Non-constructive methods for avoiding repetitions in words

Narad Rampersad

Department of Mathematics and Statistics
University of Winnipeg

Avoidance in words

Many problems in combinatorics on words have the following form:

- Let S be a given set of words over an alphabet Σ.
- Does there exist an infinite word over the alphabet Σ that avoids S?
- That is, does there exist an infinite word $\mathbf{w}$ such that no factor of $\mathbf{w}$ is an element of S?

Avoiding squares

- e.g., $S = \{xx : x \in \{0, 1, 2\}^*\}$
- i.e., S is the set of squares over a 3-letter alphabet
- Thue (1906) showed that there is an infinite word over $\{0, 1, 2\}$ that avoids S.

Iterated morphisms

- Thue's demonstration of this result is constructive.
- He explicitly produces an infinite word with the desired property.
- This word is defined by iterating a morphism.

Non-constructive methods

- We focus on non-constructive methods for proving results of this type.
- for instance: the probabilistic method,
- or other counting arguments

Finitary and infinitary results

- we are looking for an infinite word avoiding a set S
- many of the techniques we will see only give arbitrarily large, finite words avoiding S
- the existence of an infinite word avoiding S can be obtained by a standard argument
- sometimes presented topologically as a compactness argument
- or derived combinatorially from König's tree lemma

König's Lemma (reformulated)

Theorem (König's Lemma)

Let X be an infinite set of finite words over an alphabet Σ. Then there is an infinite word $\mathbf{x}$ over Σ such that every factor of $\mathbf{x}$ is a factor of infinitely many words in X.

Comments on König's lemma

- König's lemma is itself a non-constructive result.
- Even if X is given effectively, the result does not give an explicit construction of the word $\mathbf{x}$.
- The result can be strengthened somewhat.

Uniformly recurrent words

- a word is recurrent if each of its factors occurs infinitely often
- it is uniformly recurrent if for each factor, the distance between consecutive occurrences of that factor is bounded
- König's result can be improved to get a uniformly recurrent word

A stronger version of König

Theorem (Furstenburg 1981)

Let X be an infinite set of finite words over an alphabet Σ. Then there is a uniformly recurrent word $\mathbf{x}$ over Σ such that every factor of $\mathbf{x}$ is a factor of infinitely many words in X.

Proved by Furstenburg using ergodic theory; combinatorial proof given by others, such as Justin and Pirillo; we present a proof due to Currie and Linek.

Proof of Furstenburg's result

Define

$$K = \{\mathbf{x} \in \Sigma^\omega : \text{every factor of } \mathbf{x} \text{ is a factor of infinitely many words in } X\}$$

Also define

$$\mathscr{S} = \{S \subseteq \Sigma^* : \exists \mathbf{x} \in K, \mathbf{x} \text{ avoids } S\}$$

and let $\mathscr{S}$ be partially ordered by inclusion.

The proof is an application of Zorn's Lemma to $\mathscr{S}$.

Proof of Furstenburg's result

- $\mathscr{S}$ is non-empty (there must exist $\mathbf{x} \in K$, $n \in \mathbb{N}$, $a \in \Sigma$ such that $\mathbf{x}$ avoids a^n).
- We show every chain in $\mathscr{S}$ has an upper bound in $\mathscr{S}$.
- Let $\{S_\alpha\}_{\alpha \in I}$ be such a chain.
- Let $S = \cup_\alpha S_\alpha$.
- Claim: $S \in \mathscr{S}$.

Proof of Furstenburg's result

- Define $S_n = \{s \in S : |s| \leq n\}$.
- Then $S_n \subseteq S_\beta$ for some $\beta \in I$.
- Since $S_\beta \in \mathscr{S}$, there exists $\mathbf{x} \in K$ such that $\mathbf{x}$ avoids S_β (and hence S_n).
- Let u_n be an arbitrary factor of $\mathbf{x}$ of length n.
- By König, there is an infinite word $\mathbf{y}$ such that every factor of $\mathbf{y}$ is a factor of infinitely many words of $\{u_n\}_{n\geq 1}$.

Proof of Furstenburg's result

- We have $\mathbf{y} \in K$.
- For every n, the word $\mathbf{y}$ avoids S_n, so $\mathbf{y}$ avoids S.
- So $S \in \mathscr{S}$.
- By Zorn's Lemma, $\mathscr{S}$ has a maximal element $\hat{S}$.
- Then there is some $\mathbf{z} \in K$ that avoids $\hat{S}$.

Proof of Furstenburg's result

- Claim: $\mathbf{z}$ is uniformly recurrent.
- Suppose to the contrary that there is some factor s such that there are arbitrarily large gaps between consecutive occurrences of s.
- Then there are arbitrarily large factors of $\mathbf{z}$ that avoid $\hat{S} \cup \{s\}$.
- By König there is an infinite word in K that avoids $\hat{S} \cup \{s\}$.
- So $\hat{S} \cup \{s\} \in \mathscr{S}$, contradicting the maximality of $\hat{S}$.

A typical use of Furstenburg's result

- given a set of forbidden words, we show the existence of arbitrarily large words avoiding the forbidden words
- by Furstenburg's result, there is an infinite, uniformly recurrent word avoiding the set of forbidden words
- next we begin to examine different methods for showing avoidability

An early use of the probabilistic method

One of the earliest uses of the probabilistic method in combinatorics on words was to prove:

Theorem (Beck 1981)

For any real $\epsilon > 0$, there exist an integer N_ϵ and an infinite binary word $\mathbf{w}$ such that for every factor x of $\mathbf{w}$ of length $n > N_\epsilon$, all occurrences of x in $\mathbf{w}$ are separated by a distance at least $(2 - \epsilon)^n$.

The Lovász local lemma

- The proof is based on a lemma from probabilistic combinatorics known as the Lovász local lemma.
- allows one to give lower bounds on the probability of an intersection of several events when there are dependencies among the events

Entropy compression

- Moser and Tardos (2010) gave an algorithmic version of the Lovász local lemma based on an argument known as entropy compression.
- led to many improvements on earlier results proved using the local lemma
- well-suited for applications to avoidability in words
- easier to use than the local lemma
- gives sharper results

An application of the method

Theorem (Grytczuk, Kozik, and Micek 2013)

For every sequence $L_1, L_2, \ldots$ of 4-element sets, there exists a squarefree word $s_1 s_2 \cdots$ such that $s_i \in L_i$ for all $i \geq 1$.

Squarefree words over 4 letters

- let's apply the method to show the existence of an infinite squarefree word over the alphabet $\{1, 2, 3, 4\}$
- there are squarefree words over a 3-letter alphabet, so this result is not optimal
- the idea is to give a randomized algorithm that attempts to generate a squarefree word
- then show that some sufficiently long execution of the algorithm must generate a long squarefree word

The algorithm

Input: n

1: $S = \epsilon, i = 1$
2: **while** $i \leq n$ **do**
3: randomly choose $r \in \{1, 2, 3, 4\}$ and append r to S
4: let $S = s_1 s_2 \cdots s_i$
5: **if** $s_1 s_2 \cdots s_i$ is squarefree **then**
6: set i to $i + 1$
7: **else** $s_1 s_2 \cdots s_i$ ends with a square xx
8: delete the second occurrence of x
9: set i to $i - |x|$
10: **end if**
11: **end while**

Analyzing the algorithm

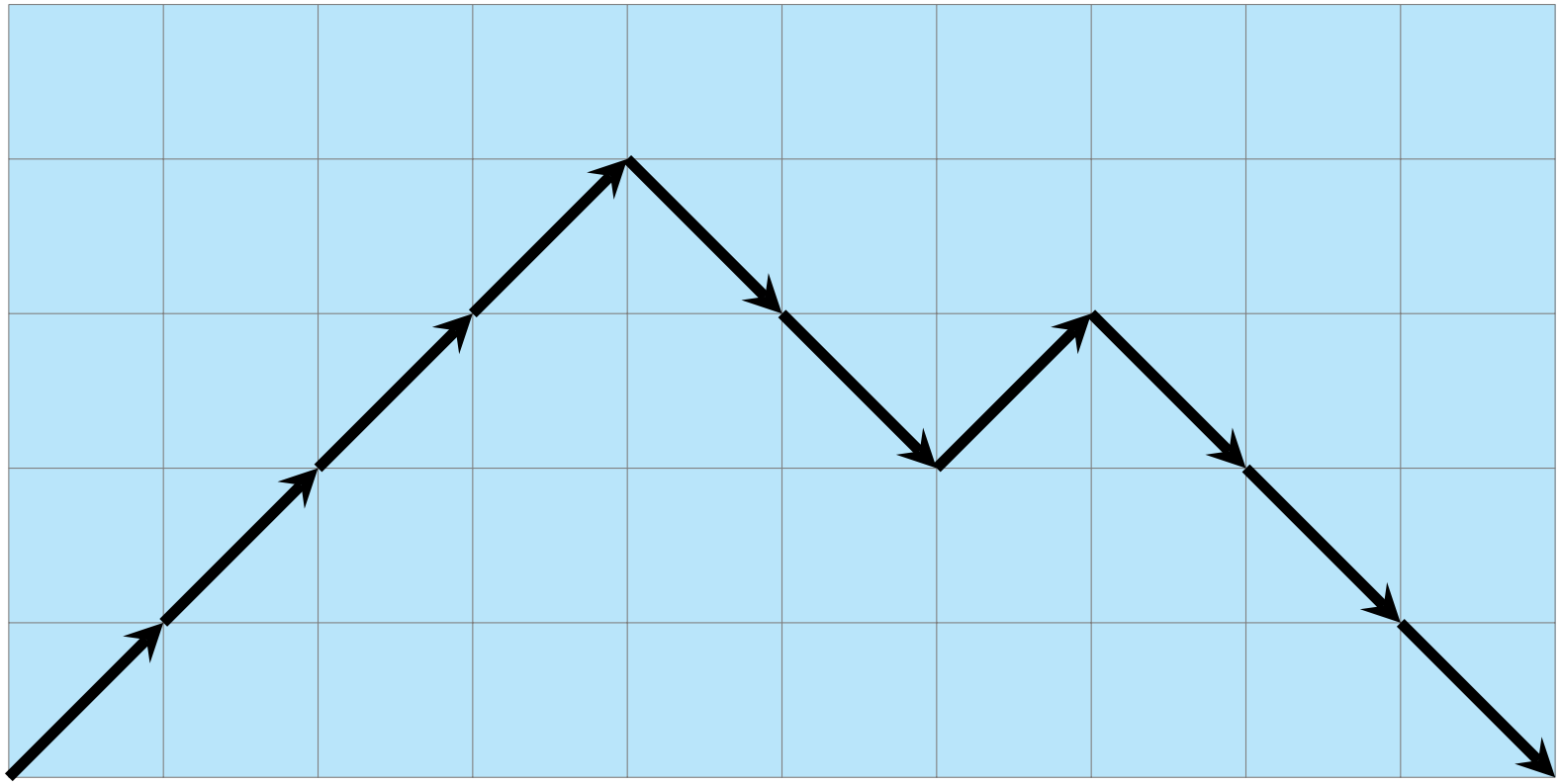
- fix n
- let M be the number of insertions (line 3) made during some execution of the algorithm
- let $r_1, r_2, \ldots, r_M$ be the sequence of random choices of letters inserted
- there are 4^M such sequences
- this sequence uniquely determines the execution of the algorithm

Another encoding of the execution

- we will describe the execution of the algorithm another way; i.e., by specifying:
  1. the sequence $S = s_1 s_2 \cdots s_i$ at the end of the execution, and
  2. the time and length of each deletion
- since each deletion consists of half of a square xx, the deleted block can be recovered from the first half still present in S
- with this information, we can describe the execution of the algorithm by running it backwards

Tracking the lengths of the deleted blocks

$$1 \to 12 \to 121 \to 1212 \to 12 \to 123$$



(if we terminate with a word of length i, we add i down-steps to the path)

Counting the number of paths

- the sequence of deletions is thus represented by a so-called Dyck path of length $2M$
- the number of such paths is the Catalan number

$$C_M = \binom{2M}{M}/(M+1)$$

- since the execution of the algorithm is uniquely determined by the final sequence S and this path, there are at most $(1/3)(4^{n+1}-1)C_M$ executions

Getting a contradiction

- suppose the algorithm fails to produce a squarefree word of length n
- so for M arbitrarily large, every execution of the algorithm fails to terminate after M steps
- from our two different counts of the number of executions, we have $4^M \leq (1/3)(4^n - 1)C_M$

The inequality fails to hold

Thus,

$$\begin{aligned} 4^M &\leq \left(\frac{4^n - 1}{3}\right) C_M \\ &\ll 4^n \left(\frac{4^M}{M^{3/2}\sqrt{\pi}}\right), \end{aligned}$$

which is not possible for M sufficiently large.

The contradiction means that some execution of the algorithm terminates after producing a squarefree word of length n.

Getting an infinite squarefree word

- so for all n there is a squarefree word of length n over $\{1, 2, 3, 4\}$
- by König's lemma, there is an infinite squarefree word over 4 letters
- we can do better, since we know that there are ternary squarefree words, but the method is very useful for showing the avoidability of more complicated patterns

A recent result using entropy compression

Theorem (Camungol and R. 2013)

Let $0 < \alpha < 1$ and let $k > 16^{1/\alpha}$ be an integer. Then there exists an infinite word $\mathbf{z}$ over a k-letter alphabet such that whenever xx' is a factor of $\mathbf{z}$ with $|x| = |x'|$, the length of the longest common subsequence of x and x' is at most $\alpha|x|$.

Generating functions

- Bell and Goh (2007) used another non-constructive method based on generating functions.
- This approach was originally due to Golod, and was used by Golod and Shafarevich to disprove several longstanding open problems in algebra.
- We give a version of the method specialized to our setting of combinatorics on words.

The combinatorial lemma of Golod

Theorem (Golod)

Let S be a set of words over a k-letter alphabet, each word of length at least 2. Suppose that for each $i \geq 2$, the set S contains at most c_i words of length i. If the power series expansion of

$$G(x) := \left(1 - kx + \sum_{i \geq 2} c_i x^i\right)^{-1}$$

has non-negative coefficients, then there are least $[x^n]G(x)$ words of length n over an k-letter alphabet that avoid S.

Patterns

- in his Ph.D. thesis, Cassaigne made a conjecture concerning the avoidability of long patterns
- a square is an instance of the pattern xx
- in general, a pattern can have more variables
- e.g., $01\,012\,01\,01\,012$ is an instance of the pattern $xyxxy$
- Which patterns are avoidable? What size of alphabet does one need to avoid a given pattern?

Cassaigne's Conjecture

The following was conjectured by Cassaigne in 1994 and proved in 2013 simultaneously and independently by Ochem and Pinlou and by Blanchet-Sadri and Woodhouse.

Theorem

Let p be a pattern with m distinct variables.

1. If $|p| \geq 3 \cdot 2^{m-1}$, then p is avoidable over a binary alphabet.
2. If $|p| \geq 2^m$, then p is avoidable over a ternary alphabet.

The proof techniques

- Pinlou and Ochem obtained this result by applying the Moser–Tardos entropy compression method.
- Blanchet-Sadri and Woodhouse proved it by using the generation function method based on Golod's lemma.

A criterion of Miller

Here is another criterion for avoidability, somewhat similar to the generating function approach.

Proposition (Miller 2011)

Let S be a set of non-empty words over a k-letter alphabet Σ. If there exists $c \in (1/k, 1)$ such that

$$\sum_{s \in S} c^{|s|} \leq kc - 1,$$

then there is an infinite word over Σ that avoids S.

Showing the existence of squarefree words

As an example, let's show that there are infinite squarefree words over a 7 letter alphabet (a very weak result!).

Let $k = 7$ and let $S = \{xx : x \in \Sigma^*\}$. Let $c \in (1/7, 1)$ be a constant to be specified later.

$$\begin{aligned}\sum_{s \in S} c^{|s|} &= \sum_{i \geq 1} c^{2i} 7^i \\ &= \sum_{i \geq 1} (7c^2)^i \\ &= \frac{1}{1 - 7c^2} - 1.\end{aligned}$$

Satisfying the inequality

We need

$$\frac{1}{1-7c^2} - 1 \leq 7c - 1$$

If $c = 0.22$ then LHS is approx. 0.512 and RHS is 0.54.

By Miller's criterion there is an infinite squarefree word over 7 letters.

Words with high Kolmogorov complexity

One can also show that there are infinite words for which every factor has fairly high Kolmogorov complexity.

Theorem (Durand, Levin, Shen 2001)

Let $0 < \alpha < 1$. There is an infinite binary word $\mathbf{x}$ such that every factor u of $\mathbf{x}$ satisfies $K(u) > \alpha|u| - O(1)$, where $K(u)$ denotes the prefix Kolmogorov complexity of u.

Proving the Kolmogorov complexity result

We do the obvious thing: avoid the low complexity words.

Define $b = -\log_2(1-\alpha) + 1$ and

$$S = \{s \in \{0,1\}^* : K(s) \leq \alpha|s| - b\}.$$

Set $c = 2^{-\alpha}$. Then

$$\sum_{s\in S} c^{|s|} = \sum_{s\in S} 2^{-\alpha|s|} \leq \sum_{s\in S} 2^{-K(s)-b} \leq 2^{-b} \sum_{s\in\{0,1\}^*} 2^{-K(s)} \leq 2^{-b},$$

where we have applied Kraft's Inequality in the last step.

It is a routine calculation to verify that $2^{-b} < 2c - 1$, so there is an infinite word avoiding S.

Avoiding sufficiently long forbidden words

Similarly, one can derive the following result, previously established by Rumyantsev and Ushakov (2006) using the Lovász local lemma.

Theorem (Rumyantsev and Ushakov 2006)

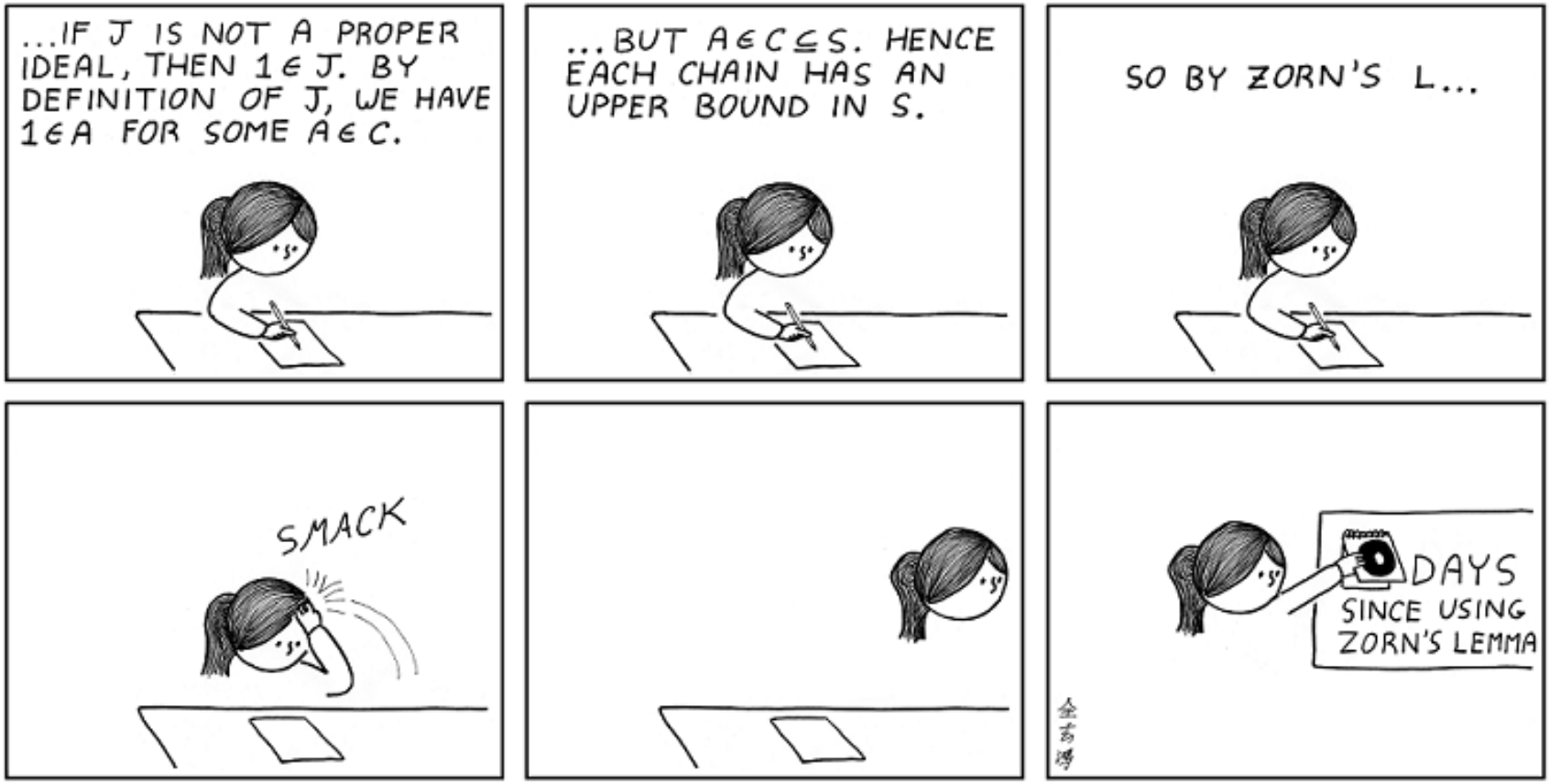
Let S be a set of non-empty words over a k-letter alphabet Σ and let $\alpha \in [0, 1)$. There is a positive integer d such that if S contains at most $k^{\alpha m}$ words of length m for each $m \geq d$, and none of length less than d, then there is an infinite word over Σ that avoids S.

Limitations of these methods

- often one does not obtain an optimal alphabet size
- none of these methods seem to be applicable to avoidance of abelian patterns
- e.g., an abelian square is a repetition xx', where x' can be obtained by rearranging the symbols of x
- the probability that a random word contains an abelian square seems to be too high for these methods to work

Conclusion

- These are just some selected examples to illustrate applications of some of these non-constructive methods.
- There are many other results on words that have been proved using these types of techniques.



("Zornaholic" – Abstruse Goose #133)

The End