DENSITY DICHOTOMY IN RANDOM WORDS

JOSHUA COOPER & DANNY RORABAUGH

Department of Mathematics, University of South Carolina
cooper@math.sc.edu & rorabaug@email.sc.edu

Abstract. Word W is said to *encounter* word V provided there is a homomorphism ϕ mapping letters to nonempty words so that $\phi(V)$ is a substring of W. For example, taking ϕ such that $\phi(h) = c$ and $\phi(u) = ien$, we see that "science" encounters "huh" since $cienc = \phi(huh)$. The density of V in W is the proportion of substrings of W that are homomorphic images of V. So the density of "huh" in "science" is $2/\binom{8}{2}$. A word is *doubled* if every letter that appears in the word appears at least twice.

The dichotomy: Let V be a word over any alphabet, Σ a finite alphabet with at least 2 letters, and $W_n \in \Sigma^n$ chosen uniformly at random. V is doubled if and only if $\delta(V, W_n) = 0$ asymptotically almost surely.

Other density results we prove include convergence for nondoubled words and concentration of the limit distribution for doubled words.

1. Definitions

We discuss here (free) words: elements of the semigroup formed from a nonempty alphabet Σ with the binary operation of concatenation, denoted by juxtaposition, and with the empty word ε as the identity element. The set of all finite words over Σ is Σ^* and the set of Σ-words of length $k \in \mathbb{N}$ is Σ^k. For alphabets Γ and Σ, a homomorphism $\phi : \Gamma^* \to \Sigma^*$ is uniquely defined by a function $\phi : \Gamma \to \Sigma^*$. We call a homomorphism *nonerasing* provided it is defined by $\phi : \Gamma \to \Sigma^* \setminus \{\varepsilon\}$; that is, no letter maps to ε.

Let V and W be words. The *length* of W, denoted $|W|$, is the number of letters in W, including multiplicity. Denote with $\mathrm{L}(W)$ the set of letters found in W and with $||W||$ the number of letter repeats in W, so $|W| = |\mathrm{L}(W)| + ||W||$. For example $|banana| = 6$, $\mathrm{L}(banana) = \{a, b, n\}$, and $||banana|| = 3$. W has $\binom{|W|+1}{2}$ *substrings*, each defined by an ordered pair (i, j) with $0 \le i < j \le |W|$. Denote with $W[i, j]$ the word found in the (i, j)-substring, which consists of $j - i$ consecutive letters of W, beginning with the $(i + 1)$-th. V is a *factor* of W, denoted $V \le W$, provided $V = W[i, j]$ for some $0 \le i < j \le |W|$; that is, $W = SUT$ for some (possibly empty) words S and T. For example, $banana[2, 6] = nana \le banana$.

W is an *instance* of V, or V*-instance*, provided there exists a nonerasing homomorphism ϕ such that $W = \phi(V)$. For example, *banana* is an instance of *cool* using homomorphism ϕ defined by $\phi(c) = b$, $\phi(o) = an$, and $\phi(l) = a$. W *encounters* V, denoted $V \preceq W$, provided W' is an instance of V for some factor $W' \le W$. For example $cool \preceq bananasplit$. For $W \ne \varepsilon$, denote with $\delta(V, W)$ the proportion of substrings of W that give instances of V. For example, $\delta(xx, banana) = 2/\binom{7}{2}$. $\delta_{sur}(V, W)$ is the characteristic function for the event that W is an instance of V.

Fix alphabets Γ and Σ. An *encounter* of V in W is an ordered triple (a, b, ϕ) where $W[a, b] = \phi(V)$ for homomorphism $\phi : \Gamma^* \to \Sigma^*$. When $\Gamma = \mathrm{L}(V)$ and $W \in \Sigma^*$, denote with $\hom(V, W)$ the number of encounters of V in W. For example, $\hom(ab, cde) = 4$ since $cde[0, 2]$ and $cde[1, 3]$ are instances of ab, each for one homomorphism $\{a, b\}^* \to \{c, d, e\}^*$, and $cde[0, 3]$ is an instance of ab under two homomorphisms. Note that the conditions on Γ and Σ are necessary for $\hom(V, W)$ to not be 0 or ∞.

Fact 1. *For fixed words V and $W \neq \varepsilon$,*

$$\binom{|W|+1}{2}\delta(V,W) \leq \hom(V,W).$$

2. Background

Word encounters have primarily been explored from the perspective of avoidance. Word W *avoids* a (pattern) word V provided $V \not\preceq W$. V is *k-avoidable* provided, from a k-letter alphabet, there are infinitely many words that avoid V. The premier result on word avoidance is generally considered to be the proof of Thue [10] that the word aa is 3-avoidable but not 2-avoidable. Two seminal papers on avoidability, by Bean, Ehrenfeucht, and McNulty [1] and Zimin [11, 12], include classification of unavoidable words–that is, words that are not k-avoidable for any k. Recently, the authors [5] and Tao [9] investigated bounds on the length of words that avoid unavoidable words. There remain a number of open problems regarding which words are k-avoidable for particular k. See Lothaire [7] and Currie [6] for surveys on avoidability results and Blanchet-Sadri and Woodhouse [3] for recent work on 3-avoidability.

A word is *doubled* provided every letter in the word occurs at least twice. Every doubled word is k-avoidable for some $k > 1$ [7]. For a doubled word V with $k \geq 2$ distinct letters and an alphabet Σ with $|\Sigma| = q \geq 4$, $(k,q) \neq (2,4)$, Bell and Goh [2] showed that there are at least $\lambda(k,q)^n$ words in Σ^n that avoid V, where

$$\lambda(k,q) = m\left(1 + \frac{1}{(m-2)^k}\right)^{-1}.$$

This exponential lower bound on the number of words avoiding a doubled word hints at the moral of the present work: instances of doubled words are rare. For a doubled word V and an alphabet Σ with at least 2 letters, the probability that a random word $W_n \in \Sigma^n$ avoids V is asymptotically 0. Indeed, the event that $W_n[b|V|,(b+1)|V|]$ is an instance of V has nonzero probability and is independent for distinct b. Nevertheless, $\delta(V,W_n)$, the proportion of substrings of W that are instances of V, is asymptotically negligible.

We find motivation for considering word densities from the central importance of graph densities in the combinatorial limit theory of graphs (see Lovász [8]).

2.1. The Dichotomy.

Theorem 2. *Let V be a word on any alphabet. Fix an alphabet Σ with $q \geq 2$ letters, and let $W_n \in \Sigma^n$ be chosen uniformly at random. The following are equivalent:*

((i)) *V is doubled (that is, every letter in V appears at least twice);*
((ii)) $\lim_{n\to\infty} \mathbb{E}(\delta(V,W_n)) = 0$*;*
((iii)) $\delta(V,W_n) = 0$ *asymptotically almost surely.*

Proof. $(ii) \iff (iii)$ since $\delta(V,W_n)$ is a nonnegative random variable.

First we prove $(i) \implies (ii)$. Note that in W_n, there are in expectation the same number of encounters of V as there are of any anagram of V. Indeed, if V' is an anagram of V and ϕ is a nonerasing homomorphism, then $|\phi(V')| = |\phi(V)|$.

Fact 3. *If V' is an anagram of V, then* $\mathbb{E}(\hom(V,W_n)) = \mathbb{E}(\hom(V',W_n))$.

Assume V is doubled and let $\Gamma = \mathrm{L}(V)$ and $k = |\Gamma|$. Given Fact 3, we consider an anagram $V' = XY$ of V, where $|X| = k$ and $\Gamma = \mathrm{L}(X) = \mathrm{L}(Y)$. That is, X comprises one copy of each letters in Γ and all the duplicate letters of V are in Y.

We obtain an upper bound for the average density of V by estimating $\mathbb{E}(\hom(V',W_n))$. To do so, sum over starting position i and length j of encounters of X in W_n that might extend to an encounter of V'. There are $\binom{j+1}{k+1}$ homomorphisms ϕ that map X to $W[i,i+j]$ and the probability that $W_n[i+j, i+j+|\phi(Y)|] = \phi(Y)$ is at most q^{-j}. Also, the series $\sum_{j=k}^{\infty} \binom{j+1}{k+1} q^{-j}$ converges (try the ratio test) to some c not dependent on n.

$$\begin{aligned}
\mathbb{E}(\delta(V, W_n)) &\leq \frac{1}{\binom{n+1}{2}} \mathbb{E}\left(\hom(V', W_n)\right) \\
&< \frac{1}{\binom{n+1}{2}} \sum_{i=0}^{n-|V|} \sum_{j=k}^{n-i} \binom{j+1}{k+1} q^{-j} \\
&< \frac{1}{\binom{n+1}{2}} \sum_{i=0}^{n-|V|} c \\
&= \frac{c(n-|V|+1)}{\binom{n+1}{2}} \\
&= O(n^{-1}),
\end{aligned}$$

We prove $(ii) \Longleftarrow (i)$ by contraposition. Assume there is a letter x that occurs exactly once in V. Write $V = TxU$ where $\mathrm{L}(V) \setminus \mathrm{L}(TU) = \{x\}$. We obtain a lower bound for $\mathbb{E}(\delta(V, W_n))$ by only counting encounters with $|\phi(TU)| = |TU|$. Note that each such encounter is unique to its instance, preventing double-counting. For this undercount, we sum over encounters with $W_n[i, i+j] = \phi(x)$.

$$\begin{aligned}
\mathbb{E}(\delta(V, W_n)) &= \mathbb{E}(\delta(TxU, W_n)) \\
&\geq \frac{1}{\binom{n+1}{2}} \sum_{i=|T|}^{n-|U|-1} \sum_{j=1}^{i-|T|} q^{-||TU||} \\
&= q^{-||TU||} \frac{1}{\binom{n+1}{2}} \sum_{i=|T|}^{n-|U|-1} (i - |T|) \\
&= q^{-||TU||} \frac{\binom{n-|UT|}{2}}{\binom{n+1}{2}} \\
&\sim q^{-||TU||} \\
&> 0.
\end{aligned}$$

□

It behooves us now to develop more precise theory for these two classes of words: doubled and nondoubled. Lemma 5 below both helps develop that theory and gives insight into the detrimental effect that letter repetition has on encounter frequency.

Proposition 4. *For $\overline{r} = \{r_1, \ldots, r_k\} \in (\mathbb{Z}^+)^k$ and $d = \gcd_{i \in [k]}(r_i)$, there exists integer $N = N_{\overline{r}}$ such that for every $n > N$ there exist coefficients $a_1, \cdots, a_k \in \mathbb{Z}^+$ such that $dn = \sum_{i=1}^k a_i r_i$ and $a_i \leq N$ for $i \geq 2$.*

Proof. For each $j \in [r_1/d]$, find integer coefficients $b_i^{(j)}$ for $i \in [k]$ so that $jd = \sum_{i=1}^k b_i^{(j)} r_i$. Let $m = 1 + \left|\min\left(b_i^{(j)}\right)\right|$, the minimum taken over all i and j. Define $a_i^{(j)} = b_i^{(j)} + m > 0$ and $R = \sum_{i=1}^k r_i$. Now for each j,

$$\sum_{i=1}^k a_i^{(j)} r_i = \sum_{i=1}^k b_i^{(j)} r_i + \sum_{i=1}^k m r_i = jd + mR.$$

Set $N = r_1 + mR$. For $n > N$, identify $j_n \in [r_1/d]$ such that $dn \equiv j_n d + mR \pmod{r_1}$. Then $a_i = a_i^{(j_n)}$ for $i > 1$ and $a_1 = \frac{1}{r_1}\left(dn - \sum_{i=2}^k a_i r_i\right)$. □

Lemma 5. *For any word V, Let $\Gamma = \mathrm{L}(V) = \{x_1, \dots, x_k\}$ where x_i has multiplicity r_i for each $i \in [k]$. Let U be V with all letters of multiplicity $r = \min_{i\in[k]}(r_i)$ removed. Finally, let Σ be any finite alphabet with $|\Sigma| = q \geq 2$ letters. Then for a uniformly randomly chosen V-instance $W \in \Sigma^{dn}$, where $d = \gcd_{i\in[k]}(r_i)$, there is asymptotically almost surely a homomorphism $\phi : \Gamma^* \to \Sigma^*$ with $\phi(V) = W$ and $|\phi(U)| < \sqrt{dn}$.*

Proof. Let a_n be the number of V-instances in Σ^n and b_n be the number of homomorphisms $\phi : \Gamma^* \to \Sigma^*$ such that $|\phi(V)| = n$. Let b_n^1 be the number of these ϕ such that $\phi(U) < \sqrt{n}$ and b_n^2 the number of all other ϕ so that $b_n = b_n^1 + b_n^2$. Similarly, let a_n^1 be the number of V-instances in Σ^n for which there exists a ϕ counted by b_n^1 and a_n^2 the number of instances with no such ϕ, so $a_n = a_n^1 + a_n^2$. Observe that $a_n^2 \leq b_n^2$.

Without loss of generality, assume $r_1 = r$ (rearrange the x_i if not). We now utilize $N = N_{\overline{r}}$ from Proposition 4. For sufficiently large n, we can undercount a_{dn}^1 by counting homomorphisms ϕ with $|\phi(x_i)| = a_i$ for the a_i attained from Proposition 4. Indeed, distinct homomorphisms with the same image-length for every letter in V produce distinct V-instances. Hence

$$\begin{aligned} a_{dn}^1 &\geq q^{\sum_{i=1}^k a_i} \\ &\geq q^{\left(\frac{dn-(k-1)N}{r} + r(k-1)\right)} \\ &= cq^{\left(\frac{dn}{r}\right)}, \end{aligned}$$

where $c = q^{(k-1)(r^2-N)/r}$ depends on V but not on n. To overcount b_n^2 (and a_{dn}^2 by extension), we consider all $\binom{n+1}{|V|+1}$ ways to partition an n-letter length and so determine the lengths of the images of the letters in V. However, for letters with multiplicity strictly greater than r, the sum of the lengths of their images must be at least $\sqrt{n}$.

$$\begin{aligned} b_n^2 &\leq \binom{n+1}{|V|+1} \sum_{i=\lceil\sqrt{n}\rceil}^{n} q^{\left(\frac{n-i}{r} + \frac{i}{r+1}\right)} \\ &= \binom{n+1}{|V|+1} \sum_{i=\lceil\sqrt{n}\rceil}^{n} q^{\left(\frac{n}{r} - \frac{i}{r(r+1)}\right)} \\ &< n^{|V|+2} q^{\left(\frac{n}{r} - \frac{\sqrt{n}}{r(r+1)}\right)} \\ &= q^{\frac{n}{r}} o(1). \end{aligned}$$

$$\begin{aligned} a_{dn}^2 &\leq b_{dn}^2 \\ &= o(a_{dn}^1). \end{aligned}$$

That is, the proportion of V-instances of length dn that cannot be expressed with $|\phi(U)| < \sqrt{dn}$ diminishes to 0 as n grows. □

3. Density of Nondoubled Words

In Theorem 2, we show that the density of nondoubled V in long random words (over a fixed alphabet with at least two letters) does not approach 0. The natural follow-up question is: Does the density converge? To answer this question, we first prove the following lemma. Fixing $V = TxU$ where x is a nonrecurring letter in V, the lemma tells us that all but a diminishing proportion of V-instances can be obtained by some ϕ with $|\phi(TU)|$ negligible.

Lemma 6. *Let $V = U_0 x_1 U_1 x_2 \cdots x_r U_r$ with $r \geq 1$, where $U = U_0 U_1 \cdots U_r$ is doubled with k distinct letters (though any particular U_j may be the empty word), the x_i are distinct, and no x_i occurs in U.*

Further, let Γ be the $(k+r)$-letter alphabet of V and let Σ be any finite alphabet with $q \geq 2$ letters. Then there exists a nondecreasing function $g(n) = o(n)$ such that, for a randomly chosen V-instance $W \in \Sigma^n$, there is asymptotically almost surely a homomorphism $\phi : \Gamma^* \to \Sigma^*$ with $\phi(V) = W$ and $|\phi(x_r)| > n - g(n)$.

Proof. Let $X_i = x_1 x_2 \cdots x_i$ for $0 \leq i \leq r$ (so $X_0 = \varepsilon$). For any word W, let Φ_W be the set of homomorphisms $\{\phi : \Gamma^* \to \Sigma^* \mid \phi(V) = W\}$ that map V onto W. Define $\mathbf{P}_i$ to be the following proposition for $i \in [r]$:

> There exists a nondecreasing function $f_i(n) = o(n)$ such that, for a randomly chosen V-instance $W \in \Sigma^n$, there is asymptotically almost surely a homomorphism $\phi \in \Phi_W$ such that $|\phi(UX_{i-1})| \leq f_i(n)$.

The conclusion of this lemma is an immediate consequence of $\mathbf{P}_r$, with $g(n) = f_r(n)$, which we will prove by induction. Lemma 5 provides the base case, with $r = 1$ and $f_1(n) = \sqrt{n}$.

Let us prove the inductive step: $\mathbf{P}_i$ implies $\mathbf{P}_{i+1}$ for $i \in [r-1]$. Roughly speaking, this says: If most instances of V can be made with a homomorphism ϕ where $|\phi(UX_{i-1})|$ is negligible, then most instances of V can be made with a homomorphism ϕ where $|\phi(UX_i)|$ is negligible.

Assume $\mathbf{P}_i$ for some $i \in [r-1]$, and set $f(n) = f_i(n)$. Let A_n be the set of V-instances in Σ^n such that $|\phi(UX_{i-1})| \leq f(n)$ for some $\phi \in \Phi_W$. Let B_n be the set of all other V-instances in Σ^n. $\mathbf{P}_i$ implies $|B_n| = o(|A_n|)$.

Case 1: $U_i = \varepsilon$, so x_i and x_{i+1} are consecutive in V. When $|\phi(UX_{i-1})| \leq f(n)$, we can define ψ so that $\psi(x_i x_{i+1}) = \phi(x_i x_{i+1})$ and $|\psi(x_i)| = 1$; otherwise, let $\psi(y) = \phi(y)$ for $y \in \Gamma \setminus \{x_i, x_{i+1}\}$. Then $|\phi(UX_i)| \leq f(n) + 1$ and $\mathbf{P}_{i+1}$ with $f_{i+1}(n) = f_i(n) + 1$.

Case 2: $U_i \neq \varepsilon$, so $|U_i| > 0$. Let $g(n)$ be some nondecreasing function such that $f(n) = o(g(n))$ and $g(n) = o(n)$. (This will be the f_{i+1} for $\mathbf{P}_{i+1}$.) Let A_n^α consist of $W \in A_n$ such that $|\phi(UX_i)| \leq g(n)$ for some $\phi \in \Phi_W$. Let $A_n^\beta = A_n \setminus A_n^\alpha$. The objective henceforth is to show that $|A_n^\beta| = o(|A_n^\alpha|)$.

For $Y \in A_n^\beta$, let Φ_Y^β be the set of homomorphisms $\{\phi \in \Phi_Y : |\phi(UX_{i-1})| \leq f(n)\}$ that disqualify Y from being in B_n. Hence $Y \in A_n$ implies $\Phi_Y^\beta \neq \emptyset$. Since $Y \notin A_n^\alpha$, $\phi \in \Phi_Y^\beta$ implies $|\phi(UX_i)| > g(n)$, so $|\phi(x_i)| > g(n) - f(n)$. Pick $\phi_Y \in \Phi_Y^\beta$ as follows:

- Primarily, minimize $|\phi(U_0 x_1 U_1 x_2 \cdots U_{i-1} x_i)|$;
- Secondarily, minimize $|\phi(U_i)|$;
- Tertiarily, minimize $|\phi(U_0 x_1 U_1 x_2 \cdots U_{i-1})|$.

Roughly speaking, we have chosen ϕ_Y to move the image of U_i as far left as possible in Y. But since $Y \notin A_n^\alpha$, we want it further left!

To suppress the details we no longer need, let $Y = Y_1 \phi_Y(x_i) \phi_Y(U_i) \phi_Y(x_{i+1}) Y_2$, where $Y_1 = \phi_Y(U_0 x_1 U_1 x_2 \cdots U_{i-1})$ and $Y_2 = \phi_Y(U_{i+1} x_{i+2} \cdots U_r)$.

Consider a word $Z \in \Gamma^n$ of the form $Y_1 Z_1 \phi_Y(U_i) Z_2 \phi_Y(U_i) \phi_Y(x_{i+1}) Y_2$, where Z_1 is an initial string of $\phi_Y(x_i)$ with $2f(n) \leq |Z_1| < g(n) - 2f(n)$ and Z_2 is a final string of $\phi_Y(x_i)$. (See Figure 1.) In a sense, the image of x_i was too long, so we replace a leftward substring with a copy of the image of U_i. Let C_Y be the set of all such Z with $|Z_1|$ a multiple of $f(n)$. For every $Z \in C_Y$ we can see that $Z \in A_n^\alpha$, by defining $\psi \in \Phi_Z$ as follows:

$$\psi(y) = \begin{cases} Z_1 & \text{if } y = x_i; \\ Z_2 \phi_Y(U_i) \phi_Y(x_{i+1}) & \text{if } y = x_{i+1}; \\ \phi_Y(y) & \text{otherwise.} \end{cases}$$

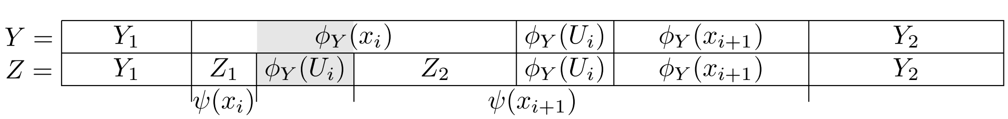

FIGURE 1. Replacing a section of $\phi_Y(x_i)$ in Y to create Z.

Claim 1: $\liminf_{|Y|=n\to\infty} |C_Y| = \infty$.

Since we want $2f(n) \le |Z_1| < g(n) - 2f(n)$, and $g(n) - 2f(n) < |\phi_Y(x_i)| - |\phi_Y(U_i)|$, there are $g(n)-4f(n)$ places to put the copy of $\phi_Y(U_i)$. To avoid any double-counting that might occur when some Z and Z' have their new copies of $\phi_Y(U_i)$ in overlapping locations, we further required that $f(n)$ divide $|Z_1|$. This produces the following lower bound:

$$|C_Y| \ge \left\lfloor \frac{g(n) - 4f(n)}{f(n)} \right\rfloor \to \infty.$$

Claim 2: For distinct $Y, Y' \in A_n^\beta$, $C_Y \cap C_{Y'} = \emptyset$.

To prove Claim 2, take $Y, Y' \in A_n^\beta$ with $Z \in C_Y \cap C_{Y'}$. Define $Y_1 = \phi_Y(U_0x_1U_1x_2\cdots U_{i-1})$ and $Y_2 = \phi_Y(U_{i+1}x_{i+2}\cdots U_r)$ as before and $Y_1' = \phi_{Y'}(U_0x_1U_1x_2\cdots U_{i-1})$ and $Y_2' = \phi_{Y'}(U_{i+1}x_{i+2}\cdots U_r)$. Now for some Z_1, Z_1', Z_2, Z_2',

$$Y_1Z_1\phi_Y(U_i)Z_2\phi_Y(U_i)\phi_Y(x_{i+1})Y_2 = Z = Y_1'Z_1'\phi_{Y'}(U_i)Z_2'\phi_{Y'}(U_i)\phi_{Y'}(x_{i+1})Y_2',$$

with the following constraints:

((i)) $|Y_1\phi_Y(U_i)| \le |\phi_Y(UX_i)| \le f(n)$;
((ii)) $|Y_1'\phi_{Y'}(U_i)| \le |\phi_{Y'}(UX_i)| \le f(n)$;
((iii)) $2f(n) \le |Z_1| < g(n) - 2f(n)$;
((iv)) $2f(n) \le |Z_1'| < g(n) - 2f(n)$;
((v)) $|Z_1\phi_Y(U_i)Z_2| = |\phi_Y(x_i)| > g(n) - f(n)$;
((vi)) $|Z_1'\phi_{Y'}(U_i)Z_2'| = |\phi_{Y'}(x_i)| > g(n) - f(n)$.

As a consequence:

- $|Y_1Z_1\phi_Y(U_i)| < g(n) - f(n) < |Z_1'\phi_{Y'}(U_i)Z_2'|$, by (i), (iii), and (vi);
- $|Y_1Z_1| \ge |Z_1| > 2f(n) > |Y_1'|$, by (iii) and (ii).

Therefore, the copy of $\phi_Y(U_i)$ added to Z is properly within the noted occurrence of $Z_1'\phi_{Y'}(U_i)Z_2'$ in Z', which is in the place of $\phi_{Y'}(x_i)$ in Y'. In particular, the added copy of $\phi_Y(U_i)$ in Z interferes with neither Y_1' nor the original copy of $\phi_{Y'}(U_i)$. Thus Y_1' is an initial substring of Y and $\phi_{Y'}(U_i)\phi_{Y'}(x_{i+1})Y_2'$ is a final substring of Y. Likewise, Y_1 is an initial substring of Y' and $\phi_Y(U_i)\phi_Y(x_{i+1})Y_2$ is a final substring of Y'. By the selection process of ϕ_Y and $\phi_{Y'}$, we know that $Y_1 = Y_1'$ and $\phi_Y(U_i)\phi_Y(x_{i+1})Y_2 = \phi_{Y'}(U_i)\phi_{Y'}(x_{i+1})Y_2'$. Finally, since $f(n)$ divides Z_1 and Z_1', we deduce that $Z_1 = Z_1'$. Otherwise, the added copies of $\phi_Y(U_i)$ in Z and of $\phi_{Y'}(U_i)$ in Z' would not overlap, resulting in a contradiction to the selection of ϕ_Y and $\phi_{Y'}$. Therefore, $Y = Y'$, concluding the proof of Claim 2.

Now $C_Y \subset A_n^\alpha$ for $Y \in A_n^\beta$. Claim 1 and Claim 2 together imply that $|A_n^\beta| = o(|A_n^\alpha|)$.

□

Observe that the choice of $\sqrt{n}$ in Lemma 5 was arbitrary. The proof works for any function $f(n) = o(n)$ with $f(n) \to \infty$. Therefore, where Lemma 6 claims the existence of some $g(n) \to \infty$, the statement is in fact true for all $g(n) \to \infty$.

Let $\mathbb{I}_n(V, \Sigma)$ be the probability that a uniformly randomly selected length-n Σ-word is an instance of V. That is,

$$\mathbb{I}_n(V, \Sigma) = \frac{|\{W \in \Sigma^n \mid \phi(V) = W \text{ for some homomorphism } \phi : \mathrm{L}(V)^* \to \Sigma^*\}|}{|\Sigma|^n}.$$

Fact 7. *For any V and Σ and for $W_n \in \Sigma^n$ chosen uniformly at random,*

$$\binom{n+1}{2}\mathbb{E}(\delta(V, W_n)) \quad = \quad \sum_{m=1}^{n}(n+1-m)\mathbb{E}(\delta_{sur}(V, W_m))$$

$$= \sum_{m=1}^{n}(n+1-m)\mathbb{I}_m(V,\Sigma).$$

Denote $\mathbb{I}(V,\Sigma) = \lim_{n\to\infty}\mathbb{I}_n(V,\Sigma)$. When does this limit exist?

Theorem 8. *For nondoubled V and alphabet Σ, $\mathbb{I}(V,\Sigma)$ exists. Moreover, $\mathbb{I}(V,\Sigma) > 0$.*

Proof. If $|\Sigma| = 1$, then $\mathbb{I}_n(V,\Sigma) = 1$ for $n \geq |V|$.

Assume $|\Sigma| = q \geq 2$. Let $V = TxU$ where x is the right-most nonrecurring letter in V. Let $\Gamma = \mathrm{L}(V)$ be the alphabet of letters in V. By Lemma 6, there is a nondecreasing function $g(n) = o(n)$ such that, for a randomly chosen V-instance $W \in \Sigma^n$, there is asymptotically almost surely a homomorphism $\phi : \Gamma^* \to \Sigma^*$ with $\phi(V) = W$ and $|\phi(x_r)| > n - g(n)$.

Let a_n be the number of $W \in \Sigma^n$ such that there exists $\phi : \Gamma^* \to \Sigma^*$ with $\phi(V) = W$ and $|\phi(x_r)| > n - g(n)$. Lemma 6 tells us that $\frac{a_n}{q^n} \sim \mathbb{I}_n(V,\Sigma)$. Note that $\frac{a_n}{q^n}$ is bounded. It suffices to show that $a_{n+1} \geq qa_n$ for sufficiently large n. Pick n so that $g(n) < \frac{n}{3}$.

For length-n V-instance W counted by a_n, let ϕ_W be a homomorphism that maximizing $|\phi_W(x_r)|$ and, of such, minimizes $|\phi_W(T)|$. For each ϕ_W and each $a \in \Sigma$, let ϕ_W^a be the function such that, if $\phi_W(x_r) = AB$ with $|A| = \lfloor|\phi_W(x_r)|/2\rfloor$, then $\phi_W^a(x) = AaB$; $\phi_W^a(y) = \phi_W(y)$ for each $y \in \Gamma \setminus \{x\}$ Roughly speaking, we are sticking a into the middle of the image of x.

Suppose we are double-counting, so $\phi_W^a(V) = \phi_Y^b(V)$. As

$$|\phi_W(x_r)|/2 > (n - g(n))/2 > n/3 > g(n) \geq |\phi_Y(TU)|$$

and vice-versa, the inserted a (resp., b) of one map does not appear in the image of TU under the other map. So $\phi_W(T)$ is an initial string and $\phi_W(U)$ a final string of $\phi_Y(V)$, and vice-versa. By the selection criteria of ϕ_W and ϕ_Y, $|\phi_W(T)| = |\phi_Y(T)|$ and $|\phi_W(U)| = |\phi_Y(U)|$. Therefore the location of the added a in $\phi_W^a(V)$ and the added b in $\phi_W^b(V)$ are the same. Hence, $a = b$ and $W = Y$.

Moreover $\mathbb{I}(V,\Sigma) \geq q^{-||V||} > 0$. □

Example 9. *Let $V = x_1x_2\cdots x_k$ have k distinct letters. Since every word of length at least k is a V-instance, $\mathbb{I}(V,\Sigma) = 1$ for every alphabet Σ. When even one letter in V is repeated, finding $\mathbb{I}(V,\Sigma)$ becomes a nontrivial task.*

Example 10. *Zimin's classification of unavoidable words is as follows* [11, 12]*: Every unavoidable word with n distinct letters is encountered by Z_n, where $Z_0 = \varepsilon$ and $Z_{i+1} = Z_ix_{i+1}Z_i$ with x_{i+1} a letter not occurring in Z_i. For example, $Z_2 = aba$ and $Z_3 = abacaba$. The authors can calculate $\mathbb{I}(Z_2,\Sigma)$ and $\mathbb{I}(Z_3,\Sigma)$ to arbitrary precision* [4].

TABLE 1. $\mathbb{I}(Z_2,\Sigma)$ and $\mathbb{I}(Z_3,\Sigma)$ computed to 7 decimal places.

$\lvert\Sigma\rvert$	*2*	*3*	*4*	*5*	*6*	*7*	$\cdots$
$\mathbb{I}(Z_2,\Sigma)$	*0.7322132*	*0.4430202*	*0.3122520*	*0.2399355*	*0.1944229*	*0.1632568*	$\cdots$
$\mathbb{I}(Z_3,\Sigma)$	*0.1194437*	*0.0183514*	*0.0051925*	*0.0019974*	*0.0009253*	*0.0004857*	$\cdots$

Corollary 11. *Let V be a nondoubled word on any alphabet. Fix an alphabet Σ, and let $W_n \in \Sigma^n$ be chosen uniformly at random. Then*

$$\lim_{n\to\infty}\mathbb{E}(\delta(V,W_n)) = \mathbb{I}(V,\Sigma).$$

Proof. Let $\mathbb{I} = \mathbb{I}(V,\Sigma)$ and $\epsilon > 0$. Pick $N = N_\epsilon$ sufficiently large so $|\mathbb{I} - \mathbb{I}_n(V,\Sigma)| < \frac{\epsilon}{2}$ when $n > N$. Applying Fact 7 for $n > \max(N, 4N/\epsilon)$,

$$|\mathbb{I} - \mathbb{E}(\delta(V,W_n))| = \left|\mathbb{I}\frac{1}{\binom{n+1}{2}}\sum_{m=1}^{n}(n+1-m) - \frac{1}{\binom{n+1}{2}}\sum_{m=1}^{n}(n+1-m)\mathbb{I}_m(V,\Sigma)\right|$$

$$\begin{aligned}
&\le \frac{1}{\binom{n+1}{2}}\sum_{m=1}^{n}(n+1-m)|\mathbb{I}-\mathbb{I}_m(V,\Sigma)| \\
&= \frac{1}{\binom{n+1}{2}}\left[\sum_{m=1}^{N}+\sum_{m=N+1}^{n}\right](n+1-m)|\mathbb{I}-\mathbb{I}_m(V,\Sigma)| \\
&< \frac{1}{\binom{n+1}{2}}\left[\sum_{m=1}^{\lfloor \epsilon n/4\rfloor}(n+1-m)1+\sum_{m=N+1}^{n}(n+1-m)\frac{\epsilon}{2}\right] \\
&< \frac{1}{\binom{n+1}{2}}\left[\frac{\epsilon n}{4}n+\binom{n+1}{2}\frac{\epsilon}{2}\right] \\
&< \epsilon.
\end{aligned}$$

□

4. Density of Doubled Words

Our main dichotomy says that the average density of a doubled word in large random words (over a fixed alphabet with at least two letters) goes to 0. Thus the expected number of instances in a random word of length n is $o(n^2)$. Perhaps we can find lower-order asymptotics for the expected number of instances of a doubled word. Hencefore, if $\binom{x}{y}$ is used with nonintegral x, we mean

$$\binom{x}{y} = \frac{\prod_{i=0}^{y-1}(x-i)}{y!}.$$

Proposition 12. *For $\overline{r} = (r_1, \ldots, r_k) \in (\mathbb{Z}^+)^k$, let $a_n(\overline{r})$ be the number of k-tuples $\overline{a} = (a_1, \cdots, a_k) \in (\mathbb{Z}^+)^k$ so that $\sum_{i=1}^{k} a_i r_i = n$. Then $a_n(\overline{r}) \le \binom{n/d+k+1}{k+1}$, where $d = \gcd_{i\in[k]}(r_i)$.*

Proof. If $d \nmid n$, then $a_n(\overline{r}) = 0$. Otherwise, for each $\overline{a}$ counted by $a_n(\overline{r})$, there is a unique corresponding $\overline{b} \in (\mathbb{Z}^+)^k$ such that $1 \le b_1 < b_2 < \cdots < b_k = n/d$ and $b_j = \frac{1}{d}\sum_{i=1}^{j} a_i r_i$. The number of strictly increasing k-tuples of positive integers with largest value n/d is $\binom{n/d+k+1}{k+1}$. □

Fix alphabet Σ with $q = |\Sigma|$. The number of instances of V in Σ^n is $q^n \mathbb{I}_n(V,\Sigma)$. Assume V is doubled. Let $\Gamma = \mathrm{L}(V) = \{x_1, \ldots, x_k\}$ and r_i be the multiplicity of x_i in V for each $i \in [k]$. Let $d = \gcd_{i\in[k]}(r_i)$ and $r = \min_{i\in[k]}(r_i)$. Note that $\mathbb{I}_n(V,\Sigma) = 0$ when $d \nmid n$. But perhaps

$$\lim_{\substack{n\to\infty \\ d|n}} \frac{q^n}{f(n)}\mathbb{I}_n(V,\Sigma)$$

exists for some function f that only depends on q and V. For inspiration, note that $q^n\mathbb{I}_n(U^m,\Sigma) = q^{n/m}\mathbb{I}_{n/m}(U,\Sigma)$ when $m \mid n$. Furthermore, using Proposition 12,

$$q^n\mathbb{I}_n(V,\Sigma) \le \mathbb{E}(\hom(V,W_n)) < \binom{n/d+k+1}{k+1}q^{n/r}. \tag{1}$$

Now select some letter x of multiplicity r and let U be V with all copies of x removed. When $r|(n-|U|)$, we can get a lower bound on the number of instances by counting homomorphism ϕ with $|\phi(U)| = |U| = |V| - r$:

$$q^n\mathbb{I}_n(V,\Sigma) \ge q^{(n-|U|)/r+(k-1)} = (q^{k-|V|/r})q^{n/r}. \tag{2}$$

Conjecture 13. *The following limit exists:*

$$\lim_{\substack{n\to\infty \\ d|n}} q^{n(1-1/r)}\mathbb{I}_n(V,\Sigma).$$

By (2), the limit (if it exists) cannot be 0. Theorem 8 is a special case of this conjecture, with $d = r = 1$.

5. Concentration

For doubled V and $|\Sigma| > 1$, we established that the expectation of the density $\delta(V, W_n)$ converges to zero. What is the concentration of the distribution of this density? By (1), we can bound the probability that randomly chosen $W_n \in \Sigma^n$ is a V-instance:

$$\mathbb{P}(\delta_{sur}(V, W_n) = 1) = \mathbb{I}_n(V, \Sigma) \leq \binom{n/d + k + 1}{k + 1} q^{n(1-r)/r}.$$

From this observation we get the following probabilistic result (which is only interesting for $q, r > 1$).

Lemma 14. *Let V be a word with k distinct letters, each occurring at least $r \in \mathbb{Z}^+$ times. Let Σ be a q-letter alphabet and $W_n \in \Sigma^n$ chosen uniformly at random. Recall that $\binom{n+1}{2}\delta(V, W_n)$ is the number substrings of W_n that are V-instances. Then for any nondecreasing function $f(n) > 0$,*

$$\mathbb{P}\left(\binom{n+1}{2}\delta(V, W_n) > n \cdot f(n)\right) < n^{k+3} q^{f(n)(1-r)/r}.$$

Proof. Since $\delta_{sur}(V, W) \in \{0, 1\}$,

$$\sum_{m=1}^{\lfloor f(n) \rfloor} \sum_{\ell=0}^{n-m} \delta_{sur}(V, W_n[\ell, \ell + m]) \quad < \quad n \cdot f(n).$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\binom{n+1}{2}\delta(V, W_n) > n \cdot f(n)\right) &= \mathbb{P}\left(\sum_{m=1}^{n} \sum_{\ell=0}^{n-m} \delta_{sur}(V, W_n[\ell, \ell + m]) > n \cdot f(n)\right) \\
&< \mathbb{P}\left(\sum_{m=\lceil f(n) \rceil}^{n} \sum_{\ell=0}^{n-m} \delta_{sur}(V, W_n[\ell, \ell + m]) > 0\right) \\
&< \sum_{m=\lceil f(n) \rceil}^{n} \sum_{\ell=0}^{n-m} \mathbb{P}\left(\delta_{sur}(V, W_n[\ell, \ell + m]) > 0\right) \\
&= \sum_{m=\lceil f(n) \rceil}^{n} (n - m + 1)\mathbb{P}\left(\delta_{sur}(V, W_m) = 1\right) \\
&\leq \sum_{m=\lceil f(n) \rceil}^{n} (n - m + 1)\binom{m/d + k + 1}{k + 1} q^{m(1-r)/r} \\
&< n(n - m + 1)\binom{n/d + k + 1}{k + 1} q^{f(n)(1-r)/r} \\
&< n^{k+3} q^{f(n)(1-r)/r}.
\end{aligned}$$

□

Theorem 15. *Let V be a doubled word, Σ an alphabet with $q \geq 2$ letters, and $W_n \in \Sigma^n$ chosen uniformly at random. Then the p^{th} raw moment and the p^{th} central moment of $\delta(V, W_n)$ are both $O\left((\log(n)/n)^p\right)$.*

Proof. Let us use Lemma 14 to first bound the p-th raw moments for $\delta(V, W_n)$, assuming $r \geq 2$. To minimize our bound, we define the following function on n, which acts as a threshold for "short" substrings of a random length-n word:

$$s_p(n) = \frac{r}{1 - r} \log_q(n^{-(k+5+p)}) = t_p \log n,$$

where $t_p = \frac{r(k+5+p)}{(r-1)\log(q)} > 0$.

$$\begin{aligned}
\mathbb{E}(\delta(V, W_n)^p) &= \sum_{i=0}^{\binom{n+1}{2}} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right)\left(\frac{i}{\binom{n+1}{2}}\right)^p \\
&< \sum_{i=0}^{\lfloor n\cdot s_p(n)\rfloor} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right)\left(\frac{i}{\binom{n+1}{2}}\right)^p \\
&\quad + \sum_{i=\lceil n\cdot s_p(n)\rceil}^{\binom{n+1}{2}} n^{k+3} q^{s_p(n)(1-r)/r}\left(\frac{i}{\binom{n+1}{2}}\right)^p \\
&< \left(\frac{n\cdot s_p(n)}{\binom{n+1}{2}}\right)^p + n^{k+5} q^{s_p(n)(1-r)/r} \\
&= \left(\frac{nt_p \log n}{\binom{n+1}{2}}\right)^p + n^{k+5} q^{\log_q\left(n^{-(k+5+p)}\right)} \\
&= O_p\left(\left(\frac{\log n}{n}\right)^p\right).
\end{aligned}$$

Setting $p = 1$, there exists some $c > 2$ such that $\mathbb{E}_n = \mathbb{E}(\delta(V, W_n)) < (c\log n)/n$. We use this upper bound on the expectation (1st raw moment) to bound the central moments.

$$\begin{aligned}
\mathbb{E}(|\delta(V, W_n) - \mathbb{E}_n|^p) &= \sum_{i=0}^{\binom{n+1}{2}} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right)\left|\frac{i}{\binom{n+1}{2}} - \mathbb{E}_n\right|^p \\
&\le \sum_{i=0}^{\lfloor n\cdot s_p(n)\rfloor} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right)\left(\frac{c\log n}{n}\right)^p \\
&\quad + \sum_{i=\lceil ns_p(n)\rceil}^{\binom{n+1}{2}} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right)(1)^p \\
&< \left(\frac{c\log n}{n}\right)^p + n^{k+5} q^{s_p(n)(1-r)/r} \\
&= O_p\left(\left(\frac{\log n}{n}\right)^p\right).
\end{aligned}$$

□

Corollary 16. *Let V be a doubled word, Σ an alphabet with $q \ge 2$ letters, and $W_n \in \Sigma^n$ chosen uniformly at random. Then*

$$\frac{1}{n} \ll \mathbb{E}(\delta(V, W_n)) \ll \frac{\log n}{n}.$$

Proof. The upper bound was stated explicitly in the proof of Theorem 15. The lower bound follows from an observation made in the Background section: "the event that $W_n[b|V|, (b+1)|V|]$ is an instance of V has nonzero probability and is independent for distinct b." Hence

$$\mathbb{E}(\delta(V, W_n)) \ge \frac{1}{\binom{n+1}{2}}\left\lfloor \frac{n}{|V|}\right\rfloor \mathbb{I}_{|V|}(V, \Sigma) = \Omega(n^{-1}).$$

□

The bound that Theorem 15 gives on the variance (2nd central moment) is not very interesting. However, we obtain nontrivial concentration using covariance and the fact that most "short" substrings in a word do not overlap.

Theorem 17. *Let V be a doubled word, Σ an alphabet with $q \geq 2$ letters, and $W_n \in \Sigma^n$ chosen uniformly at random.*

$$\mathrm{Var}(\delta(V, W_n)) = O\left(\mathbb{E}(\delta(V, W_n))^2 \frac{(\log n)^3}{n}\right).$$

Proof. Let $X_n = \binom{n+1}{2}\delta(V, W_n)$ be the random variable counting the number of substrings of W_n that are V-instances. For fixed n, let $X_{a,b}$ be the indicator variable for the event that $W_n[a, b]$ is a V-instance, so $X_n = \sum_{a=0}^{n-1}\sum_{b=a+1}^{n} X_{a,b}$. Let $(a, b) \sim (c, d)$ denote that $[a, b]$ and $[c, d]$ overlap. Note that

$$\begin{aligned} \mathrm{Cov}(X_{a,b}, X_{c,d}) &\leq \mathbb{E}(X_{a,b}X_{c,d}) \\ &\leq \min(\mathbb{E}(X_{a,b}), \mathbb{E}(X_{c,d})) \\ &= \min(\mathbb{I}_{(b-a)}(V, \Sigma), \mathbb{I}_{(b-a)}(V, \Sigma)) \\ &\leq \binom{i/d + k + 1}{k + 1} q^{i(1-r)/r}, \end{aligned}$$

for $i \in \{b - a, d - c\}$. For $i < n/3$, the number of intervals in W_n of length at most i that overlap a fixed interval of length i is less than $\binom{3i}{2}$. Let $s(n) = s_0(n) = t_0 \log n$ as defined in Theorem 15. For sufficiently large n,

$$\begin{aligned} \mathrm{Var}(X_n) &= \sum_{\substack{0\leq a<b\leq n \\ 0\leq c<d\leq n}} \mathrm{Cov}(X_{a,b}, X_{c,d}) \\ &\leq \sum_{(a,b)\sim(c,d)} \min(\mathbb{I}_{(b-a)}(V, \Sigma), \mathbb{I}_{(b-a)}(V, \Sigma)) \\ &= \left[\sum_{\substack{(a,b)\sim(c,d) \\ b-a,d-c\leq s(n)}} + \sum_{\substack{(a,b)\sim(c,d) \\ else}}\right] \min(\mathbb{I}_{(b-a)}(V, \Sigma), \mathbb{I}_{(b-a)}(V, \Sigma)) \\ &< 2\sum_{i=1}^{\lfloor s(n)\rfloor} (n + 1 - i)\binom{3i}{2} \cdot 1 \\ &\quad + \sum_{i=\lceil s(n)\rceil}^{n} (n + 1 - i)\binom{n+1}{2} \cdot \binom{i/d + k + 1}{k + 1} q^{i(1-r)/r} \\ &< 2s(n)n(3s(n))^2 + nnn^2n^{k+1}q^{s(n)(1-r)/r} \\ &= 18(t_0 \log n)^3 n + n^{5+k} q^{\log_q\left(n^{-(k+5)}\right)} \\ &= O(n(\log n)^3). \end{aligned}$$

Since $\mathbb{E}(\delta(V, W_n)) = \Omega(n^{-1})$ by Corollary 16,

$$\begin{aligned} \mathrm{Var}(\delta(V, W_n)) &= \mathrm{Var}\left(\frac{X_n}{\binom{n+1}{2}}\right) \\ &= \frac{\mathrm{Var}(X_n)}{\binom{n+1}{2}^2} \end{aligned}$$

$$= O\left(\frac{(\log n)^3}{n^3}\right)$$
$$= O\left(\mathbb{E}(\delta(V, W_n))^2 \frac{(\log n)^3}{n}\right).$$

□

Question 18. *For nondoubled word V, what is the concentration of the density distribution of V in random words?*

References

[1] D. R. Bean, A. Ehrenfeucht, and G. F. McNulty, *Avoidable Patterns in Strings of Symbols*, Pac. J. of Math. **85:2** (1979), 261–294.
[2] J. P. Bell and T. L. Goh, *Exponential lower bounds for the number of words of uniform length avoiding a pattern*, Information and Computing **205** (2007), 1295–1306.
[3] F. Blanchet-Sadri and B. Woodhouse, *Strict bounds for pattern avoidance*, Theoretical Computer Science **506** (213).
[4] J. Cooper and D. Rorabaugh, *Asymptotic Density of Zimin Words*, in preparation.
[5] ———, *Bounds on Zimin Word Avoidance*, Congressus Numerantium (2014+), arXiv:1409.3080.
[6] J. D. Currie, *Pattern avoidance: themes and variations*, Theoretical Computer Science **339** (2005).
[7] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, Cambridge, 2002.
[8] L. Lovász, *Large Networks and Graph Limits*, American Mathematical Society, Providence, 2012.
[9] J. Tao, *Pattern occurrence statistics and applications to the Ramsey theory of unavoidable patterns.*
[10] A. Thue, *Über unendliche Zeichenreihen*, Norske Vid. Skrifter I Mat.-Nat. Kl., vol. 7, Kristiania, 1906.
[11] A. I. Zimin, *Blokirujushhie mnozhestva termov*, Mat. Sb. **119** (1982), 363–375.
[12] ———, *Blocking sets of terms*, Math. USSR-Sb. **47** (1984), 353–364.