TWINS IN WORDS AND SHUFFLE SQUARES

EMILY HUANG, IHYUN NAM, AND RISHUBH THAPER
ADVISOR: XIAOYU HE

August 2021

Abstract

For a word S over an alphabet Σ, we define $f(S)$ as the largest integer m such that there are two disjoint identical subwords, called twins, of length m in S. Let $f(n,\Sigma) = \min\{f(S) \colon S \in \Sigma^n\}$. Axenovich, Person, and Puzynina (2012) showed that $2f(n, \{0,1\}) = n - o(n)$; that is, nearly perfect twins exist in all binary words. In this paper, we describe a greedy algorithm for constructing large twins that results in a tighter lower bound on $f(n)$. We also enumerate related objects called *shuffle squares*, which is are words S for which $f(S) = |S|/2$.

Contents

1 Introduction

An *alphabet* Σ of size k is a set of k letters, which are conventionally $0, 1, \ldots, k-1$. A *word* $S = s_1 s_2 \cdots s_n$ over the alphabet Σ is a sequence $s_1, s_2, \ldots, s_n$ where $s_i \in \Sigma$ for all $1 \le i \le n$. A *subword* of S is a word $T = s_{i_1} s_{i_2} \cdots s_{i_t}$, where $1 \le i_1 < i_2 < \cdots < i_t \le n$ that can be found entirely in S. The sequence $(i_1, i_2, \ldots, i_t)$ is called the *support* of T and denoted $\mathrm{supp}(T)$. Given a word $S = s_1 s_2 \cdots s_n \in \Sigma^n$, its *reverse* S^R is equal to the word $s_n s_{n-1} \cdots s_1$.

The syntactical (structural) properties of words and their associated subwords have been investigated in the combinatorics of words and formal language theory. Some characteristic problems include reconstructing a word from its subwords, mapping words to matrices, and counting subword occurrences [8, 12, 14]. One of the most studied concepts, however, is the longest common subsequence (LCS) between a pair of words, with attention given to bounding LCS length and computing the LCS for any word pair [2, 3, 6, 10]. LCS has applications in many fields such as computational biology, since DNA is edited via insertions and deletions of base pairs [16].

Building on the idea of common subsequences, this paper examines the prevalence of identical disjoint subwords in words, called *twins*, over a given alphabet. In particular, we study a closely related object called *shuffle squares*.

Definition 1.1 (Twins)**.** Let $S \in \Sigma^n$ be a word of length n over the alphabet Σ. Let $T_1, T_2 \subset S$ be subwords such that $T_1 \cap T_2 = \emptyset$ and $T_1 = T_2$; that is, T_1 and T_2 are identical and disjoint. We call such subwords *twins*.

Definition 1.2 (Shuffle Square)**.** Let $S \in \Sigma^{2n}$ be a word of length $2n$ over the alphabet Σ. If there exist twins $T_1, T_2 \subset S$ such that $T_1 \cup T_2 = S$, then T_1 and T_2 are *perfect twins*, and we call S a *shuffle square*.

Definition 1.3 (Reverse Shuffle Square)**.** Let $S \in \Sigma^{2n}$ be a word of length $2n$ over the alphabet Σ. If there exist twins $T_1, T_2 \subset S$ such that $T_2 = T_1^R$, then we call S a *reverse shuffle square*.

The first occurrence of twins in the literature is a novel result by Axenovich, Puzynina, and Person [1] on the length of maximal twins in binary words. On the other hand, shuffle squares form the basis of a 2012 expository paper by Henshall, Rampersad, and Shallit [9], who listed several open problems regarding their complexity and enumeration.

The relevant theorems will be formally introduced in the next section, but it is worth outlining the general structure of the paper here. In Section 3, we provide the proof of the main theorem in [1]. In Section 4, we move on to shuffle squares and devise a greedy algorithm for constructing large twins. This immediately gives a lower bound on the number of binary shuffle squares, which we also prove in Section 4. In Sections 5 and 6, we finish up our examination by proving two novel asymptotic formulas on the number of shuffle squares and reverse shuffle squares over large alphabets.

The final section (7) is devoted to a conjecture on the complete enumeration of binary shuffle squares that we believe to be true based on numerical evidence.

2 Preliminaries

This section is divided into three parts, treating twins, shuffle squares, and useful combinatorial identities separately. The first part introduces the primary The third part is especially instrumental in proving the theorems in Sections 4, 5, and 6.

2.1 Twins

For a word $S \in \Sigma^n$, let $f(S)$ be the largest integer m such that there are twins of length m in S. Let

$$f(n, \Sigma) = \min\{f(S) \colon S \in \Sigma^n\}.$$

It is easy to see that $f(n, \Sigma) \leq \lfloor n/2 \rfloor$ for all positive integers n since no word can have twins of length greater than $\lfloor n/2 \rfloor$. A slightly more non-trivial observation for $f(n, \{0, 1\}$ is the following.

Lemma 2.1. *For all positive integers n, $f(n, \{0, 1\}) \geq \lfloor n/3 \rfloor$.*

Proof. Consider any $S \in \{0, 1\}^n$ and split it into consecutive triples. Each triple has either two zeros or two ones, so we can build a subword S_1 by choosing one repeated element from each triple and a subword S_2 by choosing the other repeated element. This results in two twins each of length $\lfloor n/3 \rfloor$. □

For example, if $S = 100110001$, then we can find twins of length $9/3 = 3$ equal to 010 by choosing the repeated element in each consecutive triple: $S = 101110001$. Here, one twin is colored blue and the other red.

In 2012, Axenovich, Person, and Puzynina [1] proved that $2f(n) = n - o(n)$; that is, nearly perfect twins exist in all binary words.

Theorem 2.2 (Axenovich, Person, and Puzynina, 2012)**.** *There exists an absolute constant C such that*

$$\left(1 - C\left(\frac{\log n}{\log \log n}\right)^{-1/4}\right) n \leq 2f(n, \{0, 1\}) \leq n - \log n.$$

The proof of Theorem 2.2 employs a regularity lemma to show that all words can be partitioned into blocks that look random in a weak sense. The lemma is analagous to Szemeredi's regularity lemma for graphs and is proved in a similar manner, by a classical density increment argument [11]. For its beauty and simplicity, we present the proof in its full form in Section **??**.

The most important implication of this result is that all binary words contain almost perfect twins. Our work extends this idea by considering words that can, in fact, be decomposed into two disjoint identical subwords. These objects are formally called shuffle squares, as described in the previous section.

2.2 Shuffle Squares

We first devise a greedy algorithm for constructing large twins that provides valuable insight into the ubiquity of binary shuffle squares. Although Rizzi and Vialette [13] recently determined that verifying binary shuffle squares is NP-complete, their exact quantity remains a mystery. However, our greedy algorithm locates a definitive portion of binary shuffle squares, thus providing a lower bound on their exact number.

Theorem 2.3. *For all positive integers n, $|\mathrm{SS}_2(n)| \geq \binom{2n}{n}$.*

The proof of Theorem 2.3 employs a bijection from binary shuffle squares found by the greedy algorithm to lattice paths from $(0,0)$ to $(2n,0)$, where each step is of size $(1,\pm 1)$. It will be elaborated in Section 4.

In the final part of this paper, we generalize our bijective methods to larger alphabets. In particular, we prove an asymptotic formula for the number of shuffle squares of length $2n$ over an alphabet of k letters (for large k), which was conjectured by Henshall, Rampersad, and Shallit [9] based on numerical evidence.

Theorem 2.4. *For large k ($k >> 2$) and all positive integers n,*

$$|\mathrm{SS}_k(n)| = \frac{1}{n+1}\binom{2n}{n}k^n - \binom{2n-1}{n+1}k^{n-1} + O_n(k^{n-2}).$$

By adjusting the machinery, we obtain a similar asymptotic formula for reverse shuffle squares.

Theorem 2.5. *For large k ($k >> 2$) and all positive integers n,*

$$|\mathrm{RSS}_k(n)| = \frac{1}{n+1}\binom{2n}{n}k^n - \frac{2n^3+9n^2-35n+30}{n^3+3n^2+2n}\binom{2n-2}{n-1} + O_n(k^{n-2}).$$

The proofs of Theorems 2.4 and 2.5 are presented in Sections 5 and 6, respectively.

2.3 Some Useful Identities

The proofs of Theorems 2.3 and 2.4 rely on several self-contained combinatorial identities on the Catalan numbers. For completeness, we review the Catalan numbers and list the relevant identities here.

Definition 2.6 (Catalan numbers)**.** The *Catalan numbers* $\{C_n\}$ are defined as $C_0 = 1$, and for all $n \geq 1$,

$$C_n = \sum_{k=0}^{n-1} C_k C_{n-1-k}.$$

It is well-known that $C_n = \frac{1}{n+1}\binom{2n}{n}$ for all nonnegative integers n.

Catalan numbers enumerate a variety of objects. The proofs in this paper invoke Dyck paths and 123-avoiding permutations, so we define them here.

Definition 2.7 (Dyck path)**.** A *Dyck path* of semilength n is a lattice path from $(0,0)$ to $(2n,0)$, where each step is of size $(1,\pm 1)$, that never crosses below the x-axis. The number of Dyck paths of semilength n is C_n.

Definition 2.8 (123-avoiding permutation)**.** Let $\mathcal{S}_n$ be the set of permutations on $[n]$. A permutation $\pi \in \mathcal{S}_n$ is called 123-*avoiding* if there do not exist $i_1 < i_2 < i_3$ such that $\pi(i_1) < \pi(i_2) < \pi(i_3)$. The number of 123-avoiding permutations on $[n]$ is C_n.

There are two identities that will be pertinent in the later part of this paper. The first is a simple Catalan convolution, which be instrumental in the proof of Theorem 4.4.

Proposition 2.9. *For $i = 0, 1, 2, \dots$, let C_i be the ith Catalan number. Then*

$$\sum_{k=0}^{n-1}(k+1)C_kC_{n-k-1} = \frac{1}{2}\binom{2n}{n}.$$

Proof. Define the sequence $\{a_n\}$ as follows: $a_0 = \frac{1}{2}$, and for all $k \geq 1$, $a_n = \sum_{k=0}^{n-1}(k+1)C_kC_{n-k-1}$. We want to show that $a_n = \frac{1}{2}\binom{2n}{n}$. The proof is by generating functions.

Denote by $a(x)$ the generating function for $\{a_n\}$; that is,

$$a(x) = a_0 + a_1x + a_2x^2 + \cdots .$$

Also, denote by $c(x)$ the generating function for the Catalan numbers, so that

$$c(x) = 1 + x + 2x^2 + 5x^3 + \cdots .$$

It is well known that $c(x) = \frac{1-\sqrt{1-4x}}{2x}$. Note that

$$\frac{1}{\sqrt{1-4x}} = (xc(x))' = \sum_{k=0}^{\infty}(k+1)C_kx^k,$$

and so

$$\begin{aligned} x \cdot \frac{1}{\sqrt{1-4x}} \cdot \frac{1-\sqrt{1-4x}}{2x} + \frac{1}{2} &= x(xc(x))'c(x) \\ &= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n-1}(k+1)C_kC_{n-k-1}\right)x^n + \frac{1}{2} \\ &= \sum_{n=0}^{\infty} a_nx^n \\ &= a(x). \end{aligned}$$

Thus,

$$a(x) = \frac{1-\sqrt{1-4x}}{2\sqrt{1-4x}} + \frac{1}{2} = \frac{1}{2\sqrt{1-4x}}.$$

Now, it is a simple exercise (see [15], p. 53, 2.5.11, or apply the extended Binomial Theorem) to find that

$$\frac{1}{\sqrt{1-4x}} = \sum_{n=0}^{\infty}\binom{2n}{n}x^n,$$

so

$$a(x) = \frac{1}{2}\sum_{n=0}\binom{2n}{n}x^n,$$

and $a_n = \frac{1}{2}\binom{2n}{n}$, as desired. □

A *valley* in a Dyck path is an instance of an up-step (size of $(1,1)$) followed by a down-step (size of $(1,-1)$). We will require the enumeration of valleys across all Dyck paths of semilength n for our proof of Theorem 2.4. The enumeration itself is certainly not novel (see [7] nad OEIS A002054), but the proof presented here is simpler and arguably more intuitive than those in the current literature.

Proposition 2.10. *The number of valleys across all Dyck paths of semilength n is $\binom{2n-1}{n+1}$.*

Proof. For $n \geq 0$, let V_n be the number of valleys across all Dyck paths of semilength n. We will derive a recursive formula for V_n that can be solved explicitly via generating functions.

For $1 \leq k \leq n$, let $\mathcal{D}_{n,k}$ be the set of Dyck paths of semilength n that return to the x-axis for the first time at the point $(2k, 0)$. Furthermore, let $V_{n,k}$ be the number of valleys across all such paths.

Each path $P \in \mathcal{D}_{n,k}$ looks like $uAdB$, where A is a Dyck path of semlength $k-1$ and B is a Dyck path of semilength $n-k$. Each of the valleys across all $A \in \mathcal{D}_{k-1}$ is counted C_{n-k} times, while each of the valleys across all $\beta \in \mathcal{D}_{n-k}$ is counted C_{k-1} times. Moreover, since $B = u\cdots$, there is another valley between the end of A and the start of B. This is counted $C_{k-1}C_{n-k}$ times. However, we must be careful to note that this valley only occurs for $k \leq n-1$, as the sub-path B is empty in $\mathcal{D}_{n,n}$.

Thus, we have the recursion

$$\begin{aligned} V_n &= \sum_{k=1}^{n-1} (V_{k-1}C_{n-k} + V_{n-k}C_{k-1} + C_{k-1}C_{n-k}) + (V_{n-1}C_0 + V_0C_{n-1}) \\ &= 2\sum_{k=0}^{n-1} V_k C_{n-1-k} + C_n - C_{n-1}. \end{aligned}$$

Let $v(x) = \sum_{n=0}^{\infty} V_n x^n$ be the generating function of the sequence $\{V_n\}$. Applying the "Snake Oil" method described in [15], we multiply both sides of the above recursion by x^n and sum over all $n \geq 1$ to obtain

$$v(x) = 2xv(x)c(x) + (1-x)c(x) - 1,$$

where $c(x)$ is the generating function of the Catalan numbers. Hence,

$$v(x) = \frac{c(x)(1-x)-1}{1-2xc(x)}.$$

Plugging in the closed form of $c(x)$ gives

$$v(x) = \frac{1}{\sqrt{1-4x}}\left(\frac{1-\sqrt{1-4x}}{2x}\right)(1-x) - \frac{1}{\sqrt{1-4x}}.$$

It is known that $\frac{1}{\sqrt{1-4x}}\left(\frac{1-\sqrt{1-4x}}{2x}\right) = \sum_{n=0}^{\infty}\binom{2n+1}{n}x^n$ and $\frac{1}{\sqrt{1-4x}} = \sum_{n=0}^{\infty}\binom{2n}{n}x^n$ ([15], p. 53-54, 2.5.11 and 2.5.15), so

$$\begin{aligned} V_n &= \binom{2n+1}{n} - \binom{2n-1}{n-1} - \binom{2n}{n} \\ &= \binom{2n-1}{n+1}, \end{aligned}$$

as desired. $\square$

We will refer back to these identities in Sections 4 and 5. For now, we return to the terminology of twins to prove Theorem 2.2, which is the primary literature background for our new results.

3 Maximal Twins in Binary Words

The main idea behind the proof of Theorem 2.2 is a regularity lemma for words, analagous to Szemeredi's regularity lemma for graphs. Before stating this lemma, we require some further definitions about word structure.

Definition 3.1 (Factor)**.** A *factor* of a word $S \in \Sigma^n$ is a subword of S consisting of consecutive elements of S, i.e., $s_i s_{i+1} \dots s_{i+m}$ for some $1 \leq i \leq n$ and $0 \leq m \leq n-i$. We denote such a subword $S[i, i+m]$, indicating that we are extracting the interval of elements $[s_i, s_{i+m}]$ from S.

Definition 3.2 (Density)**.** If S is a word over the alphabet Σ and $q \in \Sigma$, then we denote $|S|_q$ the number of elements in S equal to q. The *density* $d_q(S)$ of q in S is defined to be $|S|_q/|S|$, the fraction of elements in S equal to q.

For two subwords S' and S'' of S , we say that S' is contained in S'' if $\operatorname{supp}(S') \subseteq \operatorname{supp}(S'')$. Finally, if $S = s_1 s_2 \dots s_n$, $S[1, i] = A$, and $S[i+1, n] = B$, then we write $S = AB$ and call S a *concatenation* of A and B.

3.1 Regularity Lemma for Words

Density provides a natural way of defining regularity for words.

Definition 3.3 (ε-regular word)**.** Call a word $S \in \Sigma^n$ ε-regular if for every i, $\varepsilon n+1 \leq i \leq n-2\varepsilon n+1$ and every $q \in \Sigma$ it holds that

$$|d_q(S) - d_q(S[i, i+\varepsilon n - 1])| < \varepsilon. \tag{1}$$

Notice that in the case $\Sigma = \{0, 1\}$, $d_0(S) = 1 - d_1(S)$, so

$$\begin{aligned} |d_0(S) - d_0(S[i, i+\varepsilon n - 1)]| < \varepsilon &\iff |(1 - d_1(S)) - (1 - d_1(S[i, i+\varepsilon n - 1]))| < \varepsilon \\ &\iff |d_1(S) - d_1(S[i, i+\varepsilon n - 1)]| < \varepsilon. \end{aligned}$$

Thus, when $\Sigma = \{0, 1\}$, we shall let $d(S) = d_1(S)$.

Informally, regularity means that we can select any window of letters in S and be confident that the frequencies of letters in the window will not deviate too much from their frequencies in the entire word.

Definition 3.4 (ε-regular partition)**.** We call $\mathcal{S} \coloneqq (S_1, S_2, \dots, S_n)$ a *partition* of S if $S = S_1 S_2 \cdots S_n$ ($\mathcal{S}$ is a concatenation of consecutive S_i's). A partition $\mathcal{S}$ is an *ε-regular partition* of a word $S \in \Sigma^n$ if

$$\sum_{\substack{i \in [t], \\ S_i \text{ is not } \varepsilon \text{ regular}}} |S_i| \leq \varepsilon n,$$

i.e., the total length of ε-irregular subwords is at most εn.

The regularity lemma proper states that, given a certain number of parts, all reasonably large words can be decomposed into an ε-regular partition.

Theorem 3.5 (Regularity Lemma for Words). *For every $\varepsilon > 0$ and t_0 there is an n_0 and T_0 such that any word $S \in \Sigma^n$, for any $n \geq n_0$, admits an ε-regular partition of S into $S_1, S_2, \ldots, S_t$ with $t_0 \leq t \leq T_0$. In fact, $T_0 \leq t_0 3^{1/\varepsilon^4}$ and $n_0 = t_0 \varepsilon^{-\varepsilon^{-4}}$.*

The proof of the regularity lemma for words employs a similar idea to that of Szemerédi's regularity lemma for graphs, that of an *energy increment* argument.

The "energy" function Axenovich, et al. manufacture is a quantity called the *index*, which they associate with a specific partition of a word S. The idea is that if we repeatedly partition S into smaller and smaller parts, then at some level, the constraints on the partition index will necessitate the existence of an ε-regular partition.

Definition 3.6 (Index of a Partition). Let $\mathcal{S} \coloneqq (S_1, S_2, \ldots, S_t)$ be a partition of $S \in \Sigma^n$ into consecutive factors. We define

$$\operatorname{ind}(\mathcal{S}) = \sum_{q \in \Sigma} \sum_{i \in [t]} d_q(S_i)^2 \frac{|S_i|}{n}.$$

Further, for convenience we set $\operatorname{ind}_q(\mathcal{S}) = \sum_{i \in [t]} d_q(S_i)^2 \frac{|S_i|}{n}$.

We can see that $\operatorname{ind}_q(\mathcal{S})$ is a kind of weighted mean square of the q-densities of the factors in the partition, so $\operatorname{ind}(\mathcal{S})$ is the sum of the weighted mean squares of each letter density.

Since the proof of the regularity lemma involves repeated partitioning, we need to formally define the concept of partitioning a partition.

Definition 3.7 (Refinement of a Partition). Let $\mathcal{S} = (S_1, S_2, \ldots, S_t)$ and

$$\mathcal{S}' = \left(S'_{1,1}, S'_{1,2}, \ldots, S'_{1,s_1}, S'_{2,1}, S'_{2,2}, \ldots, S'_{2,s_2}, \ldots, S'_{t,1}, S'_{t,2}, \ldots, S'_{t,s_t}\right)$$

be partitions of $S \in \Sigma^n$. We say that $\mathcal{S}'$ *refines* $\mathcal{S}$ and write $\mathcal{S}' \preceq \mathcal{S}$ if, for every $i = 1, 2, \ldots, t$, $S_i = S'_{i,1} S'_{i,2} \ldots S'_{i,s_1}$.

The most important observation about the index (energy) of a partition is that it is *nondecreasing* across refinements.

Lemma 3.8. *Let $\mathcal{S}$ and $\mathcal{S}'$ be partitions of $S \in \Sigma^n$, and suppose $\mathcal{S}' \preceq \mathcal{S}$. Then*

$$\operatorname{ind}(\mathcal{S}') \geq \operatorname{ind}(\mathcal{S}).$$

Proof. Let $\mathcal{S} = (S_1, S_2, \ldots, S_t)$ and

$$\mathcal{S}' = \left(S'_{1,1}, S'_{1,2}, \ldots, S'_{1,s_1}, S'_{2,1}, S'_{2,2}, \ldots, S'_{2,s_2}, \ldots, S'_{t,1}, S'_{t,2}, \ldots, S'_{t,s_t}\right),$$

with $S_i = S'_{i,1} S'_{i,2} \cdots S'_{i,s_i}$ for all $1 \leq i \leq t$.

Take any $q \in \Sigma$. Then,

$$\begin{aligned} \operatorname{ind}_q(\mathcal{S}') &= \sum_{i=1}^{t} \sum_{j=1}^{s_i} d_q(\mathcal{S}'_{i,j})^2 \frac{|S'_{i,j}|}{n} \\ &= \sum_{i=1}^{t} \frac{|S_i|}{n} \sum_{j=1}^{s_i} d_q(\mathcal{S}'_{i,j})^2 \frac{|S'_{i,j}|}{|S_i|}, \end{aligned}$$

where in the second step we multiplied the sum by $|S_i|/|S_i| = 1$. Now, let $g(x) = x^2$, and let X_i be a random variable taking on the value $d_q(S'_{i,j})$ with probability $|S'_{i,j}|/|S_i|$, for each $j = 1, 2, \ldots, s_i$. Then, by Jensen's inequality,

$$\begin{aligned}
\operatorname{ind}_q(\mathcal{S}') &= \sum_{i=1}^{t} \frac{|S_i|}{n} \sum_{j=1}^{s_i} d_q(\mathcal{S}'_{i,j})^2 \frac{|S'_{i,j}|}{|S_i|} \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} \mathbb{E}[g(X_i)] \\
&\geq \sum_{i=1}^{t} \frac{|S_i|}{n} g\left(\mathbb{E}[X_i]\right) \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} \left(\sum_{j=1}^{s_i} d_q(S'_{i,j}) \frac{|S'_{i,j}|}{|S_i|} \right)^2 \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} \left(\sum_{j=1}^{s_i} \frac{|S'_{i,j}|_q}{|S'_{i,j}|} \cdot \frac{|S'_{i,j}|}{|S_i|} \right)^2 \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} \left(\sum_{j=1}^{s_i} \frac{|S'_{i,j}|_q}{|S_i|} \right)^2 \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} d_q\left(S_i\right)^2 \\
&= \operatorname{ind}_q(\mathcal{S}).
\end{aligned}$$

Hence,

$$\operatorname{ind}(\mathcal{S}') = \sum_{q\in\Sigma} \operatorname{ind}_q(\mathcal{S}') \geq \sum_{q\in\Sigma} \operatorname{ind}_q(\mathcal{S}) = \operatorname{ind}(\mathcal{S}),$$

as desired. □

Our main idea for the proof of the regularity lemma is repeatedly refining a given partition of a word S. We will show that at some stage of these successive refinements, there must be an ε-regular partition.

We start with a lemma that shows that if S is *not* ε-regular, then we can find a refinement (in this case, a first-level partition) of $S = (S)$ whose index exceeds the index of (S) by at least ε^3.

Lemma 3.9. *Let $S \in \Sigma^m$ be an ε-irregular word. Then there is a partition (A, B, C) of S such that $|A|, |B|, |C| \geq \varepsilon m$ and*

$$\operatorname{ind}((A, B, C)) \geq \operatorname{ind}((S)) + \varepsilon^3 = \left(\sum_{q\in\Sigma} d_q(S)^2 \right) + \varepsilon^3. \tag{2}$$

Proof. Since S is not ε-regular, there exists an element $q \in \Sigma$ and an i with $\varepsilon m+1 \leq i \leq m-2\varepsilon m+1$ such that $|d - d(S[i, i + \varepsilon m - 1]| \geq \varepsilon$, where $d \coloneqq d_q(S)$ and $d(T) \coloneqq d_q(T)$ for any subword T of S.

Assume, without loss of generality, that $d-d(S[i,i+\varepsilon m-1]\geq \varepsilon$, and set $\gamma \coloneqq d-d(S[i,i+\varepsilon m-1]$, $A \coloneqq S[1,i-1]$, $B \coloneqq S[i,i+\varepsilon m-1]$, and $C \coloneqq S[i+\varepsilon m, m]$. Furthermore, let $a \coloneqq |A| = i-1$, $b \coloneqq |B| = \varepsilon m$, and $c \coloneqq |C| = m-\varepsilon m - i + 1$. Observe that

$$|S|_q = d(A)a + d(B)b + d(C)c = dm, \quad d((A,C)) = \frac{dm-(d-\gamma)b}{a+c}, \quad d(B) = d-\gamma.$$

It is also easy to see that $a+c = m-b$ and $\mathrm{ind}_q((A,B,C)) = \mathrm{ind}_q((A,C,B))$. Note further that

$$\begin{aligned}
\mathrm{ind}_q((A,B,C)) &= d(A)^2\frac{a}{m} + +d(C)^2\frac{c}{m} + d(B)^2\frac{b}{m} \\
&= \frac{|A|_q^2}{am} + \frac{|C|_q^2}{cm} + d(B)^2\frac{b}{m} \\
&= \frac{1}{m(a+c)}\left(\frac{|A|_q^2}{a} + \frac{|C|_q^2}{c}\right)(a+c) + d(B)^2\frac{b}{m} \\
&\overset{\text{Cauchy-Schwarz}}{\geq} \frac{1}{m(a+c)}\left(|A|_q + |C|_q\right)^2 + d(B)^2\frac{b}{m} \\
&= \frac{1}{m(a+c)}|AC|_q^2 + d(B)^2\frac{b}{m} \\
&= d((A,C))^2\frac{a+c}{m} + d(B)^2\frac{b}{m}.
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{ind}_q((A,B,C)) &\geq d((A,C))^2\frac{a+c}{m} + d(B)^2\frac{b}{m} \\
&= \left(\frac{dm-(d-\gamma)b}{a+c}\right)^2\frac{a+c}{m} + (d-\gamma)^2\frac{b}{m} \\
&= \frac{(dm-(d-\gamma)b)^2}{(m-b)m} + (d-\gamma)^2\frac{b}{m} \\
&= \frac{1}{(m-b)m}\left[d^2m^2 - 2dm(d-\gamma)b + (d-\gamma)^2b^2 + (d-\gamma)^2b(m-b)\right] \\
&= \frac{1}{(m-b)m}\left[d^2m^2 - 2dm(d-\gamma)b + (d-\gamma)^2mb\right] \\
&= \frac{1}{(m-b)m}\left[d^2m^2 - 2d^2mb + 2d\gamma mb + d^2mb - 2d\gamma mb + \gamma^2mb\right] \\
&= \frac{1}{(m-b)m}\left[d^2(m^2-mb) + \gamma^2mb\right] \\
&= d^2 + \frac{\gamma^2b}{m-b} \\
&\geq d^2 + \frac{\varepsilon^3m}{(1-\varepsilon)m} \geq d^2 + \varepsilon^3.
\end{aligned}$$

The case when $d-d(S[i,i+\varepsilon m-1]) \leq -\varepsilon$ works out similarly. Indeed, set $\gamma \coloneqq d-d(S[i,i+\varepsilon m-1])$ as before and notice that $|\gamma|\geq\varepsilon$, and all the computations above are exactly the same.

So, $\mathrm{ind}_q((A,B,C)) \geq d_q^2 + \varepsilon^3$. For all other $q' \in \Sigma$, Lemma 3.8 gives that $\mathrm{ind}_{q'}((A,B,C)) \geq \mathrm{ind}_{q'}((S)) = d_{q'})^2(S)$. Thus,

$$\mathrm{ind}((A,B,C)) = \mathrm{ind}_q((A,B,C)) + \sum_{q' \in \Sigma - \{q\}} \mathrm{ind}_{q'}((A,B,C)) \geq \sum_{q' \in \Sigma} d_{q'}(S)^2 + \varepsilon^3.$$

□

Having shown that it is possible to refine an ε-irregular word to have a much larger index, we are in a position to finish the argument.

Proof of the Regularity Lemma for Words. Take $\varepsilon > 0$ and t_0 as given. We will give a bound on n_0 later. Suppose that we have word $S \in \Sigma^n$. Split it into t_0 consecutive factors $S_1, S_2, \ldots, S_{t_0}$ of the same length $\frac{n}{t_0}$. If $\mathcal{S} \coloneqq (S_1, S_2, \ldots, S_{t_o})$ is not an ε-regular partition, then let $I \subseteq [t_0]$ be the set of all indices such that, for every $i \in I$, S_i is not ε-regular. Then $\sum_{i \in I} S_i \geq \varepsilon n$. By Lemma 3.9, we can refine each S_i, $i \in I$, into factors A_i, B_i, and C_i, such that $\mathrm{ind}((A_i, B_i, C_i)) \geq \sum_{q \in \Sigma} d_q(S_i)^2 + \varepsilon^3$ (in the case that (1) is violated for several values of q, choose an arbitrary such q). We perform such refinement for each S_i, $i \in I$, obtaining a partition $\mathcal{S}' \preceq \mathcal{S}$, noticing that

$$\begin{aligned}
\mathrm{ind}(\mathcal{S}') = &\sum_{q \in \Sigma} \sum_{j \in [t_0] \setminus I} d_q(S_j)^2 \frac{|S_j|}{n} + \\
&\quad \sum_{q \in \Sigma} \sum_{i \in I} \left(d_q(A_i)^2 \frac{|A_i|}{n} + d_q(B_i)^2 \frac{|B_i|}{n} + d_q(C_i)^2 \frac{|C_i|}{n} \right) \\
&= \sum_{q \in \Sigma} \sum_{j \in [t_0] \setminus I} d_q(S_j)^2 \frac{|S_j|}{n} + \sum_{i \in I} \mathrm{ind}((A_i, B_i, C_i)) \frac{|S_i|}{n} \\
&\overset{(2)}{\geq} \sum_{q \in \Sigma} \sum_{j \in [t_0] \setminus I} d_q(S_j)^2 \frac{|S_j|}{n} + \sum_{i \in I} \left(\mathrm{ind}(S_i) + \varepsilon^3 \right) \frac{|S_i|}{n} \\
&= \mathrm{ind}(\mathcal{S}) + \varepsilon^3 \frac{\sum_{i \in I} |S_i|}{n} \\
&\geq \mathrm{ind}(\mathcal{S}) + \varepsilon^4.
\end{aligned}$$

Thus, $\mathcal{S}'$ refines $\mathcal{S}$ and has a higher index. If $\mathcal{S}'$ is not an ε-regular partition of S, then we can repeat the procedure above to obtain a refinement $\mathcal{S}'' \preceq \mathcal{S}'$, etc. But the index of any partition is bounded above by 1. Since the increment of the index that we get at each step is at least ε^4 and each word in the partition decreases in length by a factor of at most ε at each step, it follows that we can perform at most ε^{-4} steps so that the resulting factors are non-trivial, and therefore we will eventually find an ε-regular partition of S.

Such a partition consists of at most $3^{1/\varepsilon^4} t_0$ words, since at each iteration each of the words is partitioned into at most 3 new ones. Therefore, $T_0 \leq 3^{1/\varepsilon^4} t_0$ and each factor in the partition has length at least $t_0^{-1} \varepsilon^{1/\varepsilon^4} n$. □

3.2 Proof of Theorem 2.2

We are now ready to finish the proof of the main theorem. Before we do so, we show a useful claim about twins in ε-regular words.

Proposition 3.10. *If S is an ε-regular word, then $2f(S) \geq |S| - 5\varepsilon|S|$.*

Proof. Let $|S| = m$. We partition S into $t = 1/\varepsilon$ consecutive factors $S_1, \ldots, S_{1/\varepsilon}$, each of length εm. Since S is ε-regular, $|d(S) - d(S_i)| < \varepsilon$ for every $i \in \{2, \ldots, 1/\varepsilon\}$. Thus each S_i has at least $(d(S) - \varepsilon)\varepsilon m$ occurrences of 1s and at least $(1 - d(S) - \varepsilon)\varepsilon m$ occurrences of 0s. Let $S_i(1)$ be a subword of S_i consisting of exactly $(d(S) - \varepsilon)\varepsilon m$ 1s and $S_i(0)$ be a subword of S_i consisting of exactly $(1 - d(S) - \varepsilon)\varepsilon m$ 0s. Consider the following two identical disjoint subwords of S: $A = S_2(1)S_3(0)S_4(1) \cdots S_{t-2}(1)$ and $B = S_3(1)S_4(0)S_5(1) \cdots S_{t-1}(1)$. When t is odd, A and B are constructed similarly, as a kind of "delayed matching," with B always behind A by a length of $S_i(1)$.

We can find that A and B together have at least $m - 2\varepsilon^2(1/\varepsilon - 3) - 3\varepsilon m$ elements. To see why, note that we "threw away" at most

$$\varepsilon m - (\varepsilon m - 2\varepsilon^2 m) = 2\varepsilon^2 m$$

elements in each factor S_i, $i \in \{3, \ldots, t-2\}$ as well as the factors S_2 and S_{t-1} combined to obtain exactly $(d(S)-\varepsilon)\varepsilon m$ 1s and $(1-d(S)-\varepsilon)\varepsilon m$ 0s. Thus, in total, we discarded at most $2\varepsilon^2 m\,(1/\varepsilon - 3)$ elements to form the twins. Next, there are exactly $2\varepsilon m$ elements in S_1 and S_t combined, and at most εm elements in the unused subwords $S_2(0)$ and $S_{t-1}(0)$. Hence, in total, we failed to include at most

$$2\varepsilon^2(1/\varepsilon - 3) + 2\varepsilon m + \varepsilon m = 2\varepsilon^2(1/\varepsilon - 3) + 3\varepsilon m$$

elements, so

$$2f(S) \geq |A| + |B| \geq m - 2\varepsilon^2(1/\varepsilon - 3) - 3\varepsilon m \geq m - 5\varepsilon m,$$

as desired. □

Axenovich, Person, and Puzynina remark that we can slightly improve on $5\varepsilon m$ by finding twins of size $\varepsilon m/3$ each in S_1 and S_t using Lemma 2.1, but this does not give great improvement.

Proof of Theorem 2.2. Let n be at least n_0, which is as asserted by the Regularity Lemma for Words. for given $\varepsilon > 0$ and $t_0 \coloneqq \lceil \frac{1}{\varepsilon} \rceil$. Let $S \in \{0,1\}^n$. Again, Theorem 3.5 asserts an ε-regular partition of S into $S_1, S_2, \ldots, S_t$ with $1/\varepsilon \leq t \leq T_0$. We apply Proposition 3.10 to every ε-regular factor S_i of S. Furthermore, since S_is appear consecutively in S, we can put the twins from each of S_is together obtaining twins for the whole word S. This way we see:

$$2f(S) \geq \sum_{\substack{i \in [t], \\ S_i \text{ is } \varepsilon\text{-regular}}} (|S_i| - 5\varepsilon|S_i|) \geq n - 5\varepsilon n - \varepsilon n = n - 6\varepsilon n,$$

here εn corresponds to the maximum length of all ε-irregular factors. Choosing $\varepsilon = C\left(\frac{\log n}{\log\log n}\right)^{-1/4}$ and an appropriate C, we see that $n \geq \varepsilon^{-\varepsilon^{-4}}$. Therefore, by Theorem 3.5, $2f(n, \{0,1\}) \geq \left(1 - C(\log n)^{-1/4}\right)n$.

To prove the upper bound on $f(n, \{0,1\})$, we construct a binary word S such that $2f(S) \leq |S| - \log|S|$. Let $S = S_k S_{k-1} \cdots S_0$, where $|S_i| = 3^i$, S_i consists only of 1s for even i and only of 0s for odd i. In other words, S is built of iterated 0- or 1-blocks exponentially decreasing in size. Let A and B be twins in S. Assume first that A and B have the same number of elements in S_k. Since

S_k has an odd number of elements, and A, B restricted to $S' = S_{k-1} \cdots S_0$ are twins, by induction we have $|A| + |B| \leq (|S_k| - 1) + (|S'| - \log(|S'|)) = |S| - 1 - \log(|S'|) \leq |S| - \log|S|$. This last inequality is true since $|S'| = (3^k - 1)/2$, $|S| = (3^{k+1} - 1)/2$, so that

$$\begin{aligned}\log|S| - \log(|S'|) &= \log\left(\frac{|S|}{|S'|}\right)\\ &= \log\left(\frac{3^{k+1}-1}{3^k-1}\right)\\ &\leq 1,\end{aligned}$$

and $1 + \log(|S'|) \geq \log|S|$.

Now assume, without loss of generality, that A has more elements in S_k than B does in S_k. Then B cannot have any element in S_{k+1}, since S_{k+1} consists of all bits different from those in S_k. Suppose, for the sake of contradiction, that $|A|+|B| > |S|-\log|S|$. Then we have that $|A \cap S_{k-1}| \geq |S_{k-1}|/2$, otherwise $|A|+|B| \leq |S|-|S_{k-1}|/2 \leq |S|-\log|S|$. So, $s = |A \cap S_{k-1}| \geq |S_{k-1}|/2 = 3^{k-1}/2$, and s elements of B must collectively be in $S_{k-3} \cup S_{k-5} \cup \cdots$. But $|S_{k-3}| + |S_{k-5}| + \cdots \leq 3^{k-1}/2$, a contradiction, proving Theorem 2.2. $\square$

3.3 Improving the Bound

The authors also improve the power of the fraction $C\left(\frac{\log n}{\log\log n}\right)$ in the lower bound for $2f(n)$ by tightening the regularity lemma.

The argument proceeds as follows: In the proof of Theorem 2.2 we set up an index (energy function) that increased by at least ε^4 at each refinement. This increment was found rather roughly, so to improve it, let us consider the jth refinement step in the procedure, starting from the initial partition $\mathcal{S} = (S_1, S_2, \ldots, S_{t_0})$. Recall that I is in the interval consisting of all indices i such that S_i is not ε-regular. Let α_j be such that

$$\sum_{i \in I} |S_i| = \alpha_j n.$$

Thus, rather than taking the obvious bound $\sum_{i \in I} |S_i| \geq \varepsilon n$, we assign a specific constant α_j to the fraction of the total length of the word S consisting of ε-irregular parts.

In the proof in the previous section, we iterate as long as $\alpha_j \geq \varepsilon$ holds. And by performing an iteration step we merely use the fact that $\alpha_j \geq \varepsilon$ to get that the index increases by at least ε^4. Recall that

$$\operatorname{ind}(\mathcal{S}) = \sum_{q \in \Sigma} \sum_{j \in [|\mathcal{S}|]} d_q(S_j)^2 \frac{|S_j|}{n},$$

and for each further refinement $\mathcal{S}' \preceq \mathcal{S}$ it holds that

$$\begin{aligned}\operatorname{ind}(\mathcal{S}) &\leq \operatorname{ind}(\mathcal{S}') \\ &= \frac{(1-\alpha_j)n}{n} \operatorname{ind}(\mathcal{S}_1) + \frac{\alpha_j n}{n} \operatorname{ind}(\mathcal{S}_1) \\ &\leq \sum_{q \in \Sigma} \sum_{j \in [|\mathcal{S}|] \setminus I} d_q(S_j)^2 \frac{|S_j|}{n} + \alpha_j,\end{aligned} \tag{3}$$

where $\mathcal{S}_1$ consists of ε-regular words from $\mathcal{S}$ (these are not refined/partitioned anymore) and $\mathcal{S}_2$ consists of ε-irregular words from $\mathcal{S}$ (and their lengths add up to $\alpha_j n$).

Let ℓ be the total number of steps until we arrive at an ε-regular partition. Let $\alpha_1, \alpha_2, \ldots, \alpha_\ell$ be the numbers, where $\alpha_j n$ is the sum of the lengths of all ε-irregular words in the partition at step j, $j \in [\ell]$.

By our discussion above we have

$$1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_\ell \geq \varepsilon.$$

Next, we partition $(\varepsilon, 1]$ into $\log_2 \frac{1}{\varepsilon}$ consecutive intervals $(y_{i+1}, y_i]$ where $y_1 = 1$ and $y_{i+1} = y_i/2$. We claim that each interval $(y_{i+1}, y_i]$ contains at most $\frac{2}{\varepsilon^3}$ α_js. Indeed, the increase of the index during step j where $\alpha_j \in (y_{i+1}, y_i]$ is at least

$$\alpha_j \varepsilon^3 > y_{i+1}\varepsilon^3.$$

Further, let j' be the smallest index such that $\alpha_{j'} \leq y_i$ and j'' be the largest index such that $\alpha_{j''} \geq y_{i+1}$. Let ind_j be the index before the jth refinement step. Then by (3) the following holds for $j' + 1 \leq j \leq j''$:

$$\mathrm{ind}_{j'+1} \leq \mathrm{ind}_j \leq \mathrm{ind}_{j''} \leq \mathrm{ind}_{j'+1} + y_i.$$

This implies that the number of α_js in the interval $(y_{i+1}, y_i]$ cannot be bigger than

$$\frac{y_i}{y_{i+1}\varepsilon^3} = \frac{2}{\varepsilon^3}.$$

Thus, we obtain the following upper bound on ℓ:

$$\ell \leq \frac{2\log_2 \frac{1}{\varepsilon}}{\varepsilon^3},$$

which leads to $T_0 \leq t_0 3^{(-2\log\varepsilon)/\varepsilon^3}$, $n_0 = t_0 \varepsilon^{-(2\log 1/\varepsilon)/\varepsilon^3}$, and thus we can regularize with $\varepsilon = \left(\frac{(\log\log n)^2}{\log n}\right)^{1/3}$.

4 The Greedy Algorithm

Axenovich, Person, and Puzynina demonstrated that all binary words have large twins. A natural question to ask, then, is if many binary words have perfect twins. In other words, how ubiquitous are binary shuffle squares? While we do not have a definitive answer to this question, our methods give interesting insight for further research.

We enumerate binary shuffle squares through a greedy algorithm that attempts to construct perfect twins. The algorithm traverses through a binary word and attempts to allocate each bit into either one twin or another, except possibly some bits at the end. It proceeds as follows:

1. Place the current (first) bit into A. Let A_0 and B_0 be the states of the twins A and B after this step.

2. Let A_i and B_i be the twins we have constructed after iteration i of the algorithm. While $|A_i| > |B_i|$, let $m = |A_i| - |B_i|$. By construction, the last m bits in A will all be the same, so let each bit be $b \in \{0, 1\}$. Continue traversing the string, placing each instance of b into B until m b's have been placed. Each time the opposite bit, $\overline{b}$, is encountered, place it into A.

3. If, at any point, $|A| = |B|$, restart the algorithm by returning to step 1.
4. Once all bits have been allocated, remove any extraneous bits from A to ensure that $A = B$.

We call a single iteration of the greedy algorithm an *epoch*. For example, it is easy to check that on the word $S = 10010110$, the greedy algorithm finds the twins 100 and 100, with the bits s_6 and s_7 unused. We can see that unused bits occur at the final step of the algorithm; these are, in fact, the extraneous bits in A that must be removed. There is only one epoch in this implementation, as the greedy algorithm does not restart at any point.

The main benefit of the greedy algorithm is that it is easy to find the exact number of words on which the algorithm does produce perfect twins. This enables us to prove Theorem 2.3.

4.1 Proof of Theorem 2.3

The proof is twofold. We first find the number of words on which the algorithm produces perfect twins *only at the final step*, which we call *prefix-free shuffle squares.*

Definition 4.1 (Prefix)**.** For $1 \leq i \leq n$, the ith prefix of a word $S \in \Sigma^n$ is the subword $S[1, i]$.

Definition 4.2 (Prefix-free shuffle square)**.** A *prefix-free shuffle square* is a shuffle square for which the greedy algorithm produces perfect twins, but no perfect twins in any prefix.

It turns out that there is a bijection between prefix-free shuffle squares and Dyck paths, as evidenced by the following integral lemma.

Lemma 4.3. *The number of words in $\{0,1\}^{2n}$ for $n = 1, 2, \dots$ on which the greedy algorithm produces perfect twins but no perfect twins for any prefix is $2C_{n-1}$, the $(n-1)$st Catalan number.*

Proof. Let $\mathcal{W}_n$ be the family of binary words of length $2n$ on which the greedy algorithm produces perfect twins but no perfect twins for any prefix. We exhibit a two-to-one correspondence between $\mathcal{W}_n$ and $\mathcal{D}_{n-1}$, the family of Dyck paths of semilength $n-1$.

Let $S = s_1 s_2 \cdots s_{2n} \in \mathcal{W}_n$ be a word of length $2n$ on which the greedy algorithm gives perfect twins but no perfect twins for any of its prefixes; that is, the last R_k decays to 0, but no previous R_k equals 0. First, apply the greedy algorithm on S to produce twins $A = s_{i_1} s_{i_2} \cdots s_{i_n}$ and $B = s_{j_1} s_{j_2} \cdots s_{j_n}$, where $A \cup B = S$.

By construction, $s_1 \in A$ and $s_{2n} \in B$. Thus, we can identify $S = s_1 s_2 \cdots s_{2n}$ with a path $P = p_1 p_2 \cdots p_{2n-2} \in \mathcal{D}_{n-1}$ as follows: For each $2 \leq i \leq 2n-1$,

$$p_{i-1} = \begin{cases} (1,1) & \text{if } s_i \in A, \\ (1,-1) & \text{if } s_i \in B. \end{cases}$$

Observe that $p_1 = (1,1)$, since otherwise $s_1 \in A, s_2 \in B$ and we would have perfect twins within the prefix $S[1,2]$. Moreover, P can never cross below the x-axis, since that would mean that S contained a prefix with perfect twins. Finally, since $|A \cap S[2, 2n-1]| = n-1$ and $|B \cap S[2, 2n-1]| = n-1$, P ends at the point $(2n-2, 0)$, so it is indeed a valid Dyck path of semilength $n-1$.

Now, the final bit $s_{2n} \in S$ is fixed since $s_{2n} = s_{j_n} = s_{i_n}$, and $i_n < 2n$. Thus, choosing $s_1 = 0$ or $s_1 = 1$ leads to words that correspond to the same path P.

At the same time, for every path $P \in \mathcal{D}_{n-1}$, we can construct a word $S \in \mathcal{W}_n$ by identifying a step of size $(1,1)$ with an element in twin A and a step of size $(1,-1)$ with an element in twin B. The values of the twins are then determined as follows:

Let $\text{supp}(A) = (i_1, i_2, \ldots, i_n)$ and $\text{supp}(B) = (j_1, j_2, \ldots, j_n)$. By construction, $i_k < j_k$ for all $1 \leq k \leq n$ (if not, then the path P must have crossed above the main diagonal). Now, apply the greedy algorithm in reverse to fill in the bits:

1. Without loss of generality, let $s_{i_1} = 1$ (we multiply by 2 at the end to account for the string's complement). Now, set a counter ℓ to 2.
2. For $k = 1, 2, \ldots, n$, while $i_\ell < j_k$, let $s_{i_\ell} = 0$ if k is odd and 1 if k is even.
3. Since $s_{j_1} s_{j_2} \cdots s_{j_n} = s_{i_1} s_{i_2} \cdots s_{i_n}$, once we have filled in the values of $s_{i_1} s_{i_2} \cdots s_{i_n}$, we are done.

It is easy to verify that the greedy algorithm, which is deterministic, produces the twins obtained from the inverse procedure above. Since we were free to choose the value of s_{i_1} as 0 or 1, each path P corresponds to two distinct words in $\mathcal{W}_n$.

We have thus shown that $\mathcal{W}_n \leftrightarrow \mathcal{D}_{n-1}$ is indeed a two-to-one correspondence. Since $|\mathcal{D}_{n-1}| = C_{n-1}$, the total number of binary words on which the greedy algorithm produces perfect twins, but not perfect twins for any prefix, is $2C_{n-1}$. □

Lemma 4.3 immediately implies Theorem 2.3.

Lemma 4.4. *The number of words in $\{0,1\}^{2n}$ for $n = 1, 2, \ldots$ on which the greedy algorithm produces perfect twins, but which may contain more than one epoch, is $\binom{2n}{n}$.*

To prove Lemma 4.4, we need the Catalan identity in Proposition 2.9.

Proof of Lemma 4.4. For each $n = 1, 2, \ldots$, let W_n be the number of words in $\{0,1\}^{2n}$ on which the greedy algorithm produces perfect twins, but not necessarily only at the last bit.

Let $2k$ be the size of the first epoch; that is, the prefix $S[1, 2k]$ contains perfect twins for the first time. There are $2C_{k-1}$ choices for the value of $S[1, k]$, after which $S[2k+1, 2n]$ can be constructed in W_{n-k} ways. Thus, we have the recursion

$$W_n = 2 \sum_{k=1}^{n} C_{k-1} W_{n-k}.$$

It is easy to check that $W_1 = 2$ and $W_2 = 6$. Using the above recursion, we prove by induction that $W_n = \binom{2n}{n}$ for all positive integers n.

Suppose $W_k = \binom{2k}{k}$ for all $k < n$. Then

$$\begin{aligned}
W_n &= 2\sum_{k=1}^{n} C_{k-1}W_{n-k} \\
&= 2\sum_{k=1}^{n}(n-k+1)C_{k-1}C_{n-k} \\
&= 2(n+1)\sum_{k=1}^{n} C_{k-1}C_{n-k} - 2\sum_{k=1}^{n} kC_{k-1}C_{n-k} \\
&= 2(n+1)C_n - 2\sum_{k=0}^{n}(k+1)C_kC_{n-k-1} \\
&= 2\binom{2n}{n} - 2\cdot\frac{1}{2}\binom{2n}{n} && \text{(by Proposition 2.9)} \\
&= 2\binom{2n}{n} - \binom{2n}{n} \\
&= \binom{2n}{n},
\end{aligned}$$

completing the induction and proving the lemma. □

Since the greedy algorithm finds $\binom{2n}{n}$ binary shuffle squares, Theorem 2.3 follows easily.

Remark. Given that the number of binary shuffle squares found by the greedy algorithm is $\binom{2n}{n}$, a bijection between these words and *all* lattice paths from $(0,0)$ to $(2n,0)$. The idea is that each epoch can begin with a 1 or 0, and we can equate this to a sub-path starting with an up-step or down-step. No matter what the first step is, the sub-path will stay on the same side of the x-axis. Thus, binary shuffle squares found by the greedy algorithm correspond to a path from $(0,0)$ to $(2n,0)$ with no restriction, and there are $\binom{2n}{n}$ of these.

Lemma 4.4 shows that the number of binary shuffle squares is at least $\binom{2n}{n}$. There may be more binary shuffle squares than $\binom{2n}{n}$ because the greedy algorithm obviously does not find all of them; for example, in the shuffle square $S = 001001$, the greedy algorithm only locates twins $T_1 = s_1s_3$ and $T_2 = s_2s_6$ with value 01. We state a conjecture on the total number of binary shuffle squares in Section 7.

Next, we will extend this result by considering shuffle squares over larger alphabets.

5 Shuffle Squares Over Large Alphabets

In this section, we prove Theorem 2.4, which states that

$$|\mathrm{SS}_k(n)| = \frac{1}{n+1}\binom{2n}{n}k^n - \binom{2n-1}{n+1}k^{n-1} + O_n(k^{n-2}).$$

The top coefficient is easily recognizable as the Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n}$ enumerating the number of Dyck paths of semilength n. The second coefficient is also seen to be the total number of valleys summed over Dyck paths of semilength n by Proposition 2.10. As mentioned before, this will be important for the proof.

We start with a simple lemma (which is certainly not new, see e.g. Bukh and Zhou, Lemma 17 [4]) that explains the first-order term. If $s \in [k]^\ell$ and $I \subseteq [\ell]$, write s_I for the subsequence of s indexed by I. Thus $s \in [k]^{2n}$ is a shuffle square if and only if there exists $I \in \binom{[2n]}{n}$ such that $s_I = s_{[2n]\setminus I}$.

Lemma 5.1. *If $s \in [k]^{2n}$ is a shuffle square, then there exists $I = \{i_1, \ldots, i_n\}$ such that $s_I = s_{[2n]\setminus I}$, and furthermore if $J = [2n]\setminus I$ consists of the indices $\{j_1, \ldots, j_n\}$, then $i_r < j_r$ for all r.*

Proof. The first part is just the definition of a shuffle square. For the second part, suppose I is a set of indices such that $s_I = s_{[2n]\setminus I}$, and $J = [2n]\setminus I$. If $i_r > j_r$ for some smallest r, then we may modify I by replacing i_r with j_r, so that $I' = I \cup \{j_r\}\setminus\{i_r\}$ and $s_{I'} = s_{[2n]\setminus I'}$ still holds. Continuing in this way we can swap out all the out-of-order elements of I with those of J, proving the claim. ☐

We say that a partition $[2n] = I \sqcup J$ is a *monotone pair* if $|I| = |J| = n$ and the r-th smallest element of I is smaller than the r-th smallest element of J. The number of monotone pairs in $[2n]$ is exactly the Catalan number C_n; form (I, J) from a Dyck path by taking I to be the set of indices on which the path moves upwards by $(+1, +1)$. Let $\mathrm{MP}(n)$ denote the set of all monotone pairs in $[2n]$.

If $s \in [k]^{2n}$ is a shuffle square, we say that (I, J) *is a monotone pair for s* if $I \sqcup J = [2n]$, they satisfy the properties $s_I = s_J$, and the corresponding indices in I are smaller than those in J. Lemma 5.1 implies the existence of monotone pairs for all shuffle squares. It follows that $|\mathrm{SS}_k(n)| \leq |\mathrm{MP}(n)| \cdot k^n = C_n \cdot k^n$, since this latter expression counts the number of ways to choose a monotone pair (I, J) and then the value of s_I, which together determine s completely. Now this is an overcount because a single word s may have many different monotone pairs. For example, constant words have C_n monotone pairs. We must correct for this.

5.1 Proof of Theorem 2.4

In order to determine the second-order term, we must compute how much this bound is overcounting via inclusion-exclusion. This would then complete the proof.

Proof of Theorem 2.4. First, we identify $\mathrm{MP}(n)$ with the family of *non-nesting perfect matchings* on $[2n]$ (this notion is defined in [13]). A perfect matching on V is a graph whose vertex set is V and where each vertex lies in exactly one edge. We say that a perfect matching on $[2n]$ it is *non-nesting* if there do not exist two edges (i, j) and (i', j') satisfying $i < i' < j' < j$. Thus, a perfect matching m on $[2n]$ is non-nesting if and only if there exists $(I, J) \in \mathrm{MP}(n)$ such that the edges in m are exactly the pairs (i_r, j_r) where i_r (resp. j_r) is the r-th smallest element of I. To avoid introducing too much notation, we slightly abuse notation and write $m \in \mathrm{MP}(n)$ to mean that m is an non-nesting perfect matching corresponding to some monotone pair in $\mathrm{MP}(n)$.

Let $\mathrm{comp}(G)$ denote the number of connected components of a graph G. We claim that

$$|\mathrm{SS}_k(n)| = \sum_{m_1} k^{\mathrm{comp}(m_1)} - \sum_{m_1 \neq m_2} k^{\mathrm{comp}(m_1 \cup m_2)} + \cdots + (-1)^r \sum_{m_1, \ldots, m_r} k^{\mathrm{comp}(m_1 \cup \cdots \cup m_r)} + \cdots \quad (4)$$

by inclusion-exclusion, where the r-th sum is over all choices of an unordered r-tuple of distinct non-nesting perfect matchings $m_i \in \mathrm{MP}(n)$. Formula (4) holds because the number of shuffle squares

s which have $m_1, \ldots, m_r$ simultaneously as its monotone pairs is $k^{\mathrm{comp}(m_1 \cup \cdots \cup m_r)}$, since the value of s on every vertex of a given connected component must be the same. But the total number of terms in this inclusion-exclusion is $O_n(1)$, and so for the purposes of proving Theorem 2.4 it suffices to select only the terms from (4) with $\mathrm{comp}(m_1 \cup \cdots \cup m_r) \geq n-1$, as all other terms summed together will be $O_n(k^{n-2})$.

It is not hard to see that the only terms in (4) with $\mathrm{comp}(m_1 \cup \cdots \cup m_r) = n$ are exactly the terms of the first summation $r = 1$, which adds up to $C_n \cdot k^n$, the desired leading term. As for $\mathrm{comp}(m_1 \cup \cdots \cup m_r) = n-1$, one can check that $r = 2$ is the only possibility. It remains to count the number of pairs $m_1 \neq m_2$ in $\mathrm{MP}(n)$ such that $\mathrm{comp}(m_1 \cup m_2) = n-1$. Since m_1 and m_2 themselves each have n components (i.e. edges) of size 2, for $\mathrm{comp}(m_1 \cup m_2) = n-1$ to hold, m_1 must share all but two of its edges with m_2, and the two remaining edges must form a four-cycle with the two corresponding edges of m_2. If the vertices of this four-cycle are $a < b < c < d$, then since m_1 and m_2 are both non-nesting they cannot contain the edges (a, d) and (b, c). We may thus assume without loss of generality that $(a, b), (c, d) \in m_1$ and $(a, c), (b, d) \in m_2$.

We claim that in order for $\mathrm{comp}(m_1 \cup m_2) = n-1$, the four indices must satisfy the additional property $c = b+1$. If not, there exists some x between b and c, and x is matched to the same vertex y in both m_1 and m_2 since m_1 and m_2 are identical outside $\{a, b, c, d\}$. If $y < a$ or $y > d$, then m_1 is not non-nesting, while if $a < y < d$ then m_2 is not non-nesting. This is a contradiction in all cases, so no such x can exist and $c = b+1$.

We are now ready to prove that the pairs $\{m_1, m_2\}$ satisfying $\mathrm{comp}(m_1 \cup m_2) = n-1$ are in bijection with pairs (P, v) of a Dyck path of semilength n and a valley in the path. Using the bijection between Dyck paths and monotone pairs, the path P is in bijection with monotone pairs (I, J) by writing down the indices of the up and down paths. Thus, (P, v) is in bijection with a choice of a monotone pair (I, J) and an element $b \in J$ such that $b+1 \in I$, since such a b corresponds exactly to going down, then up, to form a valley in P.

Let m_2 be the non-nesting perfect matching corresponding to (I, J), and let m_1 be the matching corresponding to $(I \cup \{b\} \backslash \{b+1\}, J \cup \{b+1\} \backslash \{b\})$. It is easy to see that $\mathrm{comp}(m_1 \cup m_2) = n-1$, and this gives a bijection between pairs $((I, J), b)$ and pairs (m_1, m_2) with $\mathrm{comp}(m_1 \cup m_2) = n-1$ as desired.

Proposition 2.10 tells us that the number of valleys across all Dyck paths of semilength n is $\binom{2n-1}{n+1}$. Thus, this is also the number of terms in (4) equal to $-k^{n-1}$. By (4), we find that

$$|\mathrm{SS}_k(n)| = C_n k^n - \binom{2n-1}{n+1} k^{n-1} + O(k^{n-2}),$$

completing the proof. $\square$

6 Reverse Shuffle Squares

Here, we tackle a second, and closely related, conjecture from [9] on reverse shuffle squares and prove Theorem 2.5. A *reverse shuffle square* is a word $s \in [k]^{2n}$ which can be decomposed into two subsequences of length n which are reverses of each other. Let $\mathrm{RSS}_k(n)$ denote the family of all reverse shuffle squares in $[k]^{2n}$. The conjecture is as follows.

Conjecture 6.1. *The number of reverse shuffle squares in $[k]^{2n}$ satisfies*

$$|\mathrm{RSS}_k(n)| = \frac{1}{n+1}\binom{2n}{n} k^n - \left(\binom{2n-1}{n-1} - 2^{n-1}\right) k^{n-1} + O_n(k^{n-2}).$$

Again, the top coefficient is the Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n}$. However, we will show that the conjecture is actually false, and the correct second-order term is slightly different. It is equal to B_n, which counts the number of unordered pairs of 123-avoiding permutations of length n that differ by a single transposition and satisfies $B_1 = 0$, $B_n = 2\binom{2n-2}{n-2} + 2C_{n+1} - 8C_n + 5C_{n-1}$ for $n \geq 2$. It is easy to check that the right-hand side is equal to the second-order coefficient in Theorem 2.5.

Note that the first four terms ($n = 2, 3, 4, 5$) of $\binom{2n-1}{n-1} - 2^{n-1}$ and B_n are both $1, 6, 27, 110$, which explains why the incorrect expression was guessed by [9] based on numerical evidence. However, for $n = 6$, $\binom{2n-1}{n-1} - 2^{n-1} = 430$ while $B_6 = 432$.

This time, instead of interpreting the Catalan numbers in terms of Dyck paths, we will interpret it in terms of 123-avoiding permutations.

Recall that a permutation $\pi \in S_n$ is 123-avoiding if there do not exist $i_1 < i_2 < i_3$ for which $\pi(i_1) < \pi(i_2) < \pi(i_3)$, and that the total number of 123-avoiding permutations of length n is exactly C_n. It will also be helpful to note that π is 123-avoiding if and only if it can be partitioned into two decreasing subsequences. [Remark: this is closely related to the "partition into two towers" notion in [13]].

6.1 Proof of Theorem 2.5

We begin, as before, with a characterization of reverse shuffle squares that explains the first-order term. Given a permutation $\pi \in S_n$ and a word $s \in [k]^n$, we write $\pi(s)$ for the word obtained by shuffling the letters according to π, i.e. $\pi(s)_i \coloneqq s_{\pi(i)}$. We also write s_I for the subword of s indexed by a set I of indices, and s^R for the reverse of s.

Lemma 6.2. *Suppose $s \in [k]^{2n}$ and we split $s = s's''$ into two equal halves, so that s', s'' are both words in $[k]^n$. Then, s is a reverse shuffle square if and only if $s'' = \pi(s')$ for some 123-avoiding permutation π.*

Proof. We first prove the only-if direction in the special case that $k \geq n$ and every letter in s appears exactly twice.

It was shown by [9] that if s is a reverse shuffle square, then s is an *abelian square*, which is a word where the second half is a permutation of the first. Thus, $s'' = \pi(s')$ for some permutation π. Since every letter in s appears exactly twice, this π is unique. We show that it is 123-avoiding. If not, there are three indices $i_1 < i_2 < i_3$ for which $\pi(i_1) < \pi(i_2) < \pi(i_3)$. Thus, $s'_{i_1}, s'_{i_2}, s'_{i_3}$ appear in the same relative order in s' as they do in s''. These six letters appear at positions $i_1 < i_2 < i_3 < n + \pi(i_1) < n + \pi(i_2) < n + \pi(i_3)$ in the original word s.

Since s is a reverse shuffle square, so must its restriction to the six positions above, as the three letters there do not appear elsewhere in s. But the restriction to these six positions of s is a word of the form *abcabc*, which cannot be a reverse shuffle square. This proves the special case.

For the general case, suppose $t \in [k]^{2n}$ is any reverse shuffle square, which means that there exists two index subsets $I, J \in \binom{[2n]}{n}$ partitioning $[2n]$ such that the restrictions t_I and t_J are reverses of each other. Then, t is a homomorphic image of the word $s \in [n]^{2n}$ defined so that $s_I = 1...n$ and $s_J = n...1$, and s is a reverse shuffle square where every letter appears exactly twice. Applying the special case above to s, we obtain a 123-avoiding permutation π such that the second half of s is π applied to the first half. As t is a homomorphic image of s, this holds for t as well, so this proves the only-if direction.

To prove the if direction, note that a permutation π is 123-avoiding if and only if it can be partitioned into two decreasing subsequences. Suppose s satisfies $s'' = \pi(s')$ for such a π, and let

$[n] = I_\pi \sqcup J_\pi$ be a partition of the index set of π for which $\pi|_{I_\pi}$ and $\pi|_{J_\pi}$ are both decreasing. Define $I \coloneqq I_\pi \cup (n + \pi(J_\pi))$ and $J \coloneqq J_\pi \cup (n + \pi(I_\pi))$, we see that I and J partition $[2n]$. Because π is decreasing when restricted to both I_π and J_π, it follows that the part of s_I in s'' is the reverse of the part of s_J in s', and similarly the part of s_J in s'' is the reverse of the part of s_I in s'. This means that $s_I = s_J^R$, completing the proof that s is a reverse shuffle square. □

Let $\mathrm{Av}_n(123)$ denote the family of all 123-avoiding permutations of length n. We obtain an upper bound $|\mathrm{RSS}_k(n)| \le C_n k^n$ by sending each reverse shuffle square s to an ordered pair (π, s') of a 123-avoiding permutation π corresponding to s and the first half s' of s. The full word s can be reconstructed from this data by taking $s'' = \pi(s')$. It remains to understand the overcounting to get at the second-order term.

To each $\pi \in \mathrm{Av}_n(123)$, associate the matching $m(\pi)$ on $[2n]$ whose edges are $(i, n + \pi(i))$. We define S_π to be the set of k^n words of the form $s = s'\pi(s')$ in $[k]^{2n}$. We obtain that $s \in S_\pi$ exactly if $s_i = s_j$ whenever $i \sim j$ in $m(\pi)$. As a result, for multiple permutations $\pi_1, \cdots, \pi_r$, the intersection $S_{\pi_1} \cap \cdots \cap S_{\pi_r}$ is exactly the set of words $s \in [k]^{2n}$ which are constant on every connected component of $m(\pi_1) \cup m(\pi_2) \cup \cdots \cup m(\pi_r)$. By inclusion-exclusion, we obtain

$$|\mathrm{RSS}_k(n)| = \sum_{\pi} k^n - \sum_{\pi_1,\pi_2} k^{\mathrm{comp}(m(\pi_1)\cup m(\pi_2))} + \cdots + (-1)^r \sum_{\pi_1,\ldots,\pi_r} k^{\mathrm{comp}(m(\pi_1)\cup\cdots\cup m(\pi_r))} + \cdots,$$

where the r-th sum is a sum over unordered r-tuples of distinct $\pi_i \in \mathrm{Av}_n(123)$. We find that all k^n terms appear in the first sum, and that all k^{n-1} terms appear in the second (this latter fact follows from the observation that $m(\pi)$ is precedence-free (doesn't include two edges $(i_1, j_1), (i_2, j_2)$ with $i_1 < j_1 < i_2 < j_2$). Thus, we have

$$|\mathrm{RSS}_k(n)| = C_n k^n - B_n k^{n-1} + O_n(k^{n-2}),$$

where B_n is the number of unordered pairs $\pi_1, \pi_2 \in \mathrm{Av}_n(123)$ satisfying $\mathrm{comp}(m(\pi_1) \cup m(\pi_2)) = n-1$. The only way for $\mathrm{comp}(m(\pi_1) \cup m(\pi_2)) = n-1$ to occur is if π_1 and π_2 differ by exactly one transposition (i.e. $\pi_1 = (ij) \circ \pi_2$ in cycle notation for some $i, j \in [n]$), so that $m(\pi_1) \cup m(\pi_2)$ has exactly one component of size 4. Thus B_n enumerates the pairs claimed in the theorem, and it remains to show

$$B_n = 2\binom{2n-2}{n-2} + 2C_{n+1} - 8C_n + 5C_{n-1} \tag{5}$$

for $n \ge 2$. This is attempted in the next section.

6.2 A Closed Form for B_n

In this section, we prove the following formula for B_n, which is defined for $n \ge 1$ as the number of unordered pairs of elements of $\mathrm{Av}_n(123)$ which differ by a single transposition, which is almost all the way towards (5). Define the Catalan convolutions

$$C_{n,k} \coloneqq \frac{k}{2n-k}\binom{2n-k}{n},$$

which enumerate (see [5]) the number of 123-avoiding permutations π of length n with $\pi(k) = n$.

Lemma 6.3. *For all $n \geq 2$,*

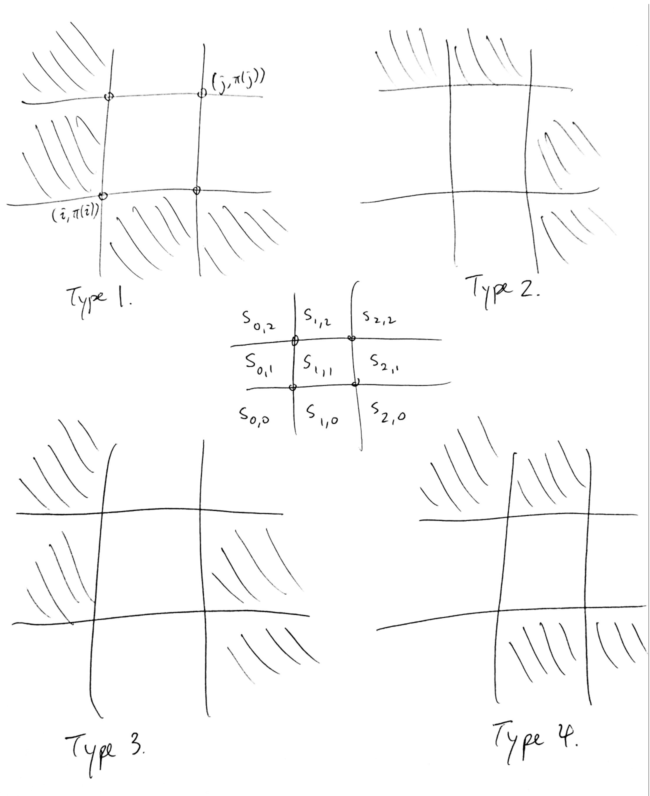
$$B_n = 2A_n + 2C_{n+1} - 8C_n + 5C_{n-1},$$

where

$$A_n = \sum_{a+b+c+d=n-2} \binom{a+c}{a} C_{a+b+1,a+1} C_{c+d+1,c+1}, \tag{6}$$

the sum over all 4-compositions $a+b+c+d = n-2$.

This formula will appear from another application of inclusion-exclusion, which will depend on the following diagrams.



Recall that every permutation π can be represented in the plane by plotting all the points $(i, \pi(i))$, and π is 123-avoiding if and only if the plot doesn't contain three points in increasing order. Suppose $\pi \in \mathrm{Av}_n(123)$ and there is a transposition (ij) for which $(ij) \circ \pi \in \mathrm{Av}_n(123)$ as well. By swapping π with $(ij) \circ \pi$ if necessary, we may assume $\pi(i) < \pi(j)$ as in the diagram. Then, the four vertical and horizontal lines through the two points $(i, \pi(i))$ and $(j, \pi(j))$ divide the plane into nine rectangular sectors, as shown. We say that the pair $(\pi, (ij))$ is of *type* t (for $t \in [4]$) if all the remaining points in the plot of π fall into only the shaded regions in the picture labelled "Type t." For example, $(\pi, (ij))$ is of type 1 if and only if for all $i' \notin \{i, j\}$, either $i' < i$ and $\pi(i') > i$, or $i' > i$ and $\pi(i') < i$. Note that it's possible for a pair to be of more than one type.

Lemma 6.4. *If $\pi \in \mathrm{Av}_n(123)$, $1 \le i < j \le n$, and $\pi(i) < \pi(j)$, and $(ij) \circ \pi \in \mathrm{Av}_n(123)$, then $(\pi, (ij))$ is in (at least) one of the four types.*

Proof. Label the nine sectors as $s_{x,y}$ in the middle diagram in the figure, so that $x = 0$ if the sector is left of i, $x = 1$ if it is between i and j, and $x = 2$ if it is to the right of y, and similarly for y. Since $\pi \in \mathrm{Av}_n(123)$, $s_{0,0}, s_{1,1}$ and $s_{2,2}$ must be empty, since any point in any of them would form a 123-pattern with $\pi(i)$ and $\pi(j)$. Thus these three sectors are always empty, as in the diagram.

Next, note that $s_{0,1}$ and $s_{1,2}$ cannot both be nonempty, since a point in $s_{0,1}$ and a point in $s_{1,2}$ would form a 123-pattern with $(i, \pi(j))$ in $(ij) \circ \pi$. Similarly, at least one of $s_{1,0}$ and $s_{2,1}$ may be nonempty if $(j, \pi(i))$ appears in the diagram for $(ij) \circ \pi$. This completes the proof. $\square$

Let $P_{n,t}$ denote the collection of pairs $(\pi, (ij))$ of $\pi \in \mathrm{Av}_n(123)$ and $1 \le i < j \le n$ for which $1 \le i < j \le n$ of type t for $t = 1, 2, 3, 4$. Clearly, $\cup_{t=1}^4 P_{n,t}$ is in bijection with the set of pairs $\{\pi_1, \pi_2\} \in \binom{\mathrm{Av}_n(123)}{2}$ differing by a transposition, so it suffices to enumerate this union. We proceed by inclusion-exclusion.

Lemma 6.5. *For $n \ge 2$, collections $P_{n,t}$ satisfy*

$$|P_{n,1}| = |P_{n,2}| = C_{n+1} - 2C_n, \tag{7}$$

$$|P_{n,3}| = |P_{n,4}| = A_n, \tag{8}$$

$$|P_{n,1} \cap P_{n,2}| = |P_{n,3} \cap P_{n,4}| = C_{n-1}, \tag{9}$$

$$|P_{n,1} \cap P_{n,3}| = |P_{n,1} \cap P_{n,4}| = |P_{n,2} \cap P_{n,3}| = |P_{n,2} \cap P_{n,4}| = C_n - C_{n-1}, \tag{10}$$

$$\begin{aligned} |P_{n,1} \cap P_{n,2} \cap P_{n,3}| &= |P_{n,1} \cap P_{n,2} \cap P_{n,4}| \\ &= |P_{n,1} \cap P_{n,3} \cap P_{n,4}| = |P_{n,2} \cap P_{n,3} \cap P_{n,4}| = C_{n-1}, \end{aligned} \tag{11}$$

$$|P_{n,1} \cap P_{n,2} \cap P_{n,3} \cap P_{n,4}| = C_{n-1}, \tag{12}$$

where A_n is defined by (6).

Before we prove the lemma, note that it implies Lemma 6.3 by inclusion-exclusion. Indeed, we have

$$B_n = \Big|\bigcup_{t=1}^4 P_{n,t}\Big| = [2(C_{n+1} - 2C_n) + 2A_n] - [2C_{n-1} + 4(C_n - C_{n-1})] + [4C_{n-1}] - [C_{n-1}]$$

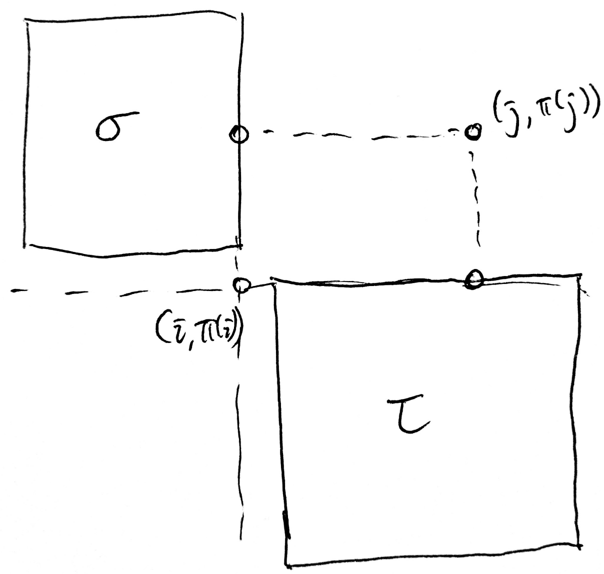
by inclusion-exclusion and reading off the values in Lemma 6.5.

Proof of Lemma 6.5. The system of equations can really be reduced to four distinct cases: $P_{n,1}$, $P_{n,4}$, $P_{n,1} \cap P_{n,2}$ (with only three allowed regions), and $P_{n,1} \cap P_{n,3}$ (with only two allowed regions). Any other set of shaded regions can be reflected to obtain one of these four.

We save $P_{n,4}$ to the end, and handle the other three that are immediately representable in terms of Catalan numbers. We start by proving (7), which will follow from

$$|P_{n,1}| = \sum_{i=1}^{n-1} C_i C_{n-i}. \tag{13}$$

The proof is by bijection: take two nonempty 123-avoiding permutations σ and τ with $|\sigma| + |\tau| = n$. Let $i = |\sigma|$, and $\pi(i) = n - i = |\tau|$. Given (σ, τ), we obtain $(\pi, (ij))$ of type 1 as follows.



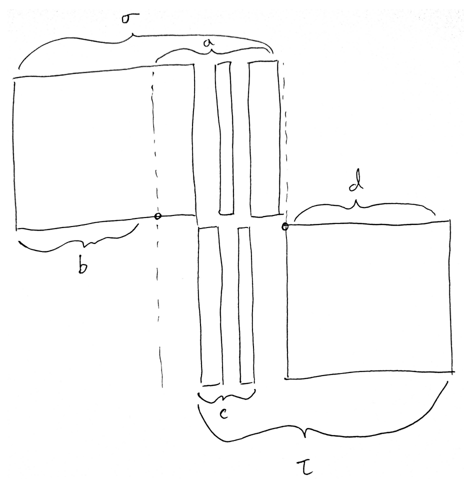
Place a copy of σ in the upper-left $i \times i$ square of the $n \times n$ grid, and a copy of τ in the lower-right $(n-i) \times (n-i)$ square, and insert the point $(i, n-i)$. As there are $n+1$ points in total now, this is not a valid permutation. The offending points are those in σ and τ which get placed on the horizontal and vertical lines through $(i, \pi(i))$. Define $(j, \pi(j))$ such that j is the x-coordinate of the offending point in τ, and $\pi(j)$ is the y-coordinate of the offending point in σ. Remove the two offending points and insert $(j, \pi(j))$ to obtain an honest permutation $\pi \in \mathrm{Av}_n(123)$.

This exhibits a bijection between $P_{n,1}$ and ordered pairs (σ, τ) of nonempty 123-avoiding permutations whose lengths sum to n, thus proving the convolution formula (13). This implies (7) by the standard convolution identity $C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}$.

The remaining identities $|P_{n,1} \cap P_{n,2}| = C_{n-1}$ and $|P_{n,1} \cap P_{n,3}| = C_n - C_{n-1}$ are obtained via very similar arguments and left as an exercise. This proves equations (9) through (12), leaving only (8), which expands as

$$|P_{n,4}| = \sum_{a+b+c+d=n-2} \binom{a+c}{a} C_{a+b+1,a+1} C_{c+d+1,c+1}.$$

This is again a bijection argument, illustrated by the below diagram.



Let σ and τ be two 123-avoiding permutations with sizes $|\sigma| = a+b+1$ and $|\tau| = c+d+1$, such that $\sigma(b+1) = 1$ and $\tau(c+1) = c+d+1$. The number of such pairs is exactly $C_{a+b+1,a+1}C_{c+d+1,c+1}$ by [5]. Place σ to the top left of τ as before, but notice that σ ends in a decreasing sequence of length a, and τ begins with a decreasing sequence of length c. Thus, these two parts may be horizontally interleaved arbitrarily in the middle in $\binom{a+c}{a}$ ways. This completes the proof. □

The only thing left to prove Theorem 2.5 is to prove the identity

$$\binom{2n-2}{n-2} = \sum_{a+b+c+d=n-2} \binom{a+c}{a} C_{a+b+1,a+1}C_{c+d+1,c+1} \tag{14}$$

for all $n \geq 2$. Recall from above that $C_{n,k} \coloneqq \frac{k}{2n-k}\binom{2n-k}{n}$ is a Catalan convolution, thus named because $C_{n,k}$ satisfies the identity

$$C_{n,k} = \sum_{a_1+\cdots+a_k=n-k} \prod_{i=1}^{k} C_{a_i}. \tag{15}$$

In both equations above, the sum is over all nonnegative compositions, i.e. choices of the summands from nonnegative integers.

We first note that $C_{n,k}$ is exactly the number of Dyck paths from $(0,0)$ to $(2n,0)$ which touch the x-axis exactly $k-1$ times internally; this is because such a path breaks down into k subpaths of lengths $a_1+1, \ldots, a_k+1$ which each stay on or above the line $y = 1$ internally, hence (15).

Lemma 6.6. *The Catalan convolutions satisfy (14) for all $n \geq 2$.*

Proof. The proof is by double-counting. We claim that both sides enumerate the family F of paths between $(0,0)$ and $(2n,0)$ where each step is a $U = (+1,+1)$ or $D = (+1,-1)$, such that the path starts and ends with a U (note that such paths are certainly not Dyck paths, as the second-to-last point on the path is $(2n-1,-1)$). The left side clearly enumerates such paths, because there are $\binom{2n-2}{n-2}$ strings over the binary alphabet $\{U, D\}$ of length $2n$ with n U's and n D's which start and end with U. As for the right side, take any $p \in F$ and suppose it touches the line $x = 0$ a total of $t \geq 1$ times internally. These t points break p up into $t+1 \geq 2$ segments, which are themselves either strict Dyck paths (strict meaning staying entirely above the diagonal internally) or else the reflections of strict Dyck paths over the x-axis. Let there be $a+1$ of the positive segments and $b+1$ of the negative segments. Then, we map p to the pair (p_+, p_-) of Dyck paths where p_+ is obtained by concatenating all the positive segments together, and p_- by concatenating all the negative segments.

It is easy to check that this is a surjective map from F to the union $\cup_{a+b+c+d} D_{a+b+1,a+1} \times D_{c+d+1,c+1}$, where $D_{n,k}$ is the family of Dyck paths of semi-length n with exactly $k-1$ internal points, so $|D_{n,k}| = C_{n,k}$. Furthermore, the preimage of (p_+, p_-) has size exactly $\binom{a+c}{a}$, because this is the number of ways to interleave the $a+1$ segments of p_+ and the $c+1$ segments of p_-, excepting the first segment of p_+ which must go at the beginning of $p \in F$,and the last segment of p_- which must go at the end. This completes the proof of (14). □

7 Future Work

Previously, we examined the problem of just how many words with perfect twins, or shuffle squares, there are. We know that there are at least $\binom{2n}{n}$ binary shuffle squares of length $2n$. However, numerical evidence suggests that the actual number is significantly larger.

Conjecture 7.1.

$$|SS_2(n)| = \left(\frac{1}{2} - o(1)\right) 4^n$$

While the previous approaches have found a closed formula for the number of shuffle squares, Conjecture 7.1 states that almost half of all binary word have perfect twins, and that as the length of the word approaches infinity (the length of the string grows asymptotically), half of all binary word have perfect twins.

Recall that the original twins in words problem proposed by Axenovich, Person and Puzynina (2012) [1] stated that nearly perfect twins exist in all binary words. While the original twins in words problem and Conjecture 7.1 are similar but different results, Conjecture 7.1 can help us better understand the behavior of twins, especially twins of maximal length, in words.

In a similar way to how we approached the earlier formulas for the number of binary shuffle squares, the greedy algorithm may be a promising starting point for understanding Conjecture 7.1.

8 Acknowledgements

The authors would like to thank our mentor Xiaoyu He for helping us navigate the literature and for answering countless questions. We are also grateful for his patient guidance and many valuable insights. We would also like to thank Pawel Grzegrzolka and the Stanford Mathematics Department for making the Stanford Undergraduate Research Institute in Mathematics (SURIM) research program a wonderful experience, as well as for providing funding and support for the research.

References

[1] M. Axenovich, Y. Person, and S. Puzynina, *A regularity lemma and twins in words.* J. Combin. Theory Ser. A, 120 (2012), no. 4, 733–743, arXiv:1204.2180.

[2] B. Bukh and R. Hogenson. *Length of the longest common subsequence between overlapping words.* SIAM J. Discrete Math., 34 (2018), no. 1, 721-729, arXiv: 1803.03238.

[3] B. Bukh and J. Ma, *Longest common subsequence in sets of words.* SIAM J. Discrete Math., 28 (2014), no. 4, 2042-2049, arXiv:1406.7017.

[4] B. Bukh and L. Zhou, *Twins in words and long common subsequences in permutations.* Israel Journal of Mathematics, 213 (2013), no. 1, 183-209, arXiv:1307.088.

[5] S. Connolly, Z. Gabor, and A. Godbole, *The Location of the first ascent in a 123-avoiding permutation*, arXiv:1401.2691.

[6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, MIT Press and McGraw-Hill (2001), 350–355.

[7] . E. Deutsch, *Dyck path enumeration*, Discrete Mathematics 204 (1999), 167-202, https://doi.org/10.1016/S0012-365X(98)00371-9.

[8] M. Dudik. and L.J. Schulman, *Reconstruction from subsequences.* J. Comb. Theory, Ser. A 103 (2003), no. 2, 337–348, https://doi.org/10.1016/S0097-3165(03)00103-1.

[9] D. Henshall, N. Rampersad, and J. Shallit, *Shuffling and Unshuffling.* Bull. EATCS, 107 (2012), 131-142, arXiv:1106.5767.

[10] D. S. Hirschberg, *A linear space algorithm for computing maximal common subsequences*, Communications of the ACM 18 (1975), no. 6, 341–343, https://doi.org/10.1145/360825.360861.

[11] J. Komlos and M. Simonovits, *Szemerédi's regularity lemma and its applications in graph theory.* In: Combinatorics, Paul Erdós is Eighty, Vol. 2 (Keszthely, 1993), volume 2 of Bolyai Soc. Math. Stud., pp. 295352. János Bolyai Math. Soc., Budapest, 1996.

[12] A. Mateescu, A. Salomaa, and S. Yu. *Subword histories and parikh matrices.* J. Comput. Syst. Sci., 68 (2004), no. 1, 1–21, https://doi.org/10.1016/j.jcss.2003.04.001.

[13] R. Rizzi and S. Viallete, *On recognizing words that are squares for the shuffle product.* Theoretical Computer Science. International Computer Science Symposium in Russia (2013), 10.1016/j.tcs.2017.04.003.

[14] A. Salomaa. *Counting (scattered) subwords.* Bulletin of the EATCS, 81 (2003),165–179, https://doi.org/10.1142/9789812562494_0061.

[15] H.S. Wilf, generatingfunctionology, 2nd ed., Academic Press, New York, 1994, https://doi.org/10.1016/C2009-0-02369-1.

[16] X. Xia, *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Pro- teomics and Transcriptomics.* New York: Springer, 2007, https://doi.org/10.1007/978-0-387-71337-3.