On Large Alphabet Compression

Rob Gorman, Vaishakh Ravindakumar

Abstract—In this report, we present results in Large Alphabet Compression. We first show that the min-max redundancy of standard compression tends towards infinity for sufficiently large alphabets. With this, we motivate two other approaches that are employed in compressing large alphabets, namely pattern and shape compression. We then present upper and lower bounds on the min-max redundancy of the same.

I. INTRODUCTION

Shannon showed that any discrete source can be compressed no further than its entropy and if the distribution of the source is known, compression to essentially the entropy can be achieved. However in most practical situations, for instance, text/image compression, the underlying distribution is rarely known.

A common assumption in these scenarios is that the distribution belongs to some class \( \mathcal{P} \), such as the class of i.i.d., Markov or stationary distributions. The objective is then to perform compression almost as well as an oracle which knows the underlying distribution.

Formally, compression of a discrete random variable \( X \in \mathcal{X} \) consists of a prefix-free 1-1 mapping \( \phi : \mathcal{X} \rightarrow \{0,1\}^* \), where \( \{0,1\}^* \) is the collection of all finite binary strings. Every such encoding of \( x \in \mathcal{X} \) is approximately \( \log(1/q(x)) \).

The problem of compression can thus be essentially be reduced to finding such a \( q \) and the resulting compression redundancy is measured by the log-loss \( \log p(x)/q(x) \). Thus, we henceforth refer to compression as finding a probability assignment \( q \).

Now, for a given class \( \mathcal{P} \), the worst case redundancy is defined as

\[
\hat{R}(\mathcal{P}) = \inf_q \sup_{p \in \mathcal{P}} \sum_{x \in \mathcal{X}} p(x) \log p(x)/q(x)
\]

and characterizes the lowest number of additional bits required in the worst case by any possible encoder \( q \).

Characterizing \( \hat{R}(\mathcal{P}) \) for various classes \( \mathcal{P} \) has been the subject of a long line of work. In particular, when \( \mathcal{P} = \mathcal{I}_m^n \) or the class of i.i.d. distributions over \( m \)-alphabet strings of length \( n \), \( [6],[1],[11],[2],[9],[12],[13],[4] \) have shown that

\[
\hat{R}(\mathcal{I}_m^n) = \frac{m-1}{2} \log \frac{n}{\pi} + \log \left( \frac{\Gamma(\frac{1}{2}m)}{\Gamma(\frac{1}{2})} \right) + o_m(1).
\]

For any fixed alphabet size \( m \), the per-symbol redundancy is thus \( O(\log n/n) \) and which diminishes rapidly with \( n \). However, in many applications the alphabet size is very large or even unknown (infinite). An example of this is language compression, when we consider compression over words and especially when we incorporate their Markov structure. Another example is image compression, if we take into account the inter-pixel dependencies.

In such situations, if either \( m/n = \Omega(1) \) or \( m/n \rightarrow \infty \), the redundancy \( \hat{R}(\mathcal{I}_m^n) \) is either a constant or blows up, as detailed in Section III. Thus in the rest of the report (Sections III and IV), alternate approaches taken to compress large alphabets are elucidated. Towards this, the notion of pattern and shape compression are introduced and bounds on their worst-case redundancies are provided. Finally, we summarize in Section V.

II. STANDARD COMPRESSION

Here, again, we let \( \mathcal{P} = \mathcal{I}_m^n \) but consider the case when the alphabet size \( m \), is allowed to grow with the sequence length \( n \).

A. Bounds

We first characterize the resulting redundancy of the class \( \mathcal{I}_m^n \) in various regimes of \( m \) and \( n \). For this, we first note that in [10], Shtarkov showed that for finite \( \mathcal{X} \), \( \hat{R}(\mathcal{P}) \) is achieved by the distribution \( q^*(x) \) defined by

\[
\sum_{x' \in \mathcal{X}} \sup_{p \in \mathcal{P}} p(x')
\]

and the resulting redundancy is

\[
\hat{R}(\mathcal{P}) = \log \left( \sum_{x \in \mathcal{X}} \sup_{p \in \mathcal{P}} p(x) \right).
\]

For the case when \( \mathcal{P} = \mathcal{I}_m^n \), manipulating the above expression as shown in [7], give us the following theorems.

Theorem 1. For all \( m \) and \( n \),

\[
\hat{R}(\mathcal{I}_m^n) \leq \frac{l-1}{2} \log \frac{n}{2\pi} + \log \left( \frac{\Gamma(\frac{1}{2}l)}{\Gamma(l/2)} \right) + \log \left( \frac{m}{l} \right) + O(1).
\]

Theorem 2. For all \( m \), \( n \),

\[
\hat{R}(\mathcal{I}_m^n) \geq \log \left( \sum_{l=1}^{\min(m,n)} \left( \frac{m}{l} \right)^{\frac{1}{2}} \frac{\sqrt{2\pi n}}{\Gamma(l)} \right) \frac{n-1}{l} \left( \frac{l}{2\pi n} \right)^{\frac{1}{2}}.
\]

Simplifying these theorems, again, as in [7], gives the following corollary.

Corollary 1. When \( m = o(n) \),

\[
\hat{R}(\mathcal{I}_m^n) \sim \frac{m-1}{2} \log \frac{n}{m}.
\]
when \( m = \Theta(n) \),
\[
\hat{R}(I_m^m) = \Theta(n)
\]
and when \( n = o(m) \),
\[
\hat{R}(I_m^m) \sim n \log \frac{m}{n}.
\]

B. Interpretation

It can thus be seen from Corollary \([1]\) that for the case when \( m = \Omega(n) \), the per-symbol redundancy is either a constant or blows up with \( n \) and thus standard compression fails.

Now it can be shown that the lower bound on redundancy in Theorem \([2]\) is dominated by the term

\[
\log \left( \sum_{l=1}^{\min\{m,n\}} \binom{m}{l} \binom{n-1}{l-1} \right)
\]

which allows for a simple interpretation of these results.

As is shown in Section \([IV]\) every sequence can be conveyed by describing its 'type' \([IV]\) and then identifying the sequence among all those of this 'type'. Since i.i.d. distributions assign all sequences of a given 'type' the same probability, using the same number of bits to identify all sequences within the 'type' is optimal, and thus the optimal redundancy is approximately the number of bits needed to describe the 'type'.

Now there are precisely \( \sum_{l=1}^{\min\{m,n\}} \binom{m}{l} \binom{n-1}{l-1} \) 'types' where in each summand, the first term corresponds to specifying the symbols appearing in the sequence, and the second to the number of times each symbol appears. Thus, the 'log' of this quantity, is as expected, the dominating term in the lower bound of Theorem \([2]\).

This also motivates the definition of such 'profile' and 'type' based compression approaches. These are shown next in Sections \([III]\) and \([IV]\).

### III. Pattern Compression

In this section, we introduce the notion of pattern compression and provide bounds on its redundancy.

A. Definitions

Let \( A \) be any alphabet. For \( \bar{x} = x_1, \ldots, x_n \in A^n \), let
\[
A(\bar{x}) \equiv \{x_1, \ldots, x_n\}
\]
denote the set of symbols in \( \bar{x} \). We define the index of \( x \in A(\bar{x}) \) as
\[
l_x(\bar{x}) \equiv \min\{|A(x_i)| : 1 \leq i \leq n \text{ and } x_i = x\},
\]
or one more than the number of distinct symbols before \( x \)’s first appearance in the sequence. We then define the pattern of \( \bar{x} \) as the concatenation of all indexes or
\[
\Psi(\bar{x}) \equiv l_x(\bar{x})_1 l_x(\bar{x})_2 \ldots l_x(\bar{x})_n.
\]

For example the pattern of the sequence “abracadabra” is
\[
\Psi(\text{abracadabra}) = 12314151231.
\]

Now any sequence can be compressed by first conveying the pattern and then separately sending the alphabet-dictionary. Note that the pattern here is independent of the alphabet size and this allows us to derive bounds on its redundancy independent of the underlying alphabet size \( m \).

B. Pattern Redundancy

Let \( \Psi^n \) denote the set of all length-\( n \) patterns and \( \Psi^* \) be the set of all patterns. Let \( A^* \) denote the collection of all finite length sequences. Then, every probability distribution \( p \) over \( A^* \) induces a distribution \( p_\Psi \) over patterns in \( \Psi^* \), where
\[
p_\Psi(\tilde{\psi}) \equiv p(\{\bar{x} \in A^* : \Psi(\bar{x}) = \tilde{\psi}\})
\]
is the probability that a string generated according to \( p \) has pattern \( \tilde{\psi} \). For a collection \( \mathcal{P} \) of distributions over \( A^* \) let
\[
\mathcal{P}_\Psi \equiv \{p_\Psi : p \in \mathcal{P}\}
\]
denote the collection of distributions over \( \Psi^* \) induced by probability distributions in \( \mathcal{P} \). The worst-case redundancy of patterns generated according to an unknown distribution in \( \mathcal{P}_\Psi \) is then
\[
\hat{R}(\mathcal{P}_\Psi) \equiv \inf_{q} \sup_{p \in \mathcal{P}} \sup_{\tilde{\psi}} \log \frac{p(\tilde{\psi})}{q(\tilde{\psi})}.
\]

While the redundancy of any class can be characterized via the Shtarkov’s sum as \( \hat{R}(\mathcal{P}) = \log (\sum_{x \in X} \sup_{p \in \mathcal{P}} p(x)) \), this expression is hard to evaluate for the class \( \mathcal{P}_\Psi \) that we consider here. Thus towards proving bounds on this redundancy, it’s instructive to define the notion of ‘profiles’.

C. Profiles

For a pattern \( \tilde{\psi} \), the multiplicity of \( \psi \in \mathbb{Z}^+ \) in \( \tilde{\psi} \) is
\[
\mu_\psi \equiv \mu_\psi(\tilde{\psi}) \equiv \{|1 \leq i \leq |\tilde{\psi}| : \psi_i = \psi\},
\]
the number of times \( \psi \) appears in \( \tilde{\psi} \). The prevalence of a multiplicity \( \mu \in \mathbb{N} \) in \( \psi \) is
\[
\varphi_\mu \equiv \varphi_\mu(\tilde{\psi}) \equiv \{|\psi : \mu_\psi = \mu\},
\]
the number of symbols appearing \( \mu \) times in \( \tilde{\psi} \). Finally, the profile of \( \tilde{\psi} \) is then
\[
\varphi_\tilde{\psi} \equiv \varphi_\tilde{\psi}(\psi) \equiv (\varphi_{|\tilde{\psi}|}, \ldots, \varphi_1),
\]
the vector of prevalences of \( \mu \) in \( \tilde{\psi} \) for \( 1 \leq \mu \leq |\tilde{\psi}| \). For instance, if \( \tilde{\psi} = 12131, \mu_1 = 3, \mu_2 = \mu_3 = 1 \) and \( \varphi = (0, 0, 1, 0, 2) \).
Let the number of patterns of a given profile \( \hat{\psi} \) be denoted by \( N(\hat{\psi}) \). It can be shown through standard counting arguments that
\[
N(\hat{\psi}) = \frac{n!}{\prod_{\mu=0}^{n} (\mu!)^{p_{\mu} \cdot \hat{\psi}^{\mu}}. \tag{3}
\]

We now present an upper bound on pattern redundancy as shown in [8].

**D. Upper Bound**

We say that \( \hat{\mathcal{P}} \) dominates \( \mathcal{P} \) if \( \forall x \in \mathcal{X}, \sup_{p \in \mathcal{P}} p(x) \leq \sup_{p \in \hat{\mathcal{P}}} p(x) \). Then, it’s easy to see that \( \hat{R}(\mathcal{P}) \leq \hat{R}(\hat{\mathcal{P}}) \). We further have the following Lemma from [8].

**Lemma 1.** For any class \( \mathcal{P} \), \( \hat{R}(\mathcal{P}) \leq \log(|\mathcal{P}|) \) where |\( \mathcal{P} \)| denotes the number of elements in \( \mathcal{P} \).

For a class \( \hat{\mathcal{P}} \) which dominates \( \mathcal{P} \), the above lemma implies \( \hat{R}(\hat{\mathcal{P}}) \leq \hat{R}(\mathcal{P}) \leq \log(|\mathcal{P}|) \). Thus, a small dominating set \( \hat{\mathcal{P}} \) allows us to upper bound \( \hat{R}(\mathcal{P}) \). Towards this, we have the following lemma.

**Lemma 2.** If \( \mathcal{P} = \mathcal{I}_{n} \) (in which case we denote \( \mathcal{P}_{\psi} = \mathcal{I}_{n}^{\psi} \)) and \( N(\hat{\psi}) \) is the number of patterns of a given profile \( \hat{\psi} \), for all patterns in the profile, \( p_{\psi}(\hat{\psi}) \leq \frac{1}{N(\hat{\psi})} \).

The lemma follows because if the underlying distribution is i.i.d., all patterns \( \hat{\psi} \) of a given profile have the same induced probability \( p_{\psi}(\hat{\psi}) \Rightarrow p_{\hat{\psi}}(\hat{\psi}) \leq \frac{1}{N(\hat{\psi})} \).

Now consider \( \hat{\mathcal{P}} = \{p_{\hat{\psi}}\} \), generated by distributions which give probability \( \frac{1}{N(\hat{\psi})} \) to patterns in \( \hat{\psi} \) and probability 0 to other patterns. Clearly \( \hat{\mathcal{P}} \) dominates \( \mathcal{I}_{n}^{\psi} \) \( \Rightarrow \hat{R}(\mathcal{I}_{n}^{\psi}) \leq \log(|\mathcal{P}|) = \log(|\#n\text{-length profiles}|) \).

Further, it can be easily shown that the number of \( n \)-length profiles is the same as the number of \( n \)-partitions where an \( n \)-partition of a positive integer \( n \) is a multiset of positive integers whose sum is \( n \). In the following lemma from [5], the number of such partitions is bounded.

**Lemma 3.** The number of \( n \)-partitions of a positive integer \( n \) is \( \leq \exp(\pi \sqrt{\frac{2}{3}} \sqrt{n}) \).

Lemma [3] thus implies the following upper bound on the pattern redundancy of \( \mathcal{I}_{n}^{\psi} \).

**Theorem 3.** For all \( n \),
\[
\hat{R}(\mathcal{I}_{n}^{\psi}) \leq \left( \pi \sqrt{\frac{2}{3}} \log(e) \right) \sqrt{n}.
\]

**E. Lower Bound**

From [8], we also have the following lower bound on \( \hat{R}(\mathcal{I}_{n}^{\psi}) \).

**Theorem 4.** As \( n \) increases,
\[
\hat{R}(\mathcal{I}_{n}^{\psi}) \geq \log \left( \frac{e^{23/12}}{\sqrt{2\pi}} \right) n^{3/2} (1 + o(1)).
\]

This bound is obtained via simplification of the Shtarkov sum by using Equation 3. We skip details of the proof here.

**IV. Shape Compression**

In this section, we introduce an alternate notion of shape compression and provide bounds on its redundancy.

**A. Definitions**

Let \( \mathcal{A} \) be a possibly infinite alphabet with an order “\(<\)”. Also recall that the symbols appearing in a sequence \( \bar{x} \), is denoted by \( \mathcal{A}(\bar{x}) \). Then the rank \( \rho_{\bar{x}}(x) \) of \( x \in \mathcal{A}(\bar{x}) \) is the number
\[
\rho_{\bar{x}}(x) \overset{\text{def}}{=} |\{y \in \mathcal{A}(\bar{x}) : y \leq x\}|
\]

of distinct symbols in \( \mathcal{A}(\bar{x}) \) not larger than \( x \). The shape of \( \bar{x} \) is defined as
\[
S(\bar{x}) \overset{\text{def}}{=} \rho_{\bar{x}}(x_1)\rho_{\bar{x}}(x_2)\ldots\rho_{\bar{x}}(x_n)
\]
the concatenation of all ranks. For example, consider the Roman alphabet with standard order \( a < b < c < \ldots < z \) and the string \( \bar{x} = \text{"abracadabra"} \). Then \( \rho_{\bar{x}}(a) = 1, \rho_{\bar{x}}(b) = 2, \rho_{\bar{x}}(c) = 3, \rho_{\bar{x}}(d) = 4, \rho_{\bar{x}}(r) = 5 \), and
\[
S(\text{"abracadabra"}) = 125131412512341.
\]

**B. Shape Redundancy**

Let \( S^n \) denote the set of all length-\( n \) shapes and \( S^* \) be the set of all shapes. Then every probability distribution \( p_{S} \) over \( \mathcal{A}^* \) induces a distribution \( p_{S} \) over shapes in \( S^* \), where
\[
p_{S}(\bar{s}) \overset{\text{def}}{=} p(\{\bar{x} \in \mathcal{A}^* : S(\bar{x}) = \bar{s}\})
\]
is the probability that a sequence generated according to \( p \) has shape \( \bar{s} \). For a collection \( \mathcal{P} \) of distributions over \( \mathcal{A}^* \) let
\[
\mathcal{P}_{S} \overset{\text{def}}{=} \{p_{S} : p \in \mathcal{P}\}
\]
denote the collection of distributions over \( S^* \) induced by probability distributions in \( \mathcal{P} \). Similar to pattern redundancy, the worst-case redundancy of shapes generated according to an unknown distribution in \( \mathcal{P}_{S} \) is
\[
\hat{R}(\mathcal{P}_{S}) \overset{\text{def}}{=} \inf_{q} \sup_{\bar{s} \in S^*} \sup_{p \in \mathcal{P}} \log \frac{p(\bar{s})}{q(\bar{s})}.
\]
To help in proving shape redundancy bounds, we introduce the notion of ‘types’ analogous to profiles for patterns as in section III-C.
C. Types

Let $S_m^n$ be the set of length-$n$ shapes with $m$ symbols. The multiplicity of $i$ in $\bar{s} \in S_m^n$ is $$\mu_i \overset{\text{def}}{=} \mu_i(\bar{s}) \overset{\text{def}}{=} |\{1 \leq j \leq n : s_j = i\}|$$ the number of times $i$ occurs in $\bar{s}$. The type of $\bar{s} \in S_m^n$ is $$\bar{u}(\bar{s}) \overset{\text{def}}{=} (\mu_1, \ldots, \mu_m)$$ or the $m$-tuple of multiplicities of the symbols in $\bar{s}$.

It’s easy to see that the type of any length-$n$ shape corresponds to an ordered partition of $n$, where the number of partition parts equals the number of shape symbols. This implies that the number of types of length-$n$ shapes with $m$ symbols is the number of such partitions, which can be shown to be $\binom{n-1}{m-1}$. Summing over all values of $m \leq n$, the total number of types of length-$n$ shapes is then $2^{n-1}$. Also, a standard counting argument shows that for a collection of shapes of type $\bar{u}$, denoted by $S_\bar{u}$, we have that $$|S_\bar{u}| = \binom{n}{u_1, \ldots, u_m} = \frac{n!}{u_1! \cdots u_m!}.$$ 

We now present an upper bound on shape redundancy as shown in [7].

D. Upper Bound

As before, every i.i.d. class $\mathcal{P}$ induces the same probability on all shapes in $S_\bar{u}$. Therefore the probability of every shape in $S_\bar{u}$ is at most $1/|S_\bar{u}|$. Now let $\bar{p}_\bar{u}$ be the uniform distribution on shapes of a given type $\bar{u}$ so that every shape in $S_\bar{u}$ has probability $1/|S_\bar{u}|$, and let $$I^n_S = \{\bar{p}_\bar{u} : \bar{u} \in \mathcal{U}^n\}$$ be the collection of all such uniform distributions. Recall from Lemma 1, that for a class $\mathcal{P}$ which dominates $\mathcal{P}$, $\hat{R}(\mathcal{P}) \leq \log(|\mathcal{P}|)$. In this case, since $I^n_S$ dominates $I^n_S$, the below theorem follows.

Theorem 5. For all $n$, $$\hat{R}(I^n_S) \leq n - 1.$$ 

E. Lower Bound

The shape per symbol redundancy lower bound, as shown in [7], can be proven by lower-bounding the maximum shape probabilities and incorporating them into Shtarkov’s sum. The following theorem results.

Theorem 6. For all $n > 1$ $$\hat{R}(I^n_S) \geq \frac{n}{4} \cdot \log \frac{4}{\pi e^{1/6}} > 0.027n$$

V. Conclusion

We summarize our report as follows. We first saw that the redundancy of i.i.d. distributions increases to infinity with alphabet size. Unlike fixed and small alphabets where per-symbol redundancy diminishes with increasing block length, alphabets with a size proportional to the block length entail constant per-symbol redundancy and alphabets with size much larger than the block length have per-symbol redundancy trending toward infinity. 

Therefore we investigated the redundancy of patterns and shapes instead. Again, looking at i.i.d. distributions, the shape redundancy $\hat{R}(I^n_S)$ is bounded by $$\frac{n}{4} \cdot \log \frac{4}{\pi e^{1/6}} + O(1) \leq \hat{R}(I^n_S) \leq n.$$ 

So between 0.027 and 1 additional bits per symbol are required to convey shapes. Furthermore, pattern redundancy $\hat{R}(I^n_P)$ is bounded by $$\left(\frac{3}{2} \log(e)\right) n^{1/2 + O(1)} \leq \hat{R}(I^n_P) \leq \left(\sqrt{\frac{3}{4}} \log(e)\right) \sqrt{n}.$$ 

So that the per-symbol redundancy of encoding patterns diminishes to zero as the block length increases, regardless of the alphabet size.

References