On the Number of Distinct Subpalindromes in Words

Mikhail V. Rubinchik, Arseny M. Shur

Ural Federal University, Ekaterinburg, Russia

September 18, 2014

Palindromes and Squares

A palindrome is a word which is equal to its reversal, like

a	*i*	*b*	*o*	*h*	*p*	*h*	*o*	*b*	*i*	*a*

Palindromes and Squares

A palindrome is a word which is equal to its reversal, like

a	*i*	*b*	*o*	*h*	*p*	*h*	*o*	*b*	*i*	*a*

Palindromes are one of the most simple and common repetitions in words, along with squares, which are words consisting of two equal parts, like

c	*o*	*u*	*s*	*c*	*o*	*u*	*s*

Palindromes and Squares

A palindrome is a word which is equal to its reversal, like

a	i	b	o	h	p	h	o	b	i	a

Palindromes are one of the most simple and common repetitions in words, along with squares, which are words consisting of two equal parts, like

c	o	u	s	c	o	u	s

Palindromes are in some sense counterparts of squares:

- in a sequence of states of some finite-state machine, a square indicates repeated behaviour, while a palindrome shows that the machine reversed back to front;
- among the basic data structures, palindromes correspond to stacks, while squares correspond to queues; as a consequence, the language of all palindromes is context-free, while the language of all squares is not.

Counting Factors

We consider finite words over finite (k-letter) alphabets; we write $w = w[1..n]$ for a word of length n; words of the form $w[i..j]$ are factors of w.

A lot of results on the possible number of distinct palindromic factors and square factors in a word:

- max number of palindromes is n (Droubay, Pirillo, 2001);
- max number of squares is between $n - O(\sqrt{n})$ and $2n - O(\log n)$ (Ilie, 2007);
- min number of palindromes is k for $k \geq 3$ and 8 for $k = 2$;
- min number of squares is 0 for $k \geq 3$ (Thue, 1912) and 3 for $k = 2$ (Fraenkel, Simpson, 1995).

Counting Factors

We consider finite words over finite (k-letter) alphabets; we write $w = w[1..n]$ for a word of length n; words of the form $w[i..j]$ are factors of w.

A lot of results on the possible number of distinct palindromic factors and square factors in a word:

- max number of palindromes is n (Droubay, Pirillo, 2001);
- max number of squares is between $n - O(\sqrt{n})$ and $2n - O(\log n)$ (Ilie, 2007);
- min number of palindromes is k for $k \geq 3$ and 8 for $k = 2$;
- min number of squares is 0 for $k \geq 3$ (Thue, 1912) and 3 for $k = 2$ (Fraenkel, Simpson, 1995).

Counting Factors

We consider finite words over finite (k-letter) alphabets; we write $w = w[1..n]$ for a word of length n; words of the form $w[i..j]$ are factors of w.

A lot of results on the possible number of distinct palindromic factors and square factors in a word:

- max number of palindromes is n (Droubay, Pirillo, 2001);
- max number of squares is between $n - O(\sqrt{n})$ and $2n - O(\log n)$ (Ilie, 2007);
- min number of palindromes is k for $k \geq 3$ and 8 for $k = 2$;
- min number of squares is 0 for $k \geq 3$ (Thue, 1912) and 3 for $k = 2$ (Fraenkel, Simpson, 1995).

Problem

Find the expected number of distinct palindromic factors in a random k-ary word.

Simple Answer

Theorem

The expected number of distinct palindromic factors in a random word of length n over a fixed nontrivial alphabet is $\Theta(\sqrt{n})$.

Simple Answer

Theorem

The expected number of distinct palindromic factors in a random word of length n over a fixed nontrivial alphabet is $\Theta(\sqrt{n})$.

As a by-product of the technique used, we also get

Theorem (seems to be known before)

The expected number of distinct square factors in a random word of length n over a fixed nontrivial alphabet is $\Theta(\sqrt{n})$.

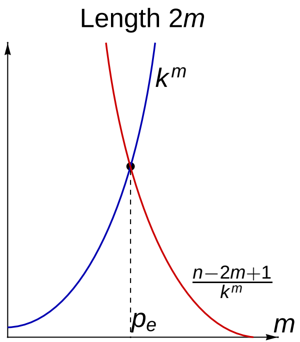
Some Explanations

Let k (alphabetic size) be fixed; $E(n)$ is the expectation studied. The expected number $E_m(n)$ of distinct palindromic factors of length m in a random word of length n is not greater than

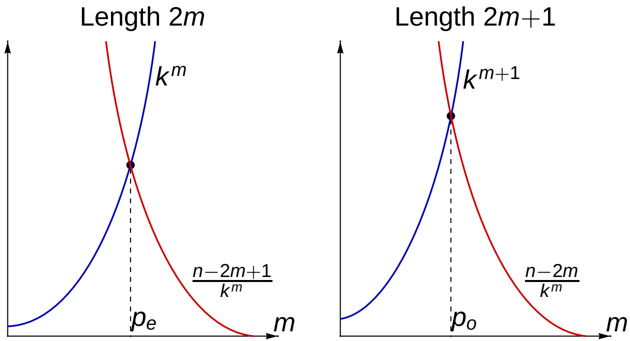
- ⋆ the total number of k-ary palindromes of length m;
- ⋆ the expected number of occurrences of palindromic factors of length m in a random word of length n.

Let k (alphabetic size) be fixed; $E(n)$ is the expectation studied. The expected number $E_m(n)$ of distinct palindromic factors of length m in a random word of length n is not greater than

- ⋆ the total number of k-ary palindromes of length m; blue
- ⋆ the expected number of occurrences of palindromic factors of length m in a random word of length n. red

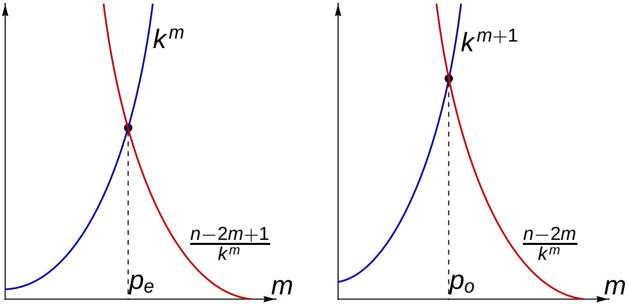


Some Explanations (Ctd)



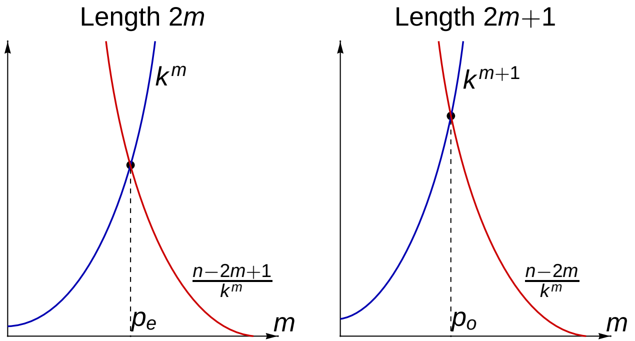
- $E(n) = \sum E_m(n)$ is bounded by the total area under the graphs;
- since all graphs are those of exponents, the area under each pair of graphs equals to the height of the highest point up to a constant multiple; thus, $E(n) = O(\sqrt{n})$;
- some additional considerations show that the upper bound is sharp up to a constant multiple, implying $E(n) = \Theta(\sqrt{n})$.

Some Explanations (Ctd)



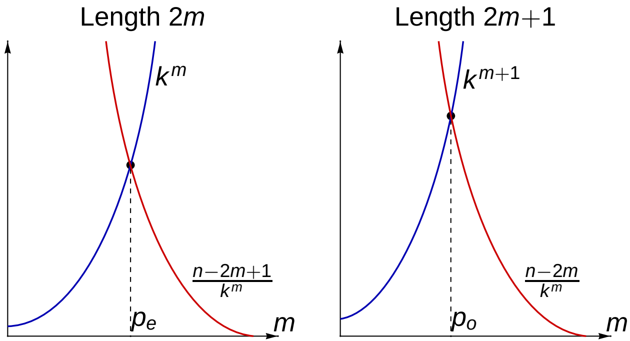
- $E(n) = \sum E_m(n)$ is bounded by the total area under the graphs;
- since all graphs are those of exponents, the area under each pair of graphs equals to the height of the highest point up to a constant multiple; thus, $E(n) = O(\sqrt{n})$;
- some additional considerations show that the upper bound is sharp up to a constant multiple, implying $E(n) = \Theta(\sqrt{n})$.

Some Explanations (Ctd)



- $E(n) = \sum E_m(n)$ is bounded by the total area under the graphs;
- since all graphs are those of exponents, the area under each pair of graphs equals to the height of the highest point up to a constant multiple; thus, $E(n) = O(\sqrt{n})$;
- some additional considerations show that the upper bound is sharp up to a constant multiple, implying $E(n) = \Theta(\sqrt{n})$.

Some Explanations (Ctd)



- $E(n) = \sum E_m(n)$ is bounded by the total area under the graphs;
- since all graphs are those of exponents, the area under each pair of graphs equals to the height of the highest point up to a constant multiple; thus, $E(n) = O(\sqrt{n})$;
- some additional considerations show that the upper bound is sharp up to a constant multiple, implying $E(n) = \Theta(\sqrt{n})$.

Dependence on k

Refinement of the obtained result: *consider* $E(n,k)$ *instead of* $E(n)$ *and find the dependence of the constant in the* $\Theta(\sqrt{n})$ *expression on* k.

Dependence on k

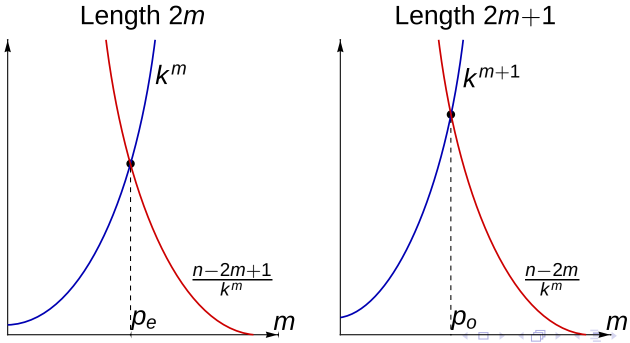
Refinement of the obtained result: *consider* $E(n,k)$ *instead of* $E(n)$ *and find the dependence of the constant in the* $\Theta(\sqrt{n})$ *expression on* k.

- intuition: more letters – more luck to get a palindrome;

Dependence on k

Refinement of the obtained result: *consider $E(n,k)$ instead of $E(n)$ and find the dependence of the constant in the $\Theta(\sqrt{n})$ expression on k.*

- intuition: more letters – more luck to get a palindrome;
- broken by the picture: the peak on the right graph is $\approx \sqrt{kn}$;



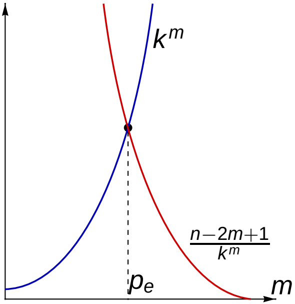
Dependence on k

Refinement of the obtained result: *consider* $E(n,k)$ *instead of* $E(n)$ *and find the dependence of the constant in the* $\Theta(\sqrt{n})$ *expression on* k.

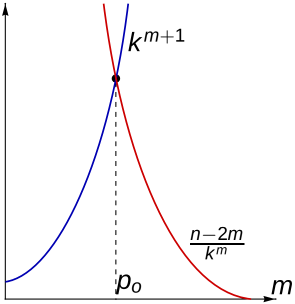
- intuition: more letters – more luck to get a palindrome;
- broken by the picture: the peak on the right graph is $\approx \sqrt{kn}$;
- is $E(n,k) = \Theta(\sqrt{kn})$? Not so easy.

Dependence on k (Ctd)

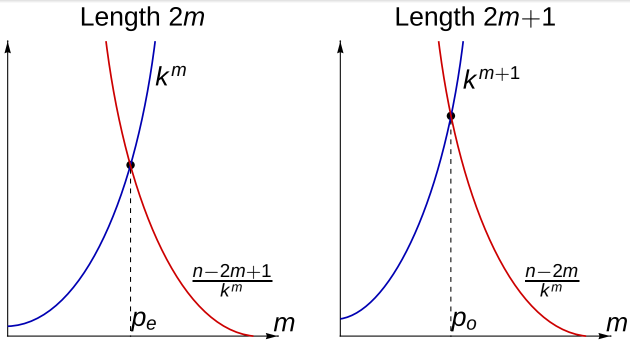
Length $2m$



Length $2m+1$

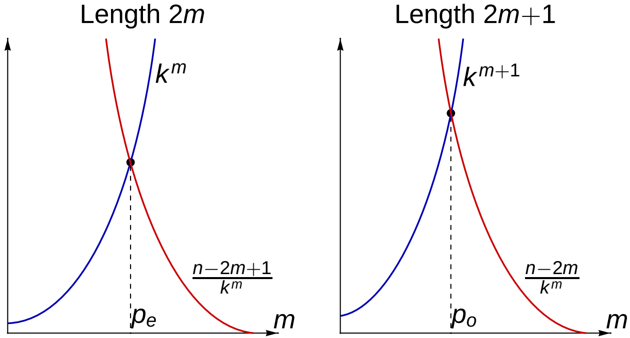


- If p_o is an integer , we get $\sqrt{kn}$;



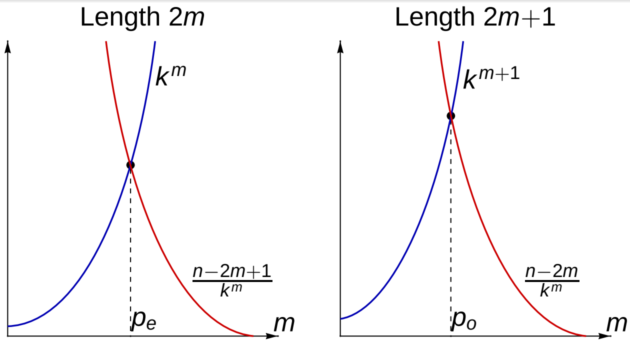
- If p_o is an integer [half-integer], we get $\sqrt{kn}$ $[2\sqrt{n}]$;

Dependence on k (Ctd)



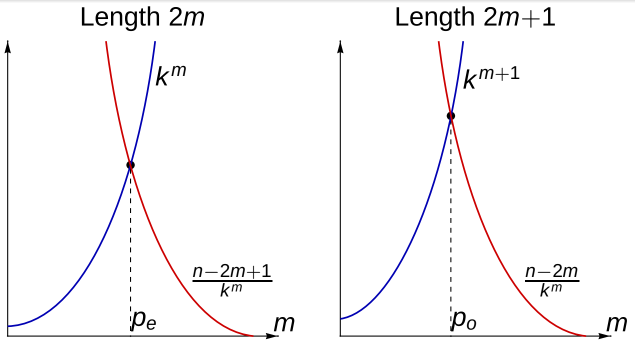
- If p_o is an integer [half-integer], we get $\sqrt{kn}$ $[2\sqrt{n}]$;
- similar for p_e, but the values are $\sqrt{k}$ times less;

Dependence on k (Ctd)

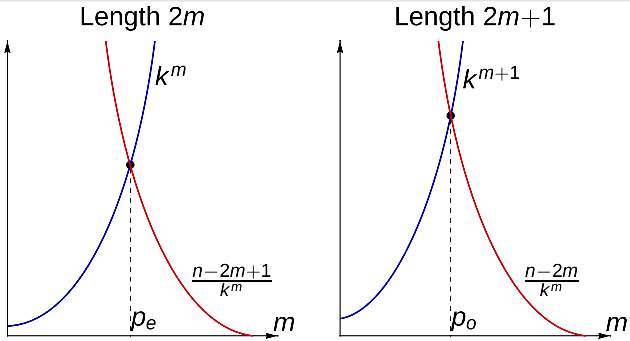


- If p_o is an integer [half-integer], we get $\sqrt{kn}$ [$2\sqrt{n}$];
- similar for p_e, but the values are $\sqrt{k}$ times less;
- note that $p_e \approx p_o + 1/2$;

Dependence on k (Ctd)



- If p_o is an integer [half-integer], we get $\sqrt{kn}$ $[2\sqrt{n}]$;
- similar for p_e, but the values are $\sqrt{k}$ times less;
- note that $p_e \approx p_o + 1/2$;
- ▶ our bound oscillates between the orders of $\sqrt{n}$ and $\sqrt{kn}$;



- If p_o is an integer [half-integer], we get $\sqrt{kn}$ $[2\sqrt{n}]$;
- similar for p_e, but the values are $\sqrt{k}$ times less;
- note that $p_e \approx p_o + 1/2$;
- our bound oscillates between the orders of $\sqrt{n}$ and $\sqrt{kn}$;
- more precisely, between $(3+\frac{4}{k-1})\sqrt{n}$ for $n \approx k^{2l}$ and $(1+\frac{4}{k-1})\sqrt{kn}$ for $n \approx k^{2l+1}$. What next?

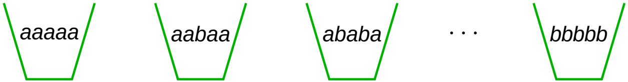
Suppose (even if this is not true) that for a random k-ary word of length n all events of type “to contain a given palindrome of length m” are independent and equiprobable.

Balls and Bins

Suppose (even if this is not true) that for a random k-ary word of length n all events of type "to contain a given palindrome of length m" are independent and equiprobable.

Balls: palindromic factors of length m of our random word

Bins: distinct palindromes of length m



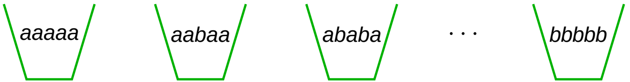
Balls and Bins

Suppose (even if this is not true) that for a random k-ary word of length n all events of type “to contain a given palindrome of length m” are independent and equiprobable.

Balls: palindromic factors of length m of our random word

$w[i_1..j_1]$ $w[i_2..j_2]$ $w[i_3..j_3]$ $w[i_4..j_4]$ $w[i_5..j_5]$ $\cdots$ $w[i_s..j_s]$

Bins: distinct palindromes of length m



Folklore Proposition

For N bins and CN balls, the expected number of empty bins is Ne^{-C}.

Testing The Model

Theorem (or not?)

The function $E(n,k)$ oscillates between its maximums

$$E(n,k) = \left(1 - \tfrac{1}{e} + \tfrac{4}{k-1} - \tfrac{k+1}{2(k^3-1)} - O\left(\tfrac{1}{ke^k}\right)\right)\sqrt{kn} + O\left(\tfrac{\sqrt{k}\log n}{\sqrt{n}}\right)$$

for p_o nearly integer and minimums

$$E(n,k) = \left(3 - \tfrac{1}{e} + \tfrac{4}{k-1} - \tfrac{k^2+1}{2(k^3-1)} - O\left(\tfrac{1}{e^k}\right)\right)\sqrt{n} + O\left(\tfrac{\sqrt{k}\log n}{\sqrt{n}}\right)$$

for p_e nearly integer.

Testing The Model

Theorem (or not?)

The function $E(n,k)$ oscillates between its maximums

$$E(n,k) = \left(1 - \frac{1}{e} + \frac{4}{k-1} - \frac{k+1}{2(k^3-1)} - O\left(\frac{1}{ke^k}\right)\right)\sqrt{kn} + O\left(\frac{\sqrt{k}\log n}{\sqrt{n}}\right)$$

for p_o nearly integer and minimums

$$E(n,k) = \left(3 - \frac{1}{e} + \frac{4}{k-1} - \frac{k^2+1}{2(k^3-1)} - O\left(\frac{1}{e^k}\right)\right)\sqrt{n} + O\left(\frac{\sqrt{k}\log n}{\sqrt{n}}\right)$$

for p_e nearly integer.

Experimental data for $C(k) = E(n,k)/\sqrt{n}$:

k	$C(k)$ for $p_e \approx$ integer		$C(k)$ for $p_o \approx$ integer	
	by Thm	experimental	by Thm	experimental
2	6.140	6.129 for $N = 2^{16}$	6.152	6.164 for $N = 2^{17}$
3	4.390	4.393 for $N = 3^{12}$	4.397	4.408 for $N = 3^{13}$
10	3.026	3.023 for $N = 10^6$	3.387	3.388 for $N = 10^7$
50	2.704	2.702 for $N = 50^4$	5.046	5.038 for $N = 50^3$

Bad news: our assumption was totally wrong, because the events "to contain a given palindrome of length m" are dependent and have different probabilities.

$$aaa \cdots aaa \text{ is less probable than } baa \cdots aab,$$

and each of them "suppresses" the other.

Bad news: our assumption was totally wrong, because the events "to contain a given palindrome of length m" are dependent and have different probabilities.

$$aaa \cdots aaa \text{ is less probable than } baa \cdots aab,$$

and each of them "suppresses" the other.

Why the predictions with balls and bins were so good?

Bad news: our assumption was totally wrong, because the events "to contain a given palindrome of length m" are dependent and have different probabilities.

$$aaa\cdots aaa \text{ is less probable than } baa\cdots aab,$$

and each of them "suppresses" the other.

Why the predictions with balls and bins were so good?

Good news: the probabilities for all palindromes of length m are quite close; moreover, for a $(\frac{k-2}{k})$th share of them the probability is exactly the same; the dependencies are also quite weak.

Still, this does not allow us to prove the theorem by the balls-and-bins method.

Further Analysis

A correct proof:

- find the probability that a random word contains a “typical” palindrome and estimate the probabilities for “rare” palindromes

Further Analysis

A correct proof:

- find the probability that a random word contains a “typical” palindrome and estimate the probabilities for “rare” palindromes
  - possible, thanks to the method of Guibas, Odlyzko (1981);

Further Analysis

A correct proof:

- find the probability that a random word contains a “typical” palindrome and estimate the probabilities for “rare” palindromes
  - possible, thanks to the method of Guibas, Odlyzko (1981);
  - $P(n,k,m) \approx 1 - \frac{\left(k - \frac{1}{k^{m-1}} - \frac{m-k-1}{k^{2m-1}}\right)^n}{\left(1 - \frac{m-1}{k^m} + \frac{(m-1)(m-2k)}{k^{2m}}\right) \cdot k^n}$

A correct proof:

- find the probability that a random word contains a "typical" palindrome and estimate the probabilities for "rare" palindromes
  - possible, thanks to the method of Guibas, Odlyzko (1981);
  - $P(n,k,m) \approx 1 - \frac{\left(k - \frac{1}{k^{m-1}} - \frac{m-k-1}{k^{2m-1}}\right)^n}{\left(1 - \frac{m-1}{k^m} + \frac{(m-1)(m-2k)}{k^{2m}}\right) \cdot k^n}$
- use the linearity of expectation to compute the expected number of palindromes for m close to p_e and p_o.

Further Analysis

A correct proof:

- find the probability that a random word contains a "typical" palindrome and estimate the probabilities for "rare" palindromes
  - possible, thanks to the method of Guibas, Odlyzko (1981);
  - $P(n,k,m) \approx 1 - \frac{\left(k - \frac{1}{k^{m-1}} - \frac{m-k-1}{k^{2m-1}}\right)^n}{\left(1 - \frac{m-1}{k^m} + \frac{(m-1)(m-2k)}{k^{2m}}\right) \cdot k^n}$
- use the linearity of expectation to compute the expected number of palindromes for m close to p_e and p_o.

Theorem

The function $E(n,k)$ oscillates between its maximums

$$E(n,k) = \left(1 - \frac{1}{e} + \frac{4}{k-1} - \frac{k+1}{2(k^3-1)} - O\left(\frac{1}{ke^k}\right)\right)\sqrt{kn} + O\left(\frac{\sqrt{k}\log n}{\sqrt{n}}\right)$$

for p_o nearly integer and minimums

$$E(n,k) = \left(3 - \frac{1}{e} + \frac{4}{k-1} - \frac{k^2+1}{2(k^3-1)} - O\left(\frac{1}{e^k}\right)\right)\sqrt{n} + O\left(\frac{\sqrt{k}\log n}{\sqrt{n}}\right)$$

for p_e nearly integer.

Thank you for your attention!