Published in final edited form as:

Proc Data Compress Conf. 2015 April ; 2015: 413–422. doi:10.1109/DCC.2015.75.

Compression for Similarity Identification: Computing the Error Exponent

Amir Ingber* and Tsachy Weissman†

*Yahoo! Labs, 701 First Ave., Sunnyvale, CA 94089

†Stanford University, 350 Serra Mall, Stanford, CA 94305

Abstract

We consider the problem of compressing discrete memoryless data sequences for the purpose of similarity identification, first studied by Ahlswede et al. (1997). In this setting, a source sequence is compressed, where the goal is to be able to identify whether the original source sequence is similar to another given sequence (called the query sequence). There is no requirement that the source will be reproducible from the compressed version.

In the case where no false negatives are allowed, a compression scheme is said to be reliable if the probability of error (false positive) vanishes as the sequence length grows. The minimal compression rate in this sense, which is the parallel of the classical rate distortion function, is called the *identification rate*. The rate at which the error probability vanishes is measured by its exponent, called the identification exponent (which is the analog of the classical excess distortion exponent). While an information-theoretic expression for the identification exponent was found in past work, it is uncomputable due to a dependency on an auxiliary random variable with unbounded cardinality.

The main result of this paper is a cardinality bound on the auxiliary random variable in the identification exponent, thereby making the quantity computable (solving the problem that was left open by Ahlswede et al.). The new proof technique relies on the fact that the Lagrangian in the optimization problem (in the expression for the exponent) can be decomposed by coordinate (of the auxiliary random variable). Then a standard Carathéodory - style argument completes the proof.

1 Introduction

Consider the problem of compressing a length n source sequence $\mathbf{x} \in \mathcal{X}^n$, where the goal is to identify whether $\mathbf{x}$ is similar to another sequence $\mathbf{y} \in \mathcal{Y}^n$, knowing the compressed representation of $\mathbf{x}$ only. Such a setting is depicted in Fig. 1.

Possible application of such a scheme include efficient similarity search in databases, where a large database is compressed for similarity identification, reducing its size significantly. Then, given a query sequence, it is now possible to answer whether a similar sequence exists

in the database. If indeed such a sequence exists, it is then retrieved from the original database (which can be stored on a remote location/slower media).

We assume that the alphabets $\mathcal{X}$, $\mathcal{Y}$ are finite. Let $\rho: \mathcal{X} \times \mathcal{Y} \rightarrow [0, \infty)$ denote a distortion measure between symbols. This measure is naturally extended to n -sequences as

$$d(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^n \rho(x_i, y_i), \quad (1)$$

where x_i and y_i are the elements of $\mathbf{x}$ and $\mathbf{y}$, respectively. For a similarity threshold $D \geq 0$, we say that $\mathbf{x}$ and $\mathbf{y}$ are similar whenever $d(\mathbf{x}, \mathbf{y}) \leq D$.

When similarity queries are answered from compressed data, it is expected that errors will occur. We make the distinction between false negatives (FNs), where the returned answer is “the sequences are not similar” but the sequence are similar in reality, and false positives, where the returned answer is “similar”, but in reality they are not. A good compression scheme, therefore, is a scheme where both error probabilities are small.

The problem was first studied from an information-theoretic perspective by Ahlswede et al. [1]. First, the authors defined an achievable compression rate as the minimal rate s.t. there exist compression schemes with vanishing FP and FN probabilities, as the sequence length grows. Surprisingly, they noticed that there exist schemes with vanishing FP and FN probabilities, with rate that is arbitrarily small, so the achievable rate in this sense is always zero (see [1, Thm. 1]). Recently, it was found that if FNs are not allowed, the problem of achievable rate becomes interesting again and the minimal achievable rate it is not always zero (see [2], which also studies the connection to lossy compression schemes).

After discovering that requiring the probabilities of FP and FN to go to zero leads to degeneracy, the authors in [1] moved on to the case where the error probabilities are required to vanish *exponentially*. Specifically, they studied the minimal compression rate where the FP and FN probabilities are required to vanish exponentially with prescribed exponents α and β , respectively. For i.i.d. source and query sequences that are statistically independent, the minimal rate was found (as a function of the exponents α and β). However, the expression for computing the rate depends on an auxiliary random $U \in \mathcal{U}$ variable that has unbounded cardinality. Indeed, the authors in [1] state that “*It seems to the authors that there is no easy way to apply the support lemma ... to upper bound the cardinality of the set $\mathcal{U}$.*”, and moved on to try and approximate the expression with growing alphabet size. However, even for the binary alphabet the computation is infeasible, since it requires solving an optimization problem of prohibitively high dimensions (involving probability distributions of cardinality at least 2^{64} ; more details in the next section).

In this work we make the exponent computable (for the case of no false negatives), by providing a new cardinality bound on the alphabet size of the auxiliary random variable. The proof relies on fundamental results in nonlinear programming, and specifically on Lagrange multipliers. We show that the Lagrangian in the optimization problem (in the expression for

the exponent) can be decomposed by coordinate (of the auxiliary random variable). This makes the problem slightly more structured, allowing for a standard Carathéodory - style argument to give a bound on the cardinality.

In the next section we formally describe the setting and state the main result. The proof follows in Sec. 3.

2 Main result and proof outline

2.1 Notation

We generally follow the notation of [3]. Random variables are denoted with capital letters (e.g. X, Y, U). Finite alphabets are denoted by $\mathcal{X}, \mathcal{Y}, \mathcal{U}$. Probability distributions are denoted by P with a subscript denoting the random variable(s): for example, P_X denotes the distribution of X , and $P_{Y|X}$ denotes the conditional probability of Y given X . We shall also use the term “channels” interchangeably to refer to conditional distributions. For example, a channel V from $\mathcal{X}$ for $\mathcal{Y}$ denotes a conditional probability distribution of a random variable taking values in $\mathcal{Y}$, given another random variable taking values in $\mathcal{X}$. Also, the set of distributions over an alphabet $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$.

We use the standard notation for information-theoretic quantities: the entropy of a random variable X is given by¹ $H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$, and is equivalently denoted by $H(P_X)$. The mutual information between random variables X and Y is denoted by $I(X; Y)$. The Kullback-Leibler divergence between two random variables over the same alphabet (or between two distributions) is denoted by $D(X\|Y) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{P_Y(x)}$. For two channels (conditional distributions) V and W from $\mathcal{X}$ to $\mathcal{Y}$, and a random variable X with distribution P_X taking values in $\mathcal{X}$, the conditional KL divergence is given by

$$D(V\|W|P_X) = \sum_{x \in \mathcal{X}} P_X(x) D(V(\cdot|x)\|W(\cdot|x)) \quad (2)$$

$$= \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} V(y|x) \log \frac{V(y|x)}{W(y|x)}, \quad (3)$$

where $V(\cdot|x)$ denotes the output distribution of the channel V , with the random variable X at its input (and similarly for $W(\cdot|x)$).

With slight abuse of notation, we denote by $D(V\|P_Y|P_X)$ the conditional KL divergence (given P_X) between the channel V and a channel whose output distribution is P_Y regardless of its input. It can be shown that $D(V\|P_Y|P_X) = I(X; \tilde{Y}) + D(\tilde{Y}\|Y)$, where $\tilde{Y}$ is a random

¹Logarithms are taken with base 2, and information rates are given in bits.

variable (independent of X and Y) that results in the output of the channel V with X at its input.

2.2 Problem formulation

Let $\mathbf{X} \in \mathcal{X}^n$ be an i.i.d. vector with distribution P_X , and let $\mathbf{Y} \in \mathcal{Y}^n$ be distributed *i.i.d.* with distribution P_Y . A compression scheme is a function $T: \mathcal{X}^n \rightarrow B$, where $B \subseteq \{0, 1\}^*$ is a prefix-free set, and a decision function $g: \mathcal{Y}^n \times B \rightarrow \{\mathbb{Y}, \mathbb{N}\}$. A source sequence $\mathbf{x} \in \mathcal{X}^n$ is compressed into a bit sequence $T(\mathbf{x})$, and then the similarity query of whether $\mathbf{x}$ and $\mathbf{y}$ are similar is answered by $g(\mathbf{y}, T(\mathbf{x}))$.

The compression rate of the scheme is given by²

$$R = \frac{1}{n} [\text{length of } T(\mathbf{x})]. \quad (4)$$

We do not allow false negatives for the scheme. Formally, this is equivalent to the requirement that for all $\mathbf{x}, \mathbf{y}$ s.t. $d(\mathbf{x}, \mathbf{y}) \leq D$, it will hold that $g(\mathbf{y}, T(\mathbf{x})) = \mathbb{Y}$.

The false positive probability, which we wish to minimize, is given by

$$\Pr(\text{FP}) = \Pr\{g(\mathbf{Y}, T(\mathbf{X})) = \mathbb{Y} | d(\mathbf{X}, \mathbf{Y}) > D\}. \quad (5)$$

The first question to be asked is the following: what is the minimal rate for which there exists a scheme with vanishing $\Pr(\text{FP})$? This question is not answered in [1], where the authors allowed $\Pr(\text{FN}) \rightarrow 0$ (as opposed to $\Pr(\text{FN}) = 0$). However, they did show that $\Pr(\text{FP})$ can be made to vanish exponentially fast if and only if R is above a certain threshold. In [2], it is shown that this threshold, which we call $R_{\text{ID}}(D)$, is indeed the fundamental limit in the sense that the probability of false positives cannot be made to vanish (even not exponentially) if $R < R_{\text{ID}}(D)$. The fundamental limit, called the identification rate, is given by

$$R_{\text{ID}}(D) \triangleq \min_{P_{U|X}: \sum_{u \in \mathcal{U}} P_U(u) \bar{\rho}(P_{X|U}(\cdot|u), P_Y) \geq D} \min_{P_U} I(X; U), \quad (6)$$

where the $\bar{\rho}(\cdot, \cdot)$ distance is defined by

$$\rho(P_X, P_Y) \triangleq \min_{X, Y: \text{random variables with marginals } P_X, P_Y} [\rho(X, Y)]. \quad (7)$$

²The symbol $\mathbb{E}[\cdot]$ denotes statistical expectation, where the standard $E(\cdot)$ denotes the error exponent function.

As mentioned above, $\Pr\{\text{FP}\}$ can be made to vanish exponentially if $R > R_{\text{ID}}(D)$. We denote this exponent by $E(R)$, and formally define it by

$$E(R) \triangleq \limsup_n \frac{1}{n} \log p_n^*, \quad (8)$$

where p_n^* is the optimal false positive probability at rate R and sequence length n . This exponent was derived in [1], but remains essentially uncomputable, as we elaborate next.

2.3 The expression for the false-positive exponent

The false positive error exponent can be extracted from Ahlswede et al. [1] and is given by

$$E(R) \triangleq \limsup_{k \rightarrow \infty} E_k(R), \quad (9)$$

where

$$E_k(R) \triangleq \max_{W: I(X;U) \leq R} \min_{\Phi: \mathbb{E}[\rho(X,\tilde{Y})] \leq D} I(U; \tilde{Y}) + D(\tilde{Y} \| Y), \quad (10)$$

and $|\mathcal{U}| = k$. In (10), W denotes a channel from $\mathcal{X}$ to $\mathcal{U}$, and Φ denotes a channel from $\mathcal{X} \times \mathcal{U}$ to $\mathcal{Y}$. The output of the channel Φ , denoted $\tilde{Y}$, takes values in $\mathcal{Y}$. Note that Y is independent of $(X, U, \tilde{Y})$. See Fig. 2 for an illustration of these relationships.

What makes (9) uncomputable is the absence of any cardinality bound on $\mathcal{U}$, the alphabet of the auxiliary random variable. This fact was of course noticed by the authors of [1], which attempted to quantify how fast the convergence in (9) takes place. We repeat the result here:

Proposition 1 (Eq. (4.41) in [1])—For any $m \geq 2^{16|\mathcal{X}|^2}$, the exponent $E(R)$ is bounded by

$$E_k(R) \leq E(R) \leq E_k(R) + \frac{\log^2 m}{m}, \quad (11)$$

where k is equal to the size of m -types on the alphabet $\mathcal{X}$, given by $\binom{m+|\mathcal{X}|-1}{m}$.

For example, in the case of binary sequences, $|\mathcal{X}| = 2$, we take the minimum value of $m = 2^{64}$ for which (11) holds. For this value, we have $k = 2^{64} + 1$. The optimization problem (10) now involves maximization of a (non-convex) function of the channel W , which has $2k$ free parameters - obviously a prohibitive dimension for optimization [and this is even without

taking into account the inner minimization w.r.t. Φ (which happens to be convex)]. Therefore, a bound on the cardinality of $\mathcal{U}$ is highly desirable.

2.4 Standard cardinality bounds and why they do not work

Standard cardinality bounds on auxiliary random variables in information theory are based mainly on Carathéodory's theorem in convex analysis (see, e.g. [4]). An example for such an argument can be applied for the identification rate $R_{\text{ID}}(D)$, given in (6): for each channel W that is a candidate for optimization, we write a finite number of quantities that are of interest, all of which follow the form of $\sum_{u \in \mathcal{U}} P_U(u) \phi(P_{X|U}(\cdot|u))$, where the function $\phi(\cdot)$ does not depend on the distribution P_U . In the specific example, these values are:

$$\sum_{u \in \mathcal{U}} P_U(u) P_{X|U}(x|u) = P_X(x), \quad \text{for all } x \in \mathcal{X}, \quad (12)$$

$$\sum_{u \in \mathcal{U}} P_U(u) D(P_{X|U}(\cdot|u) \| P_X) = I(X; U), \quad (13)$$

$$\sum_{u \in \mathcal{U}} P_U(u) \bar{\rho}(P_{X|U}(\cdot|u), P_Y). \quad (14)$$

The first $|\mathcal{X}|$ quantities correspond to the fact that the marginal w.r.t. X of $(P_U, P_{X|U})$ is exactly P_X . The next quantity is the mutual information $I(X; U)$, which is what we want to minimize. The last quantity is the expression that appears in the constraint of the optimization. So far, we have $|\mathcal{X}| + 1$ quantities of interest, which are expressed as a convex combination of other quantities that do not depend on P_U (note that we can restrict our attention to $|\mathcal{X}| - 1$ elements since the first $|\mathcal{X}|$ equations must sum to 1). The coefficients of the convex combination are the values $P_U(u)$. By Carathéodory's theorem, the quantities can be expressed as a convex combination of at most $|\mathcal{X}| + 2$ elements. In other words, there exists another prior on U , denoted $P_{U'}$, with at most $|\mathcal{X}| + 2$ elements with nonzero probability, for which the quantities in (12) – (14) are the same. Therefore this $P_{U'}$, along with the same $P_{X|U}$ as before, determine a value of a new channel W' , for which the constraint is met (due to (14)), and that attains the same mutual information (due to (13)). Eq. (12) assures that the marginal of $P_{U'}$, $P_{U'X}$ w.r.t. X is still P_X . Therefore it suffices to search over all auxiliary random variables with cardinality $|\mathcal{X}| + 2$, and the search space will be exhausted.

We briefly note that more delicate arguments can improve this cardinality bound (see [2, Sec. VII]). Also, it is worth mentioning that this technique was originally used for classical problems such as Wyner-Ziv coding [5], and is sometimes referred to as the “support lemma” (see [3]).

The key point with this argument is that the quantities of interest can be represented as expectations (w.r.t. U) of functions that do not depend on P_U . Unfortunately, in the case of the error exponent at hand, the situation is different. Specifically, the minimizer Φ of the inner minimization in (10) depends on P_U through the constraint. Therefore, a more refined argument is required.

2.5 Main result and proof outline

As stated above, the main result is a new cardinality bound:

Theorem 2—*For the evaluation of $E(R)$ in (9), it suffices to consider $\mathcal{U}$ with $|\mathcal{U}| = |\mathcal{X}| + 3$. In other words,*

$$E(R) = E_{|\mathcal{X}|+3}(R). \quad (15)$$

The proof can be divided into three parts:

Reformulation. We reformulate the inner optimization in (10) so that the minimization is now on a channel V from $\mathcal{U}$ to $\mathcal{Y}$. It follows that the problem is convex and that there must be a unique solution V^* .

Lagrange. We use a Lagrange multiplier to characterize the optimal solution V^* , and to show its dependency on the channel W is only via the reverse channel $P_{X|U}$ which we denote $\bar{W}$. The key point is that the Lagrangian in the inner minimization in (10) decomposes (w.r.t. the elements of $\mathcal{U}$).

Carathéodory. Finally, since the minimizer V^* does not depend on P_U , we can use the standard Carathéodory-style argument to arrive at the cardinality bound.

The proof follows in the next (and final) section.

3 Proof of the main result

In this section we give the proof of the main result. We present standard derivations and arguments in propositions, and state them without proofs due to space limitations.

3.1 The $\bar{\rho}$ distance and its properties

The $\bar{\rho}$ distance between P_X, P_Y (as appears in $R_{\text{ID}}(D)$) is defined by

$$\bar{\rho}(P_X, P_Y) \triangleq \min_Q \mathbb{E}_Q [\rho(X, Y)], \quad (16)$$

where Q is any joint distribution on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X and P_Y .

The $\bar{\rho}$ distance goes by many names, such as the Wasserstein (or Vaserstein) distance, the Kantorovich distance and also the Transport distance (see [6] for a survey). We mention

(without proof) that if $P_X \in \mathcal{P}(\mathcal{X})$ is fixed, then $\bar{\rho}(P_X, P_Y)$ is a convex function of $P_Y \in \mathcal{P}(\mathcal{Y})$. It follows that the set $\{P \in \mathcal{P}(\mathcal{Y}) : \bar{\rho}(P_X, P) \leq D\}$ is convex.

3.2 Reformulating the Identification Exponent with $\bar{\rho}$

Proposition 3 (Reformulation with $\bar{\rho}$)—The error exponent can be rewritten as

$$E_k(R) \triangleq \max_{W: I(X;U) \leq R} \min_{V: \sum_{u \in \mathcal{U}} P_U(u) \cdot \bar{\rho}(\bar{W}(\cdot|u), V(\cdot|u)) \leq D} I(U; \tilde{Y}) + D(\tilde{Y} \| Y), \quad (17)$$

where $(P_U, \bar{W})$ denotes the joint distribution of U and the output of the reverse channel from $\mathcal{U}$ to $\mathcal{X}$, that is determined by P_X and W . V is a channel from $\mathcal{U}$ to $\mathcal{Y}$. The relationship between the channels and the RVs is depicted in Fig. 3.

The proof of the proposition follows directly from the definition of the $\bar{\rho}(\cdot, \cdot)$ distance and is omitted. We also note that the connection to $\bar{\rho}(\cdot, \cdot)$ has already been made in [1] in the context of the identification rate (see [1, Appendix III]).

Next, consider a fixed channel W from $\mathcal{X}$ to $\mathcal{U}$ as a candidate for the optimization in (17). Define the exponent it attains by

$$\tilde{E}(W) \triangleq \min_{V: \sum_{u \in \mathcal{U}} P_U(u) \cdot \bar{\rho}(\bar{W}(\cdot|u), V(\cdot|u)) \leq D} D(V \| P_Y | P_U), \quad (18)$$

where $(P_U, \bar{W})$ denote the marginal distribution of U and the reverse channel from $\mathcal{U}$ to $\mathcal{X}$, respectively, that are determined by P_X and W .

Proposition 4 (Properties of $\tilde{E}(W)$)—For any channel W with $P_U(u) > 0$ for all $u \in \mathcal{U}$, if (18) is feasible (i.e. there exist a channel V that satisfies the constraints), then the minimum in (18) is finite, and is attained at a unique channel V^* .

Proof: First note that the optimal value is bounded from below since the KL divergence is nonnegative. Next, consider the domain

$$\Omega(W) \triangleq \left\{ V: \sum_{u \in \mathcal{U}} P_U(u) \cdot \bar{\rho}(\bar{W}(\cdot|u), V(\cdot|u)) \leq D \right\}. \quad (19)$$

Following the convexity property of $\bar{\rho}(\cdot, \cdot)$, it is easy to see that $\Omega(W)$ is nothing but a level set of a convex function, and therefore convex. Since the KL divergence is a strictly convex function (given one of the distributions), and that $P_U(u) > 0$ for all u , it follows that the conditional KL divergence is also strictly convex. As an optimization program of a strictly convex function over a convex domain, (18) must have a unique minimizer V^* .

Next, define the Lagrangian of the optimization program (18):

$$\mathcal{L}(V, \lambda) \triangleq D(V \| P_Y | P_U) + \lambda \left[\sum_{u \in \mathcal{U}} P_U(u) \cdot \bar{\rho} \left(V(\cdot | u), \overleftarrow{W}(\cdot | u) \right) - D \right] \quad (20)$$

$$= \sum_{u \in \mathcal{U}} P_U(u) \left[D(V(\cdot | u) \| P_Y) + \lambda \bar{\rho} \left(V(\cdot | u), \overleftarrow{W}(\cdot | u) \right) - \lambda D \right]. \quad (21)$$

Following the proof of Prop. 4, it is easy to see that $\mathcal{L}(V, \lambda)$ is a strictly convex function of V .

For any constant $\lambda \geq 0$, define the function $F_\lambda : \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{Y})$:

$$F_\lambda(\Psi) \triangleq \arg \min_P D(P \| P_Y) + \lambda \bar{\rho}(\Psi, P). \quad (22)$$

Note that since the minimizer here is unique, $F_\lambda(\cdot)$ is well defined. Typically, the argument Ψ of the function F_λ will be the conditional distribution $\overleftarrow{W}(\cdot | u)$ for some $u \in \mathcal{U}$, and the output will be the conditional distribution $V(\cdot | u)$ for the same $u \in \mathcal{U}$.

Given a channel W and, and a $\lambda \geq 0$, we define the minimizer of the Lagrangian as:

$$V_\lambda \triangleq \arg \min_{V: \mathcal{P}(\mathcal{U}) \rightarrow \mathcal{P}(\mathcal{Y})} \mathcal{L}(V, \lambda). \quad (23)$$

With the definition of F_λ , and the fact that the Lagrangian decomposes as in (21), we know that

$$V_\lambda(\cdot | u) = F_\lambda(\overleftarrow{W}(\cdot | u)), \text{ for all } u \in \mathcal{U}. \quad (24)$$

We pause here to recall the following concepts from nonlinear programming ([7, Sec. 5]). Since we only have one inequality, we slightly simplify the notations.

Consider the general optimization program $\min_{x \in \mathcal{D}: g(x) \leq 0} f(x)$, where $\mathcal{D} \subseteq \mathbb{R}^n$, and assume its optimal value is f^* . A vector x^* is called an optimal solution if $x^* \in \mathcal{D}$ and $f(x^*) = f^*$. We say that $\lambda^* \geq 0$ is a *Lagrange multiplier* if $f^* = \inf_{x \in \mathcal{D}} f(x) + \lambda^* g(x)$. We utilize the following important result:

Proposition 5 (Prop. 5.1.5 in [7])— *(x^*, λ^*) is an optimal solution-Lagrange multiplier pair if and only if:*

$$x^* \in \mathcal{D}; g(x) \leq 0 \quad (\text{Primal feasibility}), \quad (25)$$

$$\lambda^* \geq 0 \quad (\text{Dual feasibility}), \quad (26)$$

$$x^* = \arg \min_{x \in \mathcal{D}} f(x) + \lambda^* g(x) \quad (\text{Lagrangian Optimality}), \quad (27)$$

$$\lambda^* g(x^*) = 0 \quad (\text{Complementary slackness}). \quad (28)$$

Note that Prop. 5 is only useful when Lagrange multipliers exist. For our case, the problem at hand is (18), where there is one constraint, which can be rewritten as a set of linear constraints. In addition, the optimization domain is the probability simplex which is polyhedral. Combined with the fact that the target function is convex, we can use Prop. 5.2.1 of [7] and deduce that if (18) is feasible, then there must exist at least one Lagrange multiplier λ^* (in other words, there is no duality gap; see [7, Ch. 5]).

Let W be a given channel where the optimization (18) is feasible so the attained exponent $\tilde{E}(W)$ is finite. Let V^* be the minimizing channel for $\tilde{E}(W)$. By the discussion above, there exists at least one Lagrange multiplier λ^* . Therefore (V^*, λ^*) are an optimal solution-Lagrange multiplier pair. By Prop. 5, we know that

$$\sum P_U(u) \bar{\rho}(\tilde{W}(\cdot|u), V^*(\cdot|u)) \leq D \quad (\text{Primal feasibility}), \quad (29)$$

$$\lambda^* \geq 0 \quad (\text{Dual feasibility}), \quad (30)$$

$$V^* = \arg \min_V \mathcal{L}(V, \lambda^*) \quad (\text{Lagrangian Optimality}), \quad (31)$$

$$\lambda^* \left[\sum P_U(u) \bar{\rho}(\tilde{W}(\cdot|u), V^*(\cdot|u)) - D \right] = 0 \quad (\text{Complementary slackness}). \quad (32)$$

By (31) and (23)–(24), we get that

$$V_{\lambda^*}(\cdot|u) = F_{\lambda^*}(\overleftarrow{W}(\cdot|u)), \text{ for all } u \in \mathcal{U}. \quad (33)$$

In other words, the minimizer V^* does not depend on the prior P_U at all, but only on $\overleftarrow{W}(\cdot|u)$ and the Lagrange multiplier λ^* .

Let us summarize what is known about $\overleftarrow{W}(\cdot|u)$:

$$\sum_{u \in \mathcal{U}} P_U(u) \overleftarrow{W}(x|u) = P_X(x) \quad \forall x \in \mathcal{X}, \quad (34)$$

$$\sum_{u \in \mathcal{U}} P_U(u) H(\overleftarrow{W}(\cdot|u)) = H(P_X) - R, \quad (35)$$

$$\sum_{u \in \mathcal{U}} P_U(u) \bar{\rho}(\overleftarrow{W}(\cdot|u), F_{\lambda^*}(\overleftarrow{W}(\cdot|u))) \triangleq D_0 \leq D, \quad (36)$$

$$\sum_{u \in \mathcal{U}} P_U(u) D(F_{\lambda^*}(\overleftarrow{W}(\cdot|u)) \| P_Y) = \tilde{E}(W). \quad (37)$$

Consider the $(|\mathcal{X}| + 2)$ -dimensional vector (assuming w.l.o.g. that $\mathcal{X} = \{1, 2, \dots, |\mathcal{X}|\}$):

$$[P_X(1), P_X(2), \dots, P_X(|\mathcal{X}|-1), H(P_X) - R, D_0, \tilde{E}(W)]. \quad (38)$$

Note that we only need $|\mathcal{X}| - 1$ equations of the type (34). By the equations above, this vector is a convex combinations of $|\mathcal{U}|$ points. By Carathéodory's theorem, the point can be represented as a convex combination of $|\mathcal{X}| + 3$ of these points. In other words, there exists a prior $P_{U'}$, with at most $|\mathcal{X}| + 3$ nonzero elements, for which (34)–(37) hold (with P_U replaced by $P_{U'}$). Let us interpret these new equations. Eq. (34) for $P_{U'}$ means that the pair $P_{U'}$, $\overleftarrow{W}$ induces a joint distribution on $\mathcal{X}, \mathcal{U}$ that can be decomposed to P_X, W' , where W' is a channel from $\mathcal{X}$ to $\mathcal{U}$. Eq. (35) for $P_{U'}$ says that the conditional entropies $H(X|U)$ and $H(X|U')$ are equal, and therefore $I(P_X, W) = I(P_X, W')$.

The next desirable step is to claim that $\tilde{E}(W) = \tilde{E}(W')$. However, this is not immediate from (37) since we do not know yet who is the minimizer V^* for the new channel W' . For this, observe that because (36) holds for $P_{U'}$, it is easy to verify that (29)–(32) (now with P_U replaced with $P_{U'}$) hold for V^* and λ^* . Then, we can invoke Prop. 5 (this time at the opposite direction), and conclude that (V^*, λ^*) is an optimal solution-Lagrange multiplier

pair for the minimization problem that results in $\tilde{E}(W')$. Therefore the same V^* is a minimizer for W' , and therefore the exponent that attained by W' , $\tilde{E}(W')$ is given by (37), and equal to $\tilde{E}(W)$.

To summarize, given a channel W with an arbitrary size of the alphabet $\mathcal{U}$, we have shown an alternative channel W' with at most $|\mathcal{X}| + 3$ elements of $\mathcal{U}$ that have nonzero probability, for which $\mathcal{I}(P_X, W) = \mathcal{I}(P_X, W')$, and $\tilde{E}(W) = \tilde{E}(W')$. Therefore, for the evaluation of the exponent $\tilde{E}(R)$ in (9), it suffices to consider $\mathcal{U}$ with $|\mathcal{U}| = |\mathcal{X}| + 3$. This completes the proof of the theorem.

References

1. Ahlswede R, Yang E-h, Zhang Z. Identification via compressed data. IEEE Trans on Information Theory. Jan; 1997 43(1):48–70.
2. Ingber, A., Weissman, T. The minimal compression rate for similarity identification. IEEE Trans on Information Theory. 2013. (under revision)[Online]. Available: <http://arxiv.org/abs/1312.2063>
3. Csiszár, I., Körner, J. Information Theory - Coding Theorems for Discrete Memoryless Systems. Cambridge; 2011.
4. Rockafellar, RT. Convex Analysis. Princeton University Press; 1970.
5. Wyner AD, Ziv J. The rate-distortion function for source coding with side information at the decoder. IEEE Trans on Information Theory. Jan; 1976 22(1):1–10.
6. Villani, C. Optimal transport: old and new. Vol. 338. Springer; 2009.
7. Bertsekas, DP. Nonlinear Programming. 2. Athena Scientific; 1999. [Online]. Available: <http://www.athenasc.com/nonlinbook.html>

Figure 1.
Detecting similarity from compressed data. $T(\cdot)$ denotes the compression function.

Figure 2.
Relationships between the RVs X , Y , U , $\tilde{Y}$ and the channels W and Φ .

Y : independent of $(X, U, \tilde{Y})$

Figure 3. Relationships between the RVs X , Y , U , $\tilde{Y}$ and the channels $\overleftarrow{W}$ and V .