Upper and Lower Bounds on the Reliability of Content Identification

Gautam Dasarathy
Electrical and Computer Engineering
University of Wisconsin - Madison
dasarathy@wisc.edu

Stark C. Draper
Electrical and Computer Engineering
University of Toronto
stark.draper@utoronto.ca

Abstract—In this paper we quantify upper and lower bounds on the reliability function for the problem of content identification from a large database based on noisy queries.

I. INTRODUCTION

We consider the problem of content identification from a database. The database houses quantized representations of $2^{nR_1}$ length-$n$ “enrollment” vectors, where $R_1 > 0$ is what we call the identification rate (we will not worry about integer effects in this paper). When presented with a noisy observation of one of the (non-quantized) enrollment vectors, the goal is to identify (using only this observation and the stored data) which enrollment vector generated the noisy observation (or the “query”).

The information-theoretic limits of this problem were found in [1] (see also [2] for related problem formulations) under the following model. The enrollment vectors, $X_m$ for $m = 1, \ldots, 2^{nR_1}$ are chosen in an i.i.d. manner according to $p_X$. An index of a codeword in a pre-defined rate-$R_C$ codebook $C$ is used to represent each of these vectors in a database, where $R_C > 0$ is what we call the compression rate. As it takes $nR_C$ bits to store the index of the representation of each enrollment, and $2^{nR_1}$ representations are stored, the entire database is of size $nR_C2^{nR_1}$ bits. The query $Y$ presented during the identification phase is a length-$n$ observation of one of the $2^{nR_1}$ (unquantized) enrollment vectors observed via the discrete memoryless channel (DMC) $p_{Y|X}(\cdot | \cdot)$. The decoder’s objective is to identify reliably the codeword corresponding to the enrollment vector from which the query was generated.

In [1], the capacity region of this problem was shown to be parameterized by a (rate-distortion) test channel $p_{U|X}$. Given the joint distribution $p_{U|X}(u|x)p_X(x)p_{Y|X}(y|x)$, the “compression/identification” rate pair $(R_C, R_I)$ is achievable if

$$R_C > I(U;X) \quad R_I < I(U;Y).$$

The achievable rate region is the convex hull of the union of achievable rate pairs over all test channels. The achievability is closely related the the Wyner-Ziv problem. The codebook $C$ needs to have good covering properties $(R_C > I(U;X))$ to ensure reliable encoding. The union of codewords stored in the database (which is a subset of $C$) needs to form a code that has good packing properties $(R_I < I(U;Z))$ for reliable decoding.

In a previous paper [3], we quantified an achievable error exponent tradeoff for a particular encoding and decoding strategy, in other words, we derived a lower bound on the reliability function for the content identification problem. Our results used a novel lemma [3, Lemma 2] that characterized the end-to-end statistical relationship between the codeword and the channel output. In this paper, we will first quantify an upper bound on the error exponent over all encoding/decoding strategies. This upper bound can be shown to be strictly tighter than the naive sphere packing upper bound. We also derive a modification of the lower bound from [3] whose form parallels the upper bound expression obtained here.

II. PROBLEM FORMULATION

In this section we formally state the problem setting and define the notation that we will use throughout the paper.

Environment: We suppose that there are $M = 2^{nR_1}$ items to be represented in the database. To each item is associated a length-$n$ “feature vector” or “enrollment vector” $X(m) \in \mathcal{X}^n$, $m = 1, 2, \ldots, M$ which are drawn independently from $p_X$ in an i.i.d manner where $\mathcal{X}$ is a finite alphabet.

Enrollment Phase: In the enrollment phase, each feature vector is mapped to a codeword selected from a pre-defined rate-$R_C$ codebook $C \triangleq \{u(1), u(2), \ldots, u(2^{nR_C})\}$ ignoring integer effects. The codewords are made up of symbols from the finite alphabet $\mathcal{U}$. We represent the operation of assigning a codeword $u$ to the feature vector $x$ by the function $f : \mathcal{X}^n \to \{1, 2, \ldots, L\}$, where we define $L = 2^{nR_C}$. The notation $J(m)$ will be used to denote the (random) quantity $f(X(m))$. Observe that the codeword that gets assigned to object $m$ is $u(J(m))$. The set of all (possibly non-distinct) codewords corresponding to the enrolled items $\{u(J(1)), u(J(2)), \ldots, u(J(M))\}$ will henceforth be called the database and will be denoted as $\mathcal{D}$. The database $\mathcal{D} \subset C$ can be thought of as an analogue to the random bin of codewords in the Wyner-Ziv problem.

Identification Phase: An index $W$ is selected uniformly at random from $\{1, 2, \ldots, M\}$. This corresponds to the item that the user wishes to query for. A noisy version of $\mathcal{X}^n(W)$, $\mathcal{Y}^n \in \mathcal{Y}^n$ is then observed at the database where the conditional distribution of $\mathcal{Y}^n$ is given by $\Pr[Y = y|X = x] = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$, i.e., we model the noise as a DMC $p_{Y|X} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is also assumed to be a finite
alphabet. The objective at the identification phase is to produce an estimate \( \hat{W} \) from the observed random vector \( Y \) and the stored sequences \( D \). This estimation operation, which we will call decoding, is denoted by the function \( g : Y^n \times U^nM \rightarrow \{1, 2, \ldots, M\} \). Our aim is to design an encoding function \( f(.) \) and a decoding function \( g(.) \) such that, with high probability, \( g(.) \) returns the correct value of \( W \).

Given a finite alphabet \( S \), we write \( \Pi(S) \) to denote the set of all distributions on \( S \). For other notation, we mainly follow [4].

### III. MAIN RESULTS

We will now state the main results of our paper. Given some joint distribution \( p_{XY} \in \Pi(X \times Y) \), let \( P_i(f, g, p_{XY}) \) denote the probability that a particular choice \( (f, g) \) of encoding and decoding function does not estimate \( W \) correctly. Then, we can define the content identification reliability function as follows

\[
\rho(p_{XY}, R_I, R_C) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \frac{\min_{f, g} P_i(f, g, p_{XY})}{P_i(0, 0, p_{XY})} \right).
\]

where the minimization is over all encoding and decoding function pairs \( (f, g) \) such that \( \log M \leq n(R_I - \epsilon) \) and \( \log L \leq n(R_C + \epsilon) \).

The following theorem provides an upper bound for the reliability function \( \rho \). That is, it gives us a lower bound on the probability of error over all choices of valid encoding and decoding functions for the problem.

**Theorem 1.** Given \( p_{XY} \in \Pi(X \times Y) \) and \( R_I, R_C > 0 \), the reliability function \( \rho(p_{XY}, R_I, R_C) \) is upper bounded by

\[
\rho_{\inf}(p_{XY}, R_I, R_C) = \inf_{q_X} \inf_{q_{XY} : I(X, S) \leq R_I} \inf_{q_{XY} : I(Y, S) \leq R_C} D(q_{XY} \| p_{XY}).
\]

(1)

Here \( Y, X, S \) have the joint distribution \( q_{YS} \mid q_{S} \mid X \) and the cardinality of \( S \) satisfies

\[ |S| \leq |X| + |Y| + 2 \]

Next, we will state a theorem that provides a lower bound to the reliability function \( \rho \) and therefore upper bounds the probability of error over all choices of encoders and decoders.

**Theorem 2.** Given \( p_{XY} \in \Pi(X \times Y) \) and \( R_I, R_C > 0 \), the reliability function \( \rho(p_{XY}, R_I, R_C) \) is lower bounded by the quantity

\[
\rho_{\inf}(p_{XY}, R_I, R_C) = \inf_{q_X} \sup_{q_{XY} : I(X, U) < R_C} D(q_{XYU} \| p_{XYU}q_{U} \mid X, U) + |I(U, Y) - R_I|^+,
\]

where the joint distribution of \( X, Y, U \) is \( q_{XYU} q_{U} \mid X, U \).

The rest of the paper will be devoted to proving these two results.

### IV. PROOFS

**A. Proof of Theorem 1**

In order to prove Theorem 1, we begin by picking an arbitrary encoder-decoder pair \( f, g \) that satisfies the rate constraints. That is, \( f \) and \( g \) are such that

\[
\log M \leq nR_I \quad \log L \leq nR_C.
\]

Let us define the “error set”

\[
\mathcal{E} = \{ (x(1), \ldots, x(M), w, y) : g(y, x(1), \ldots, x(M)) \neq w \}.
\]

Our proof will follow these steps:

1. **Step (A):** Let \( q_{XY} \in \Pi(X \times Y) \) be a distribution such that \( I(Y, f(X)) \leq n(R_I - \epsilon) \). Recall that \( g \) is an estimator for the random variable \( W \) and it has access to the random variables \( \{J(i)\}_{i \in [M]} \) and \( Y \). Therefore, Fano’s inequality (see e.g., [5]) tells us that

\[
q^n_{XY}(\mathcal{E}) \geq \frac{H(W \mid J(1), \ldots, J(M), Y) - 1}{\log M}.
\]

Notice here that the conditional entropy is computed with respect to the joint distribution \( q^n_{XY} \).

To get a lower bound on \( q^n_{XY}(\mathcal{E}) \), we will lower bound the conditional entropy term above.

\[
H(W \mid J(1), \ldots, J(M), Y)
\]

\[
= H(W) - I(W; J(1), \ldots, J(M), Y)
\]

\[
\geq H(W) - I(W; Y \mid J(1), \ldots, J(M))
\]

\[
= H(W) - H(Y \mid J(1), \ldots, J(M)) + H(Y \mid J(1), \ldots, J(M), W)
\]

\[
\geq H(W) - H(Y) + H(Y \mid J(W))
\]

\[
\geq nR_I - I(f(X); Y)
\]

In (a), we have used the definition of conditional mutual information and that fact that \( I(W; J(1), \ldots, J(M)) = 0 \) since these quantities are independent. In (b) we expand out the conditional mutual information term and in (c) we use
the fact that conditioning cannot increase entropy (i.e., \(H(Y \mid \{J(i)\}_{i=1}^{M}) \leq H(Y)\)) and that \(Y\) only depends on \(J(W)\) and therefore \(H(Y \mid \{J(i)\}_{i=1}^{M}, W) = H(Y \mid J(W))\). Finally, in (d) we use the fact that \(W\) is a uniform random variable over a set of size \(M = 2^nR_t\) and the fact that, by definition, \(J(W) = f(X)\). The final step follows from our assumption on \(q_{XY}\). Let us assume that \(n \geq 2^{-1}\). Substituting this lower bound back in (4), we have

\[
q_{XY}(\varepsilon) \geq \frac{p_{XY}(\varepsilon \cap D)}{p_{XY}(\varepsilon)} \geq \frac{p_{XY}(\varepsilon \cap D)}{p_{XY}(\varepsilon)} \geq e^{-n[D(q_{XY} \| p_{XY}) + \delta]}.
\]

**Step (B):** In order to lower bound the probability of error, \(p_{XY}(\varepsilon)\), we will use the so-called “change of measure” argument typically used in establishing large deviation rate functions (see e.g., [6] and [4, Page 268, Problem 13]) for a version that is more applicable to an information theoretic setting. The idea is to lower bound the mass of a set (in this case \(\varepsilon\)) under one measure (in this case \(p_{XY}\)) using the mass of the same set under a different measure (in this case \(q_{XY}\)). For \(\delta > 0\), let us define the set of all \((q_{XY}, p_{XY}) – divergence typical sequences as follows

\[
\Delta = \left\{(x, y) : \frac{1}{n} \log q_{XY}(x, y) - D(q_{XY} \| p_{XY}) \leq \delta \right\}.
\]

Notice that by Chebyshev’s inequality, we have that \(q_{XY}(\Delta^c)\) is no greater than \(\frac{1}{n} \log \frac{p_{XY}(\varepsilon)}{p_{XY}(\varepsilon \cap D)}\). Assuming that \(p_{XY} > 0\), this quantity can be uniformly upper bounded by some \(\Delta > 0\) over all \(q_{XY}\) when \(X\) and \(Y\) are finite. Of course, one could obtain a tighter bound for \(q_{XY}(\Delta^c)\) using, for instance, Chernoff bounds. However, this weaker bound suffices for our current purpose. We can now proceed to bound the probability of error as follows for \(n \geq \frac{4R_t}{\delta}\),

\[
\mathbb{P}[\text{error}] = p_{XY}(\varepsilon) \geq p_{XY}(\varepsilon \cap D) \geq \sum_{(x, y) \in \varepsilon \cap D} q_{XY}(x, y) \frac{p_{XY}(x, y)}{q_{XY}(x, y)} \geq \frac{p_{XY}(\varepsilon \cap D)}{p_{XY}(\varepsilon)} \geq e^{-n[D(q_{XY} \| p_{XY}) + \delta]}.
\]

In (a) we use the definition of the divergence typical set \(D\) and in (b) we use the fact that \(P(A \cap B) \geq |P(A) - P(B')|\)^\#. Finally in (c) we use our assumption that \(n \geq \frac{4R_t}{\delta}\). Since this bound holds for any \(q_{XY}\) such that \(f(X, Y) \leq n(R_t - \delta)\), it has to hold even if we take the supremum over all such \(q_{XY}\).’s. Also, since we want our lower bound to hold irrespective of the quality of the encoder/decoder pair picked, we can take the infimum of right side with respect to all valid coding functions. Therefore, the probability of error can be lower bounded by the following quantity

\[
\inf_{q_{XY}} \sup_{f} \inf_{q_{XY} : I(X_f; Y) \leq n(R_t - \epsilon)} \frac{\epsilon}{4R_t} \exp \left( - \frac{n[D(q_{XY} \| p_{XY}) + \delta]}{4R_t} \right).
\]

**Step (C):** Before we proceed, let us restrict our attention to the exponent. Notice that our objective will now be to upper bound the subsequent quantities.

\[
\inf_{q_X} \sup_{f} \inf_{q_{XY} : I(X_f; Y) \leq n(R_t - \epsilon)} D(q_{XY} \| p_{XY}) \leq \inf_{q_X} \sup_{f} \inf_{q_{XY} : I(X_f; Y) \leq n(R_t - \epsilon)} D(q_{XY} \| p_{XY})
\]

In (a) we changed the order of minimization and maximization, and we used the fact that \(\log L = H(f(X)) \geq I(f(X); X)\). In (b), we are replacing the deterministic mapping \(f(X)\) by a random variable \(U\) whose alphabet has cardinality no less than the range of \(f(X)\). Since deterministic functions are a special case of such random mappings, we are increasing the domain of maximization, and hence the inequality (this does not affect the inner minimization since the domain of that infimum is over \(q_{XY}\)). Towards “single-letterizing” the arguments of the optimizations above, we do the following two calculations. First, we observe that

\[
I(Y; U) = \sum_{i=1}^{n} H(Y_i | Y_1^{i-1}) - H(Y_i | U, X_1^{i-1}) \leq \sum_{i=1}^{n} H(Y_i) - H(Y_i | U, X_1^{i-1}) = \sum_{i=1}^{n} I(Y_i; U, X_1^{i-1}) = \sum_{i=1}^{n} I(Y_i; V_i) \leq I(Y_T; V_T, T)
\]

where (a) follows from observing that \(H(Y_i | Y_1^{i-1}) \leq H(Y_i)\) and that \(H(Y_i | U, X_1^{i-1}) = H(Y_i | U, X_1^{i-1}) \leq H(Y_i | U, X_1^{i-1}) \leq H(Y_i | U, X_1^{i-1})\) because of the Markov chain \(Y_1 - U - X_1^{i-1}\). In (b), we set \(V_i \triangleq U, X_1^{i-1}\) and in (c) we merely introduce a random variable \(T\) that is uniformly distributed in \([1, \ldots, n]\) (this is the so called “time sharing” random variable.) Second, using a similar calculation, we have

\[
I(X; U) = \sum_{i=1}^{n} I(X_i; U | X_1^{i-1}) = \sum_{i=1}^{n} I(X_i; X_1^{i-1}, U) = \sum_{i=1}^{n} I(X_i; V_i) = I(X_T; V_T, T)
\]

Using these two calculations and defining \(W \triangleq (V_T, T)\), we can upper bound (7) as

\[
\inf_{q_X} \sup_{f} \inf_{q_{XY} : I(X_f; W) \leq n(R_t + \epsilon)} D(q_{XY} \| p_{XY})
\]
Finally, using continuity arguments, we can show that

\[ \inf_{p_X} \sup_{q_{S|X}} D(q_X \parallel p_X) + D\left( \tilde{J}_{Y|S;X|S|X} \parallel q_{Y|S|X} \right) \]

and that

\[ f(q_X, q_{S|X}) \equiv \inf_{q_{Y|S|X}} D(q_{XY} \parallel p_{XY}). \]

Also, since the optimization problem is convex, we know that \( q_{Y|X} \) satisfies the KKT conditions [7]. To conclude this step, we use the fact that the KKT conditions are essentially \([X] \parallel [Y]\) linear equalities involving \( q_{Y|X} \) (which correspond to the gradient condition) and a set of \([X] \parallel [Y]\) linear equalities corresponding to the mutual information constraint. Therefore, by Caratheodory’s theorem, we have that the cardinality \( |S| \leq |X| \parallel [Y] + |X| + 2 \) and that \( q_{S|Y} \) satisfies the necessary constraints.

**Step (D):** Finally, using continuity arguments, we can show that

\[ \lim_{\epsilon \to 0} \tilde{p}_U(p_{XY}, R_1 + \epsilon, R_C + \epsilon) \to p_U(p_{XY}, R_1, R_C). \]

This concludes the proof.

**B. Proof of Theorem 2**

In [3], we show, using an important lemma about the end-to-end behavior across a Markov chain, that the reliability function \( \rho \) of the problem is lower bounded by

\[ \tilde{p}_L \equiv \inf_{q_X} \sup_{q_{S|X}} D(q_X \parallel p_X) + D\left( J^{\ast}_{Y|S;X|S|X} \parallel q_{S|X} \right) \]

where \( q_{S|Y}(\cdot | x) = \sum_{s \in S} q_X(x) q_{S|X}(\cdot | x) \) is the distribution induced on \( S \) by \( q_X \) and \( q_{S|X} \).

It can be shown that this error exponent is positive in the capacity region indicated in [1]. We will now show that \( \tilde{p}_L \) equals \( \rho_L \) so that the expression matches \( \rho_U \) more closely. Towards this end, consider the first two terms in the expression for \( \tilde{p}_U \) and observe that

\[ D(q_X \parallel p_X) + D\left( J^{\ast}_{Y|S;X|S|X} \parallel q_{S|X} \right) \]

where \( E(q_{Y|S}) \) is defined as the set

\[ \{ J : \sum_{x \in X} J(x, y | s) = q_{Y|S}(y | s), \forall (s, y) \in S \times Y \}. \]

It can be shown that this error exponent is positive in the capacity region indicated in [1]. We will now show that \( \tilde{p}_L \) equals \( \rho_L \) so that the expression matches \( \rho_U \) more closely. Towards this end, consider the first two terms in the expression for \( \tilde{p}_U \) and observe that

\[ D(q_X \parallel p_X) + D\left( J^{\ast}_{Y|S;X|S|X} \parallel q_{S|X} \right) \]

\[ = D(q_X \parallel p_X) + D\left( q_X q_{S|X} q_{Y|S|X} \parallel q_X q_{S|X} q_{Y|S|X} \right) \]

\[ = D(q_X q_{S|X} q_{Y|S|X} \parallel q_X q_{S|X} q_{Y|S|X}) \]

\[ \equiv \inf_{q_{Y|S|X}} D(q_{XY} \parallel p_{XY}). \]

This concludes the proof.

**References**


