Compression for Exact Match Identification

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

***Abstract*—In this paper, we consider the problem of determining whether sequences $\mathbf{X}$ and $\mathbf{Y}$, generated i.i.d. according to $P_X \times P_Y$, are equal given access only to the pair $(\mathbf{Y}, T(\mathbf{X}))$, where $T(\mathbf{X})$ is a rate-R compressed version of $\mathbf{X}$. In general, the rate R may not be sufficiently large to reliably determine whether $\mathbf{X} = \mathbf{Y}$. We precisely characterize this reliability – i.e., the exponential rate at which an error is made – as a function of R. Interestingly, the exponent turns out to be related to the Bhattacharyya distance between the distributions P_X and P_Y. In addition, the scheme achieving this exponent is universal, i.e. does not depend on P_X, P_Y.**

I. Introduction

We study the problem of sequence identification via compressed data. To this end, consider random n-vectors $\mathbf{X}$ and $\mathbf{Y}$ drawn i.i.d. according to the discrete product distribution $P_X \times P_Y$. Given a *signature* $T(\mathbf{X})$ consisting of nR bits, and a sequence $\mathbf{Y}$, we would like to determine whether $\mathbf{X} = \mathbf{Y}$. This setting is illustrated in Figure 1. If R is very small, it may be impossible to reliably determine whether $\mathbf{X} = \mathbf{Y}$ from the pair $(T(\mathbf{X}), \mathbf{Y})$. Hence, there is a tradeoff between R and the reliability at which we can identify $\mathbf{X}$ (given $T(\mathbf{X})$ and $\mathbf{Y}$).

This setup is motivated by the problem of identifying entries of interest in a database. By querying a compressed version of the data, query latency can be reduced and the communication requirements between a client making a query and a server hosting the database can be reduced. In many practical settings, false negatives (i.e., declaring that $\mathbf{X} \neq \mathbf{Y}$ when they are equal in truth) are not permitted. For example, failing to identify a suspect's fingerprint in a forensics database can be extremely costly. Therefore, we are interested in the problem of sequence identification via compressed data when false negatives are not permitted.

The primary contribution of this paper is a complete characterization of the identification reliability (in terms of an appropriately defined error exponent) as a function of compression rate R and the distributions P_X, P_Y. Roughly speaking, our results describe the fundamental tradeoff between the length of the signature (that can be thought of as a hash key) and the exponential decay in uncertainty about the sequence which produced the signature. Despite the significant body of work related to hash functions (and mostly the related Bloom filters, see [1] and references therein), this basic tradeoff appears to have not been studied previously.

This work is supported in part by the Rothschild Postdoctoral Fellowship, by the NSF Center for Science of Information under grant agreement CCF-0939370, and by a Google Research Award.

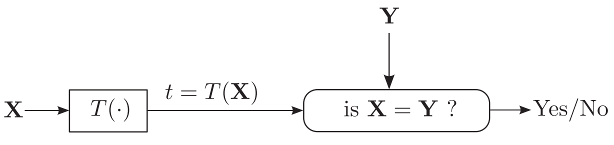

Fig. 1. Identifying whether $\mathbf{X} = \mathbf{Y}$ from the compressed version $T(\mathbf{X})$

Perhaps most closely related to this paper is the work by Ahlswede et al. [2], where the authors consider the problem of determining whether $d(\mathbf{x}, \mathbf{y}) \leq D$ from $T(\mathbf{x})$ and $\mathbf{y}$, where $d(\cdot,\cdot)$ is a distortion measure and D is given. Unlike the present paper, [2] considers variable length coding, and also permits nonzero false-negative probability. In contrast, we consider the case of exact match (i.e., determining whether $\mathbf{x} = \mathbf{y}$ from $T(\mathbf{x})$ and $\mathbf{y}$), and consider fixed-length compression. While our work seems to be related to a special case (of Hamming distortion and $D = 0$) of the work in [2], it should be noted that the main result in [2] is parameterized by an auxiliary random variable with unbounded alphabet size, and is therefore of limited practical value[1].

This paper is structured as follows. In Section II, we define the relevant notation used throughout the paper. Section III formally defines our problem setting and delivers the main results. All proofs are given in Section IV.

II. Notation

Throughout this paper we use boldface notation (e.g. $\mathbf{x}$) to denote a vector of elements $\mathbf{x} = [x_1, ..., x_n]$. Capital letters denote random variables (e.g. X) or vectors (e.g. $\mathbf{X}$). We use calligraphic fonts (e.g. $\mathcal{X}$) to represent a discrete set, and $\mathcal{P}(\mathcal{X})$ to denote all the probability distributions on the alphabet $\mathcal{X}$. We assume w.l.o.g. that the alphabet consists of integer numbers, i.e. $\mathcal{X} = \{1, 2, ..., |\mathcal{X}|\}$. Throughout, we will use the notation $[1:k]$ to denote the set of integers $\{1, 2, \dots, k\}$. Logarithms are taken to base 2, and rates are given in bits.

For two distributions P and Q on $\mathcal{X}$, $D(P||Q)$ denotes the usual Kullback-Leibler divergence (see, e.g. Cover and Thomas [3]). We also define the (base-2) Bhattacharya distance [4] between P and Q as

$$d_B(P, Q) = -\log \left(\sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)} \right). \tag{1}$$

[1] In [2, Remark 3], the authors comment about a method that would potentially lead to an approximation of the solution using an auxiliary RV with *bounded* cardinality. However, this method is computationally intractable, even for non-symmetric binary sources.

For a random variable X, we denote by $H_2(X)$ the Rényi entropy of order 2, i.e.

$$H_2(X) \triangleq -\log \sum_{x\in\mathcal{X}} P_X^2(x). \tag{2}$$

$H_2(X)$ is also known as the *collision entropy*, since whenever $X, \tilde{X}$ are independent with the same distribution P_X, we have

$$\Pr\{X = \tilde{X}\} = 2^{-H_2(X)}. \tag{3}$$

For two independent random variables X and Y, both taking values in the alphabet $\mathcal{X}$ and distributed according to P_X and P_Y respectively, we extend (2) to

$$H_2(X,Y) \triangleq -\log \sum_{x\in\mathcal{X}} P_X(x)P_Y(x). \tag{4}$$

The operational significance of $H_2(X,Y)$ is given by the following property, whose proof is trivial: for $\mathbf{X},\mathbf{Y} \in \mathcal{X}^n$, drawn i.i.d. according $P_X \times P_Y$,

$$\Pr\{\mathbf{X} = \mathbf{Y}\} = 2^{-nH_2(X,Y)}. \tag{5}$$

III. Main Results

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

$$g(T(\mathbf{x}), \mathbf{x}) = \texttt{maybe}. \tag{6}$$

In words, a scheme (T, g) is admissible only if it does not produce false negatives (i.e. makes an error where the query $g(T(\mathbf{x}), \mathbf{x})$ returns no). As discussed in Section I, false negatives are not acceptable for many systems, thus motivating the definition of admissible schemes.

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 $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}|\mathbf{X} \neq \mathbf{Y}\}$ because $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}|\mathbf{X} = \mathbf{Y}\} = 1$. In other words, we desire to minimize the probability of a false positive event.

Let $\mathbf{X},\mathbf{Y} \in \mathcal{X}^n$ be drawn i.i.d. according to $P_X \times P_Y$, i.e.

$$\Pr\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} = \prod_{i=1}^{n} P_X(x_i)P_X(y_i). \tag{7}$$

For a given compression rate R, we shall be interested in the speed at which the probability of the event $\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$ goes to zero. In general it will vanish exponentially fast, so we therefore define:

Definition 1. *Let the identification exponent for rate $R \geq 0$ be given by*

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

where $g^{(n)}, T^{(n)}$ denote the optimal rate-R identification schemes of length n.

Theorem 1. *For given distributions P_X and P_Y,*

$$\mathbf{E}_{\mathsf{ID}}(R) = \min\{R + 2d_B(P_X, P_Y), H_2(X,Y)\}. \tag{8}$$

Moreover, this exponent can be attained universally, i.e. without knowing the distributions P_X, P_Y.

As an immediate Corollary, we have a simple characterization of $\mathbf{E}_{\mathsf{ID}}(R)$ when $P_X = P_Y$.

Corollary 1. *For $P_X = P_Y$,*

$$\mathbf{E}_{\mathsf{ID}}(R) = \min\{R, H_2(X)\}. \tag{9}$$

While Theorem 1 gives a complete characterization of the identification exponent $\mathbf{E}_{\mathsf{ID}}(R)$, it reveals the following counterintuitive fact: the exponent $\mathbf{E}_{\mathsf{ID}}(R)$ is bounded by $H_2(X,Y)$ even as $R \to \infty$. For R sufficiently large, we can set $T(\mathbf{x}) = \mathbf{x}$, hence there should be no ambiguity in the identification. Ideally, this should be reflected by the exponent $\mathbf{E}_{\mathsf{ID}}(R)$. This is explained by the type of possible answers by the scheme – either no or maybe. This type of positive certainty that can be obtained (only at high rates) cannot be reflected by this set of answers.

We rectify this by defining an *augmented identification system* $(T, g^\star)$. To this end, an augmented rate-R identification system $(T, g^\star)$ consists of a signature assignment $T : \mathcal{X}^n \to [1 : 2^{nR}]$, and a query function $g^\star : [1 : 2^{nR}] \times \mathcal{X}^n \to \{\texttt{yes}, \texttt{no}, \texttt{maybe}\}$. An augmented system is said to be admissible, if for any $\mathbf{x} \in \mathcal{X}^n$ we have

$$\{g^\star(T(\mathbf{x}), \mathbf{y}) = \texttt{no}\} \Rightarrow \mathbf{x} \neq \mathbf{y} \tag{10}$$

$$\{g^\star(T(\mathbf{x}), \mathbf{y}) = \texttt{yes}\} \Rightarrow \mathbf{x} = \mathbf{y}. \tag{11}$$

In words, an admissible augmented scheme does not produce false negatives when $g^\star(T(\mathbf{x}), \mathbf{y})$ returns no, and also does not produce false positives when $g^\star(T(\mathbf{x}), \mathbf{y})$ returns yes.

Analogous to Definition 1, we define the *augmented identification exponent* $\mathbf{E}^\star_{\mathsf{ID}}(R)$ as follows:

Definition 2. *Fix a rate $R \geq 0$. Define the augmented identification exponent as*

$$\mathbf{E}^\star_{\mathsf{ID}}(R) \triangleq \limsup_{n\to\infty} -\frac{1}{n}\log\Pr\left\{g^{\star(n)}\left(T^{(n)}(\mathbf{X}), \mathbf{Y}\right) = \texttt{maybe}\right\},$$

where $g^{\star(n)}, T^{(n)}$ denote the optimal rate-R augmented identification schemes of length n.

A complete characterization of the augmented identification exponent is given by the following theorem.

Theorem 2.

$$\mathbf{E}^\star_{\mathsf{ID}}(R) = R + \min_{Q:H(Q)\geq R} D(Q||P_X) + D(Q||P_Y), \tag{12}$$

where we define a minimization over the empty set to equal infinity. This exponent can be attained universally, i.e. without knowing the distributions P_X, P_Y.

Note that calculating $\mathbf{E}^\star_{\mathsf{ID}}(R)$ is an easy task since it is a convex minimization program with convex constraints. Similar

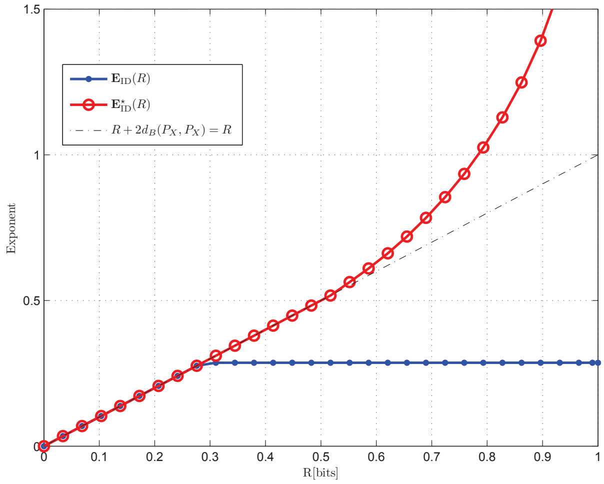

Fig. 2. Identification exponents for $X, Y \sim \text{Ber}(.1)$.

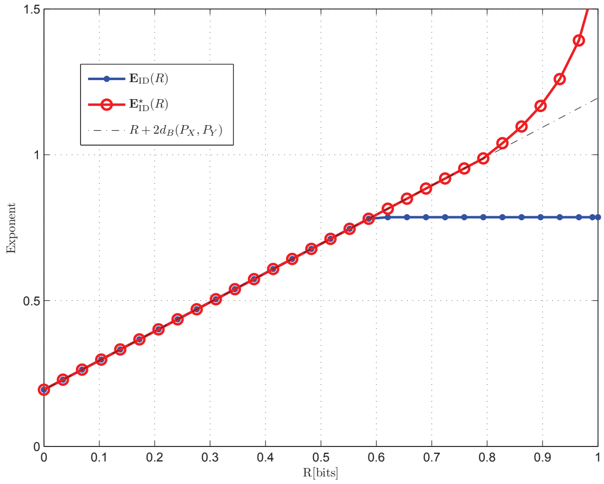

Fig. 3. Identification exponents for $X \sim \text{Ber}(.1)$ and $Y \sim \text{Ber}(.4)$.

to Corollary 1, we can specialize Theorem 2 to the setting where $P_X = P_Y$:

Corollary 2. *For $P_X = P_Y$,*

$$\mathbf{E}^{\star}_{\text{ID}}(R) = R + 2\mathbf{E}_S(P_X, R), \tag{13}$$

where $\mathbf{E}_S(P_X, R)$ is the error exponent in source coding (cf. [5, Theorem 2.15]), i.e.

$$\mathbf{E}_S(P, R) \triangleq \min_{Q:H(Q)\geq R} D(Q||P). \tag{14}$$

As an example, the identification exponents for the Bernoulli case are plotted in Figures 2 and 3. In Fig. 2, X and Y have the same distribution, and the identification exponents are zero at $R = 0$ (cf. Corollary 1). The case of $P_X \neq P_Y$ can be seen in Fig. 3. In both figures the augmented exponent $\mathbf{E}^{\star}(R) \to \infty$ when $R \to \log|\mathcal{X}| = 1$bit, as expected.

IV. Proofs

A. Preliminaries

In the proofs we rely on the method of types [5]. For $\mathbf{x} \in \mathcal{X}^n$ and $a \in \mathcal{X}$, let $N(a|\mathbf{x})$ denote the number of occurrences of a in $\mathbf{x}$. The *type* of the sequence $\mathbf{x}$, denoted $\mathbb{P}_{\mathbf{x}}$, is defined as the vector $\frac{1}{n}[N(1|\mathbf{x}), N(2|\mathbf{x}), ..., N(|\mathcal{X}||\mathbf{x})]$. For any sequence length n, let $\mathcal{P}_n(\mathcal{X})$ denote the set of possible n-types, i.e.

$$\mathcal{P}_n(\mathcal{X}) \triangleq \{P \in \mathcal{P}(\mathcal{X}) | \forall x \in \mathcal{X},\ nP(x) \in \mathbb{Z}_+\}. \tag{15}$$

For a type $P \in \mathcal{P}_n(\mathcal{X})$, the *type class* $\mathbb{T}_P$ is defined as the set of all sequences $\mathbf{x} \in \mathcal{X}^n$ with type P, i.e.

$$\mathbb{T}_P \triangleq \{\mathbf{x} \in \mathcal{X}^n : \mathbb{P}_{\mathbf{x}} = P\}. \tag{16}$$

Definition 3. *Let P and Q be two distributions over the alphabet $\mathcal{X}$. The normalized Schur product, denoted $P \diamond Q$, is given by the following distribution*

$$P \diamond Q\ (x) \triangleq \frac{P(x)Q(x)}{\sum_{x' \in \mathcal{X}} P(x')Q(x')}. \tag{17}$$

Proposition 1. *For any two i.i.d. random variables $X, \tilde{X}$ taking values in $\mathcal{X}$, we have*

$$\Pr\{X = \tilde{X}\} \geq \frac{1}{|\mathcal{X}|}. \tag{18}$$

Proof: Jensen's inequality implies $H_2(X) \leq H(X) \leq \log|\mathcal{X}|$. Therefore, the claim follows from (3). ■

Proposition 2. *For $\mathbf{X}, \mathbf{Y} \in \mathcal{X}^n$ drawn i.i.d. according to $P_X \times P_Y$,*

$$\Pr\{\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} \leq (n+1)^{|\mathcal{X}|} 2^{-n2d_B(P_X, P_Y)}. \tag{19}$$

Moreover, this is sharp in the sense that for any $\epsilon > 0$,

$$\Pr\{\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} \geq 2^{-n(2d_B(P_X, P_Y)+\epsilon)}. \tag{20}$$

for n sufficiently large.

It is interesting to note that Proposition 2 lends an operational significance to Bhattacharyya distance. That is, if $\mathbf{X}, \mathbf{Y}$ are drawn i.i.d. according to $P_X \times P_Y$, then they have the same type with probability roughly $2^{-n2d_B(P_X, P_Y)}$. This interpretation was unknown to the authors.

Proof of Proposition 2: Recalling [5, Lemma 2.6], we have $\Pr\{\mathbf{X} \in Q\} \leq 2^{-nD(Q||P_X)}$ for $Q \in \mathcal{P}_n(\mathcal{X})$. Thus, it follows that

$$\begin{aligned}
\Pr\{\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} &= \sum_{Q \in \mathcal{P}_n(\mathcal{X})} \Pr\{\mathbf{X} \in Q\} \Pr\{\mathbf{Y} \in Q\} && (21)\\
&= \sum_{Q \in \mathcal{P}_n(\mathcal{X})} 2^{-n(D(Q||P_X) + D(Q||P_Y))} && (22)\\
&\leq (n+1)^{|\mathcal{X}|} 2^{-n(\min_{Q\in\mathcal{P}(\mathcal{X})} D(Q||P_X) + D(Q||P_Y))}, && (23)
\end{aligned}$$

where the final inequality holds since $|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}$. Straightforward calculus reveals that

$$\min_{Q \in \mathcal{P}(\mathcal{X})} D(Q||P_X) + D(Q||P_Y) = 2d_B(P_X, P_Y). \tag{24}$$

The converse statement follows easily since [5, Lemma 2.6] also states that $\Pr\{\mathbf{X} \in Q\} \geq (n+1)^{-|\mathcal{X}|} 2^{-nD(Q||P_X)}$ for $Q \in \mathcal{P}_n(\mathcal{X})$. ■

B. Proof of Theorem 1

We now prove Theorem 1. As usual, the proof is divided into two parts; the direct part and the converse part.

Proof of Theorem 1: **Direct part:** Fix $\epsilon > 0$. Define the signature of a sequence $\mathbf{x}$ as $T(\mathbf{x}) = [\mathbb{P}_{\mathbf{x}}, T_0(\mathbf{x})]$, where $T_0(\mathbf{x})$ is drawn randomly and uniformly from the set $[1 : 2^{nR_0}]$, where $R_0 < R$. Note that although formally we defined $T(\cdot)$ to return an integer number, we allow ourselves to write $T(\mathbf{x})$ as above, where the interpretation is that of an equivalent representation by integers, which exists as long as $|\mathcal{P}_n(\mathcal{X})|2^{nR_0} < 2^{nR}$. This, in turn, is guaranteed, for large enough n, since the number of types is polynomial in n. The function $g(\cdot,\cdot)$ returns `maybe` if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Note that the scheme does not depend on P_X, P_Y and is therefore universal.

We now analyze the probability of the event $\{g(T(\mathbf{X}), \mathbf{Y})) = \texttt{maybe}\}$, averaged over the random selection of the signature function. As usual, one scheme exists with performance as good as the ensemble average.

$$\begin{aligned}
&\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\} && (25)\\
&= \Pr\{T(\mathbf{X}) = T(\mathbf{Y})\} && (26)\\
&\le \Pr\{T(\mathbf{X}) = T(\mathbf{Y}) | \mathbf{X} \neq \mathbf{Y}\} + \Pr\{\mathbf{X} = \mathbf{Y}\}. && (27)
\end{aligned}$$

We continue with

$$\begin{aligned}
&\Pr\{T(\mathbf{X}) = T(\mathbf{Y}) | \mathbf{X} \neq \mathbf{Y}\} && (28)\\
&\quad = \Pr\{T(\mathbf{X}) = T(\mathbf{Y}) | \mathbf{X} \neq \mathbf{Y}, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\}\\
&\qquad \times \Pr\{\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}} | \mathbf{X} \neq \mathbf{Y}\} && (29)\\
&\quad \overset{(a)}{=} 2^{-nR_0} \Pr\{\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}} | \mathbf{X} \neq \mathbf{Y}\} && (30)\\
&\quad \overset{(b)}{\le} 2^{-nR_0} \Pr\{\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} / \Pr\{\mathbf{X} \neq \mathbf{Y}\} && (31)\\
&\quad \overset{(c)}{\le} 2^{-n(R_0 + 2d_B(P_X, P_Y) - \epsilon)}, && (32)
\end{aligned}$$

where (a) follows since any two sequences sharing the same type will have the same signature with probability 2^{-nR_0}, (b) follows from the Bayes rule and (c) follows (5) and from (19) in Proposition 2 and holds for n sufficiently large. Combining (27), (32), and (5), we have

$$\begin{aligned}
&\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\} && (33)\\
&\le 2^{-n(R_0 + 2d_B(P_X, P_Y) - \epsilon)} + 2^{-nH_2(X,Y)} && (34)\\
&\le 2 \cdot 2^{-n \times \min\{R_0 + 2d_B(P_X, P_Y) - \epsilon, H_2(X,Y)\}}, && (35)
\end{aligned}$$

which completes the proof of the direct part, as ϵ is arbitrarily small and R_0 is arbitrarily close to R.

Converse part: Let $T : \mathcal{X}^n \to [1 : 2^{nR}]$ be given. We aim to lower bound the probability for `maybe`:

$$\begin{aligned}
&\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\} && (36)\\
&= \Pr\{T(\mathbf{X}) = T(\mathbf{Y})\} && (37)\\
&= \Pr\{T(\mathbf{X}) = T(\mathbf{Y}) | \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} \cdot \Pr\{\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\}. && (38)
\end{aligned}$$

Since both $\mathbf{X}$ and $\mathbf{Y}$ are i.i.d., when $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}$ we get that both $\mathbf{X}$ and $\mathbf{Y}$ are distributed uniformly on the same type class. Therefore $T(\mathbf{X})$ and $T(\mathbf{Y})$ are two (independent) random variables with the same distribution, each taking values in $[1 : 2^{nR}]$. It follows from Proposition 1 that

$$\Pr\{T(\mathbf{X}) = T(\mathbf{Y}) | \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} \ge 2^{-nR}. \qquad (39)$$

It therefore follows from the second statement of Proposition 2 that

$$\mathbf{E}_{\mathsf{ID}}(R) \le 2d_B(P_X, P_Y) + R. \qquad (40)$$

In addition, the probability for `maybe` is lower bounded by the quantity

$$\Pr\{\mathbf{X} = \mathbf{Y}\} = 2^{-nH_2(X,Y)}. \qquad (41)$$

Hence the converse is complete. ■

C. Proof of Theorem 2

To prove Theorem 2, we require the following lemma.

Lemma 1.

$$\begin{aligned}
&\mathbf{E}_S(P_X \diamond P_Y, R) + H_2(X,Y)\\
&\qquad \ge R + \min_{Q : H(Q) \ge R} D(Q || P_X) + D(Q || P_Y). && (42)
\end{aligned}$$

Proof: For an arbitrary distribution Q, we have

$$\begin{aligned}
&D\left(Q || P_X \diamond P_Y\right) && (43)\\
&\quad = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log P_X \diamond P_Y(x) && (44)\\
&\quad = -H(Q) - \sum_x Q(x) \log P_X(x)\\
&\qquad - \sum_x Q(x) \log P_Y(x) - \sum_x Q(x) H_2(X,Y) && (45)\\
&\quad = H(Q) + D(Q || P_X) + D(Q || P_Y) - H_2(X,Y). && (46)
\end{aligned}$$

It follows that

$$\begin{aligned}
&H_2(X,Y) + \mathbf{E}_S(R, P_X \diamond P_Y) && (47)\\
&\quad = H_2(X,Y) + \min_{Q : H(Q) \ge R} D\left(Q || P_X \diamond P_Y\right) && (48)\\
&\quad = \min_{Q : H(Q) \ge R} H(Q) + D(Q || P_X) + D(Q || P_Y) && (49)\\
&\quad \ge R + \min_{Q : H(Q) \ge R} D(Q || P_X) + D(Q || P_Y). && (50)
\end{aligned}$$

■

Proof of Theorem 2: **Direct part:** Fix $\epsilon > 0$. Define the mapping T as:

$$T(\mathbf{x}) \triangleq \begin{cases} (\mathbb{P}_{\mathbf{x}}, \mathbf{x}), & H(\mathbb{P}_{\mathbf{x}}) < R; \\ (\mathbb{P}_{\mathbf{x}}, T_0(\mathbf{x})), & \text{otherwise}, \end{cases} \qquad (51)$$

where, similar to the proof of Theorem 1, the function $T_0(\mathbf{x})$ assigns a random signature in the set $[1 : 2^{nR_0}]$ for $R_0 < R$.

Define the query function $g^\star$ by:

$$g^\star(T(\mathbf{x}), \mathbf{y}) = \begin{cases} \texttt{yes}, & \text{if } \max\{H(\mathbb{P}_{\mathbf{x}}), H(\mathbb{P}_{\mathbf{y}})\} < R \\ & \quad \text{and } \mathbf{x} = \mathbf{y}; \\ \texttt{no}, & \text{if } T(\mathbf{x}) \neq T(\mathbf{y}); \\ \texttt{maybe}, & \text{if } T(\mathbf{y}) = T(\mathbf{x}), H(\mathbb{P}_{\mathbf{x}}) \ge R. \end{cases}$$

It can easily be verified that, by construction, the scheme defined by $(T, g^{\star})$ is admissible, and also universal. Next, we analyze the probability for `maybe`:

$$\Pr\{g^{\star}(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\} \tag{52}$$
$$= \Pr\{T(\mathbf{X}) = T(\mathbf{Y}), H(\mathbb{P}_{\mathbf{X}}) \geq R\} \tag{53}$$
$$\leq \Pr\{T(\mathbf{X}) = T(\mathbf{Y}), H(\mathbb{P}_{\mathbf{X}}) \geq R, \mathbf{X} \neq \mathbf{Y}\} \tag{54}$$
$$+ \Pr\{H(\mathbb{P}_{\mathbf{X}}) \geq R, \mathbf{X} = \mathbf{Y}\}. \tag{55}$$

The first term of (55) can be bounded as follows:

$$\begin{aligned}
&\Pr\{T(\mathbf{X}) = T(\mathbf{Y}), H(\mathbb{P}_{\mathbf{X}}) \geq R, \mathbf{X} \neq \mathbf{Y}\}\\
&= \Pr\left\{T_0(\mathbf{X}) = T_0(\mathbf{Y}), \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}, H(\mathbb{P}_{\mathbf{X}}) \geq R, \mathbf{X} \neq \mathbf{Y}\right\}\\
&= \Pr\left\{T_0(\mathbf{X}) = T_0(\mathbf{Y}) \middle| \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}, H(\mathbb{P}_{\mathbf{X}}) \geq R, \mathbf{X} \neq \mathbf{Y}\right\}\\
&\qquad \times \Pr\{H(\mathbb{P}_{\mathbf{X}}) \geq R, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}, \mathbf{X} \neq \mathbf{Y}\}\\
&\leq 2^{-nR_0} \times 2^{-n(\min_{Q:H(Q)\geq R} D(Q||P_X)+D(Q||P_Y)-\epsilon)}.
\end{aligned}$$

The last inequality can be proved in a manner similar to Proposition 2.

Next, the second term of (55) can be bounded as:

$$\Pr\{H(\mathbb{P}_{\mathbf{X}}) \geq R, \mathbf{X} = \mathbf{Y}\} \tag{56}$$
$$= \Pr\{H(\mathbb{P}_{\mathbf{X}}) \geq R | \mathbf{X} = \mathbf{Y}\} \Pr\{\mathbf{X} = \mathbf{Y}\} \tag{57}$$
$$\leq 2^{-n(\mathbf{E}_S(P_X \diamond P_Y, R) + H_2(X,Y) - \epsilon)}, \tag{58}$$

where (58) follows from (5) and since for sufficiently large n,

$$\Pr\{H(\mathbb{P}_{\mathbf{X}}) \geq R\} \leq 2^{-n(\mathbf{E}_S(P_X, R) - \epsilon)} \tag{59}$$

(see, e.g. [5, Thm. 2.15]). An application of Lemma 1 establishes the direct part.

Converse part: Let $(T, g^{\star})$ be a rate-R, admissible, augmented identification scheme. Define U_T to be the set of sequences that T maps to unique signatures, i.e.

$$\mathsf{U}_T \triangleq \left\{\mathbf{x} \in \mathcal{X}^n : T^{-1}(T(\mathbf{x})) = \{\mathbf{x}\}\right\}. \tag{60}$$

Note that $|\mathsf{U}_T| \leq 2^{nR}$. With this notation, we have that

$$\Pr\{g^{\star}(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\} \tag{61}$$
$$\geq \Pr\{T(\mathbf{X}) = T(\mathbf{Y}), \mathbf{X} \notin \mathsf{U}_T\} \tag{62}$$
$$= \Pr\{T(\mathbf{X}) = T(\mathbf{Y}), \mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T\} \tag{63}$$
$$\geq \Pr\{T(\mathbf{X}) = T(\mathbf{Y}), \mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} \tag{64}$$
$$= \Pr\{T(\mathbf{X}) = T(\mathbf{Y}) | \mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\}$$
$$\times \Pr\{\mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\}. \tag{65}$$

Note that given $(\mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}})$, $\mathbf{X}$ and $\mathbf{Y}$ are still independent, and have the same distribution (uniformly on the same type class). Therefore the random variables $T(\mathbf{X})$ and $T(\mathbf{Y})$ are independent and have the same distribution. Proposition 1 implies that

$$\Pr\{T(\mathbf{X}) = T(\mathbf{Y}) | \mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} \geq 2^{-nR}. \tag{66}$$

As for the other term in (65), write the following:

$$\Pr\{\mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T, \mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y}}\} \tag{67}$$
$$= \sum_{Q \in \mathcal{P}_n(\mathcal{X})} \Pr\{\mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T | \mathbf{X} \in \mathbb{T}_Q, \mathbf{Y} \in \mathbb{T}_Q\}$$
$$\times \Pr\{\mathbf{X} \in \mathbb{T}_Q, \mathbf{Y} \in \mathbb{T}_Q\} \tag{68}$$
$$\geq \sum_{\substack{Q \in \mathcal{P}_n(\mathcal{X});\\ H(Q) \geq R+\epsilon}} \Pr\{\mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T | \mathbf{X} \in \mathbb{T}_Q, \mathbf{Y} \in \mathbb{T}_Q\}$$
$$\times 2^{-n[D(Q||P_X)+D(Q||P_Y)]} \tag{69}$$

for some $\epsilon > 0$. We then write

$$\Pr\{\mathbf{X} \notin \mathsf{U}_T, \mathbf{Y} \notin \mathsf{U}_T | \mathbf{X} \in \mathbb{T}_Q, \mathbf{Y} \in \mathbb{T}_Q\} \tag{70}$$
$$= \Big(\Pr\{\mathbf{X} \notin \mathsf{U}_T | \mathbf{X} \in \mathbb{T}_Q\}\Big)^2 \tag{71}$$
$$= \left(\frac{|\mathbb{T}_Q \setminus \mathsf{U}_T|}{|\mathbb{T}_Q|}\right)^2. \tag{72}$$

In the above, (71) follows since given $\mathbf{X} \in \mathbb{T}_Q, \mathbf{Y} \in \mathbb{T}_Q$, $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution and are still independent, and (72) follows since $\mathbf{X}$ is distributed uniformly within the type class $\mathbb{T}_Q$. For any type $Q \in \mathcal{P}(\mathcal{X})$ such that $H(Q) \geq R + \epsilon$, we have

$$|\mathbb{T}_Q \cap \mathsf{U}_T| \leq |\mathsf{U}_T| \leq 2^{nR}, \tag{73}$$

which is exponentially smaller than $|\mathbb{T}_Q|$, for which (for large enough n) $|\mathbb{T}_Q| \geq 2^{n(H(Q)-\epsilon/2)} \geq 2^{n(R+\epsilon/2)}$. Therefore $|\mathbb{T}_Q \setminus \mathsf{U}_T| \geq \frac{1}{2}|\mathbb{T}_Q|$ holds for large enough n and (72) is lower bounded by $\frac{1}{4}$.

Replacing the summation with a maximization with respect to Q in (69), and combining with (65),(66) we have (for large enough n):

$$\Pr\{g^{\star}(T(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$$
$$\geq 2^{-nR} \max_{\substack{Q \in \mathcal{P}_n(\mathcal{X});\\ H(Q) \geq R+\epsilon}} \tfrac{1}{4} \cdot 2^{-n[D(Q||P_X)+D(Q||P_Y)]}. \tag{74}$$

The converse finally follows by the continuity of $D(Q||P)$ and $H(Q)$ with respect to Q and by taking ϵ to be arbitrarily small. ■

Acknowledgment

We would like to thank Golan Yona for stimulating discussions that motivated this work, and also to the anonymous reviewers who, among other things, pointed to us that the achieving schemes are universal.

References

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

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

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

[4] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52 –60, February 1967.

[5] I. Csiszár and J. Körner, *Information Theory - Coding Theorems for Discrete Memoryless Systems*. Cambridge, 2011.