



# LGP Lumieres LGPL

## Problem 10 Play on words

1\*  
OSOUDIOT  
ZOO\*

# PROBLEM 10

## PLAY ON WORDS

### SUMMARY

In this problem, we introduce the effectiveness  $\text{Eff}(S) = \lim_{n \rightarrow \infty} |S^{\leq n}|^{1/n}$  and use submultiplicativity and Fekete's lemma to prove it exists with the exact bounds  $1 \leq \text{Eff}(S) \leq |S|$ , then construct codes of size  $k$  attaining both extremes. For any word  $v$ , we compute the minimal concatenation length  $\ell_S(v)$  by reducing to a shortest-path problem in a directed acyclic graph on the positions of  $v$ . Equipping code sets with a natural prefix metric, we show that convergence in this metric forces  $\text{Eff}(S_n) - \text{Eff}(T_n) \rightarrow 0$ , establishing continuity.

Question	Result
1	Solved
2	Solved
3	Solved
4	Solved
5	Solved
6	Solved
7	Solved

We then define the density  $D(S) = \lim_{n \rightarrow \infty} \frac{|S^{\leq n}|}{|S|^n}$ , prove it exists and equals  $\frac{|S|}{|S|-1}$  exactly for uniquely decodable codes (and 0 otherwise for sets of cardinality greater than 2), and use this to study the normalized counts  $\omega_{\ell,k}$ , showing  $\omega_0 = \infty$  and  $\omega_\ell = 0$  for  $\ell \in (0, 1]$ . Finally, we extend all results to arbitrary finite alphabets and incorporate length constraints via antichains in the prefix poset.

### CONTENTS

Summary	1
Contents	1
1 Bounds for the Effectiveness of $S$	3
2 Maximizing and Minimizing $S$	5
3 Computation of $\ell_S(v)$	7
4 Distances Between Sets	10
5 Density of a Set $S$	12
a) Well Definedness of The Density . . . . .	12
b) Limit Behavior of Effectiveness Differences Assuming $\mathcal{D}(S_n) = \mathcal{D}(T_n) = 1$ . . . . .	16
c) Limit Behavior of Effectiveness Differences Assuming $\mathcal{D}(S_n) = \mathcal{D}(T_n) = \ell$ for $\ell \in [0, 1)$ . . . . .	16
6 $\ell$ -sparse Subsets $S$	17
a) Existence of $\omega_\ell$ . . . . .	17
b) An example of a 0-Sparse Subset . . . . .	18
c) $\omega_0$ for $k \geq 2$ . . . . .	18
d) Estimation of $\omega_\ell$ for $\ell \in (0, 1]$ . . . . .	19
7 Further Extensions and Generalisations	20
a) Generalisation to an arbitrary alphabet . . . . .	21
b) Reinterpretation of $\omega_{0,k}$ with a Length Constraint . . . . .	22

c)	Cost Weighted Sets . . . . .	25
i)	Weighted Effectiveness . . . . .	25
ii)	Limit Behavior of Weighted Effectiveness Under a Cost Weighted Distance Metric	27

**Problem Statement:** Misha has the set  $\{0, 1\}^*$  of all finite words over the alphabet  $\{0, 1\}$ , meaning the set  $\{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \dots\}$  with  $\varepsilon$  being the empty word. Misha notices the following, sometimes if you have a finite subset  $S \subset \{0, 1\}^*$ , you can represent some elements of  $\{0, 1\}^*$  as concatenation of elements of  $S$ . We say that the set of all concatenations of words of  $S$  is generated by  $S$  and we denote it by  $S^*$ . For example if  $S = \{00, 011\}$  we can concatenate the words of  $S$  to obtain the word  $01100011 \in S^*$ .

## 1 Bounds for the Effectiveness of $S$

**Question 1:** When  $S = \{0, 1\}$  we can represent all elements of  $\{0, 1\}^*$ . Since it is easy to come up with these sets, Misha wants to understand a characteristic of this set. Define the effectiveness of  $S$  as:

$$\text{Eff}(S) = \lim_{n \rightarrow \infty} |S^{\leq n}|^{\frac{1}{n}}$$

Where given a finite set  $A$ ,  $|A|$  is its cardinal.  $S^{\leq n}$  denotes the set of words that are concatenation of at most  $n$  elements of  $S$ , that is to say:

$$S^{\leq n} = \{w \mid \exists u_1, \dots, u_k \in S, k \leq n \text{ such that } w = u_1 u_2 \dots u_k\}$$

Show that for every  $S$ ,  $\text{Eff}(S)$  is well defined. Is there an upper bound on  $\text{Eff}(S)$ ? Find a lower bound for  $\text{Eff}(S)$ ?

**Theorem 1** Let  $a_n = |S^{\leq n}|$ . Then the limit

$$\text{Eff}(S) = \lim_{n \rightarrow \infty} a_n^{1/n}$$

exists. If  $k = |S|$ , one has the sharp inequalities

$$1 \leq \text{Eff}(S) \leq k.$$

*Proof.* We first show that the numbers  $(a_n)$  satisfy submultiplicativity. Given any two nonnegative integers  $m, n$ , every pair of words  $(x, y)$  with  $x \in S^{\leq m}$  and  $y \in S^{\leq n}$  yields by concatenation a word  $xy$  which lies in  $S^{\leq(m+n)}$ . Thus the image of the map

$$S^{\leq m} \times S^{\leq n} \longrightarrow S^{\leq(m+n)}, \quad (x, y) \mapsto xy,$$

is contained in  $S^{\leq(m+n)}$ . Since the size of the image cannot exceed the size of its domain in a non-injective function, we obtain

$$a_{m+n} = |S^{\leq(m+n)}| \leq |S^{\leq m} \times S^{\leq n}| = a_m a_n,$$

We now introduce a key lemma.

**Lemma 1** Let  $(a_n)_{n \geq 0}$  be a sequence of positive real numbers satisfying

$$a_{m+n} \leq a_m a_n \quad \text{for all integers } m, n \geq 0.$$

Then the limit

$$L = \lim_{n \rightarrow \infty} a_n^{1/n}$$

exists (finite or zero) and in fact

$$L = \inf_{n \geq 1} a_n^{1/n}.$$

*Proof.* Because each  $a_n > 0$ , we set  $b_n = \ln a_n$ , which makes  $b_n$  a well-defined real number. The inequality  $a_{m+n} \leq a_m a_n$  becomes

$$b_{m+n} = \ln a_{m+n} \leq \ln(a_m a_n) = b_m + b_n,$$

so  $(b_n)$  is a subadditive sequence. We invoke Fekete's Subadditive lemma:

*If a sequence  $(b_n)_{n \geq 1}$  of real numbers satisfies  $b_{n+m} \leq b_n + b_m$  for all  $n, m \geq 1$ , then the limit  $\lim_{n \rightarrow \infty} b_n/n$  exists and equals  $\inf_{n \geq 1} (b_n/n)$ .*

Exponentiating both sides gives

$$\lim_{n \rightarrow \infty} a_n^{1/n} = \exp\left(\lim_{n \rightarrow \infty} \frac{b_n}{n}\right) = \exp\left(\inf_{n \geq 1} \frac{b_n}{n}\right) = \inf_{n \geq 1} a_n^{1/n}.$$

Hence the limit  $L$  exists and equals the stated infimum. □

Thus, since  $(a_n)$  satisfies the lemma, the limit  $\lim a_n^{1/n}$  exists and defines  $\text{Eff}(S)$ .

Next, to see  $\text{Eff}(S) \geq 1$ , note that the empty word  $\varepsilon$  belongs to  $S^{\leq n}$  for every  $n$ , so  $a_n \geq 1$  and hence  $a_n^{1/n} \geq 1$ . Taking  $n \rightarrow \infty$  yields  $\text{Eff}(S) \geq 1$ .

Finally, to establish the upper bound, observe that in the case where the concatenation map is injective, concatenating exactly  $j$  words from  $S$  corresponds to choosing an ordered  $j$ -tuple in  $S^j$ . There are  $|S|^j = k^j$  such tuples and therefore at most  $k^j$  distinct results of length  $j$ . Summing over  $j = 0, \dots, n$  shows

$$a_n \leq \sum_{j=0}^n k^j = \frac{k^{n+1} - 1}{k - 1},$$

and taking the  $n$ th root gives

$$a_n^{1/n} \leq \left(\frac{k^{n+1} - 1}{k - 1}\right)^{1/n} \longrightarrow k.$$

Thus  $\text{Eff}(S) \leq k$ , completing the proof. □

## 2 Maximizing and Minimizing $S$

**Question 2:** Let  $k \geq 2$  be an integer. Out of every subset  $S$  satisfying  $|S| = k$ , maximise (and minimise)  $\text{Eff}(S)$ .

**Theorem 2** We show that

$$\min_{|S|=k} \text{Eff}(S) = 1 \quad \text{and} \quad \max_{|S|=k} \text{Eff}(S) = k,$$

and moreover for each  $k \geq 2$  there exist specific sets  $S$  of size  $k$  attaining these extremal values.

*Proof.* First we show that for every  $S$  with  $|S| = k$  one has  $\text{Eff}(S) \geq 1$ . Since  $S$  is nonempty and finite, it contains at least one nonempty word  $w$ . Then for each positive integer  $n$  the words

$$w, w^2, \dots, w^n$$

are distinct and each lies in  $S^{\leq n}$ , so

$$|S^{\leq n}| \geq n.$$

Taking  $n$ th roots gives  $|S^{\leq n}|^{1/n} \geq n^{1/n}$ , and since  $\lim_{n \rightarrow \infty} n^{1/n} = 1$ , it follows that  $\text{Eff}(S) \geq 1$ . To see that this lower bound is sharp, choose an arbitrary nonempty word  $w$  and let

$$S = \{\varepsilon, w, w^2, \dots, w^{k-1}\}.$$

Then any concatenation of up to  $n$  elements of  $S$  must be one of

$$\varepsilon, w, w^2, \dots, w^{(k-1)n},$$

so  $|S^{\leq n}| = (k-1)n + 1$  and hence  $\text{Eff}(S) = \lim((k-1)n + 1)^{1/n} = 1$ .

From Theorem 1, we note that the universal upper bound is  $k$ . We now show that this bound is also attained. Choose an integer  $\ell$  with  $2^\ell \geq k$ , and select any  $k$  distinct binary strings of length  $\ell$ , say  $u_1, \dots, u_k$ . Because all  $u_i$  have the same length, none can be a prefix of another, so the map

$$S^n \rightarrow S^{\leq n}, \quad (u_{i_1}, \dots, u_{i_n}) \mapsto u_{i_1} u_{i_2} \cdots u_{i_n}$$

is injective. We define this as a prefix free subset. A prefix free subset is also Uniquely Decodable, a word where any sequence of words can be decoded into a single, unambiguous concatenation of letters from  $S$ . Hence

$$|S^{\leq n}| \geq |S^n| = k^n,$$

and combined with the bound  $|S^{\leq n}| \leq \sum_{j=0}^n k^j$  we conclude

$$\lim_{n \rightarrow \infty} |S^{\leq n}|^{1/n} = \lim_{n \rightarrow \infty} k^{(\frac{1-k^{-n-1}}{k-1})^{1/n}} = k.$$

Thus that choice of  $S$  of size  $k$  indeed achieves  $\text{Eff}(S) = k$ , completing the proof.  $\square$

**Theorem 3** Let  $S \subset \{0, 1\}^*$  be a finite, non-uniquely-decodable code with  $|S| = k$ . Define

$$b_n = |S^{\leq n}|, \quad \text{Eff}(S) = \lim_{n \rightarrow \infty} b_n^{1/n}.$$

Then  $\text{Eff}(S) < k$ .

*Proof.* Write  $a_n = |S^n|$ . Because concatenating an  $m$ -block word with an  $n$ -block word surjects onto all  $(m+n)$ -block words, one has

$$a_{m+n} \leq a_m a_n \quad \text{for all } m, n \geq 0.$$

By Lemma 1 the sequence  $(a_n)$  satisfies

$$L := \lim_{n \rightarrow \infty} a_n^{1/n} = \inf_{n \geq 1} a_n^{1/n},$$

so in particular  $L$  exists and is finite. We later also show that  $\lim b_n^{1/n} = L$ , hence  $\text{Eff}(S) = L$ . We will prove  $L < k$ .

Since  $S$  fails unique-decodability, there exist two distinct sequences of blocks

$$(u_1, \dots, u_r) \neq (v_1, \dots, v_s) \quad \text{with} \quad u_1 \cdots u_r = v_1 \cdots v_s =: w.$$

Setting  $N = r + s$ , consider the two length- $N$  block-words

$$(u_1, \dots, u_r, v_1, \dots, v_s) \quad \text{and} \quad (v_1, \dots, v_s, u_1, \dots, u_r).$$

Both concatenate to  $w w$ , so among the  $k^N$  possible sequences of length  $N$  there must be a collision. Thus

$$a_N = |S^N| \leq k^N - 1,$$

and taking the  $N$ th root gives

$$a_N^{1/N} < k.$$

Because  $L = \inf_{n \geq 1} a_n^{1/n}$ , we conclude  $L < k$ .

It remains only to justify that  $\lim_{n \rightarrow \infty} b_n^{1/n} = L$ . Since every term of the sum defining  $b_n$  is positive, we immediately get

$$b_n = \sum_{i=0}^n a_i \geq a_n \implies b_n^{1/n} \geq a_n^{1/n} \xrightarrow{n \rightarrow \infty} L.$$

Thus  $\liminf_{n \rightarrow \infty} b_n^{1/n} \geq L$ .

On the other hand, let  $M_n = \max_{0 \leq i \leq n} a_i$ . Then each  $a_i \leq M_n$ , so

$$b_n = \sum_{i=0}^n a_i \leq (n+1) M_n.$$

Since  $L > 1$ , choose  $\varepsilon > 0$  with  $L - \varepsilon > 1$ . By definition of the limit, there exists  $N$  such that for all  $n \geq N$ ,

$$a_n^{1/n} > L - \varepsilon > 1 \implies a_n > (L - \varepsilon)^n.$$

Let  $C = \max_{0 \leq i < N} a_i$ , which is finite. Then for all  $n \geq N$ ,

$$a_n > (L - \varepsilon)^n,$$

and since  $(L - \varepsilon)^n \rightarrow \infty$ , there is  $N' \geq N$  so large that for all  $n \geq N'$ ,

$$(L - \varepsilon)^n > C \implies a_n > C \geq a_i \quad \forall 0 \leq i < N.$$

But for  $N \leq i < n$ ,  $a_i \leq a_i^{1/i \cdot i} = (a_i^{1/i})^i$ , and since  $a_i^{1/i} \rightarrow L$ , all these finitely many  $a_i$  are eventually bounded by some constant  $C'$ , which is again exceeded by  $(L - \varepsilon)^n$  for large  $n$ . Hence for all sufficiently large  $n$ ,

$$a_n > \max_{0 \leq i < n} a_i \implies M_n = a_n,$$

Hence for all large  $n$  we have  $M_n = a_n$ , giving

$$b_n \leq (n+1) a_n \implies b_n^{1/n} \leq (n+1)^{1/n} a_n^{1/n}.$$

But  $(n+1)^{1/n} \rightarrow 1$  and  $a_n^{1/n} \rightarrow L$ , so  $\limsup_{n \rightarrow \infty} b_n^{1/n} \leq L$ .

Combining  $\liminf \geq L$  and  $\limsup \leq L$  forces  $\lim b_n^{1/n} = L$ , as required.

This completes the proof. □

### 3 Computation of $\ell_S(v)$

**Question 3:** Given a word  $v$ , we define  $\ell_S(v)$  by:

$$\ell_S(v) = \min \{n \mid \exists u_1, \dots, u_n \in S, \text{ such that } v = u_1 u_2 \dots u_k\}$$

Let  $S$  be a fixed subset. Describe a way to compute  $\ell_S(v)$  for any  $v \in S^*$ .

**Theorem 4** For a given word  $v$  of length  $L$ , define a directed acyclic graph  $G = (V, E)$  by setting

$$V = \{0, 1, 2, \dots, L\},$$

and for any integers  $i, j$  with  $0 \leq i < j \leq L$ , include the edge  $(i, j)$  in  $E$  if and only if the substring

$$v[i+1:j] = v[i+1]v[i+2]\cdots v[j]$$

belongs to  $S$ . Then the number  $\ell_S(v)$  is equal to the length (i.e. the number of edges) of a shortest path from vertex 0 to vertex  $L$  in  $G$ .

*Proof.* We prove that the length of the shortest path from 0 to  $L$  in the graph  $G = (V, E)$  defined above is exactly  $\ell_S(v)$ .

Let  $v$  be a word in  $S^*$  with length  $L$ . By definition, since  $v \in S^*$ , there exists at least one decomposition of  $v$  as a concatenation of words from  $S$ ; that is, there exists some positive integer  $n$  and words  $u_1, u_2, \dots, u_n \in S$  such that

$$v = u_1 u_2 \cdots u_n.$$

Define indices  $i_0, i_1, i_2, \dots, i_n$  by setting  $i_0 = 0$  and letting  $i_k = |u_1 u_2 \cdots u_k|$  for  $1 \leq k \leq n$ . In particular, we have  $i_n = L$ . By the very definition of concatenation, for each  $k = 1, 2, \dots, n$  the substring of  $v$  given by

$$v[i_{k-1} + 1 : i_k]$$

is precisely the word  $u_k$ , and by assumption  $u_k \in S$ . Therefore, for every  $k = 1, 2, \dots, n$  there is an edge from  $i_{k-1}$  to  $i_k$  in  $G$ , because the condition  $v[i_{k-1} + 1 : i_k] \in S$  is satisfied. This shows that the sequence of vertices

$$0 = i_0, i_1, i_2, \dots, i_n = L$$

forms a directed path in  $G$  from 0 to  $L$  that uses exactly  $n$  edges. Since  $n$  is an arbitrary number for which such a decomposition exists, the minimal number of words in any decomposition of  $v$  is exactly the minimum number of edges in a path from 0 to  $L$  in  $G$ . In other words,

$$\ell_S(v) = \min\{n \mid \exists 0 = i_0 < i_1 < \cdots < i_n = L \text{ with } v[i_{k-1} + 1 : i_k] \in S \text{ for each } k = 1, \dots, n\}.$$

Conversely, suppose that there exists a path in  $G$  from 0 to  $L$  given by vertices

$$0 = i_0, i_1, i_2, \dots, i_n = L,$$

with each edge  $(i_{k-1}, i_k)$  satisfying  $v[i_{k-1} + 1 : i_k] \in S$ . Then by concatenating these substrings we obtain

$$v = v[1 : i_1] v[i_1 + 1 : i_2] \cdots v[i_{n-1} + 1 : i_n],$$

and since each factor belongs to  $S$  it follows that  $v$  has a decomposition as a concatenation of  $n$  words from  $S$ . Thus,  $\ell_S(v) \leq n$ . Taking the minimum over all such paths shows that  $\ell_S(v)$  is exactly the length (in number of edges) of a shortest path from 0 to  $L$ .

It remains to note that the graph  $G$  is acyclic. This is because the vertex set is  $\{0, 1, \dots, L\}$  and by construction every edge  $(i, j)$  satisfies  $i < j$ ; hence, no cycle can occur. In a directed acyclic graph,

the problem of finding a shortest path is well-defined and can be solved using, for example, a simple recursive method or by processing the vertices in increasing order.

Therefore, the mathematical procedure to compute  $\ell_S(v)$  is as follows:

- (1) Construct the graph  $G$  with vertices  $\{0, 1, \dots, L\}$  and add an edge from  $i$  to  $j$  if  $v[i+1:j] \in S$ .
- (2) Compute the length of a shortest path from 0 to  $L$  in  $G$ ; this length is precisely  $\ell_S(v)$ .

It might seem at first that because the vertices are labeled  $0, 1, 2, \dots, L$  (where  $L$  is the length of  $v$ ), the shortest path from 0 to  $L$  would always consist of  $L$  edges (i.e. each edge corresponding to a single letter). However, this is not the case because we include an edge from vertex  $i$  to vertex  $j$  not only when  $j = i + 1$ , but whenever the substring

$$v[i+1:j]$$

belongs to  $S$ .

Thus, if  $S$  contains words longer than one letter, then there will be edges that “skip” intermediate vertices. For instance, if  $v = 010101$  and  $S$  contains the word 01, then there is an edge from 0 to 2 (since  $v[1:2] = 01$ ) and so on. In fact, if  $S$  contains a word that is equal to  $v$  itself, then there is an edge from 0 directly to  $L$ , and the shortest path would consist of just one edge. Hence, the length of the shortest path in the graph (i.e. the number of edges in the path) is exactly the minimum number of concatenated words from  $S$  needed to form  $v$ , and it is not necessarily  $L$ .  $\square$

**Example.** Let

$$S = \{0, 10, 110, 011\}, \quad v = 0110110, \quad |v| = 7.$$

We construct a directed acyclic graph  $G = (V, E)$  whose vertices

$$V = \{0, 1, 2, 3, 4, 5, 6, 7\}$$

mark the cut-points before each symbol of  $v$ . We include  $(i \rightarrow j) \in E$  precisely when the substring  $v[i+1:j] \in S$ . A quick check yields Figure 1.

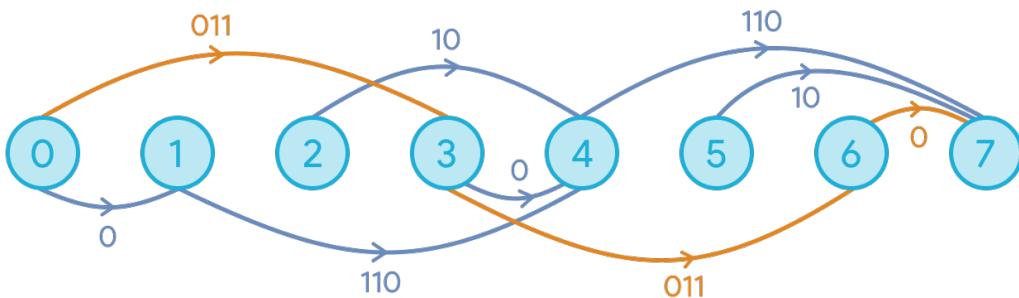


Figure 1: Directed acyclic graph for the code  $S = \{0, 10, 110, 011\}$  and word  $v = 0110110$ , with vertices  $0, 1, \dots, 7$  marking the cut-points in  $v$  and each edge  $(i \rightarrow j)$  labeled by the substring  $v[i+1:j] \in S$ . The length of the shortest path from vertex 0 to vertex 7 (highlighted) equals  $\ell_S(v) = 3$ .

Having drawn  $G$  with these eight vertices and nine labeled arcs, we see that any path from 0 to 7 must use at least three edges, for example

$$0 \xrightarrow{011} 3 \xrightarrow{011} 6 \xrightarrow{0} 7.$$

Since no path of length two exists, we conclude

$$\ell_S(v) = \min\{\#\text{edges in a path } 0 \rightarrow v\} = 3.$$

## 4 Distances Between Sets

**Question 4:** Now we can define a notion of how similar two sets are. First, let us define the distance between two elements of  $\{0, 1\}^*$ . For  $u, v \in \{0, 1\}^*$ . The distance between the words  $u$  and  $v$  is  $d(u, v) = 2^{-|P|}$ , where  $P$  is the length of the longest word that appears at the beginning of both  $u$  and  $v$ , if  $u \neq v$ , and 0 if  $u = v$ . For example, let  $u = 001$  and  $v = 000$ , then  $P = 2$  and  $d(u, v) = 1/4$ . For two finite sets  $S_1$  and  $S_2$ , we define:

$$d(S_1, S_2) = \max \left\{ \max_{x \in S_1} \min_{y \in S_2} d(x, y), \max_{x \in S_2} \min_{y \in S_1} d(x, y) \right\}$$

Let  $(S_n)_{n \in \mathbb{N}}$  and  $(T_n)_{n \in \mathbb{N}}$  be two sequences of subsets of  $\{0, 1\}^*$  such that

$$\lim_{n \rightarrow \infty} d(S_n, T_n) = 0$$

What can you say about the sequence  $(\text{Eff}(S_n) - \text{Eff}(T_n))_n$ ?

**Theorem 5** Let  $\{S_n\}_{n=1}^\infty$  and  $\{T_n\}_{n=1}^\infty$  be sequences of finite subsets of  $\{0, 1\}^*$  such that

$$\lim_{n \rightarrow \infty} d(S_n, T_n) = 0.$$

Then

$$\lim_{n \rightarrow \infty} (\text{Eff}(S_n) - \text{Eff}(T_n)) = 0.$$

*Proof.* Convergence in the prefix-metric,  $d(S_n, T_n) \rightarrow 0$ , means that for each fixed  $k$  and all large  $n$ , every block in  $S_n$  has a “partner” in  $T_n$  sharing a very long prefix (and vice versa). By replacing each block in a concatenation of at most  $k$  blocks by its partner, one obtains injections

$$S_n^{\leq k} \hookrightarrow T_n^{\leq k} \quad \text{and} \quad T_n^{\leq k} \hookrightarrow S_n^{\leq k},$$

which force  $|S_n^{\leq k}| = |T_n^{\leq k}|$  for all sufficiently large  $n$ .

Every finite set of words is either *prefix-free* or *non-prefix-free*. We consider both cases.

**Case 1: Prefix-free.** If infinitely many  $S_n$  (or  $T_n$ ) are prefix-free and distinct from their partner sequence, then no matter how long a prefix one requires, a prefix-free set cannot “approximate” another distinct prefix-free set arbitrarily well: any two different prefix-free words differ in some bit, giving a lower bound  $d \geq 2^{-L_0} > 0$ . Thus the only way  $d(S_n, T_n) \rightarrow 0$  can hold is if for some  $N$ ,  $S_n = T_n$  for all  $n \geq N$ . In that eventual-equality subcase we have  $\text{Eff}(S_n) = \text{Eff}(T_n)$  for  $n \geq N$ , so

$$\text{Eff}(S_n) - \text{Eff}(T_n) = 0 \quad (n \geq N),$$

and the difference clearly tends to 0.

**Case 2: Non-prefix-free.** Suppose instead that for all sufficiently large  $n$ , both  $S_n$  and  $T_n$  are non-prefix-free (so each contains at least two blocks, one a prefix of another). We now fix an arbitrary concatenation-length bound  $k \geq 1$ . By definition,  $S_n^{\leq k}$  is the set of all words obtained by concatenating at most  $k$  blocks from  $S_n$ . Each such word has total length at most  $k \cdot \max_{u \in S_n} |u|$ ; denote this maximum block-length by  $\ell_{\max}(n)$ . Choose  $L > k \ell_{\max}(n)$ . For all sufficiently large  $n$ , prefix-metric convergence gives  $d(S_n, T_n) < 2^{-L}$ . Therefore we can assign to each block  $u \in S_n$  a “partner”  $\pi(u) \in T_n$  whose first  $L$  bits agree with those of  $u$ .

Defining  $\pi$  on concatenations by

$$\pi(u_1 u_2 \cdots u_\ell) = \pi(u_1) \pi(u_2) \cdots \pi(u_\ell) \quad (\ell \leq k),$$

we obtain a map  $\pi: S_n^{\leq k} \rightarrow T_n^{\leq k}$ . To see that  $\pi$  is injective, observe that two distinct words  $w \neq w'$  in  $S_n^{\leq k}$  must differ in at least one position within their first  $k \ell_{\max}(n)$  bits. But since each  $\pi(u_i)$  agrees with  $u_i$  on the first  $L$  bits and  $L > k \ell_{\max}(n)$ , the images  $\pi(w)$  and  $\pi(w')$  also differ in that same position. Hence  $\pi(w) \neq \pi(w')$ . The same construction, applied symmetrically, yields an injection  $T_n^{\leq k} \rightarrow S_n^{\leq k}$ . Two finite sets admitting injections both ways must have the same cardinality, so for all large  $n$ ,

$$|S_n^{\leq k}| = |T_n^{\leq k}|.$$

Finally, by Theorem 1,

$$\text{Eff}(S) = \inf_{m \geq 1} |S^{\leq m}|^{1/m},$$

equality of  $|S_n^{\leq k}|$  and  $|T_n^{\leq k}|$  for each fixed  $k$  implies  $\text{Eff}(S_n) = \text{Eff}(T_n)$  for all sufficiently large  $n$ . Thus  $\text{Eff}(S_n) - \text{Eff}(T_n) = 0$  eventually, and the sequence of differences converges to 0.  $\square$

## 5 Density of a Set $S$

**Question 5:** We define the density of a set  $S$  as:

$$\mathcal{D}(S) = \lim_{n \rightarrow \infty} \frac{|S^{\leq n}|}{|S|^n}$$

We wish to study this notion of density. We first ask if  $\mathcal{D}(S)$  is well-defined for every subset  $S$ ? Next, we let  $(S_n)_{n \in \mathbb{N}}, (T_n)_{n \in \mathbb{N}}$  be two sequences of subsets of  $\{0, 1\}^*$  such that:  $\lim_{n \rightarrow \infty} d(S_n, T_n) = 0$  and  $\mathcal{D}(S_n) = \mathcal{D}(T_n) = 1$  and ask if  $\text{Eff}(S_n) - \text{Eff}(T_n)$  has a limit? We then ask the same question for  $\mathcal{D}(S) = \ell$  for  $\ell \in [0, 1]$ .

### a) Well Definedness of The Density

**Theorem 6** Let  $a_n = |S^{\leq n}|$ . Then, the *density* is well defined whenever  $S = \{\varepsilon\}$  where  $\mathcal{D}(S) = 1$  and for  $k \geq 2$ . Moreover,

$$0 \leq \mathcal{D}(S) \leq \frac{k}{k-1}.$$

Specifically,

$$\mathcal{D}(S) = \begin{cases} \frac{|S|}{|S|-1} & \text{if } S \text{ is Prefix-Free} \\ 0 & \text{otherwise} \end{cases}$$

*Proof.* If  $k = 1$ , and  $S = \{\varepsilon\}$  then  $a_n = 1$  and  $\mathcal{D}(S) = 1$ ; if  $S = \{w \neq \varepsilon\}$  then  $a_n = n + 1$  and the ratio  $\frac{a_n}{1^n} = n + 1$  diverges, so  $\mathcal{D}(S)$  fails to exist finitely. Hence we restrict to  $k \geq 2$ .

First, we show the limit exists. From the combinatorial definition, any concatenation of at most  $m+n$  blocks from  $S$  can be formed by first concatenating at most  $m$  blocks and then at most  $n$  blocks. Thus

$$S^{\leq m} \cdot S^{\leq n} \subseteq S^{\leq(m+n)},$$

so

$$a_{m+n} \leq a_m a_n.$$

Thus, the sequence  $a_n$  is submultiplicative. Define  $b_n = a_n/k^n$ , where  $k = |S|$ . Then

$$b_{m+n} = \frac{a_{m+n}}{k^{m+n}} \leq \frac{a_m a_n}{k^m k^n} = b_m b_n,$$

so the sequence  $(b_n)$  is also submultiplicative.

Moreover, since  $a_n$  is bounded above and non-negative, it follows that  $(b_n)$  is a bounded, nonnegative, submultiplicative sequence.

We now introduce a key lemma.

**Lemma 2 (Limit of a Bounded Submultiplicative Sequence):** Let  $(b_n)_{n \geq 0}$  be a sequence of nonnegative real numbers satisfying

$$b_{m+n} \leq b_m b_n \quad \text{for all integers } m, n \geq 0.$$

Suppose moreover that there exists a constant  $M > 0$  such that  $b_n \leq M$  for every  $n$ . Then the limit

$$\lim_{n \rightarrow \infty} b_n$$

exists and is finite.

*Proof.* First, because each  $b_n \geq 0$ , the sequence is bounded below by 0. By hypothesis it is also bounded above by  $M$ . For each integer  $n \geq 0$ , define

$$c_n = \sup\{b_k : k \geq n\}.$$

Since for every  $n$  the set  $\{b_k : k \geq n+1\}$  is contained in  $\{b_k : k \geq n\}$ , we have  $c_{n+1} \leq c_n$ . Thus the sequence  $(c_n)$  is nonincreasing. Moreover, each  $c_n$  is at least 0, so  $(c_n)$  is bounded below. A nonincreasing sequence that is bounded below converges in the real numbers. Therefore there is a real number

$$L = \lim_{n \rightarrow \infty} c_n = \inf_{n \geq 0} c_n.$$

By definition of the supremum, for every  $n$  we have  $b_n \leq c_n$ . Hence

$$\limsup_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} c_n = L.$$

It remains to show that also  $\liminf_{n \rightarrow \infty} b_n = L$ , for then the two-sided limit exists and equals  $L$ .

Fix any  $\varepsilon > 0$ . Since  $c_n \rightarrow L$  and  $c_n \geq c_{n+1}$  for all  $n$ , there exists an index  $N$  such that for all  $n \geq N$ ,

$$c_n < L + \varepsilon.$$

In particular, for every  $n \geq N$  we have  $b_n \leq c_n < L + \varepsilon$ , which shows

$$\limsup_{n \rightarrow \infty} b_n \leq L + \varepsilon.$$

On the other hand, for each  $n$  the defining property of the supremum  $c_n$  guarantees there is some  $k \geq n$  with

$$b_k > c_n - \varepsilon,$$

otherwise  $c_n$  would not be the least upper bound. Because  $c_n$  converges to  $L$ , for all sufficiently large  $n$  we also have  $c_n - \varepsilon > L - \varepsilon$ . Thus infinitely many terms of  $(b_n)$  satisfy  $b_k > L - \varepsilon$ , forcing

$$\liminf_{n \rightarrow \infty} b_n \geq L - \varepsilon.$$

Since  $\varepsilon > 0$  was arbitrary, we conclude  $\liminf_{n \rightarrow \infty} b_n = L$ . Therefore

$$\lim_{n \rightarrow \infty} b_n = \limsup_{n \rightarrow \infty} b_n = \liminf_{n \rightarrow \infty} b_n = L,$$

and the sequence converges to the finite limit  $L$ .  $\square$

By Lemma 2, the limit  $\lim_{n \rightarrow \infty} b_n$  exists and is finite. That limit is exactly the density  $\mathcal{D}(S)$ .

We now establish the bounds on  $\mathcal{D}$ . Since each  $a_n$  is a nonnegative integer and  $k \neq 0$ , we have

$$b_n = \frac{a_n}{k^n} \geq 0 \quad \text{for every } n.$$

Because the sequence  $\{b_n\}$  converges, its limit cannot be negative. Hence

$$\mathcal{D}(S) = \lim_{n \rightarrow \infty} b_n \geq 0,$$

establishing the lower bound  $0 \leq \mathcal{D}(S)$ .

We provide an example showing the lower bound is attainable.

Define

$$S = \{0, 00\} \subset \{0, 1\}^*.$$

Then  $|S| = 2$ , so we look at

$$S^{\leq n} = \{w : w \text{ is a concatenation of at most } n \text{ blocks from } S\}.$$

Every such  $w$  is a string of zeros of length at most  $2n$ , and every string of zeros up to length  $2n$  arises. Hence

$$S^{\leq n} = \{\varepsilon, 0, 00, 000, \dots, 0^{2n}\},$$

so

$$a_n = |S^{\leq n}| = 2n + 1.$$

Therefore the density is

$$\mathcal{D}(S) = \lim_{n \rightarrow \infty} \frac{a_n}{2^n} = \lim_{n \rightarrow \infty} \frac{2n + 1}{2^n} = 0,$$

because the numerator grows only linearly while the denominator grows exponentially. Thus  $S = \{0, 00\}$  is a concrete finite set whose density  $\mathcal{D}(S)$  equals zero.

To obtain the upper bound, note that concatenating exactly  $j$  words from  $S$  yields at most  $k^j$  possibilities. Hence

$$a_n = \sum_{j=0}^n |\{\text{concatenations of length } j\}| \leq \sum_{j=0}^n k^j = \frac{k^{n+1} - 1}{k - 1}.$$

Dividing by  $k^n$  gives

$$\frac{a_n}{k^n} \leq \frac{k^{n+1} - 1}{(k-1)k^n} = \frac{k}{k-1} \left(1 - k^{-(n+1)}\right) \longrightarrow \frac{k}{k-1}.$$

Therefore  $\mathcal{D}(S) \leq k/(k-1)$ .

Combining these facts, for every finite  $S$  with  $|S| \geq 2$ , the density  $\mathcal{D}(S)$  exists and satisfies

$$0 \leq \mathcal{D}(S) \leq \frac{|S|}{|S|-1}.$$

Now we show that no intermediate values can occur. Suppose first that  $\text{Eff}(S) = k$ . This happens, for example, when  $S$  is prefix-free — that is, no word in  $S$  is a prefix of another — because then all  $j$ -fold concatenations yield distinct words. In this maximal case, for all  $n$ , we have

$$a_n = \sum_{j=0}^n k^j = \frac{k^{n+1} - 1}{k-1},$$

and hence

$$\frac{a_n}{k^n} = \frac{k}{k-1} - \frac{1}{(k-1)k^n}.$$

Taking the limit as  $n \rightarrow \infty$ , we obtain

$$\mathcal{D}(S) = \frac{k}{k-1}.$$

Now suppose  $\text{Eff}(S) = r < k$ . We know, by Theorem 3, that this happens when  $S$  is not uniquely decodable — that is, there exist distinct sequences  $u_1 u_2 \cdots u_m$  and  $v_1 v_2 \cdots v_\ell$  with  $u_i, v_j \in S$ , such that the two concatenations yield the same word. In such cases, redundancy or collision among concatenated forms causes the total number  $a_n$  of distinct words to grow more slowly than the maximal case.

To quantify this, fix  $\psi > 0$  such that  $r + \psi < k$ . Then since  $\lim_{n \rightarrow \infty} a_n^{1/n} = r$ , by definition of the limit, there exists  $N$  such that for all  $n \geq N$ ,

$$a_n < (r + \psi)^n.$$

This inequality implies that for sufficiently large  $n$ , the size  $a_n$  is asymptotically bounded above by an exponential of base strictly less than  $k$ .

Now we compute the density:

$$\frac{a_n}{k^n} < \left(\frac{r + \psi}{k}\right)^n.$$

Since  $\frac{r+\psi}{k} < 1$ , this upper bound converges to zero as  $n \rightarrow \infty$ . Thus,

$$\lim_{n \rightarrow \infty} \frac{a_n}{k^n} = 0,$$

and so  $\mathcal{D}(S) = 0$ .

Thus we have concluded that

$$\mathcal{D}(S) = \lim_{n \rightarrow \infty} \frac{a_n}{k^n} = \begin{cases} \frac{k}{k-1}, & S \text{ prefix-free,} \\ 0, & \text{otherwise,} \end{cases}$$

with no other possibilities.

This completes the proof.  $\square$

**b) Limit Behavior of Effectiveness Differences Assuming  $\mathcal{D}(S_n) = \mathcal{D}(T_n) = 1$**

**Theorem 7** Let  $(S_n)$  and  $(T_n)$  be sequences of finite subsets of  $\{0,1\}^*$  with  $|S_n| = |T_n| \geq 1$ . Suppose

$$\lim_{n \rightarrow \infty} d(S_n, T_n) = 0 \quad \text{and} \quad \mathcal{D}(S_n) = \mathcal{D}(T_n) = 1$$

for all  $n$ . Then

$$\lim_{n \rightarrow \infty} (\text{Eff}(S_n) - \text{Eff}(T_n)) = 0.$$

*Proof.* By Theorem 6, the only finite set  $S \subset \{0,1\}^*$  with density 1 is  $S = \{\varepsilon\}$ . Hence for each  $n$  we have

$$S_n = \{\varepsilon\} \quad \text{and} \quad T_n = \{\varepsilon\}.$$

But then for any  $m$ ,  $|S_n^{\leq m}| = |T_n^{\leq m}| = 1$ , so by Theorem 1

$$\text{Eff}(S_n) = \lim_{m \rightarrow \infty} 1^{1/m} = 1, \quad \text{Eff}(T_n) = 1.$$

Thus  $\text{Eff}(S_n) - \text{Eff}(T_n) = 0$  for all  $n$ , and the limit of the difference is 0.  $\square$

**c) Limit Behavior of Effectiveness Differences Assuming  $\mathcal{D}(S_n) = \mathcal{D}(T_n) = \ell$  for  $\ell \in [0, 1)$**

**Theorem 8** Let  $(S_n)$  and  $(T_n)$  be sequences of finite subsets of  $\{0,1\}^*$ . Suppose

$$\lim_{n \rightarrow \infty} d(S_n, T_n) = 0 \quad \text{and} \quad \mathcal{D}(S_n) = \mathcal{D}(T_n) = \ell \quad \text{for } \ell \in [0, 1)$$

for all  $n$ . Then

$$\lim_{n \rightarrow \infty} (\text{Eff}(S_n) - \text{Eff}(T_n)) = 0.$$

*Proof.* By Theorem 6 we know that if the density  $D(S)$  lies in the interval  $[0, 1)$  then in fact  $D(S) = 0$ . In particular each  $S_n$  and  $T_n$  must be non-prefix-free (since every prefix-free set of size at least 2 has density strictly greater than 1).

We showed in Theorem 4 that  $\lim_{n \rightarrow \infty} (\text{Eff}(S_n) - \text{Eff}(T_n))$  must be 0 for any well behaving two non uniquely decodable sequences.

This completes the proof.  $\square$

## 6 $\ell$ -sparse Subsets $S$

**Question 6:** A subset is said to be  $\ell$ -sparse if its density is a given  $\ell \in [0, 1]$ . We denote by

$$\omega_{\ell,k} = \frac{|\{S \subset \{0, 1\}^*, \ell\text{-sparse, } |S| \leq k\}|}{2^k}$$

and  $\omega_\ell = \lim_{k \rightarrow \infty} \omega_{\ell,k}$ . We ask the following questions:

- a) Does  $\omega_\ell$  exist?
- b) Give an example of 0-sparse subset.
- c) For  $k \geq 2$ , compute  $\omega_0$ .
- d) Estimate as precisely as possible  $\omega_\ell$  for any  $\ell \in (0, 1]$ .

### a) Existence of $\omega_\ell$

**Theorem 9** For each  $\ell \in [0, 1]$ ,  $\omega_\ell$  exists in the extended real numbers  $\mathbb{R} \cup \{\infty\}$ .

*Proof.* Consider cases based on  $\ell$ :

**Case 1:**  $\ell \in (0, 1)$  By the Theorem 6, no finite non-empty set satisfies  $\mathcal{D}(S) = \ell$ . Thus:

$$\{S : \mathcal{D}(S) = \ell, |S| \leq k\} = \emptyset \quad \forall k \in \mathbb{N}$$

Therefore  $\omega_{\ell,k} = 0$  for all  $k$ , so  $\lim_{k \rightarrow \infty} \omega_{\ell,k} = 0$ .

**Case 2:**  $\ell = 1$  The only set with  $\mathcal{D}(S) = 1$  is  $S = \{\epsilon\}$ . Thus:

$$\{S : \mathcal{D}(S) = 1, |S| \leq k\} = \begin{cases} \emptyset & k = 0 \\ \{\{\epsilon\}\} & k \geq 1 \end{cases}$$

So:

$$\omega_{1,k} = \begin{cases} 0 & k = 0 \\ 2^{-k} & k \geq 1 \end{cases}$$

Since  $\lim_{k \rightarrow \infty} 2^{-k} = 0$ , the limit exists and equals 0.

**Case 3:**  $\ell = 0$  Sets with  $\mathcal{D}(S) = 0$  are exactly non-prefix-free sets with  $|S| \geq 2$ . Define:

$$\mathcal{S}_k = \{S \subset \{0, 1\}^* : \mathcal{D}(S) = 0, |S| \leq k\}$$

**Subcase 3.1:**  $k < 2$

No set satisfies  $|S| \geq 2$  and  $\mathcal{D}(S) = 0$ , so  $\mathcal{S}_k = \emptyset$ , thus  $\omega_{0,k} = 0$ .

**Subcase 3.2:**  $k \geq 2$ 

- ( $k = 2$  gives non-UD examples) For each  $n \geq 1$ , the set  $S_n = \{0^n, 0^{2n}\}$  fails unique decodability. Thus there are infinitely many non-UD subsets of size 2, so  $\omega_{0,2} > 0$ .
- ( $k \geq 3$  also gives non-UD examples) For each word  $w$ , the set  $T_w = \{w, w0, 0\}$  fails unique decodability. Hence  $\omega_{0,k} > 0$  for all  $k \geq 3$ .
- (Monotonicity) Adding one more allowed element (increasing  $k$ ) can only increase the numerator or leave it unchanged, so  $\omega_{0,k}$  is non-decreasing in  $k$ .
- (Existence of limit) Every non-decreasing sequence in the extended reals admits a limit (finite or  $+\infty$ ).

Therefore  $\lim_{k \rightarrow \infty} \omega_{0,k}$  exists in  $[0, \infty]$ .

In all cases, the limit exists in  $\mathbb{R} \cup \{\infty\}$ . □

**b) An example of a 0-Sparse Subset**

**Theorem 10** Any non-uniquely decodable subset is 0-sparse.

*Proof.* Recall that a set is called  $\ell$ -sparse if its density  $\mathcal{D} = \ell$ . Thus a 0-sparse subset must have density  $\mathcal{D}=0$ . By Theorem 6, we know that this occurs when a set  $S$  has cardinality  $|S| \geq 2$  and the effectiveness  $\text{Eff}(S) < |S|$ . Equivalently, this happens when  $S$  is non-uniquely decodable. □

**c)  $\omega_0$  for  $k \geq 2$** 

We now provide a more rigorous proof of that which was shown in Theorem 8.

**Theorem 11 (Computation of  $\omega_{0,k}$ ):** For integers  $k \geq 2$ ,

$$\omega_{0,k} = \infty,$$

*Proof.* Define the collection:

$$\mathcal{S}_k = \{S \subset \{0, 1\}^*: \mathcal{D}(S) = 0, |S| \leq k\}.$$

We wish to show  $\omega_{0,k} = \infty$ .

First we show  $|\mathcal{S}_2| = \aleph_0$ . For each  $n \in \mathbb{N}$ , let

$$S_n = \{0^n, 0^{2n}\}.$$

Since  $0^{2n} = 0^n \cdot 0^n$ , the set  $S_n$  fails unique decodability. Moreover, if  $S_n = S_m$  then the shorter word in each set satisfies  $0^n = 0^m$ , whence  $n = m$ . Thus  $n \mapsto S_n$  is an injective map from  $\mathbb{N}$  into  $\mathcal{S}_2$ , proving  $|\mathcal{S}_2| \geq \aleph_0$ . On the other hand, there are only countably many 2-element subsets of the countable set  $\{0, 1\}^*$ , so  $|\mathcal{S}_2| \leq \aleph_0$ . Therefore  $|\mathcal{S}_2| = \aleph_0$ .

Next we handle  $k \geq 3$ . For each word  $w \in \{0, 1\}^*$ , define

$$T_w = \{w, w0, 0\}.$$

Here  $w$  is a strict prefix of  $w0$ , so  $T_w$  is not prefix-free and hence not uniquely decodable. If  $T_w = T_v$  then comparing the two words of greater length shows  $w = v$ . Thus  $w \mapsto T_w$  is injective, giving  $|\mathcal{S}_k| \geq \aleph_0$  for all  $k \geq 3$ . Again, since there are only countably many finite subsets of a countable set,  $|\mathcal{S}_k| \leq \aleph_0$ . Hence  $|\mathcal{S}_k| = \aleph_0$  whenever  $k \geq 3$ .

Combining both cases, we conclude that for every integer  $k \geq 2$ ,

$$|\mathcal{S}_k| = \aleph_0.$$

Consequently, in the extended real sense,

$$\omega_{0,k} = \frac{|\mathcal{S}_k|}{2^k} = \frac{\aleph_0}{2^k} = \infty,$$

and therefore  $\lim_{k \rightarrow \infty} \omega_{0,k} = \infty$ .

For any real number  $M > 0$ , choose  $k \geq 2$  fixed. Since  $\aleph_0$  represents countable infinity, and  $2^k$  is finite, the quotient exceeds any real  $M$ . Thus by definition of divergence to infinity,  $\frac{\aleph_0}{2^k} = \infty$ .

Thus,

$$\omega_{0,k} = \frac{|\mathcal{S}_k|}{2^k} = \frac{\aleph_0}{2^k} = \infty.$$

and thus,

$$\omega_0 = \lim_{k \rightarrow \infty} \frac{|\mathcal{S}_k|}{2^k} = \lim_{k \rightarrow \infty} \infty = \infty$$

This completes the proof. □

#### d) Estimation of $\omega_\ell$ for $\ell \in (0, 1]$

**Theorem 12** For every  $\ell \in (0, 1]$  we have

$$\omega_\ell = 0.$$

*Proof.* We begin by recalling from our previous results that if  $S \subset \{0, 1\}^*$  is a finite set then the density

$$\mathcal{D}(S) = \lim_{m \rightarrow \infty} \frac{|S^{\leq m}|}{|S|^m}$$

can only take one of three types of values. In fact, we have proven that:

- If  $S = \{\varepsilon\}$  then  $\mathcal{D}(S) = 1$ .
- If  $S$  is a singleton and  $S \neq \{\varepsilon\}$  (i.e. it contains a nonempty word), then  $\mathcal{D}(S) = +\infty$ .
- If  $|S| \geq 2$  then either

$$\mathcal{D}(S) = 0 \quad \text{or} \quad \mathcal{D}(S) = \frac{|S|}{|S|-1} > 1.$$

Thus, it is clear that the only density values attainable in the interval  $[0, 1]$  are 0 and 1; in particular, for any fixed  $\ell \in (0, 1)$  no finite set  $S$  can have  $\mathcal{D}(S) = \ell$ . Consequently, for any  $k$  and for  $\ell \in (0, 1)$  the set

$$\{S \subset \{0, 1\}^*: S \text{ is } \ell\text{-sparse and } |S| \leq k\}$$

is empty, implying that

$$\omega_{\ell,k} = 0 \quad \text{for all } k.$$

Hence,

$$\omega_\ell = \lim_{k \rightarrow \infty} \omega_{\ell,k} = 0 \quad \text{for all } \ell \in (0, 1).$$

It remains to consider the case  $\ell = 1$ . In this case the only candidate finite subsets  $S \subset \{0, 1\}^*$  satisfying  $\mathcal{D}(S) = 1$  are, essentially, the trivial ones. Indeed, if  $S = \{\varepsilon\}$  then  $\mathcal{D}(S) = 1$ , while any other singleton  $S = \{w\}$  with  $w \neq \varepsilon$  satisfies  $\mathcal{D}(S) = +\infty$ . Furthermore, as shown above, if  $|S| \geq 2$  then either  $\mathcal{D}(S) = 0$  or  $\mathcal{D}(S) > 1$ . Thus, apart from the possibility of including the empty set (if one chooses to consider  $S = \emptyset$  as a candidate), the only finite set for which the density is 1 is  $S = \{\varepsilon\}$ . Therefore, for every  $k \geq 1$  the number of 1-sparse sets (with nonzero cardinality) is at most a constant (in fact, there is exactly one such set, namely  $\{\varepsilon\}$ ). Hence,

$$\omega_{1,k} \leq \frac{1}{2^k}.$$

Taking the limit as  $k \rightarrow \infty$  yields

$$\omega_1 = \lim_{k \rightarrow \infty} \omega_{1,k} = 0.$$

Combining the two cases, we conclude that for every  $\ell \in (0, 1]$ ,

$$\omega_\ell = 0.$$

This completes the proof. □

## 7 Further Extensions and Generalisations

**Question 7:** We suggest and study other research directions.

a) Generalisation to an arbitrary alphabet

**Theorem 13** (Effectiveness bounds for fixed  $|S| = k$  over a general alphabet of size  $q$ ): Let  $A$  be an alphabet with  $|A| = q$  and let  $S \subseteq A^*$  be any finite generator set of cardinality  $|S| = k$ . Define

$$\text{Eff}(S) = \lim_{n \rightarrow \infty} |S^{\leq n}|^{1/n},$$

where

$$S^{\leq n} = \{ w : \exists u_1, \dots, u_m \in S, m \leq n, w = u_1 u_2 \cdots u_m \}.$$

Then one always has

$$1 \leq \text{Eff}(S) \leq k.$$

*Proof.* First, note that every concatenation of exactly  $j$  elements of  $S$  produces at most  $k^j$  distinct words. Hence

$$|S^{\leq n}| = \sum_{j=0}^n |S^j| \leq \sum_{j=0}^n k^j = \frac{k^{n+1} - 1}{k - 1},$$

and taking the  $n$ th root of each side gives

$$|S^{\leq n}|^{1/n} \leq \left( \frac{k^{n+1} - 1}{k - 1} \right)^{1/n}.$$

As  $n \rightarrow \infty$  the right-hand side converges to  $k$ , establishing the upper bound  $\text{Eff}(S) \leq k$ . On the other hand, since  $S^{\leq n}$  always contains at least one word (namely the empty concatenation), we have  $|S^{\leq n}| \geq 1$  and hence  $|S^{\leq n}|^{1/n} \geq 1$ , which yields the lower bound  $\text{Eff}(S) \geq 1$ .  $\square$

**Theorem 14** (Density bounds and characterization for fixed  $|S| = k$  over a general alphabet): Let  $A$  be an alphabet with  $|A| = q$  and let  $S \subseteq A^*$  satisfy  $|S| = k$ . Suppose the limit

$$\mathcal{D}(S) = \lim_{n \rightarrow \infty} \frac{|S^{\leq n}|}{k^n}$$

exists. Then one always has

$$0 \leq \mathcal{D}(S) \leq \frac{k}{k-1}.$$

Moreover,  $\mathcal{D}(S) = k/(k-1)$  if and only if  $S$  is uniquely decodable (equivalently, prefix-free when  $S$  is finite); otherwise, collisions among distinct concatenations force exponential growth strictly below  $k^n$ , yielding  $\mathcal{D}(S) = 0$ .

*Proof.* Since  $|S^{\leq n}| = \sum_{j=0}^n |S^j|$  and each term satisfies  $|S^j| \leq k^j$ , we have

$$\frac{|S^{\leq n}|}{k^n} \leq \sum_{j=0}^n k^{j-n} = \sum_{i=0}^n k^{-i} = \frac{1 - k^{-(n+1)}}{1 - 1/k},$$

and as  $n \rightarrow \infty$  the right-hand side converges to  $1/(1 - 1/k) = k/(k-1)$ , proving the upper bound.

Trivially  $|S^{\leq n}| \geq 0$  gives the lower bound. To see the dichotomy, observe that if  $S$  is uniquely decodable then every word in  $S^{=n}$  arises from a unique length- $n$  concatenation of generators, so  $|S^{=n}| = k^n$  and hence

$$|S^{\leq n}| = \sum_{j=0}^n k^j = \frac{k^{n+1} - 1}{k - 1},$$

whence

$$\mathcal{D}(S) = \lim_{n \rightarrow \infty} \frac{k^{n+1} - 1}{(k - 1)k^n} = \frac{k}{k - 1}.$$

Now suppose  $\text{Eff}(S) = r < k$ . This happens when  $S$  is non uniquely decodable. In such cases, redundancy or collision among concatenated forms causes the total number  $a_n$  of distinct words to grow more slowly than the maximal case.

To quantify this, fix  $\psi > 0$  such that  $r + \psi < k$ . Then since  $\lim_{n \rightarrow \infty} a_n^{1/n} = r$ , by definition of the limit, there exists  $N$  such that for all  $n \geq N$ ,

$$a_n < (r + \psi)^n.$$

This inequality implies that for sufficiently large  $n$ , the size  $a_n$  is asymptotically bounded above by an exponential of base strictly less than  $k$ .

Now we compute the density:

$$\frac{a_n}{k^n} < \left(\frac{r + \psi}{k}\right)^n.$$

Since  $\frac{r + \psi}{k} < 1$ , this upper bound converges to zero as  $n \rightarrow \infty$ . Thus,

$$\lim_{n \rightarrow \infty} \frac{a_n}{k^n} = 0,$$

and so  $\mathcal{D}(S) = 0$ .

This completes the proof

□

### b) Reinterpretation of $\omega_{0,k}$ with a Length Constraint

**Definition 1 (Prefix Poset):** The *prefix poset* on  $\{0, 1\}^{\leq k}$  is the partially ordered set  $(P_k, \preceq)$  where:

- $P_k = \{w \in \{0, 1\}^*: |w| \leq k\}$  is the set of binary strings of length at most  $k$
- $u \preceq v$  if and only if  $u$  is a prefix of  $v$

A set  $S \subseteq P_k$  is an *antichain* if no two distinct elements are comparable:  $\forall u, v \in S (u \neq v \implies u \not\preceq v \text{ and } v \not\preceq u)$ .

**Theorem 15 (Computation of  $\omega_{0,k}$  with Length Constraint):** For any integer  $k \geq 2$ , the redefined density parameter is given by:

$$\omega_{0,k} = \frac{1}{2^k} \left( \sum_{j=2}^k \binom{|P_k|}{j} - \sum_{j=2}^k a_j^{(k)} \right)$$

where:

- $P_k = \{w \in \{0,1\}^* : |w| \leq k\}$
- $|P_k| = 2^{k+1} - 1$
- $a_j^{(k)} = |\{S \subseteq P_k : |S| = j, S \text{ is an antichain}\}|$

This expression is well-defined and computable for each fixed  $k \geq 2$ .

*Proof.* The redefined  $\omega_{0,k}$  is:

$$\omega_{0,k} = \frac{|\{S \subseteq \{0,1\}^* : \mathcal{D}(S) = 0, |S| \leq k, \forall w \in S (|w| \leq k)\}|}{2^k}$$

The condition  $\forall w \in S (|w| \leq k)$  restricts  $S$  to subsets of the finite set:

$$P_k = \{w \in \{0,1\}^* : |w| \leq k\}$$

The cardinality of  $P_k$  is:

$$|P_k| = \sum_{i=0}^k 2^i = 2^{k+1} - 1$$

since there are  $2^i$  binary strings of length  $i$  for  $0 \leq i \leq k$ .

By the Theorem 6,  $\mathcal{D}(S) = 0$  if and only if:

- $S$  is non-empty and not prefix-free
- $|S| \geq 2$  (since singletons have density 1 or  $\infty$ )

Moreover, a set  $S \subseteq \{0,1\}^*$  is *not prefix-free* if and only if there exist distinct  $u, v \in S$  such that  $u$  is a prefix of  $v$ . In the prefix poset  $(P_k, \preceq)$ , this means  $S$  is *not an antichain*.

Define the collection of valid sets:

$$\mathcal{S}_k = \{S \subseteq P_k : \mathcal{D}(S) = 0, |S| \leq k\}$$

Conditions simplify to:

$$\mathcal{S}_k = \{S \subseteq P_k : 2 \leq |S| \leq k, S \text{ is not an antichain}\}$$

since  $|S| \leq k$  is automatically satisfied for subsets of  $P_k$  when we enforce  $|S| \leq k$ , and  $\mathcal{D}(S) = 0$  iff  $|S| \geq 2$  and  $S$  not prefix-free iff not antichain.

The collection  $\mathcal{S}_k$  decomposes as:

$$\mathcal{S}_k = \bigcup_{j=2}^k (\{S \subseteq P_k : |S| = j\} \setminus \{S \subseteq P_k : |S| = j, S \text{ antichain}\})$$

These are disjoint unions, so:

$$|\mathcal{S}_k| = \sum_{j=2}^k |\{S \subseteq P_k : |S| = j\}| - \sum_{j=2}^k |\{S \subseteq P_k : |S| = j, S \text{ antichain}\}|$$

By definition:

- $|\{S \subseteq P_k : |S| = j\}| = \binom{|P_k|}{j} = \binom{2^{k+1}-1}{j}$
- $|\{S \subseteq P_k : |S| = j, S \text{ antichain}\}| = a_j^{(k)}$

Thus:

$$|\mathcal{S}_k| = \sum_{j=2}^k \binom{|P_k|}{j} - \sum_{j=2}^k a_j^{(k)}$$

Substituting into  $\omega_{0,k}$ :

$$\omega_{0,k} = \frac{|\mathcal{S}_k|}{2^k} = \frac{1}{2^k} \left( \sum_{j=2}^k \binom{|P_k|}{j} - \sum_{j=2}^k a_j^{(k)} \right)$$

This expression is finite and computable for each fixed  $k \geq 2$  because:

- $P_k$  is finite ( $|P_k| = 2^{k+1} - 1 < \infty$ )
- Binomial coefficients are finite integers
- $a_j^{(k)}$  counts antichains in a finite poset, which is computable by brute-force enumeration

Thus the theorem is proved.  $\square$

We provide an example computation for when  $k = 2$ .

We begin by constructing the set  $P_2$

$$P_2 = \{\epsilon, 0, 1, 00, 01, 10, 11\}, \quad |P_2| = 7$$

We then list all the antichains

- $\{0, 1\}$
- $\{0, 10\}$
- $\{0, 11\}$

- $\{1, 00\}$
- $\{1, 01\}$
- $\{00, 01\}$
- $\{00, 10\}$
- $\{00, 11\}$
- $\{01, 10\}$
- $\{01, 11\}$
- $\{10, 11\}$

Total  $a_2^{(2)} = 11$

We now compute the binomial sum

$$\sum_{j=2}^2 \binom{7}{j} = \binom{7}{2} = 21$$

$$\sum_{j=2}^2 a_j^{(2)} = 11$$

Finally, we compute  $|\mathcal{S}_2|$

$$|\mathcal{S}_2| = 21 - 11 = 10$$

And finally, we compute  $\omega_{0,2}$

$$\omega_{0,2} = \frac{10}{2^2} = \frac{10}{4} = \frac{5}{2}$$

### c) Cost Weighted Sets

#### i) Weighted Effectiveness

We consider finite sets of binary strings  $S \subseteq \{0, 1\}^*$ , called codes. We are interested in the growth behavior of the number of distinct words formed by concatenating codewords from  $S$ , subject to a given total cost.

To formalize this, we fix a cost function  $c : S \rightarrow \mathbb{R}_{>0}$ , which assigns a positive real cost to each codeword. For example, we may take  $c(w) = |w|^d$  for some fixed real  $d > 0$ , though our argument will work for arbitrary positive costs.

Define:

$$S_c^{\leq n} = \left\{ w_1 w_2 \dots w_k : w_i \in S, \sum_{i=1}^k c(w_i) \leq n \right\},$$

which is the set of all strings obtained by concatenating zero or more codewords from  $S$ , such that

the total cost of the chosen codewords is at most  $n$ .

We define the weighted effectiveness of the code  $S$  as:

$$\text{Eff}(S) := \limsup_{n \rightarrow \infty} |S_c^{\leq n}|^{1/n}.$$

This value reflects the exponential rate at which the number of such cost-bounded strings grows.

**Theorem 12** Let  $S \subseteq \{0, 1\}^*$  be a finite code with  $|S| = k \geq 1$ , and let  $c : S \rightarrow \mathbb{R}_{>0}$  be any positive cost function. Define

$$c_{\min} := \min_{w \in S} c(w).$$

Then the weighted effectiveness satisfies:

$$1 \leq \text{Eff}(S) \leq k^{1/c_{\min}}.$$

Moreover, if  $S$  is not uniquely decodable, then

$$\text{Eff}(S) < k^{1/c_{\min}}.$$

*Proof.* Let  $a_n := |S_c^{\leq n}|$ . We first show that  $\{a_n\}$  is submultiplicative. If  $x \in S_c^{\leq n}$  and  $y \in S_c^{\leq m}$ , then the concatenation  $xy \in S_c^{\leq n+m}$ , since the total cost is at most  $n + m$ . Therefore,

$$a_{n+m} \leq a_n \cdot a_m.$$

By Fekete's Lemma, the limit

$$\text{Eff}(S) = \lim_{n \rightarrow \infty} a_n^{1/n}$$

exists and equals  $\inf_{n \geq 1} a_n^{1/n}$ .

We now prove the upper and lower bounds. First, since the empty word  $\varepsilon$  has cost zero, it belongs to  $S_c^{\leq n}$  for every  $n$ , so  $a_n \geq 1$  for all  $n$ . Hence,

$$\text{Eff}(S) \geq \lim_{n \rightarrow \infty} 1^{1/n} = 1.$$

For the upper bound, note that any word formed from  $S$  of total cost at most  $n$  must use at most  $\lfloor n/c_{\min} \rfloor$  codewords, since each codeword costs at least  $c_{\min}$ . There are at most  $k^m$  sequences of  $m$  codewords, so:

$$a_n \leq \sum_{m=0}^{\lfloor n/c_{\min} \rfloor} k^m \leq \frac{k^{n/c_{\min}+1}}{k-1}.$$

Now take the  $n$ th root:

$$a_n^{1/n} \leq \left( \frac{k^{n/c_{\min}+1}}{k-1} \right)^{1/n} = k^{1/c_{\min}} \cdot k^{1/n} \cdot (k-1)^{-1/n}.$$

As  $n \rightarrow \infty$ , both  $k^{1/n} \rightarrow 1$  and  $(k-1)^{-1/n} \rightarrow 1$ , so

$$\text{Eff}(S) \leq k^{1/c_{\min}}.$$

This completes the general bounds:

$$1 \leq \text{Eff}(S) \leq k^{1/c_{\min}}.$$

Now suppose  $S$  is not uniquely decodable. Then there exist two different sequences of codewords  $u_1 u_2 \cdots u_r$  and  $v_1 v_2 \cdots v_s$ , with each  $u_i, v_j \in S$ , such that

$$u_1 u_2 \cdots u_r = v_1 v_2 \cdots v_s,$$

but the two sequences are distinct (either  $r \neq s$ , or  $u_i \neq v_i$  for some  $i$ ).

Let

$$C := \sum_{i=1}^r c(u_i) = \sum_{j=1}^s c(v_j)$$

be the common cost of these two sequences. In a code with no such ambiguity, each sequence of codewords would correspond to a unique concatenated word, so we would have  $k^r$  distinct words of cost  $C$ . But due to the ambiguity, at least one such word is lost.

So:

$$a_C \leq k^r - 1.$$

Hence:

$$a_C^{1/C} \leq (k^r - 1)^{1/C} < k^{r/C} \leq k^{1/c_{\min}}.$$

Therefore,

$$\text{Eff}(S) = \inf_{n \geq 1} a_n^{1/n} \leq a_C^{1/C} < k^{1/c_{\min}}.$$

This strict inequality shows that the effectiveness of a non-uniquely decodable code is strictly smaller than the maximal bound  $k^{1/c_{\min}}$ . This completes the proof.  $\square$

## ii) Limit Behavior of Weighted Effectiveness Under a Cost Weighted Distance Metric

We now also define a distance between codewords that takes into account both structural similarity and cost. Given two words  $u, v \in \{0, 1\}^*$ , let  $P$  denote the length of their longest common prefix. We define the **cost-weighted prefix distance** by:

$$d_c(u, v) := \begin{cases} 0 & \text{if } u = v, \\ \frac{2^{-|P|}}{\max(c(u), c(v))} & \text{if } u \neq v. \end{cases}$$

This metric reflects two things:

- Words that share a long prefix are structurally close.

- Words with higher cost are treated as semantically heavier — so if they agree, the similarity is given more weight. Low-cost codewords diverging even slightly are considered significantly different.

The distance between two finite sets  $S, T \subseteq \{0, 1\}^*$  is then defined as:

$$d_c(S, T) := \max \left\{ \sup_{x \in S} \inf_{y \in T} d_c(x, y), \sup_{y \in T} \inf_{x \in S} d_c(x, y) \right\}.$$

This distance measures how structurally and cost-wise close the two sets are: small distance implies that every codeword in one set is closely approximated by one in the other, both in terms of prefix structure and cost.

**Theorem 12** Let  $(S_n)_{n \in \mathbb{N}}$  and  $(T_n)_{n \in \mathbb{N}}$  be sequences of finite subsets of  $\{0, 1\}^*$ , and suppose a fixed cost function  $c : \{0, 1\}^* \rightarrow \mathbb{R}_{>0}$  is given. Assume there exist constants  $0 < c_{\min} \leq c(w) \leq c_{\max} < \infty$  for all codewords in all  $S_n$  and  $T_n$ . If

$$\lim_{n \rightarrow \infty} d_c(S_n, T_n) = 0,$$

then

$$\lim_{n \rightarrow \infty} |\text{Eff}(S_n) - \text{Eff}(T_n)| = 0.$$

*Proof.* Fix  $\varepsilon > 0$ . Since  $d_c(S_n, T_n) \rightarrow 0$ , there exists  $N \in \mathbb{N}$  such that for all  $n \geq N$ , we have  $d_c(S_n, T_n) < \delta$ , where  $\delta > 0$  is chosen small enough to ensure both prefix and cost similarity.

Specifically, for each  $w \in S_n$ , there exists  $w' \in T_n$  such that:

$$\frac{2^{-|P|}}{\max(c(w), c(w'))} < \delta,$$

where  $P$  is the length of the longest common prefix of  $w$  and  $w'$ . Hence,  $|P|$  is large and  $|c(w) - c(w')|$  is small. This means codewords in  $S_n$  can be matched to close approximations in  $T_n$ , with only small changes in cost.

Now consider a word  $x = w_1 w_2 \cdots w_k \in S_n^*$  of total cost at most  $t$ . For each  $w_i$ , choose a corresponding approximation  $w'_i \in T_n$  as above. Then:

$$\sum_{i=1}^k c(w'_i) \leq \sum_{i=1}^k (c(w_i) + \varepsilon) = \text{cost}(x) + k\varepsilon.$$

Since each codeword has cost at least  $c_{\min}$ , we have  $k \leq t/c_{\min}$ , so:

$$\sum_{i=1}^k c(w'_i) \leq t + \frac{t\varepsilon}{c_{\min}} = t(1 + \varepsilon').$$

Thus, every string in  $S_n^*$  of cost at most  $t$  can be approximated by a string in  $T_n^*$  of cost at most  $(1 + \varepsilon')t$ , for some small  $\varepsilon' > 0$ , and vice versa.

This implies:

$$|S_n^{\leq t}| \leq |T_n^{\leq(1+\varepsilon')t}| \cdot C_1(t),$$

for some function  $C(t)$ . Similarly, by symmetry, we can show that

$$|T_n^{\leq t}| \leq |S_n^{\leq(1+\varepsilon')t}| \cdot C_2(t)$$

Thus,  $|S_n^{\leq t}| = |T_n^{\leq t}| \implies \text{Eff}_c(S_n) = \text{Eff}(T_n)$ . Thus,

$$\lim_{n \rightarrow \infty} |\text{Eff}_c(S_n) - \text{Eff}_c(T_n)| = 0.$$

□