unexplored.

Recently, we proposed a scheme for this task that is based on existing lossy compression algorithms, for the general case where the similarity measure satisfies a triangle inequality. Although it was shown that it achieves the fundamental limits for some cases, it is suboptimal in general. In this paper we propose a new scheme that also uses lossy compression algorithms as a building block, but with a carefully chosen distortion measure that is different than the one defining the similarity between sequences. The new scheme significantly improves the compression rate compared to the previously proposed scheme in many cases. For example, for binary sources and Hamming similarity measure, simulation results show a compression rate close to the fundamental limit, and an improvement over the previously proposed scheme of up to 55% (for the same reliability). The results shed light on the fact that compression for similarity identification is inherently different than classical lossy compression.

I. Introduction

The generation of new databases and the amount of data on existing ones is growing rapidly. Due to their size, performing queries on these databases can be a challenging task. With this in mind, we study the problem of compressing a database so that queries about the original data can be answered efficiently given only the compressed version. By compressing the database, it will become possible to replicate the compressed database in several locations, thus providing easier and faster access, and potentially reducing the time needed to execute a query. Specifically, we focus on queries of the form: *"which sequences in the database are similar to a given sequence* $\mathbf{y}$*?"*, which are of practical interest in many applications.

More formally, we consider schemes that generate, for each sequence $\mathbf{x}$ in the database, a short *signature* of fixed-length, denoted by $T(\mathbf{x})$, that is stored in the compressed database. Then, given a query sequence $\mathbf{y}$, we answer the question of whether $\mathbf{x}$ and $\mathbf{y}$ are similar, based only on the signature $T(\mathbf{x})$, rather than the original sequence $\mathbf{x}$.

When answering a query, there are two types of errors that can be made: a *false positive*, when a sequence is misidentified as similar to the query sequence; and a *false negative*, when a similar sequence stays undetected. We impose the restriction that false negatives are not permitted, as even a small probability of a false negative translates to a substantial probability of misdetection of some sequences in the large database, which is unacceptable in many applications. On the other hand, false positives do not cause an error *per se* as the precise level of similarity is assessed upon retrieval of the full sequence from the large database. However, they introduce a computational burden due to the need of further verification (retrieval), so we would like to reduce their probability as much as possible.

This work is supported by a grant from the Center for Science of Information (CSoI), a fellowship from the Basque government and a Google research award.

This problem has been studied from an information-theoretic perspective in [1], [2] for discrete sources, and in [3] for Gaussian sources. These papers analyze the fundamental tradeoff between compression rate, sequence length and reliability of queries performed on the compressed data. Although these limits enable to analyze the optimality of a given scheme, the achievability proofs are non-constructive, which raises the question of how to design such schemes in practice.

In this context, we recently proposed a scheme in [4] based on lossy compression algorithms, which was shown to achieve the fundamental limits for the case where i) similarity is measured by Hamming distortion, and ii) both the sequences in the database and the query sequences are i.i.d. with entries drawn independently from a Bern(0.5) distribution. However, as discussed in [2], this scheme is suboptimal in general.

With that in mind, in this paper we propose a new scheme which builds upon the one proposed in [4], that significantly improves the compression rate in many cases. Furthermore, it achieves a compression rate close to the fundamental limit for the case of general memoryless binary sources and Hamming distortion. The proposed scheme also uses lossy compression algorithms as a building block. Specifically, the signature of a sequence $\mathbf{x}$ is composed of a compressed description of a reconstruction sequence $\hat{\mathbf{x}}$ (the output of a lossy compressor), and some additional information. However, while the scheme introduced in [4] uses off-the-shell lossy compressors, the proposed scheme carefully chooses the distortion measure to be used by the lossy compressor, such that the empirical distribution of the sequences $\mathbf{x}$ and $\hat{\mathbf{x}}$ is close to the optimal one (the one required for achieving the fundamental limit of compression for similarity identification).

For general binary sources and similarity measured by Hamming distortion, we show by simulation that the proposed scheme attains a notable improvement in performance over the one introduced in [4], and a compression rate close to the fundamental limit. For the case of binary symmetric sources and Hamming distortion, both schemes coincide. Finally, these schemes are easy to analyze and implement, and they can work with any discrete database and any similarity measure satisfying the triangle inequality.

Variants of this problem have been previously considered in the literature. For example, the Bloom filter [5], which is restricted to exact matches, enables membership queries from compressed data. Another related notion is that of Locality-Sensitive Hashing (LSH) [6, Chapter 3], which is a framework for the nearest neighbor search (NNS) problem. The key idea of LSH is to hash points in such a way that the probability of collision is higher for points that are similar than for those that are far apart. Other methods for NNS include vector approximation files (VA-File) [7], that employs scalar quantization. An extension of this method is the so called compression/clustering based search [8], which performs vector quantization implemented through clustering. While these techniques trade off accuracy with computational complexity and space, and false negatives are allowed, in our setting false negatives are not allowed, but significant compression can still be achieved.

The rest of the paper is organized as follows. In Section II we formalize the problem and recall the fundamental limits. In Section III and IV we review the scheme proposed in [4] and introduce the new one, respectively. Simulation results are shown in Section V, and we provide some concluding remarks in section VI.

II. Problem Formulation and Fundamental Limits

A. Notation and Problem Description

Let upper case, lower case, and calligraphic letters denote random variables, their specific realizations, and their alphabets, respectively. Boldface notation $\mathbf{x}$ denotes a

vector of length n, i.e., $\mathbf{x} = [x_1, \ldots, x_n]^T$, and $[1:k]$ denotes the set $\{1, 2, ..., k\}$.

Given two sequences $\mathbf{x}$ and $\mathbf{y}$, we measure their *similarity* by computing the distortion $d(\mathbf{x}, \mathbf{y})$ given by $\frac{1}{n}\sum_{i=1}^{n} \rho(x_i, y_i)$, where $\rho : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}^+$ is an arbitrary distortion measure. We say that two sequences $\mathbf{x}$ and $\mathbf{y}$ are *D-similar* (or simply *similar* when clear from the context) when $d(\mathbf{x}, \mathbf{y}) \leq D$.

We consider databases consisting of M discrete sequences of length n, i.e., $\{\mathbf{x}^{(i)}\}_{i=1}^{M}$. The proposed architecture generates, for each sequence $\mathbf{x}$, a signature $T(\mathbf{x})$, so that the compressed database is $\{T(\mathbf{x}^{(i)})\}_{i=1}^{M}$. Then, given a query sequence $\mathbf{y}$, the scheme makes the decision of whether $\mathbf{x}$ is *D*-similar to $\mathbf{y}$, based only on its compressed version $T(\mathbf{x})$, rather than on the original sequence $\mathbf{x}$. Note that a scheme is completely defined given its signature assignment and the corresponding decision rule.

More formally, a rate-R identification system (T, g) consists of a signature assignment $T : \mathcal{X}^n \to [1 : 2^{nR}]$ and a decision function $g : [1 : 2^{nR}] \times \mathcal{Y}^n \to \{\texttt{no}, \texttt{maybe}\}$. We use the notation $\{\texttt{no}, \texttt{maybe}\}$ instead of $\{\texttt{no}, \texttt{yes}\}$ to reflect the fact that false positives are permitted, while false negatives are not. This is formalized next. A system is said to be *D-admissible* if

$$g(T(\mathbf{x}), \mathbf{y}) = \texttt{maybe} \;\forall\; \mathbf{x}, \mathbf{y} \; s.t. \; d(\mathbf{x}, \mathbf{y}) \leq D. \tag{1}$$

Since a *D-admissible* scheme does not produce false negatives, a natural figure of merit is the frequency at which false positives occur, that we wish to minimize.

We recall next the fundamental limits on performance in this problem, as we will refer to them in the following sections when assessing the performance of the scheme proposed in [4] and that of the new one.

B. Fundamental limits

Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors of length n, representing the sequence from the database and the query sequence, respectively. We assume $\mathbf{X}$ and $\mathbf{Y}$ are independent, with entries drawn independently from P_X and P_Y, respectively. Define the *false positive event* as $\texttt{fp} = \{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe} | d(\mathbf{X}, \mathbf{Y}) > D\}$. For a *$D$-admissible* scheme,

$$P(g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}) = P(d(\mathbf{X}, \mathbf{Y}) \leq D) + P(\texttt{fp})P(d(\mathbf{X}, \mathbf{Y}) > D). \tag{2}$$

Note that $P(\texttt{fp})$ is the only term that depends on the scheme used, as the other terms depend strictly on the probability distribution of $\mathbf{X}$ and $\mathbf{Y}$. Hence minimizing $P(\texttt{fp})$ over all D-admissible schemes is equivalent to minimizing $P(g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe})$. Thus, for a given D, the fundamental limits characterize the tradeoff between the compression rate R and the $P(g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe})$.

Note that as $n \to \infty$, $P(d(\mathbf{X}, \mathbf{Y}) \leq D)$ goes to one or to zero (according to whether D is above or below the expected level of similarity between X and Y). The problem is non-trivial only when the event of similarity is atypical, the case on which we focus. In this case, as is evident from (2), $P(\texttt{maybe}) \to 0$ iff $P(\texttt{fp}) \to 0$.

Definition 1: For given distribution P_X, P_Y 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. $\lim_{n\to\infty} P\left(g^{(n)}\left(T^{(n)}(\mathbf{X}), \mathbf{Y}\right) = \texttt{maybe}\right) = 0$.

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

For the case considered in this paper: discrete sources, fixed-length signature assignment and zero false negatives, the identification rate is characterized in [2, Theorem 1] as

$$R_{\mathrm{ID}}(D) = \min_{P_{U|X}: \sum_{u\in\mathcal{U}} P_U(u)\bar{\rho}(P_{X|U}(\cdot|u), P_Y) \geq D} I(X; U), \tag{3}$$

where U is any random variable with finite alphabet $\mathcal{U}$ ($|\mathcal{U}| = |\mathcal{X}| + 2$ suffices to obtain the true value of $R_{\mathrm{ID}}(D)$), that is independent of Y. $\bar{\rho}(P_X, P_Y) = \min \mathbb{E}[\rho(X, Y)]$ is a distance between distributions, with ρ being the distortion under which similarity is measured, and where the minimization is w.r.t. all jointly distributed random variables X, Y with marginal distributions P_X and P_Y, respectively.

Finally, we define $D_{\mathrm{ID}}(R)$ as the inverse function of $R_{\mathrm{ID}}(D)$, i.e., the similarity threshold below which any similarity level can be achieved at given rate R.

As stated in the introduction, the scheme proposed in [4] was shown to achieve these limits in some particular examples, but not in general. Next we review this scheme and its optimality, which was analyzed in [2], as this will lead to the new scheme that we propose in this paper.

The scheme proposed in [4] and the one proposed in this paper are based on Lossy Compressors (LC) and on a Type Covering lemma (TC), respectively, and they both use a decision rule based on the triangle inequality ($\triangle$). Based on this, and to be consistent with the notation used in [2], hereafter we refer to them as the $\mathrm{LC}-\triangle$ and $\mathrm{TC}-\triangle$ schemes, respectively. Note that whereas a scheme based on lossy compressors ($\mathrm{LC}-\triangle$ scheme) is straightforward to implement, as we did in [4], implementation of the type covering lemma based scheme ($\mathrm{TC}-\triangle$ scheme) in practice is more challenging.

III. The $\mathrm{LC}-\triangle$ scheme

A. Description

The signature of the $\mathrm{LC}-\triangle$ scheme is based on fixed-length lossy compression algorithms. They are characterized by an encoding function $f_n : \mathbf{x} \to [1 : 2^{nR'}]$ and a decoding function $g_n : [1 : 2^{nR'}] \to \hat{\mathbf{x}}$, where $\hat{\mathbf{x}} = g_n(f_n(\mathbf{x}))$ denotes the reconstructed sequence. Specifically, the signature of a sequence $\mathbf{x}$ is composed of the output $i \in [1 : 2^{nR'}]$ of the lossy compressor, and the distortion between $\mathbf{x}$ and $\hat{\mathbf{x}}$, i.e., $T(\mathbf{x}) = \{i, d(\mathbf{x}, \hat{\mathbf{x}})\}$ (see Fig. 1). The total rate of the system is $R = R' + \Delta R$, where R' is the rate of the lossy-compressor and ΔR represents the extra rate to represent and store the distortion value $d(\mathbf{x}, \hat{\mathbf{x}})$.

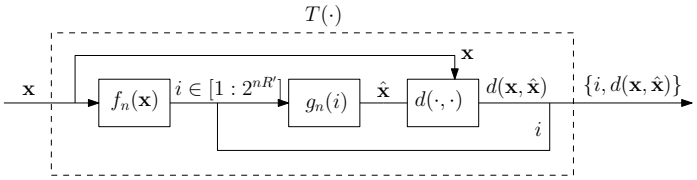


Fig. 1: Signature assignment of the $\mathrm{LC}-\triangle$ scheme for each sequence $\mathbf{x}$ in the database.

Regarding the decision function $g : [1 : 2^{nR}] \times \mathcal{Y}^n \to \{\mathtt{no}, \mathtt{maybe}\}$, recall that it must satisfy (1). Given the signature assignment described above, the decision rule for sequence $\mathbf{x}$ and query sequence $\mathbf{y}$ is based on the tuple $(T(\mathbf{x}), \mathbf{y}) = (\{i, d(\mathbf{x}, \hat{\mathbf{x}})\}, \mathbf{y})$. Notice that $\hat{\mathbf{x}}$ can be recovered from the signature as $g_n(i)$. The decision rule is given by

$$g(T(\mathbf{x}), \mathbf{y}) = \begin{cases} \mathtt{maybe}, & d(\mathbf{x}, \hat{\mathbf{x}}) - D \le d(\hat{\mathbf{x}}, \mathbf{y}) \le d(\mathbf{x}, \hat{\mathbf{x}}) + D; \\ \mathtt{no}, & \text{otherwise}, \end{cases} \tag{4}$$

which satisfies (1) for any given distortion measure satisfying the triangle inequality.

In an attempt to reduce the rate of the system (e.g., decrease the size of the compressed database) without affecting the performance, one can decrease the value of ΔR by

quantizing the distortion $d(\mathbf{x}, \hat{\mathbf{x}})$. In that case, assuming $d_0 \leq d(\mathbf{x}, \hat{\mathbf{x}}) \leq d_1$,

$$g(T(\mathbf{x}), \mathbf{y}) = \begin{cases} \texttt{maybe}, & d_0 - D \leq d(\hat{\mathbf{x}}, \mathbf{y}) \leq d_1 + D; \\ \texttt{no}, & \text{otherwise}, \end{cases} \tag{5}$$

which preserves the admissibility of the scheme. While ΔR can be arbitrary small (for $n \to \infty$), there is a tradeoff for finite n between its value and the $P(\texttt{maybe})$, as demonstrated in [4]. This will become relevant for the simulations.

B. Asymptotic analysis

Recall from rate distortion theory [9] that an optimal lossy compressor with rate R attains for long enough sequences and with high probability, a distortion between $\mathbf{x}$ and $\hat{\mathbf{x}}$ arbitrarily close to the distortion-rate function $D(R)$. Finally, consider the looser decision rule $g(T(\mathbf{x}), \mathbf{y}) = \texttt{no}$ if $d(\hat{\mathbf{x}}, \mathbf{y}) > d(\mathbf{x}, \hat{\mathbf{x}}) + D$. Note that the scheme is still admissible (zero false negatives) with this decision rule. Under these premises, as shown in [2], an $\mathrm{LC} - \triangle$ scheme of rate R can attain any similarity threshold below $D_{\mathrm{ID}}^{\mathrm{LC}-\triangle}(R)$, with

$$D_{\mathrm{ID}}^{\mathrm{LC}-\triangle}(R) \triangleq \mathbb{E}[\rho(\hat{X}, Y)] - \mathbb{E}[\rho(X, \hat{X})] = \mathbb{E}[\rho(\hat{X}, Y)] - D(R), \tag{6}$$

where $\mathbb{E}[\rho(\hat{X}, Y)]$ is completely determined by $P_{\hat{X}}$ (induced by the lossy compressor) and P_Y. Finally, let $R_{\mathrm{ID}}^{\mathrm{LC}-\triangle}(D)$ be the inverse function of $D_{\mathrm{ID}}^{\mathrm{LC}-\triangle}(R)$, i.e., the compression rate achieved for a similarity threshold D.

As shown in [2], for binary symmetric sources and Hamming distortion, $R_{\mathrm{ID}}(D) = R_{\mathrm{ID}}^{\mathrm{LC}-\triangle}(D)$, i.e., the scheme achieves the fundamental limit. However, the scheme is suboptimal in general, in the sense that $R_{\mathrm{ID}}(D) < R_{\mathrm{ID}}^{\mathrm{LC}-\triangle}(D)$.

IV. The $\mathrm{TC} - \triangle$ scheme

A. Motivation

A closer look at (6) suggests the following intuitive idea: in the distortion rate case, we wish to minimize the distortion with a constraint on the mutual information. The optimization is with respect to the transition probability $P_{\hat{X}|X}$. This is in agreement with (6), as we also want to minimize $\mathbb{E}[\rho(X, \hat{X})]$. However, the quantity $\mathbb{E}[\rho(\hat{X}, Y)]$ also depends on $P_{\hat{X}}$ (determined by $P_{\hat{X}|X}$ and P_X). This suggests optimizing both terms together. As shown in [2], this is possible, and the key is to use a type covering lemma (TC) to generate $\hat{\mathbf{x}}$ (and not just the one that minimizes the distortion between X and $\hat{X}$). Specifically, any similarity threshold below $D_{\mathrm{ID}}^{\mathrm{TC}-\triangle}(R)$ can be attained by a $\mathrm{TC} - \triangle$ scheme of rate R, where

$$D_{\mathrm{ID}}^{\mathrm{TC}-\triangle}(D) \triangleq \max_{P_{\hat{X}|X}: I(X;\hat{X}) \leq R} \mathbb{E}[\rho(\hat{X}, Y)] - \mathbb{E}[\rho(X, \hat{X})]. \tag{7}$$

As in the previous case, we denote by $R_{\mathrm{ID}}^{\mathrm{TC}-\triangle}(D)$ the inverse function of $D_{\mathrm{ID}}^{\mathrm{TC}-\triangle}(R)$. It is easy to see that $R_{\mathrm{ID}}^{\mathrm{TC}-\triangle}(D) \leq R_{\mathrm{ID}}^{\mathrm{LC}-\triangle}(D)$. Furthermore, for memoryless binary sources and Hamming distortion $R_{\mathrm{ID}}^{\mathrm{TC}-\triangle}(D) = R_{\mathrm{ID}}(D)$ and both are strictly lower than $R_{ID}^{\mathrm{LC}-\triangle}(D)$ for non-symmetric sources, the difference being particularly pronounced at low distortion, as shown in [2] (see Fig 2).

The question now is how to create a practical $\mathrm{TC}-\triangle$ scheme that achieves $R_{\mathrm{ID}}^{\mathrm{TC}-\triangle}(D)$, which will imply that the scheme achieves a smaller compression rate that an $\mathrm{LC} - \triangle$ scheme, and that it is optimal for general binary sources and Hamming distortion. While

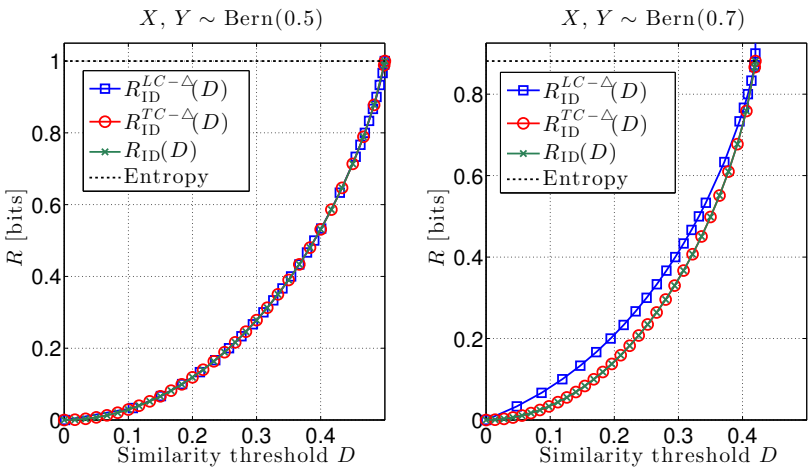


Fig. 2: Binary sources and Hamming distortion: if $P_X = P_Y = \text{Bern}(0.5)$, $R_{\text{ID}}^{\text{LC}-\triangle}(D) = R_{\text{ID}}^{\text{TC}-\triangle}(D) = R_{\text{ID}}(D)$, whereas if $P_X = P_Y = \text{Bern}(0.7)$, $R_{\text{ID}}^{\text{LC}-\triangle}(D) > R_{\text{ID}}^{\text{TC}-\triangle}(D) = R_{\text{ID}}(D)$.

creating a practical scheme that achieves $R_{\text{ID}}^{\text{LC}-\triangle}(D)$ is straightforward, as shown in [4], how to implement a $\text{TC}-\triangle$ scheme is not clear in general. In this paper, we propose a valid $\text{TC}-\triangle$ scheme, which we introduce next.

B. Description

Based on the previous results, for each sequence $\mathbf{x}$ in the database, we want to generate a signature assignment from which we can reconstruct a sequence $\hat{\mathbf{x}}$ such that the empirical distribution between $\mathbf{x}$ and $\hat{\mathbf{x}}$ is equal to the one associated with the solution to the optimization problem (7). This will imply that the scheme attains $R_{\text{ID}}^{\text{TC}-\triangle}(D)$; which is better than $R_{\text{ID}}^{\text{LC}-\triangle}(D)$, attained by the scheme proposed in [4], and even optimal for memoryless binary sources and Hamming distortion.

We propose a practical scheme for this task based on lossy compression algorithms. Specifically, we show that the desired distribution can be achieved by carefully choosing the distortion to be applied by the lossy compressor. In other words, if

$$P_{\hat{X}|X}^* = \underset{P_{\hat{X}|X}: I(X;\hat{X}) \leq R}{\arg\max} \; \mathbb{E}[\rho(\hat{X}, Y)] - \mathbb{E}[\rho(X, \hat{X})], \tag{8}$$

we are seeking a distortion measure $\rho^*(X, \hat{X})$ such that

$$P_{\hat{X}|X}^* = \underset{P_{\hat{X}|X}: I(X;\hat{X}) \leq R}{\arg\min} \; \mathbb{E}[\rho^*(X, \hat{X})], \tag{9}$$

i.e., the conditional probability induced by the lossy compressor is equal to $P_{\hat{X}|X}^*$.

We show that (9) holds if $\rho^*(X, \hat{X}) = \log \frac{1}{P_{X|\hat{X}}^*}$, where $P_{X|\hat{X}}^*$ is induced from $P_{\hat{X}|X}^*$ and P_X, and $I(X; \hat{X}) = R$. Note that $\rho^*(X, \hat{X})$ is reminiscent of logarithmic loss [10]. This is based on the following lemma [11]:

Lemma 1: Let $X \sim P_X$, and let $P_X(x) > 0$ for all $x \in \mathcal{X}$. For a channel $P_{\hat{X}|X}$, let $P_{X|\hat{X}}$ be the reversed channel, and consider a rate distortion problem with distortion measure

$$\rho(x, u) = \log \frac{1}{P_{X|\hat{X}}(x|u)}. \tag{10}$$

Then, for the rate constraint $I(X;U) \leq I(X;\hat{X})$, the optimal test channel $P^*_{U|X}$ is equal to $P_{\hat{X}|X}$.

Proof: First, note that

$$\mathbb{E}[\rho(X,U)] = \sum_{x,u} P_{X,U}(x,u) \log \frac{1}{P_{X|\hat{X}}(x|u)} \tag{11}$$

$$= \sum_{u} P_U(u) D(P_{X|U}(\cdot|u)||P_{X|\hat{X}}(\cdot|u)) + H(X|U), \tag{12}$$

and that the rate constraint implies $H(X|U) \geq H(X|\hat{X})$. Therefore,

$$\mathbb{E}[\rho(X,U)] \geq \sum_{u} P_U(u) D(P_{X|U}(\cdot|u)||P_{X|\hat{X}}(\cdot|u)) + H(X|\hat{X}). \tag{13}$$

Thus,

$$\min_{P_{U|X}: I(X;U) \leq I(X;\hat{X})} \mathbb{E}[\rho(X,U)] = H(X|\hat{X}), \tag{14}$$

and the minimum is attained if and only if $P_{X|U} = P_{X|\hat{X}}$. ■

Going back to our setting, note that the optimization problem (7) that solves for $R_{\mathrm{ID}}(D)^{\mathrm{TC}-\triangle}$, has the constraint $I(X,\hat{X}) \leq R$. The maximizing probability (8) will in general achieve $I(X,\hat{X}) = R$, and thus we can apply the lemma.

Therefore, the proposed $\mathrm{TC}-\triangle$ scheme effectively employs for the signature assignment a good lossy compressor for distortion measure $\rho(x,\hat{x}) = \log \frac{1}{P^*_{X|\hat{X}}(x|\hat{x})}$, where $P^*_{X|\hat{X}}$ is induced by $P^*_{\hat{X}|X}$, given by (8), and P_X. With an optimal lossy compressor, and assuming $I(X,\hat{X}) = R$, the joint type of the sequences $\mathbf{x}$ and $\hat{\mathbf{x}}$ will be close to $P^*_{\hat{X}|X}$, which achieves $R_{\mathrm{ID}}^{\mathrm{TC}-\triangle}$, which is optimal for the case of general binary sources and Hamming distortion. In the next section we show that the performance of the proposed scheme approaches the fundamental performance limit, and performs notably better than the $\mathrm{LC}-\triangle$ scheme.

V. Simulation results

In this section we examine the performance of both the $\mathrm{LC}-\triangle$ and the $\mathrm{TC}-\triangle$ schemes. We consider datasets composed of M binary sequences of length n, and Hamming distortion for computing the similarity between sequences. We generate the sequences in the database as $\mathbf{X} \sim \prod_{i=1}^n P_X(x_i)$, with $P_X = \mathrm{Bern}(p)$. These sequences are independent of the query sequences, generated as $\mathbf{Y} \sim \prod_{i=1}^n P_Y(y_i)$, with $P_Y = \mathrm{Bern}(q)$. With these assumptions, for each sequence $\mathbf{x}^{(i)}$ in the database, $i \in [1:M]$, given its signature $T(\mathbf{x}^{(i)})$, we can compute the probability that $g(T(\mathbf{x}^{(i)}),\mathbf{y}) = \texttt{maybe}$ (for a similarity threshold D), denoted by $P(\texttt{maybe}|T(\mathbf{x}^{(i)}))$, analytically, with the following formula:

$$P(\texttt{maybe}|T(\mathbf{x}^{(i)})) = \sum_{d=\lceil n(d_0-D)\rceil}^{\lfloor n(d_1+D)\rfloor} \sum_{i=0}^{d} \binom{n_0}{i}\binom{n-n_0}{d-i} q^{n-n_0-d+2i}(1-q)^{n_0+d-2i}, \tag{15}$$

where n_0 denotes the number of zeros of $\hat{\mathbf{x}}^{(i)}$, and d_0 and d_1 are the delimiters of the decision region to which $d(\mathbf{x}^{(i)},\hat{\mathbf{x}}^{(i)})$ belongs. If no quantization is applied, $d_0 = d_1 =$

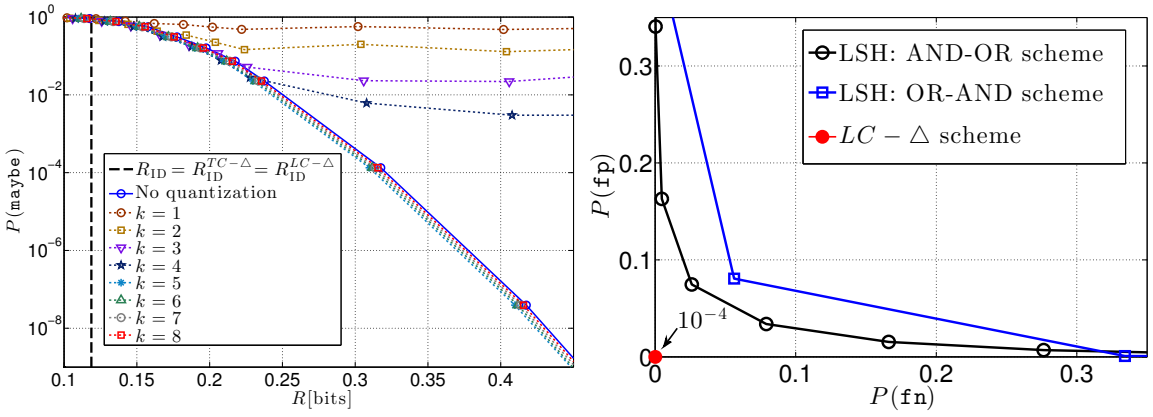
$d(\mathbf{x}(i), \hat{\mathbf{x}}^{(i)})$. Finally, we compute the probability of `maybe` for the database as the average over all the sequences it contains, i.e., $P(\texttt{maybe}) = \frac{1}{M}\sum_{i=1}^{M} P(\texttt{maybe}|T(\mathbf{x}^{(i)}))$. Note that we want this probability to be as small as possible.

Regarding the quantization of $d(\mathbf{x}, \hat{\mathbf{x}})$, we approximate the distribution of $d(\mathbf{X}, \hat{\mathbf{X}})$ as a Gaussian $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are computed empirically (for each rate). We then use the k-means algorithm to find the 2^k decision regions ($\Delta R = k/n$, i.e., k bits are allocated for the description of the quantized distortion). Thus, for each distortion, we store only the decision region to which it belongs.

A. Binary symmetric sources and Hamming distortion

The performance of the $\mathrm{LC}-\triangle$ scheme in this setting was already discussed in [4]. Therefore, since the $\mathrm{LC}-\triangle$ and the $\mathrm{TC}-\triangle$ schemes are equivalent in this case, the reader can refer to [4] for more extensive simulation results. We consider a dataset composed of $M = 1000$ binary sequences of length $n = 512$, with $p = q = 0.5$. As the fixed-length lossy compression algorithm, we use a binary-Hamming version of the successive refinement compression scheme [12].

Regarding the quantization of $d(\mathbf{x}, \hat{\mathbf{x}})$, there exists a tradeoff between the quantization level and the probability of `maybe`. Fig. 3(a) shows the results for different quantization levels (denoted by k) and a similarity threshold $D = 0.20$ (i.e., 80% similarity). As expected, no special value of k performs better than the others for any overall compression rate R. Therefore, in the subsequent figures the presented results correspond to the best value of k for each rate. As it can be observed, we can reduce the size of the database by 76% ($R = 0.24$) and retrieve on average 1% of the sequences per query. With 70% reduction we can get a $P(\texttt{maybe})$ of 10^{-4} (on average one sequence every 1000 is retrieved). One can get even more compression with the same $P(\texttt{maybe})$ for lower values of D, as shown in [4]. For example, 95% compression with a 1% average retrieval is achieved for $D = 0.05$.



(a) (b)

Fig. 3: Binary symmetric sequences and similarity threshold $D = 0.2$: (a) performance of the proposed scheme with quantized distortion (b) comparisson with LSH for rate $R = 0.3$.

Finally, we include a comparison with LSH [6]. We use the accepted family of functions $\mathcal{H} = \{h_i, i \in [1 : n]\}$, with $h_i(\mathbf{x}) = x(i)$, the i^{th} coordinate of $\mathbf{x}$, and consider both the AND-OR and the OR-AND constructions described in [6, Chapter 3]. Note that the

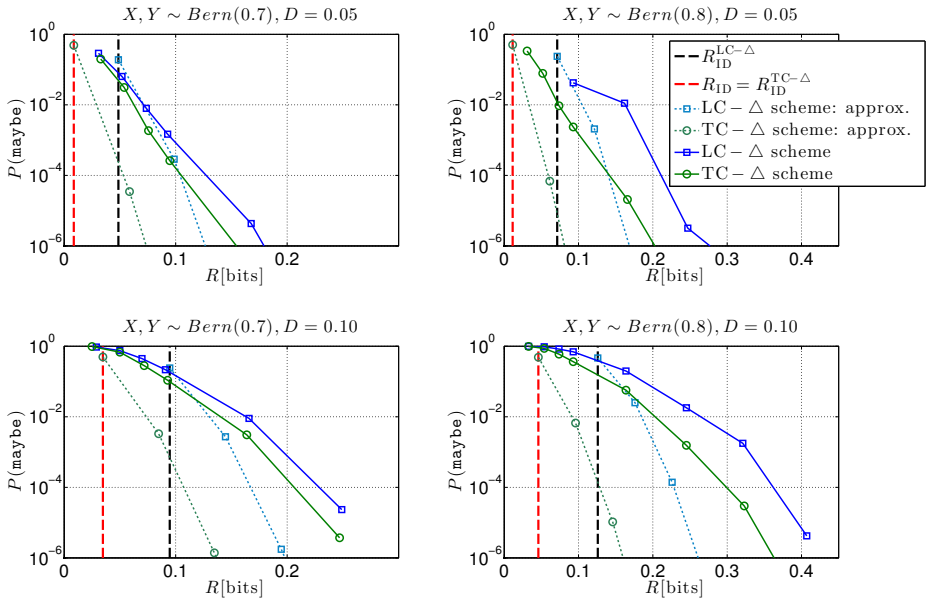


Fig. 4: Performance of the proposed schemes for sequences of length $n = 512$, similarity thresholds $D = \{0.05, 0.1\}$ and $P_X = P_Y = \text{Bern}(0.7)$ and $\text{Bern}(0.8)$.

comparison is not completely fair, as LSH allows false negatives (`fn`'s), compresses the query sequence, and its design is not optimized for the problem considered in this paper. This is reflected in Fig. 3(b), where we show the achievable probabilities of `fn`'s and false positives (`fp`'s) for both schemes, considering the database introduced above and rate $R = 0.3$. As can be observed, it is not possible to have both probabilities going to zero at the same time, whereas the proposed scheme achieves for the same rate a $P(\texttt{fp})$ close to 10^{-4} with zero `fn`'s.

B. General binary sources and Hamming distortion

We compare the performance of the $TC-\triangle$ and $LC-\triangle$ schemes, assuming $P_X = P_Y = \text{Bern}(p)$, with $p \neq 0.5$. For a fair comparison, we simulate both schemes with the lossy compressor presented in [13], that allows to specify the distortion to be used. The $LC-\triangle$ scheme uses Hamming distortion, whereas the $TC-\triangle$ scheme uses the distortion measure given by (10), with $P_{X|\hat{X}}$ computed from P_X and $P_{\hat{X}|X}$ as defined in (8) (for each rate). Note that this distortion is to be used only by the lossy compressor. The decision rule $g(T(\mathbf{x}), \mathbf{y})$ in both schemes still uses Hamming distortion to measure similarity between sequences and the triangle inequality property for computing the decision threshold.

We show simulation results in Fig. 4 for a dataset composed of $M = 1000$ sequences of length $n = 512$, and $P_X = P_Y = \text{Bern}(0.7)$ and $\text{Bern}(0.8)$. We also plot the three rates ($R_{\text{ID}} = R_{\text{ID}}^{\text{TC}-\triangle} < R_{\text{ID}}^{\text{LC}-\triangle}$) and an approximation for each scheme, computed as follows. For a given rate R, the approximation for the $\text{LC}-\triangle$ scheme assumes $P_{\hat{X}|X}$ is given as $\arg\min_{P_{\hat{X}|X}: I(X;\hat{X}) \leq R} \mathbb{E}[\rho(X, \hat{X})]$ (rate distortion optimization problem), with ρ representing Hamming distortion. On the other hand, for the $\text{TC}-\triangle$ scheme, $P_{\hat{X}|X}$ is assumed to be equal to (8). We then compute the $P(\texttt{maybe})$ of each scheme using equation (15), with $d_0 = d_1 = \mathbb{E}[\rho(X, \hat{X})]$, with ρ representing Hamming distortion, and $n_0 = nP_{\hat{X}}(\hat{x} = 0)$.

As can be observed, the $\text{TC}-\triangle$ scheme performs better than the $\text{LC}-\triangle$ scheme in all cases, as is suggested by the theory. For example, for $X, Y \sim \text{Bern}(0.7)$, $D = 0.05$

and $R = 0.13$ (87% compression), while the $\mathrm{LC}-\triangle$ scheme achieves $P(\texttt{maybe}) = 10^{-4}$, the $\mathrm{TC}-\triangle$ scheme achieves 10^{-5}. Similarly, for $D = 0.1$ and $R = 0.2$, the $P(\texttt{maybe})$ decreases from 10^{-3} to 10^{-4}, i.e., on average it retrieves 1 sequence every 10000, instead of every 1000. For the case $X, Y \sim \mathrm{Bern}(0.8)$ we observe similar results. With $D = 0.05$ (95% similarity) and $P(\texttt{maybe}) = 10^{-2}$, the $\mathrm{TC}-\triangle$ scheme attains 93% compression ($R = 0.07$), whereas the $\mathrm{LC}-\triangle$ schemes achieves only 84% compression ($R = 0.16$), i.e., a reduction in rate of 55%. Furthermore, $R = 0.07$ is close to $R_{\mathrm{ID}}^{\mathrm{LC}-\triangle}$. Similarly, for $D = 0.1$ and $P(\texttt{maybe}) = 10^{-4}$ the decrease in rate is from 0.35 to 0.3 bits, which represents an improvement in compression of 14.2%. Finally, notice that for a given rate, the smaller the similarity threshold D, the smaller the $P(\texttt{maybe})$.

VI. Concluding remarks

We investigated schemes for compressing a database so that similarity queries can be performed efficiently on the compressed database. The fundamental limits for this problem have been characterized in past work, and they serve as the basis for performance evaluation.

Recently, we proposed a scheme for this task based on lossy compression algorithms which was easy to analyze and implement. While its performance was shown to be close to the fundamental limits in some cases (e.g., binary symmetric sources and Hamming distortion), the scheme is suboptimal in general. In this paper we proposed a new scheme that builds upon the previous one and achieves a better compression rate in many cases. For example, for general memoryless binary sequences and Hamming distortion, our suggested scheme exhibits on simulated data performance approaching the fundamental limits, substantially improving over the previous scheme. The proposed scheme is also based on lossy compression algorithms, but in this case we judiciously design the distortion measure to be applied by the lossy compressor, a measure which is not Hamming despite the fact that similarity for the query is measured under Hamming. Finally, as was the case with the previously proposed scheme, the one proposed here is applicable to any discrete database and similarity measure satisfying the triangle inequality.

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] A. Ingber and T. Weissman, "The minimal compression rate for similarity identification," *In Preparation, to be submitted to the IEEE Trans. on Information Theory*, 2013.

[3] A. Ingber, T. Courtade, and T. Weissman, "Compression for quadratic similarity queries," *Submitted to the IEEE Trans. on Information Theory (http://arxiv.org/pdf/1307.6609.pdf)*, 2013.

[4] I. Ochoa, A. Ingber, and T. Weissman, "Efficient similarity queries via lossy compression," in *IEEE 51st Annual Conference on Communication, Control and Computing*, 2013.

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

[6] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2012.

[7] S. Arya and et. al, "An optimal algorithm for approximate nearest neighbor searching," in *Proceedings of the fifth annual ACM-SIAM symposium on Discrete algorithms*, 1994, pp. 573–582.

[8] S. Ramaswamy and K. Rose, "Adaptive cluster distance bounding for high-dimensional indexing," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 6, pp. 815–830, 2011.

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

[10] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 761–765.

[11] T. Courtade, "Private communication," 2013.

[12] R. Venkataramanan, T. Sarkar, and S. Tatikonda, "Lossy compression via sparse linear regression: Computationally efficient encoding and decoding," *CoRR*, vol. abs/1212.1707, 2012.

[13] A. Gupta and S. Verdú, "Nonlinear sparse-graph codes for lossy compression," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 1961–1975, 2009.