# Separating Words: A logarithmic upper bound

Mikey Watts

June 2022

## 1 Work in Progress

This paper has an error in section 5 where the probability function is introduced. I am currently considering using probabilities from the coupon's collector problem to introduce probabilities that all pairs and all machines have been sampled.

## 2 Introduction

We argue that, given any two unique words in $\{0,1\}^n$, there exists a finite state machine with $m = O(\log n)$ states that separates them.

If accurate, this proof would close this open problem.

For a good primer of the separating words problem, see Remarks on Separating Words

## 3 Prior Work

A recent proof suggests a logarithmic lower bound (suggesting this upper bound is optimal).

The state of the art is that there exists a separating DFA with $\tilde{O}(n^{1/3})$ states.

## 4 Our Approach

We show that, as $n \to \infty$, the probability approaches 1 that:

For all $a \neq b \in \{0,1\}^n$ there exists an M such that M separates a and b, where M is a DFA with $m = \log_2 n$ states.

Moreover, we show that, for large m, the converse probability is more than halved each time the number of states increases by 1.

# 5 The Probability

Let's define the probability $P(n,m)$ that, for all pairs of strings $(a,b), a \neq b$ of length $n$, there exists a machine with $m$ states that separates them.

This is equivalent to:

$$\begin{aligned} P(n,m) &= p(\text{a random pair CAN be separated})^{(\text{\# of pairs})} \\ &= [1 - p(\text{a random pair CAN'T be separated})]^{(\text{\# of pairs})} \\ &= [1 - p(\text{a randomly selected machine does not separate a random pair})^{(\text{\# of machines})}]^{(\text{\# of pairs})} \end{aligned}$$

Which, if we introduce the variables:

$$\begin{aligned} S &:= p(\text{a randomly selected machine does not separate a random pair}) \\ M &:= \text{The number of machines of size m} \\ N &:= \text{The number of pairs of unique strings of length n} \end{aligned}$$

becomes:

$$P(n,m) = (1 - S^M)^N \tag{1}$$

Now let's calculate (and bound) these variables...

# 6 M := The number of machines of size m

How many DFAs are there with m states?

Consider an m-state DFA with states $s_1, s_2, ...s_m$. Each state has 2 transitions. That's $2m$ transitions. And each transition has $m$ places it can go. So, that makes $m^{2m}$ unique transition functions. The start state can be any one of $m$ states, so that gives $m^{2m+1}$.

If we want to work with unique DFAs, then we should account for those that are isomorphic to each other. Since there are $m!$ ways to order the states, we can divide by $m!$ to get the number of unique DFAs: $\frac{m^{2m+1}}{m!}$.

For the sake of simplicity, since $m! < m^m$, we will use the lower bound:

$$M > m^m \tag{2}$$

# 7 N := The number of pairs of unique strings of length n

There are $2^n$ strings of length n, so there are $(2^n)^2 = 2^{2n}$ pairs of strings. But, $2^n$ of those combos are $(a,b)$ such that $a = b$.

For the sake of simplicity, we will use the upper bound:

$$N < 2^{2n} \tag{3}$$

# 8 S := p(a randomly selected machine doesn't separate a random pair)

Given a random pair of strings of length $n$, the probability that their final difference occurs at position $n$ is $\frac{1}{2}$, at $n-1$ is $\frac{1}{4}$, at $n-2$ is $\frac{1}{8}$, etc.

The probability that they differ in the $n^{th}$ position but are not separated is the probability that both transitions go to the same state: $\frac{1}{m}$.

Given that their final difference is in the $(n-1)^{th}$ position, the probability they are not separated is $1 - p(\text{they stay separate at each transition}) = 1 - (\frac{m-1}{m})^2$ (because if they were to transition to the same state, they'd stick together until the end since there are no more differences).

Similarly, given that their final difference is in the $(n-d)^{th}$ position, the probability they are not separated is[1] $1 - (\frac{m-1}{m})^d$

Putting this together, $S$ is:

$$\begin{gathered}\frac{1}{2}*\frac{1}{m}\\ +\\ \frac{1}{4}*(1-(\frac{m-1}{m})^2)\\ +\\ \frac{1}{8}*(1-(\frac{m-1}{m})^3)\\ +\\ \vdots\\ +\\ (\frac{1}{2})^n*(1-(\frac{m-1}{m})^n)\end{gathered}$$

Now, we can work with a simpler upper bound on this probability.

Since $(\frac{m-1}{m})^x > \frac{m-x}{m}$, then $1 - (\frac{m-1}{m})^x < 1 - \frac{m-x}{m} = \frac{x}{m}$, so we can bound by:

[1] (Actually, $\frac{m-1}{m}$ is a lower bound on the probability that the strings stay separate for a given transition. Because, knowing that the two strings have stayed separate when transitioning on digit $d_i$, we later sample transitions with replacement, i.e. there's a small chance that we see that exact transition again for later digits $d_{j>i}$, thus slightly increasing the probability of separation beyond the naive $\frac{m-1}{m}$. However, underestimating the actual probability by instead using $\frac{m-1}{m}$ is fine since it yields us an upper bound on the probability that the strings are not separated.)

$$\frac{1}{2} * \frac{1}{m}$$

$$+$$

$$\frac{1}{4} * \frac{2}{m}$$

$$+$$

$$\frac{1}{8} * \frac{3}{m}$$

$$+$$

$$\vdots$$

$$+$$

$$(\frac{1}{2})^n * \frac{n}{m}$$

Factoring out the $\frac{1}{m}$, we get $\frac{1}{m} * \sum_{i=1}^{n} \frac{i}{2^i}$, and the summation can be bounded above by 2 (see here). So, our final upper bound here is:

$$S < \frac{2}{m} \tag{4}$$

# 9 A bound on the probability of separation

Plugging in these bounds (2, 3, 4) to our probability function (1), we get a lower bound on our probability function:

$$P(n,m) > (1 - (\frac{2}{m})^{m^m})^{2^{2n}} \tag{5}$$

Now, using this lower bound on $P(n,m)$, we demonstrate that the expected DFA solution size, in terms of states, is $m = O(\log n)$

# 10 A logarithmic DFA solution size

We're going to work with a bound on the binomial expansion. But first, a little lemma:

## 10.1 A lemma

If $KX < 1$, then $(1 - X)^K < 1 - KX$

## 10.2 Proof of lemma

Consider adjacent terms in the binomial expansion $t_0 - t_1 + t_2 + t_3 - ...$:

$$1 - K * X + \binom{K}{2} * X^2 - \binom{K}{3} * X^3 ...$$

For $i > 1$,

$$\begin{aligned} \binom{K}{i} &= \binom{K}{i-1} * \frac{K-i}{i} \\ &< \binom{K}{i-1} * K \end{aligned}$$

which implies that:

$$\begin{aligned} t_i &= t_{i-1} * \frac{K-i}{i} * X \\ &< t_{i-1} * K * X \\ &< t_{i-1} * 1 \end{aligned}$$

That is, each subsequent term (ignoring the sign) is less than the previous term.

Then, since the signs of terms alternate, $1 - KX$ is a lower bound (because the rest of the expansion can be thought of as adding pairs $(t_i, t_{i+1})$ such that $t_i - t_{i+1} > 0$, and if there are an odd number of terms left, the final "pair" consists only of an addition).

## 10.3 Using the lemma & Result

Now, let $n$ and $m$ be exponentially related, such that $m = 2^n$, i.e. $m = \log_2 n$

Then, the probability function becomes:

$$P(n = 2^m, m) > (1 - (\frac{2}{m})^{m^m})^{2^{2*2^m}}$$

Now, by the above lemma, since for large $m$, $2^{4^m} * (\frac{2}{m})^{m^m} < 1$, we get the following lower bound on the probability:

$$P(n = 2^m, m) > 1 - 2^{4^m} * (\frac{2}{m})^{m^m}$$

By similar logic, since $m^m$ grows much faster than $4^m$, as $n \to \infty$, so too does $m$, and the probability approaches 1.

## 10.4 A note on the probability's rate of increase

Moreover, holding $n = 2^m$ constant but otherwise increasing the state machine size by 1 (to $m+1$), we see that the probability of NOT separation more than halves for sufficiently large $m$. This strengthens the argument that the expected solution size is $O(\log n)$:

$$\begin{aligned}
1 - P(n = 2^m, m+1) &= 2^{4^m} * (\frac{2}{m+1})^{(m+1)^{m+1}} \\
&< 2^{4^m} * (\frac{2}{m})^{m^{m+1}} \\
&= 2^{4^m} * (\frac{2}{m})^{m^m * m} \\
&= 2^{4^m} * [(\frac{2}{m})^{m^m}]^m \\
&= 2^{4^m} * (\frac{2}{m})^{m^m} * [(\frac{2}{m})^{m^m}]^{m-1} \\
&= [1 - P(n = 2^m, m)] * [(\frac{2}{m})^{m^m}]^{m-1} \\
&< [1 - P(n = 2^m, m)] * \frac{1}{2}
\end{aligned}$$