Avoiding Approximate Squares

Narad Rampersad

School of Computer Science
University of Waterloo

13 June 2007

(Joint work with Dalia Krieger, Pascal Ochem, and Jeffrey Shallit)

Repetitions in words

Definition

A square (or 2-power) is a non-empty word of the form ww (or w^2). A word is squarefree if none of its subwords are squares.

Definition

Let α be a rational number, $1 < k < 2$. An α-power is a non-empty word of the form xyx, where $|xyx|/|xy| = \alpha$. A word is α-power-free if none of its subwords are β-powers for $\beta \geq \alpha$.

Example

- `tartar` is a square.
- `tent` is a $4/3$-power.

Avoiding repetitions in words

Theorem (Thue 1906)

There exists an infinite squarefree word

$$\mathbf{x} = 210201210120210\cdots$$

over the alphabet $\{0, 1, 2\}$.

Proof (sketch).

The word $\mathbf{x}$ is obtained by iterating the map $2 \to 210$, $1 \to 20$, $0 \to 1$:

$$2 \to 210 \to 210201 \to 210201210120 \to \cdots$$

□

Morphisms

Definition

A map h like the one used to prove Thue's theorem (h sends $2 \to 210$, $1 \to 20$, $0 \to 1$) is called a morphism.

Definition

If, for some symbol a, the sequence of iterates

$$h(a), h^2(a), h^3(a), \ldots$$

converges to an infinite word $\mathbf{x}$, we say that $\mathbf{x}$ is an infinite fixed point of h, and we write $\mathbf{x} = h^\omega(a)$.

Avoiding repetitions in words

Theorem (Dejean 1972)

Over the alphabet $\{0, 1, 2\}$ *there exists an infinite word*

$$\mathbf{y} = 01202120121021202101201020120210201021\cdots$$

that is k*-power-free for all* $k > 7/4$*.*

Proof (sketch).

The word **y** is obtained by iterating the morphism
$0 \rightarrow 012021201210212021$, $1 \rightarrow 120102012021020102$1,
$2 \rightarrow 201210120102101210$2:

$$0 \rightarrow 012021201210212021\cdots$$

□

Measuring similarity of words

Definition

For words x, x' of the same length, the Hamming distance $d(x, x')$ is the number of positions in which x and x' differ.

Example

$d(\texttt{cammino}, \texttt{mattino}) = 3$.

Measuring similarity of words

Definition

Given two words x, x' of the same length, their similarity $s(x, x')$ is the fraction of the number of positions in which x and x' agree. Formally,

$$s(x, x') := \frac{|x| - d(x, x')}{|x|}.$$

Example

- $s(\mathtt{lontana}, \mathtt{ventura}) = 3/7$.
- $s(\mathtt{quelle}, \mathtt{stelle}) = 2/3$.

Similarity in finite words

Definition

The similarity of a finite word z is defined to be

$$\alpha = \max_{\substack{xx' \text{ a subword of } z \\ |x|=|x'|}} s(x, x');$$

we say such a word is α-similar.

Example

- 21020121 is $1/2$-similar.

Similarity in infinite words

Definition

We say an infinite word **z** is α-similar if

$$\alpha = \sup_{\substack{xx' \text{ a subword of } \mathbf{z} \\ |x|=|x'|}} s(x,x')$$

and there exists at least one subword xx' with $|x| = |x'|$ and $s(x,x') = \alpha$. Otherwise, if

$$\alpha = \sup_{\substack{xx' \text{ a subword of } \mathbf{z} \\ |x|=|x'|}} s(x,x'),$$

but α is not attained by any subword xx' of **z**, then we say **z** is α^--similar.

An example

Example

Recall the squarefree word constructed earlier:

$$\mathbf{x} = 2102012101202102012021012102012\cdots$$

Since $\mathbf{x}$ is squarefree it is not 1-similar. But $\mathbf{x}$ contains arbitrarily large subwords xx' where x differs from x' in only 1 position, so $\mathbf{x}$ is 1^--similar.

Computational results

The following computational results give some idea as to what the minimum similarity should be over a k-letter alphabet.

Alphabet Size k	Similarity Coefficient α	Height of Tree	Number of Leaves	Number of Maximal Words
2	1	3	4	1
3	$3/4$	41	2475	36
4	$1/2$	9	382	6
5	$2/5$	75	3902869	48
6	$1/3$	17	342356	480

Minimum similarity over a 3-letter alphabet

Theorem

There exists an infinite $3/4$*-similar word* $\mathbf{w}$ *over* $\{0, 1, 2\}$.

Let h be the 24-uniform morphism defined by

$$\begin{aligned} 0 &\rightarrow 012021201021012102120210 \\ 1 &\rightarrow 120102012102120210201021 \\ 2 &\rightarrow 201210120210201021012102. \end{aligned}$$

We claim that the fixed point $\mathbf{w} = h^{\omega}(0)$ is $3/4$-similar.

Minimum similarity over a 3-letter alphabet

- We begin by checking (with a computer) that the following lemma holds.

Lemma

Let $a, b, c \in \{0, 1, 2\}$, $a \neq b$. Let w be any subword of length 24 of $h(ab)$. If w is neither a prefix nor a suffix of $h(ab)$, then $h(c)$ and w mismatch in at least 9 positions.

- To prove our result we argue by contradiction.
- Suppose that **w** contains a minimal subword yy' with $|y| = |y'|$, and y and y' match in more than $3/4 \cdot |y|$ positions.
- We check by computer that there cannot be such a minimal counterexample with $|y| \leq 72$, so we assume that $|y| > 72$.

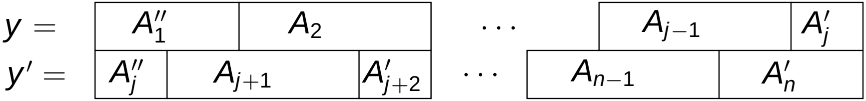
Minimum similarity over a 3-letter alphabet

- Let $w = a_1 a_2 \cdots a_n$ be a word of minimal length such that $h(w) = xyy'z$ for some x, z.
- Let us take a pictorial look at how the word $xyy'z$ decomposes into the “blocks” of the morphism h. Each A_i is a block of the morphism.

<table>
<tr><td>A_1'</td><td>A_1''</td><td colspan="6"></td><td>A_j'</td><td>A_j''</td><td colspan="6"></td><td>A_n'</td><td>A_n''</td></tr>
<tr><td colspan="2">A_1</td><td colspan="2">A_2</td><td colspan="2">$\cdots$</td><td colspan="2">A_{j-1}</td><td colspan="2">A_j</td><td colspan="2">A_{j+1}</td><td colspan="2">$\cdots$</td><td colspan="2">A_{n-1}</td><td colspan="2">A_n</td></tr>
<tr><td>x</td><td colspan="8">y</td><td colspan="8">y'</td><td>z</td></tr>
</table>

Minimum similarity over a 3-letter alphabet

- If $|A_1''| > |A_j''|$, then the picture looks like this.



- Now we look at the misaligned blocks.
- For instance, A_{j+2} in y' "straddles" A_2 and A_3 in y.
- But by the lemma, this creates at least 9 out of 24 mismatching positions between y and y'.
- This argument applies to all the misaligned blocks, and implies that y and y' mismatch in more than $1/4 \cdot |y|$ positions.
- But this contradicts our assumption that y and y' match in more than $3/4 \cdot |y|$ positions.

Minimum similarity over a 3-letter alphabet

- The same argument rules out the possibility that $|A_1''| > |A_j''|$.
- The only option left is that $|A_1''| = |A_j''|$. That is, the A_i's in y all "line up" with the A_i's in y'.
- A bit of case analysis shows that for y and y' to match in more than $3/4$ of their positions, the words $A_1 A_2 \cdots A_{j-1}$ and $A_j A_{j+1} \cdots A_{n-1}$ must match in more than $3/4$ of their positions.
- Consider the inverse images of $A_1 A_2 \cdots A_{j-1}$ and $A_j A_{j+1} \cdots A_{n-1}$ under h.
- Let

$$h(a_1 a_2 \cdots a_{j-1}) = A_1 A_2 \cdots A_{j-1},$$

and let

$$h(a_j a_{j+1} \cdots a_{n-1}) = A_j A_{j+1} \cdots A_{n-1}.$$

Minimum similarity over a 3-letter alphabet

- A quick inspection shows that any two distinct blocks mismatch in every position. Thus, a single matching position between A_1 and A_j forces $A_1 = A_j$ and $a_1 = a_j$. Similarly, a single mismatch between A_1 and A_j forces $A_1 \neq A_j$ and $a_1 \neq a_j$.
- But this implies that $a_1 a_2 \cdots a_{j-1}$ and $a_j a_{j+1} \cdots a_{n-1}$ match in at least $3/4$ of their positions.
- But $a_1 a_2 \cdots a_{n-1}$ is also a subword of **w**, and is thus a smaller counterexample than yy', contradicting minimality.
- This contradiction implies that no such counterexample exists and completes the proof.

Minimum similarity over a 4-letter alphabet

Theorem

There exists an infinite $1/2$*-similar word* $\mathbf{x}$ *over* $\{\mathtt{0}, \mathtt{1}, \mathtt{2}, \mathtt{3}\}$.

Let g be the 36-uniform morphism defined by

$$
\begin{aligned}
\mathtt{0} &\rightarrow \mathtt{012132303202321020123021203020121310} \\
\mathtt{1} &\rightarrow \mathtt{123203010313032131230132310131232021} \\
\mathtt{2} &\rightarrow \mathtt{230310121020103202301203021202303132} \\
\mathtt{3} &\rightarrow \mathtt{301021232131210313012310132313010203}.
\end{aligned}
$$

Then $\mathbf{x} = g^{\omega}(\mathtt{0})$ has the desired property.

Minimum similarity over a 4-letter alphabet

The proof is similar to that of the previous result, with the following lemma used instead.

Lemma

Let $a, b, c \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}, \mathtt{3}\}$, $a \neq b$. Let w be any subword of length 36 of $g(ab)$. If w is neither a prefix nor a suffix of $g(ab)$, then $g(c)$ and w mismatch in at least 21 positions.

We only have constructive (and optimal) results for alphabets of size 3 and 4. To say something about larger alphabets, we turn to probabilistic techniques.

The probabilistic method

- Let $A_1, A_2, \ldots, A_n$ be events in a probability space.
- We want to show $\Pr[\cap \overline{A_i}] > 0$.
- If the A_i's are mutually independent, all we need is $\Pr[A_i] < 1$.
- What do we do if the A_i's are not mutually independent?

Definition

A dependency graph on events $A_1, A_2, \ldots, A_n$ is a graph $G = \langle V, E \rangle$, where $V = \{1, 2, \ldots, n\}$, with the following property: A_i should be mutually independent of all the events A_j for which $(i, j) \notin E$.

The Lovász Local Lemma

Lemma (Lovász Local Lemma; symmetric version)

Let G be a dependency graph on events $A_1, A_2, \ldots, A_n$. Let d be the maximum degree of G. Suppose $Pr(A_i) \leq p$ for all i. If $4pd \leq 1$, then

$$\Pr\left(\bigcap_{i=1}^{n} \overline{A_i}\right) > 0.$$

- This version is applicable when the A_i's all have equal probabilities.
- When the A_i's were mutually independent, we asked that $p < 1$.
- Now we ask that $4pd \leq 1$. As long as there are not too many dependencies (i.e., d is small), this is not too much to ask.

The Lovász Local Lemma

Lemma (Lovász Local Lemma; asymmetric version)

Let G be a dependency graph on events $A_1, A_2, \ldots, A_n$. Suppose there exist real numbers $x_1, \ldots, x_n$, $0 \leq x_i < 1$, such that for all i,

$$Pr(A_i) \leq x_i \prod_{(i,j) \in E} (1 - x_j).$$

Then

$$\Pr\left(\bigcap_{i=1}^{n} \overline{A_i}\right) > 0.$$

Words with arbitrarily low similarity

Theorem

Let $c > 1$ be an integer. There exists an infinite $1/c$-similar word.

- Let Σ be a k-letter alphabet and let N be a positive integer.
- Let $w = w_1 w_2 \cdots w_N$ be a random word of length N over Σ.
- Each letter of w is chosen uniformly and independently at random from Σ.
- We now specify the "bad" events $A_1, \ldots, A_n$.
- A bad event $A_{t,r}$ is the event that two adjacent subwords y and y' of w, each of length r, beginning at positions t and $t + r$ have similarity greater than $1/c$.
- We have such events $A_{t,r}$ for all valid choices of t and r.

Bounding $\Pr(A_{t,r})$

- We need to bound from above the probability of $A_{t,r}$.
- Let us consider a subword xx', $|x| = |x'| = r$.
- We need x and x' to match in more than r/c positions.
- We will overcount the number of such words xx'.
- Let us choose $\lfloor r/c \rfloor + 1$ positions to match.
- We can do this in $\binom{r}{\lfloor r/c \rfloor + 1}$ ways.
- Now we can chose the values for these positions in $k^{\lfloor r/c \rfloor + 1}$ ways.
- With $\lfloor r/c \rfloor + 1$ positions now fixed, we have $2r - 2(\lfloor r/c \rfloor + 1)$ positions of xx' left to determine.
- We can choose the values for these positions in $k^{2r - 2(\lfloor r/c \rfloor + 1)}$ ways.

Bounding $\Pr(A_i)$

- We have actually overcounted the number of possible words xx' with more than r/c positions matching.
- An overestimate of $\mathrm{Prob}(A_i)$ is thus

$$\begin{aligned}
\mathrm{Prob}(A_i) &\leq \frac{\binom{r}{\lfloor r/c\rfloor+1} k^{\lfloor r/c\rfloor+1} k^{2r-2(\lfloor r/c\rfloor+1)}}{k^{2r}} \\
&\leq \binom{r}{\lfloor r/2\rfloor} k^{-r/c} \\
&\leq 2^r k^{-r/c}.
\end{aligned}$$

Choosing the weights x_i

- Now we must choose the x_i's.
- For all positive integers r, define $\xi_r = 2^{-2r}$.
- Note that for any real number $\alpha \leq 1/2$, we have $(1 - \alpha) \geq e^{-2\alpha}$.
- Hence, $(1 - \xi_r) \geq e^{-2\xi_r}$.
- Each event $A_{t,r}$ was associated with a pair of subwords of length r.
- We thus set $x_i = \xi_r$ for all such $A_{t,r}$.
- Let E be as in the local lemma.
- Two events share a dependency only when the corresponding subwords overlap.
- Note that a subword of length $2r$ of w overlaps with at most $2r + 2s - 1$ subwords of length $2s$.

Estimating the RHS of the local lemma

We thus have

$$
\begin{aligned}
x_i \prod_{(i,j)\in E} (1 - x_j) &\geq \xi_r \prod_{s=1}^{\lfloor N/2 \rfloor} (1 - \xi_s)^{2r+2s-1} \\
&\geq \xi_r \prod_{s=1}^{\infty} (1 - \xi_s)^{2r+2s-1} \\
&\geq \xi_r \prod_{s=1}^{\infty} e^{-2\xi_s(2r+2s-1)} \\
&\geq 2^{-2r} \prod_{s=1}^{\infty} e^{-2(2^{-2s})(2r+2s-1)}
\end{aligned}
$$

Estimating the RHS of the local lemma

$$\begin{aligned}
x_i \prod_{(i,j)\in E}(1-x_j) &\geq 2^{-2r}\exp\left[-2\left(2r\sum_{s=1}^{\infty}\frac{1}{2^{2s}}+\sum_{s=1}^{\infty}\frac{2s-1}{2^{2s}}\right)\right] \\
&\geq 2^{-2r}\exp\left[-2\left(2r\left(\frac{1}{3}\right)+\frac{5}{9}\right)\right] \\
&\geq 2^{-2r}\exp\left(-\frac{4}{3}r-\frac{10}{9}\right).
\end{aligned}$$

The hypotheses of the local lemma are met if

$$2^r k^{-r/c} \leq 2^{-2r}\exp\left(-\frac{4}{3}r-\frac{10}{9}\right).$$

Applying the local lemma

- Taking logarithms, we require

$$r \log 2 - \frac{r}{c} \log k \leq -2r \log 2 - \frac{4}{3} r - \frac{10}{9}.$$

- Rearranging terms, we require

$$c \left(3 \log 2 + \frac{4}{3} + \frac{10}{9r} \right) \leq \log k.$$

- The left side of this inequality is largest when $r = 1$, so we define

$$d_1 = 3 \log 2 + 4/3 + 10/9,$$

  and insist that $c \cdot d_1 \leq \log k$.
- For $k \geq e^{c \cdot d_1}$, the local lemma implies that with positive probability, w is $1/c$-similar.

The Infinity Lemma

- Since $N = |w|$ is arbitrary, there must exists arbitrarily large such w.
- The Local Lemma only applies to finitely many events.
- We can only use it to show the existence of finite (but arbitrarily large) words with a given property.
- To show the existence of an infinite word with the desired property we use König's Infinity Lemma.

Lemma (König)

Let A be any infinite set of finite words. There exists an infinite word $\mathbf{w}$ *such that every prefix of* $\mathbf{w}$ *is a prefix of some word in A.*

- It now follows that there exists an infinite $1/c$-similar word.

Avoiding approximate repetitions

Definition

A word xx' with $|x| = |x'|$ is a c-approximate square if $d(x, x') \leq c$.

Example

- `riffraff` is a 1-approximate square.
- `murmur` is a 0-approximate square (i.e., a square).

- In the biological sequence analysis literature, a c-approximate square is called a "c-approximate tandem repeat".
- They are typically studied from an algorithmic point of view: i.e., how to efficiently find c-approximate repeats in a string.
- We will consider questions of avoidability.

Avoiding approximate squares

Definition

A word z avoids c-approximate squares if for all its subwords xx' where $|x| = |x'|$ we have $d(x, x') \geq \min(c + 1, |x|)$.

We can prove the following over 4 letters.

Theorem

There is an infinite word over a 4-letter alphabet that avoids 1-approximate squares, and the 1 is best possible.

Avoiding approximate squares

Proof (sketch).

Let **c** be any squarefree word over $\{0, 1, 2\}$, and consider the image under the morphism h defined by

$$
\begin{array}{lcl}
0 & \rightarrow & \texttt{01203102312032103120132103201302132012301320312 3} \\
1 & \rightarrow & \texttt{012031023120321023103213021032013210312013203123} \\
2 & \rightarrow & \texttt{012031023012310213023103210231203210312013203123}
\end{array}
$$

The resulting word $\mathbf{d} = h(\mathbf{c})$ avoids 1-approximate squares. The rest of the argument is similar to that for the earlier result. □

Summary of results regarding additive similarity

We have the following results over larger alphabets.

Alphabet Size k	c	Morphism
6	2	$0 \rightarrow$ 012345 $1 \rightarrow$ 012453 $2 \rightarrow$ 012345
7	3	$0 \rightarrow$ 01234056132465 $1 \rightarrow$ 01234065214356 $2 \rightarrow$ 01234510624356
8	4	$0 \rightarrow$ 0123456071326547 $1 \rightarrow$ 0123456072154367 $2 \rightarrow$ 0123456710324765

Generalizing the construction

In fact it is possible to prove a general result.

Theorem

For all integers $n \geq 3$, there is an infinite word over an alphabet of $2n$ letters that avoids $(n-1)$-approximate squares.

Proof (sketch).

Consider the morphism h defined as follows:

$$
\begin{aligned}
0 &\rightarrow 012\cdots(n-1)n\cdots(2n-1) \\
1 &\rightarrow 012\cdots(n-1)(n+1)(n+2)\cdots(2n-1)n \\
2 &\rightarrow 012\cdots(n-1)(n+2)(n+3)\cdots(2n-1)n(n+1)
\end{aligned}
$$

If **w** is any squarefree word over $\{0,1,2\}$, then $h(\mathbf{w})$ has the desired properties. The proof is a generalization of previous arguments. □

Other similarity measures

Definition

The edit distance between two words u and v is the smallest number of insertions, deletions, or substitutions needed to transform u into v.

Theorem

There is an infinite word over 5 *letters such that all subwords* x *with* $|x| \geq 3$ *are neither squares, nor within edit distance* 1 *of any square.*

Proof (sketch).

A computer search shows that there is no such word over 4 letters. Over 5 letters we may apply the morphism

$$0 \rightarrow 01234 \qquad 1 \rightarrow 02142 \qquad 2 \rightarrow 03143.$$

to any square-free word to obtain the desired result. □

Thank you.