A strong non-overlapping Dyck code

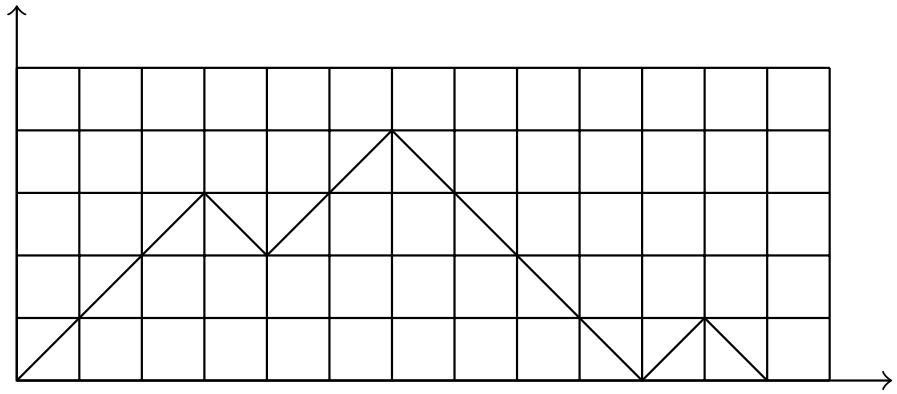
Elena Barcucci, Antonio Bernini, Renzo Pinzani

Dipartimento di Matematica e Informatica
Università degli Studi di Firenze

DLT 2021
International Conference on Developments in Language Theory
Porto (Portugal)
August 16–20, 2021

Preliminaries

A Dyck path P:



$\mathcal{D}_n = \{\text{set of Dyck paths having length } 2n,\ n \geq 0\}$

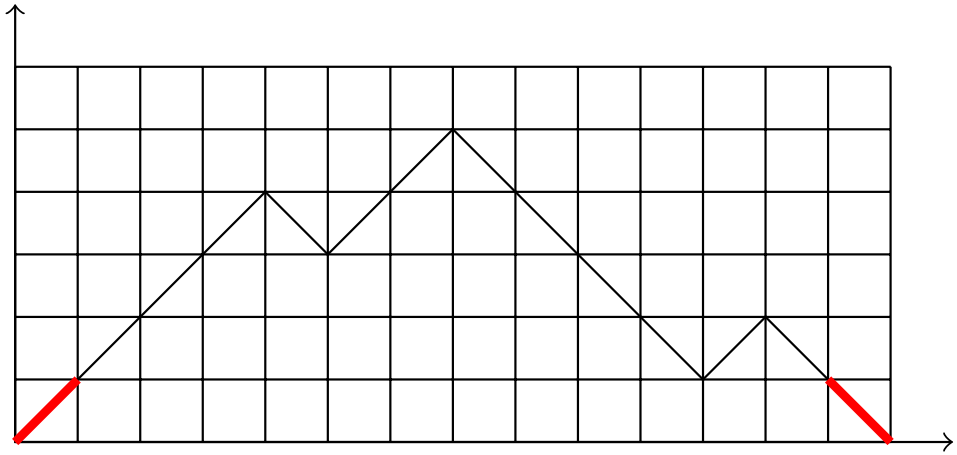
$|\mathcal{D}_n| = C_n = \frac{1}{n+1}\binom{2n}{n}$ (Catalan numbers)

P is codified into a binary words:

$$P \to 111011000010 \quad (P = 1^3 01^2 0^4 10)$$

Preliminaries

Elevated Dyck path:



The only points on the x- axis are the first one and the last one.

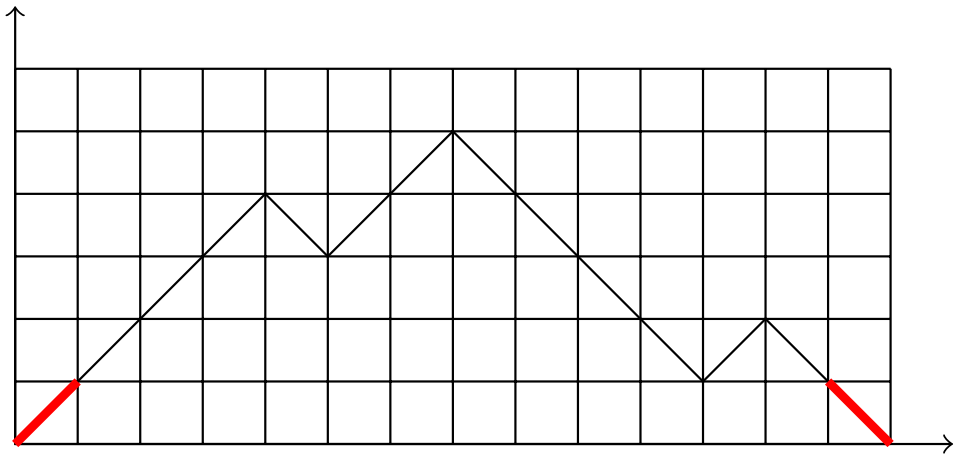
The set

$$\mathcal{ED}_n = \{1P0 | P \in \mathcal{D}_i,\ 0 \leq i \leq n-1\}$$

collect the elevated Dyck paths of length less or equal to $2n$.
$|\mathcal{E}D_n| = C_{n-1},\ n \geq 1$

Preliminaries

Elevated Dyck path:



The only points on the x- axis are the first one and the last one.

The set

$$\mathcal{ED}_n = \{1P0 | P \in \mathcal{D}_i,\ 0 \leq i \leq n-1\}$$

collect the elevated Dyck paths of length less or equal to $2n$.
$|\mathcal{E}D_n| = C_{n-1},\ n \geq 1$

Preliminaries: unbordered words

Definition

A word ω is said to be *unbordered* (or *self non-overlapping*, *bifix-free* or *self uncorrelated*) if and only if no proper prefix of ω is also a suffix of ω.

Example

111010100 is unbordered, while 101001010 contains two bifixes, 10 and 1010.

Definition

Two unbordered words v, v' are said to be *cross bifix-free* (or *non-overlapping* or *mutually uncorrelated*) if any proper prefix of v is different from any proper suffix of v', and vice versa.

Example

$v = 11011000$ and $v' = 11100100$ are cross bifix-free.

$v = \mathbf{10110}100$ and $v' = 1100\mathbf{10110}$ are not cross bifix-free.

Preliminaries: unbordered words

Definition

A word ω is said to be *unbordered* (or *self non-overlapping*, *bifix-free* or *self uncorrelated*) if and only if no proper prefix of ω is also a suffix of ω.

Example

111010100 is unbordered, while 101001010 contains two bifixes, 10 and 1010.

Definition

Two unbordered words v, v' are said to be *cross bifix-free* (or *non-overlapping* or *mutually uncorrelated*) if any proper prefix of v is different from any proper suffix of v', and vice versa.

Example

$v = 11011000$ and $v' = 11100100$ are cross bifix-free.

$v = \mathbf{10110}100$ and $v' = 1100\mathbf{110110}$ are not cross bifix-free.

Preliminaries: non-overlapping set

Definition

A set of words is said to be a *non-overlapping set* (or *cross bifix-free*) if each word is unbordered and if any two words are cross bifix-free.

Example

$\{11011000, 110100, 1100, 10\}$ is a cross bifix-free set of binary words.

$\{11011000, 110100, 1100\mathbf{10}, \mathbf{10}1100\}$ is not.

Proposition (Bilotta, 2017)

The set $\mathcal{ED}_n = \{1P0 \mid P \in \mathcal{D}_i,\ 0 \leq i \leq n-1\}$ *of elevated Dyck paths of length up to* $2n$ *is a non-overlapping set.*

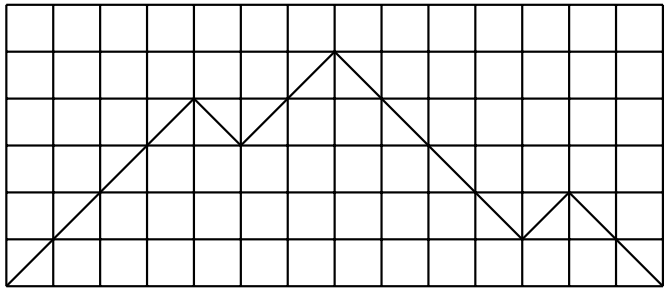
Preliminaries: strong non-overlapping set

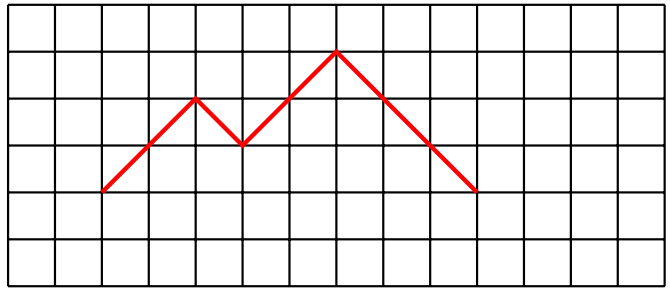
Definition

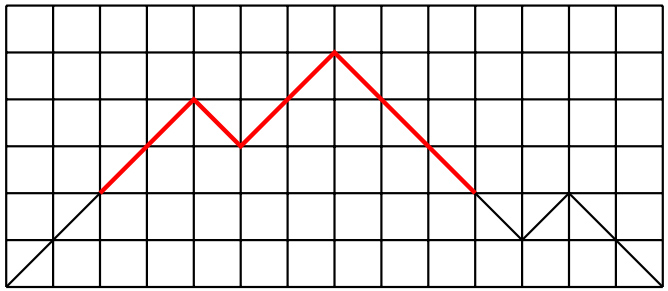
A set of words is said to be a *strong non-overlapping set* (or *cross bifix-free*) if it is a non-overlapping set of words and if every word is not an inner factor of any other word in the set.

Example

The set $\mathcal{ED}_n$ *is not* a strong non-overlapping set.







Aim and motivations

Aim

Construction of a strong non-overlapping set of binary words having variable lengths, using elevated Dyck paths.

Applications (among others...)

- telecommunication systems theory and engineering;
- study of DNA-based storage systems.

Aim and motivations

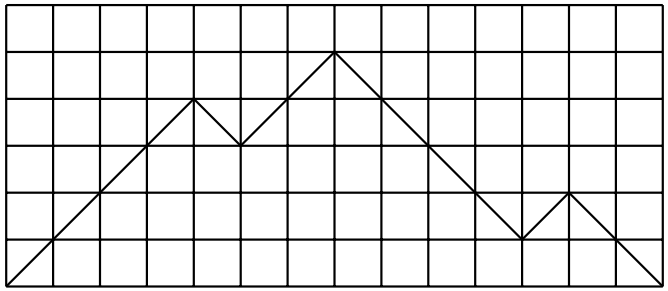
Yazdi et al. (2015) developed a random-access DNA-based storage architecture based on DNA sequences endowed with specialized address strings that may be used for selective information access. The addresses are designed to be mutually uncorrelated.

Constraints of the addresses:

1. *Constant GC content (close to 50%) for all the prefixes of the sequences of sufficiently long length.* "DNA strands with 50% GC content are more stable than DNA strands with lower or higher GC content and have better coverage during sequencing" $\rightarrow$ Mapping `A` (adenine) and `T` (thymine) by 0, and `G` (guanine) and `C` (cytosine) by 1, elevated Dyck paths have an equal number of 0's and 1's.
2. *Large mutual Hamming distance* "This reduces the probability of erroneous address selection" $\rightarrow$ Variable-length words: a word of length m and a word of length $m + d$ could be seen as two words having Hamming distance at least d.
3. *Uncorrelatedness of the addresses* "Addresses are used to provide unique identities for the blocks, their substrings should therefore not appear in "similar form" within other addresses" $\rightarrow$ Strong non-overlapping property.

The construction

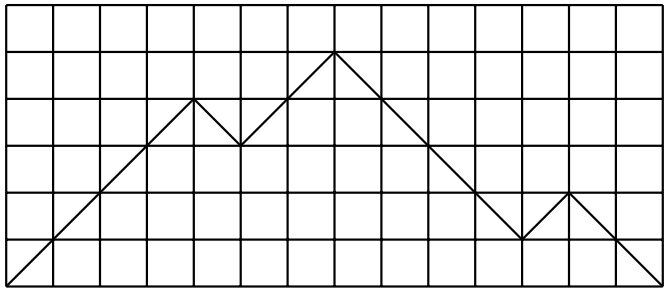
P elevated Dyck path:



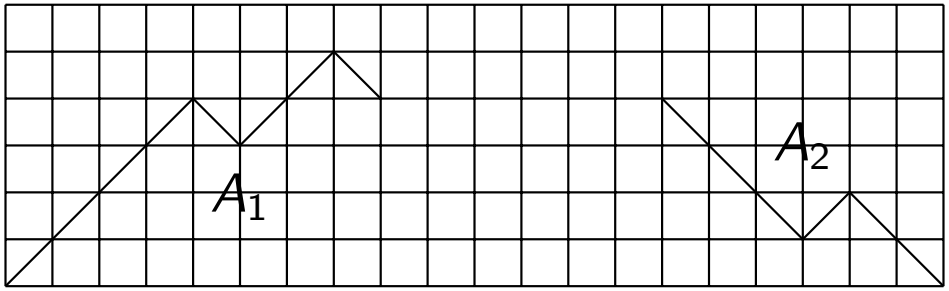
Split P in some point:

The construction

P elevated Dyck path:

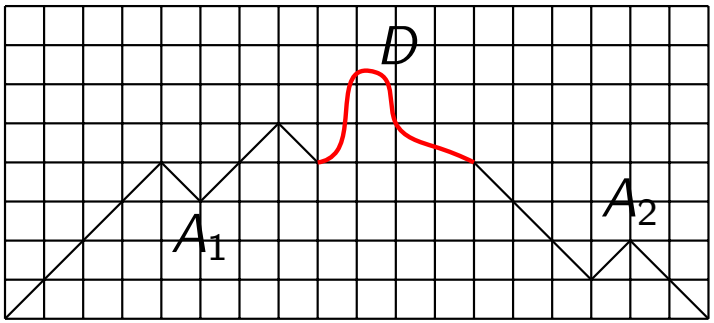


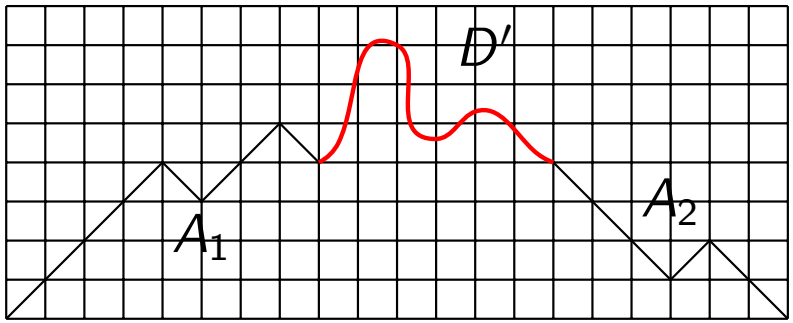
Split P in some point:



The construction

Insert two different Dyck paths in the split point:

$Q =$ 

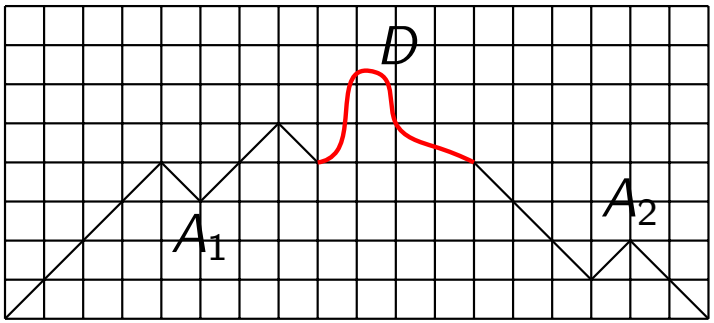
$Q' =$ 

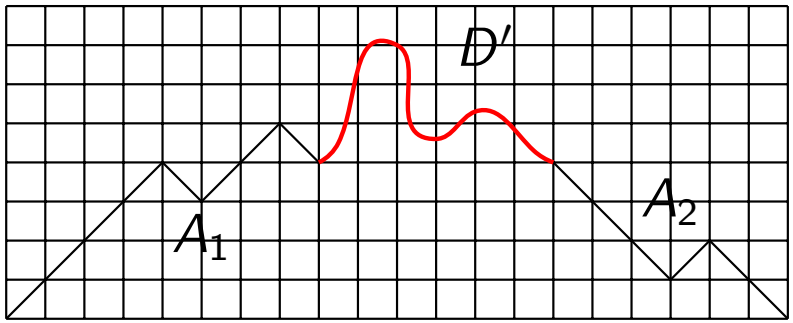
- Surely, the paths $Q = A_1 D A_2$ and $Q' = A_1 D' A_2$ are two non-overlapping paths (they are still two elevated Dyck paths);
- Most likely, they are strong non-ovelapping paths, since they have been generated by the inflation of P by means of two different Dyck paths (D and D').

Which are the hypotheses under which the strong non-overlapping property is guaranteed?

The construction

Insert two different Dyck paths in the split point:

$Q =$ 

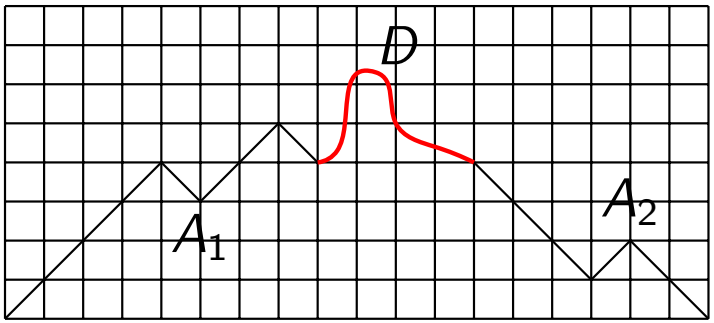
$Q' =$ 

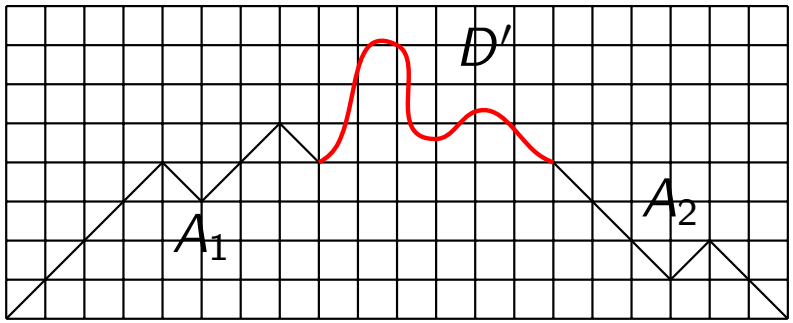
- Surely, the paths $Q = A_1 D A_2$ and $Q' = A_1 D' A_2$ are two non-overlapping paths (they are still two elevated Dyck paths);
- Most likely, they are strong non-ovelapping paths, since they have been generated by the inflation of P by means of two different Dyck paths (D and D').

Which are the hypotheses under which the strong non-overlapping property is guaranteed?

The construction

Insert two different Dyck paths in the split point:

$Q =$ 

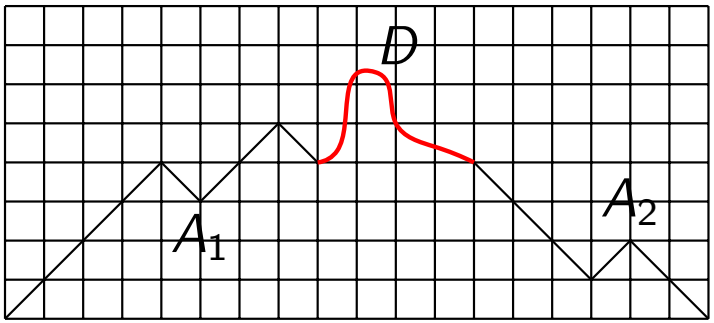
$Q' =$ 

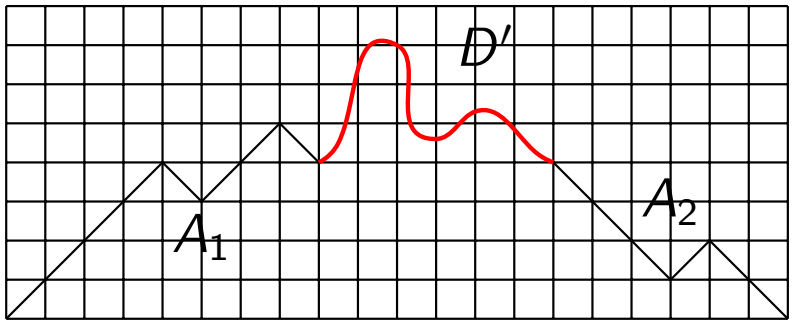
- Surely, the paths $Q = A_1 D A_2$ and $Q' = A_1 D' A_2$ are two non-overlapping paths (they are still two elevated Dyck paths);
- Most likely, they are strong non-ovelapping paths, since they have been generated by the inflation of P by means of two different Dyck paths (D and D').

Which are the hypotheses under which the strong non-overlapping property is guaranteed?

The construction

Insert two different Dyck paths in the split point:

$Q =$ 

$Q' =$ 

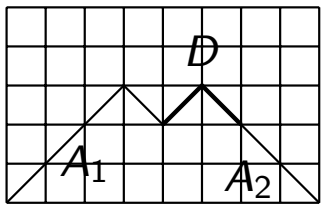
- Surely, the paths $Q = A_1 D A_2$ and $Q' = A_1 D' A_2$ are two non-overlapping paths (they are still two elevated Dyck paths);
- Most likely, they are strong non-ovelapping paths, since they have been generated by the inflation of P by means of two different Dyck paths (D and D').

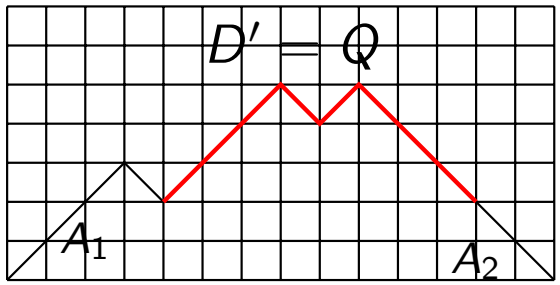
Which are the hypotheses under which the strong non-overlapping property is guaranteed?

The construction

Clearly, the path D' can not be too long: if $D' = Q$, then $Q' = A_1 Q A_2$ and Q and Q' are not strong non-overlapping.

Example

$Q =$ 

$Q' =$ 

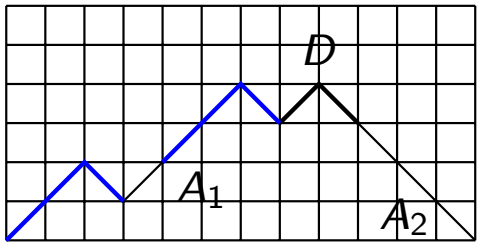
The length of D' must be less than the smallest length of the possible paths Q.

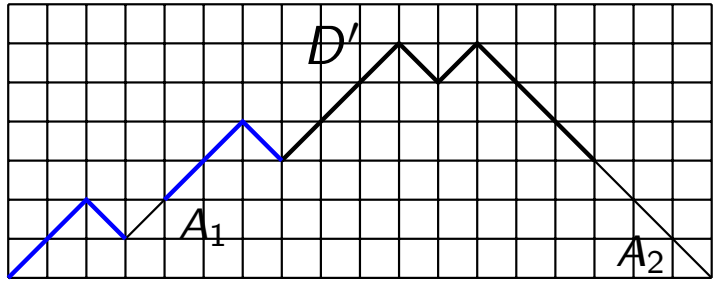
$$|D'| < \min_{Q}\{|Q|\} = |A_1\ 10\ 0^{h_{A_1}}| = |A_1| + 2 + h_{A_1}$$

(h_{A_1} is the final height of A_1)

Moreover, the prefix A_1 must be unbordered. If not, Q and Q' are not strong non-overlapping:

Example

$Q =$ 

$Q' =$ 

$Q =$

$Q' =$

With a different factorization of P (the prefix A_1 is unbordered), the paths Q and Q' are strong non-overlapping:

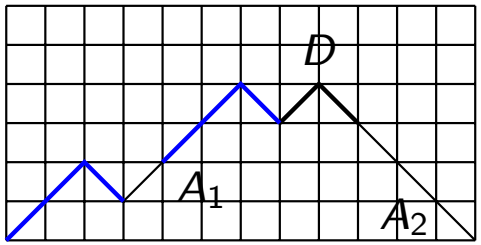
Example

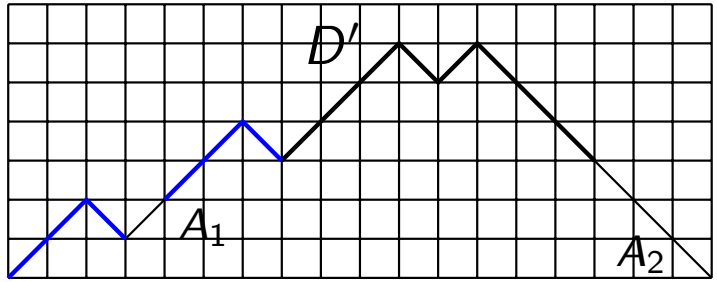
$Q =$

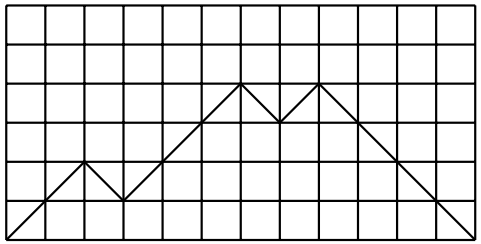
$Q' =$

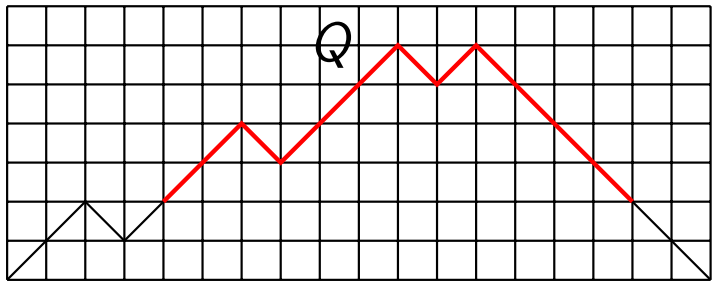
Moreover, the prefix A_1 must be unbordered. If not, Q and Q' are not strong non-overlapping:

Example

$Q =$ 

$Q' =$ 

$Q =$ 

$Q' =$ 

With a different factorization of P (the prefix A_1 is unbordered), the paths Q and Q' are strong non-overlapping:

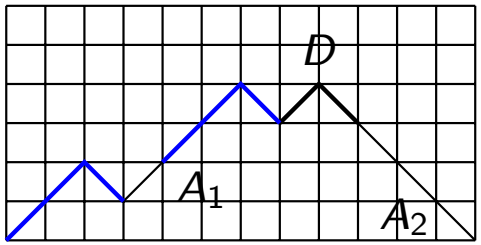
Example

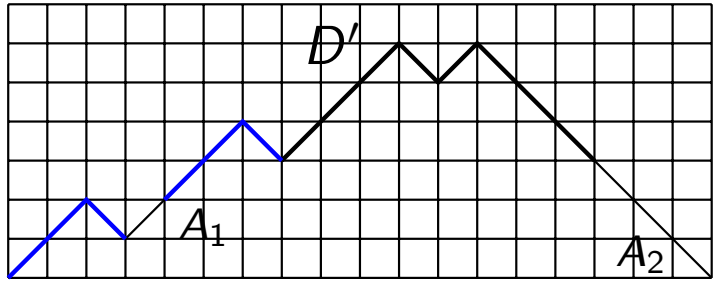
$Q =$

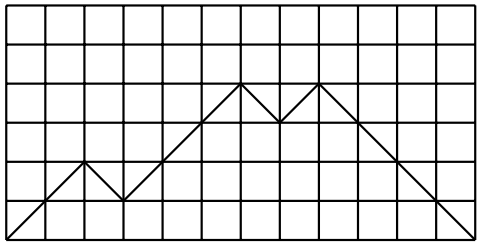
$Q' =$

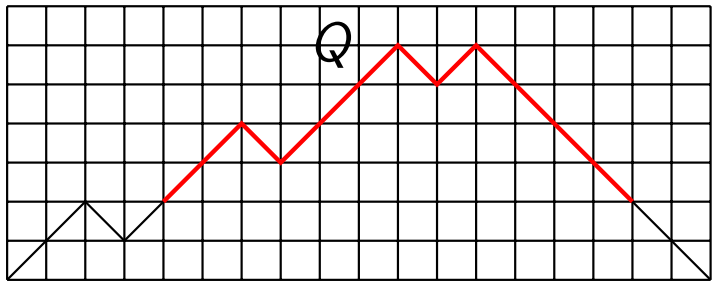
Moreover, the prefix A_1 must be unbordered. If not, Q and Q' are not strong non-overlapping:

Example

$Q =$ 

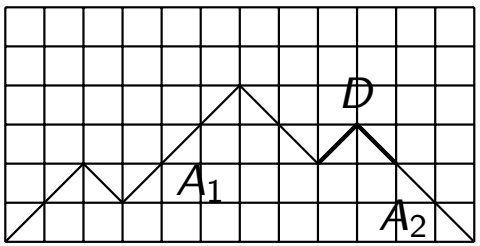
$Q' =$ 

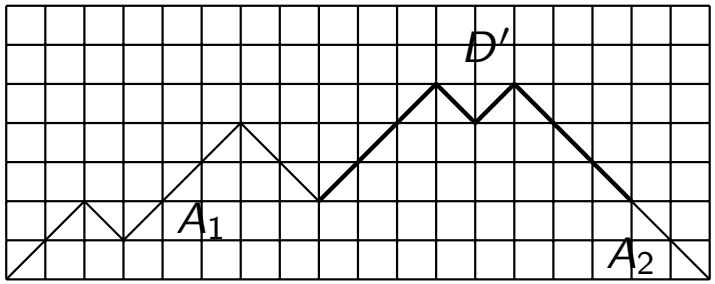
$Q =$ 

$Q' =$ 

With a different factorization of P (the prefix A_1 is unbordered), the paths Q and Q' are strong non-overlapping:

Example

$Q =$ 

$Q' =$ 

Summarizing, we have the following:

Proposition

Let $P = A_1A_2$ be an elevated Dyck path such that A_1 is an unbordered prefix of P. Let $|A_1|$ be the length of A_1 and h_{A_1} its final height. If D, D' (with $|D| < |D'|$) denote two Dyck paths such that $|D|, |D'| \leq |A_1| + h_{A_1}$, then the Dyck paths

$$Q = A_1DA_2$$

and

$$Q' = A_1D'A_2$$

are strong non-overlapping.

(Proof *ad absurdum*.)

Proposition

Let P be an elevated Dyck path and consider a factorization $P = A_1A_2$ where A_1 is an unbordered prefix with $|A_1| = \ell$ and $h_{A_1} = h$. The set

$$W_{A_1A_2} = \{A_1DA_2 | D \in \mathcal{D}_i,\ i \leq (\ell + h)/2\}$$

is a strong non-overlapping set of paths whose cardinality is $|W_{A_1A_2}| = \sum_{i=1}^{(\ell+h)/2} C_i$.

(Proof: just apply the first proposition to each couple of paths of $W_{A_1A_2}$)

Summarizing, we have the following:

Proposition

Let $P = A_1A_2$ be an elevated Dyck path such that A_1 is an unbordered prefix of P. Let $|A_1|$ be the length of A_1 and h_{A_1} its final height. If D, D' (with $|D| < |D'|$) denote two Dyck paths such that $|D|, |D'| \leq |A_1| + h_{A_1}$, then the Dyck paths

$$Q = A_1DA_2$$

and

$$Q' = A_1D'A_2$$

are strong non-overlapping.

(Proof *ad absurdum*.)

Proposition

Let P be an elevated Dyck path and consider a factorization $P = A_1A_2$ where A_1 is an unbordered prefix with $|A_1| = \ell$ and $h_{A_1} = h$. The set

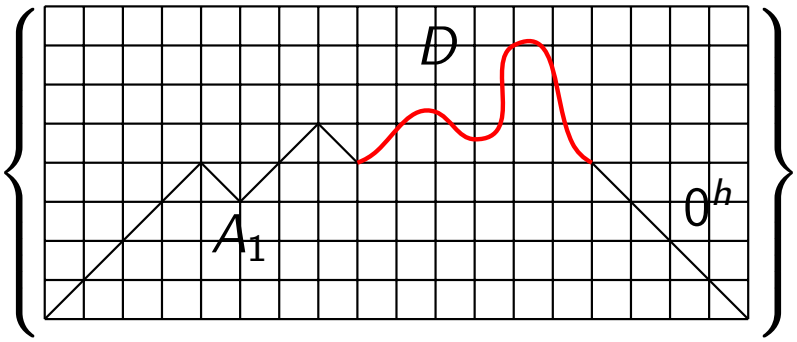
$$W_{A_1A_2} = \{A_1DA_2 | D \in \mathcal{D}_i,\ i \leq (\ell + h)/2\}$$

is a strong non-overlapping set of paths whose cardinality is $|W_{A_1A_2}| = \sum_{i=1}^{(\ell+h)/2} C_i$.

(Proof: just apply the first proposition to each couple of paths of $W_{A_1A_2}$)

Adding suffixes

The elevated Dyck paths contained in $W_{A_1A_2}$ have the same prefix A_1 and the same suffix A_2. In the following we expand the set by working at first on the suffixes, then on the prefixes.

$F =$ 

$F = \{A_1 D 0^h \mid D \in \mathcal{D}_i,\ i \leq (\ell + h)/2\}$

The length and the final heigth of A_1 are ℓ and h, respectively.

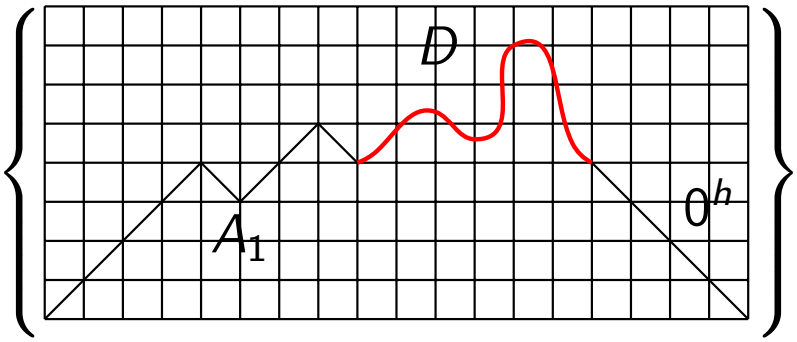
The set F is a non-overlapping set (previous proposition).

Aim: replace the suffix 0^h with suitable suffixes R in order to keep the strong non-overlapping property.

Surely, it is $|R| \geq h$, since the final path is an elevated Dyck path .

Adding suffixes

The elevated Dyck paths contained in $W_{A_1A_2}$ have the same prefix A_1 and the same suffix A_2. In the following we expand the set by working at first on the suffixes, then on the prefixes.

$F =$ 

$F = \{A_1 D 0^h \mid D \in \mathcal{D}_i,\ i \leq (\ell + h)/2\}$

The length and the final heigth of A_1 are ℓ and h, respectively.

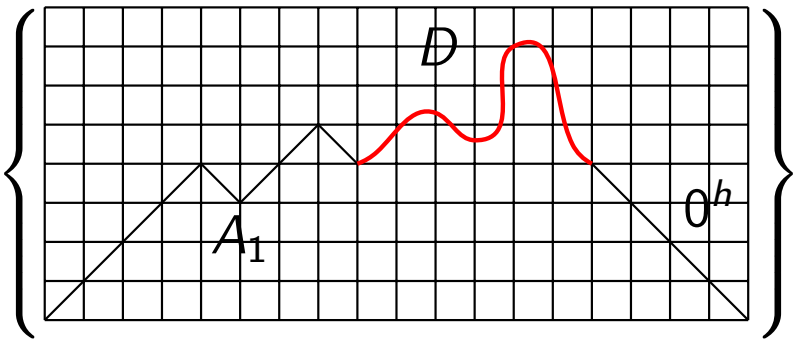
The set F is a non-overlapping set (previous proposition).

Aim: replace the suffix 0^h with suitable suffixes R in order to keep the strong non-overlapping property.

Surely, it is $|R| \geq h$, since the final path is an elevated Dyck path .

Adding suffixes

The elevated Dyck paths contained in $W_{A_1A_2}$ have the same prefix A_1 and the same suffix A_2. In the following we expand the set by working at first on the suffixes, then on the prefixes.

$F =$ 

$F = \{A_1 D 0^h \mid D \in \mathcal{D}_i,\ i \leq (\ell + h)/2\}$

The length and the final heigth of A_1 are ℓ and h, respectively.

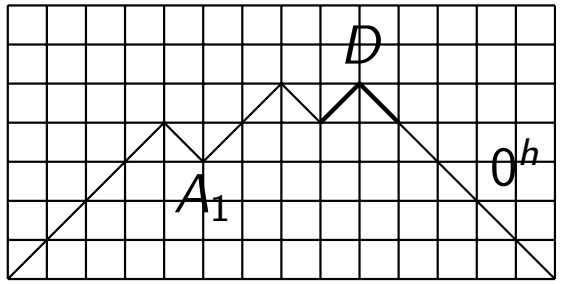
The set F is a non-overlapping set (previous proposition).

Aim: replace the suffix 0^h with suitable suffixes R in order to keep the strong non-overlapping property.

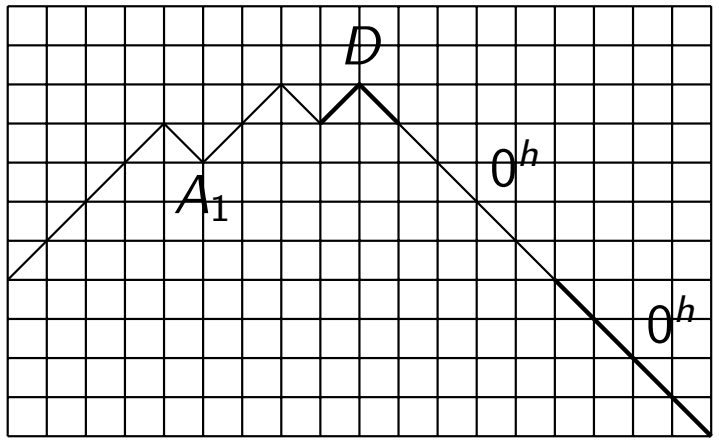
Surely, it is $|R| \geq h$, since the final path is an elevated Dyck path .

Adding suffixes

The shortest path in the set F is $Q = A_1\ 10\ 0^h$:

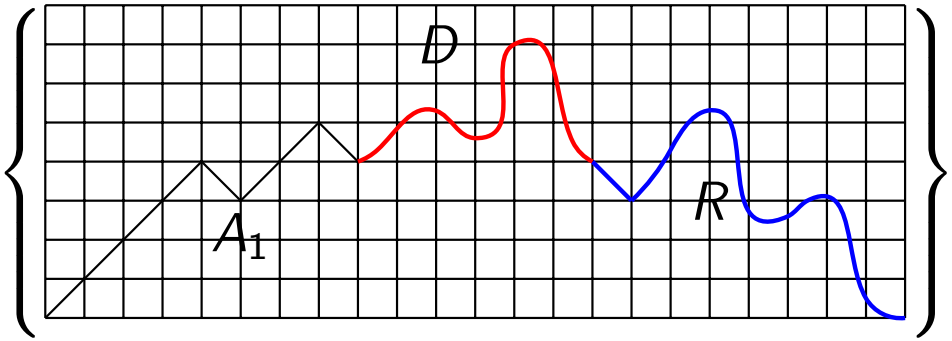
$Q =$ 

Let T be the shortest suffix containing Q (obtained by appending 0^h to Q, then $T = Q\ 0^h$):

$T =$ 

Adding suffixes

Then, the length of R must be less than the length of T: $|R| \leq \ell + 2h$.

$W_{A_1} =$ 

(The first step of R is a down step otherwise it is possible to obtain an identical path twice $\rightarrow$ problems in the enumeration of W_{A_1})

Denoting by $\mathcal{A}_2$ be the set containing the suffixes of the elevated Dyck paths starting with a down step, and defining the set:

$$W_{A_1} = \{A_1 DR | D \in \mathcal{D}_i,\ i \leq (\ell + h)/2,\ R \in \mathcal{A}_2,\ h \leq |R| \leq \ell + 2h\}\ .$$

we have the following:

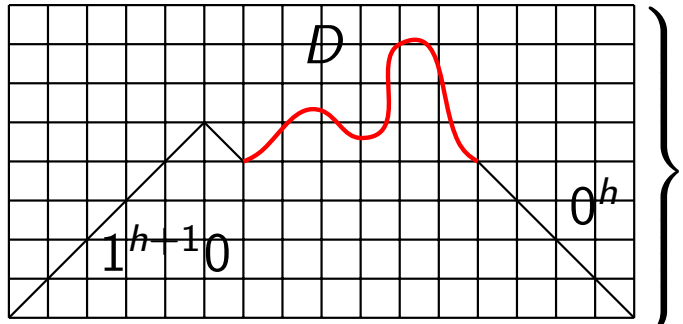
Proposition

The set W_{A_1} is strong non-overlapping.

Adding prefixes

Fix the height h reached by the prefixes.

$$F' = \left\{ \right. \quad \left. \right\} \quad F' = \{1^{h+1}0D0^h \mid D \in \mathcal{D}_i,\ i \leq (\ell + h)/2\}$$



The length and the final height of A_1 are ℓ and h, respectively.

The set F' is a non-overlapping set (previous proposition).

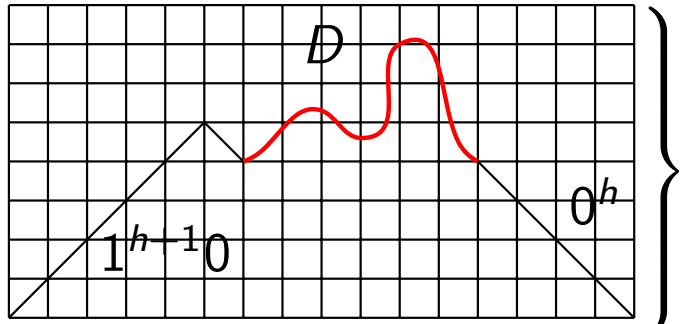
Aim: replace the prefix $1^{h+1}0$ with suitable prefixes L in order to keep the strong non-overlapping property.

Surely, it is $|L| \geq h + 2$, which is the length of the shortest undordered prefix to reach the height h.

Adding prefixes

Fix the height h reached by the prefixes.

$$F' = \left\{ \right. \quad \left. \right\} \quad F' = \{1^{h+1}0D0^h \mid D \in \mathcal{D}_i,\ i \leq (\ell + h)/2\}$$



The length and the final height of A_1 are ℓ and h, respectively.

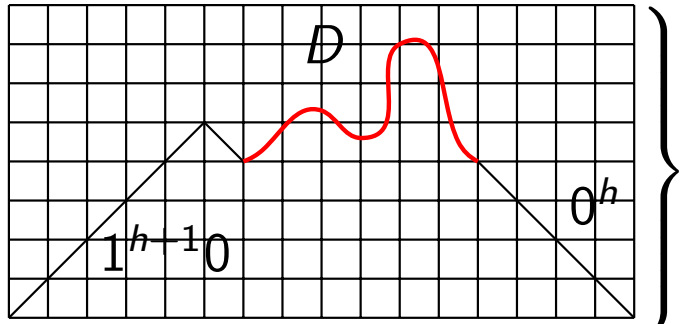
The set F' is a non-overlapping set (previous proposition).

Aim: replace the prefix $1^{h+1}0$ with suitable prefixes L in order to keep the strong non-overlapping property.

Surely, it is $|L| \geq h + 2$, which is the length of the shortest undordered prefix to reach the height h.

Adding prefixes

Fix the height h reached by the prefixes.

$$F' = \left\{ \right.$$ 

$$\left. \right\} \quad F' = \{1^{h+1}0D0^h \mid D \in \mathcal{D}_i,\ i \le (\ell + h)/2\}$$

The length and the final height of A_1 are ℓ and h, respectively.

The set F' is a non-overlapping set (previous proposition).

Aim: replace the prefix $1^{h+1}0$ with suitable prefixes L in order to keep the strong non-overlapping property.

Surely, it is $|L| \ge h + 2$, which is the length of the shortest undordered prefix to reach the height h.

Adding prefixes

It is possible to show that $|L| \leq 3h + 2$.

Moreover, all the unbordered prefixes L one can use to replace $1^{h+1}0$ must form a cross bifix-free set of elevated Dyck prefixes ending at height h.

Let $\mathcal{A}_1$ be the set containing the unbordered prefixes of the elevated Dyck paths ending at height h.

Let $X \subset \mathcal{A}_1$ be a cross bifix-free subset of $\mathcal{A}_1$.
We define the set

$$W^{(h)} = \{LDR \mid L \in X,\ h+2 \leq |L| \leq 3h+2,\ D \in \mathcal{D}_i,\ i \leq (|L|+h)/2,\\ R \in \mathcal{A}_2,\ h \leq |R| \leq |L|+2h,\}\ .$$

We have the following

Proposition

The set $W^{(h)}$ is strong non-overlapping

Enumeration (hints..)

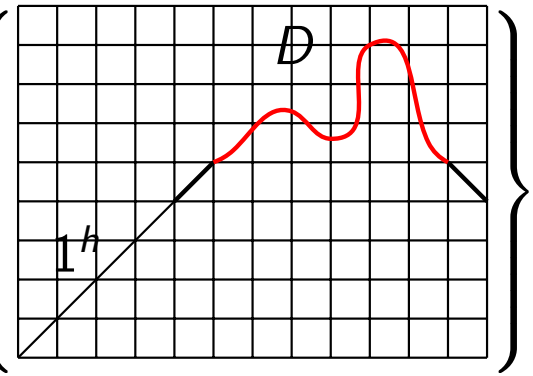
$$W^{(h)} = \{LDR \mid L \in X,\ h+2 \le |L| \le 3h+2,\ D \in \mathcal{D}_i,\ i \le (|L|+h)/2,\ R \in \mathcal{A}_2,\ h \le |R| \le |L|+2h,\}\ .$$

$$|W^{(h)}| = \sum_{k=h+2}^{3h+2} p_k^{(h)} \sum_{i=1}^{(k+h)/2} C_i \sum_{j=h}^{k+2h} s_j^{(h)}\ ,$$

- $s_j^{(h)}$ is the number of the suffixes of length j of elevated Dyck paths, starting at height h with a down step (known).
- C_i is the $i-$th Catalan number (known);
- $p_k^{(h)}$ is the number of prefixes of length k belonging to X.

A possible non-overlapping set of prefixes

$$\mathcal{A}_1 \supset X = \{1^h 1 D 0 \mid D \in \mathcal{D}_t, t \geq 0\}$$

$$X = \left\{ \right.$$ 

$$\left. \right\}$$

Proposition

The set X is a non-overlapping set.

Proposition

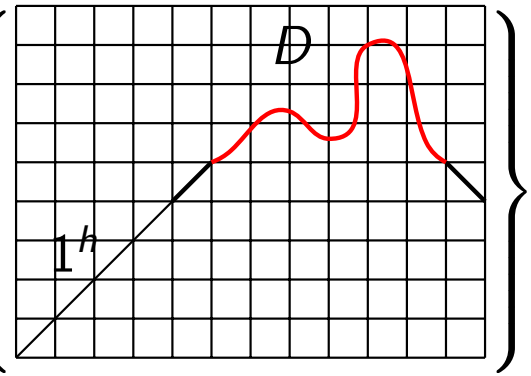
The set X is non-expandable in $\mathcal{A}_1$.

In other words, for each unbordered prefix of elevated Dyck paths $a \in \mathcal{A}_1 \setminus X$, there exists a prefix $x \in X$ such that a and x are not non-overlapping (or, equivalently, the set $X \cup a$ is not a non-overlapping set).

A possible non-overlapping set of prefixes

$$\mathcal{A}_1 \supset X = \{1^h 1 D 0 \mid D \in \mathcal{D}_t, t \geq 0\}$$

$$X = \left\{ \quad \right.$$



$$\left. \quad \right\}$$

Proposition

The set X is a non-overlapping set.

Proposition

The set X is non-expandable in $\mathcal{A}_1$.

In other words, for each unbordered prefix of elevated Dyck paths $a \in \mathcal{A}_1 \setminus X$, there exists a prefix $x \in X$ such that a and x are not non-overlapping (or, equivalently, the set $X \cup a$ is not a non-overlapping set).

Enumeration (hints...)

For our construction we have to consider prefixes with length up to $3h+2$. Therefore, we define:

$$X^{(3h+2)} = \{1^h 1 D 0 \mid D \in \mathcal{D}_t,\ 0 \le t \le h\}$$

and consider

$$W'^{(h)} = \{LDR \mid L \in X^{(3h+2)}, D \in \mathcal{D}_i,\ i \le (|L|+h)/2, \\ R \in \mathcal{A}_2,\ h \le |R| \le |L| + 2h, \}\ .$$

whose cardinality is given by:

$$|W'^{(h)}| = \sum_{t=0}^{h} C_t \sum_{i=1}^{t+h+1} C_i \sum_{j=h}^{2t+3h+2} s_j^{(h)}\ .$$

$$|W^{(h)}| = \sum_{k=h+2}^{3h+2} p_k^{(h)} \sum_{i=1}^{(k+h)/2} C_i \sum_{j=h}^{k+2h} s_j^{(h)}$$

Enumeration (hints...)

It is known:

- (Deutsch, 2011) $s_j^{(h)} = \begin{cases} 0, & \text{if either } j < h \text{ or } j-h \text{ is odd,} \\ \dfrac{h-1}{j-1}\binom{j-1}{\frac{j-h}{2}}, & \text{otherwise} \end{cases}$;
- (Dutton and Brigham, 1986) $C_t > \dfrac{2^{2t-1}}{t(t+1)\sqrt{\pi/(4t-1)}}$ for $t \geq 1$;
- (Topley, 2016) $\sum_{i=1}^{n} C_i > \dfrac{4^{n+1}}{3(n+1)\sqrt{\pi n^3}}$.

Using approximations for binomial coefficient and Stirling's approximation for factorial, we can obtain the following lower bound:

$$|W'^{(h)}| \geq \Theta_2(h) = \frac{128}{45e} \cdot (h+3)^{h-5} \cdot \frac{16^h - 1}{e^h}$$

...to be done..

Surely, the cardinality of our set should be more deeply investigated by:

- generating function of the sequence $\{|W'^{(h)}|\}_{h\geq 1}$, depending on h;
- extraction of the generic coefficient of g.f. would let to compare the cardinality of the set herein developed against the other ones in the literature.

Thank you for your attention!