

# Pattern avoidance, entropy, and hitting time

Jacob Richey

February 28, 2024

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Substring patterns</b>	<b>3</b>
2.1	Definitions . . . . .	3
2.2	Follower set representation . . . . .	6
2.2.1	Combinatorial interpretation of Theorem 2.5 . . . . .	8
2.3	Recursions . . . . .	11
2.4	Letter densities . . . . .	12
2.5	Letter densities via the MME for SFT . . . . .	17
2.6	Word counts . . . . .	19
2.7	Gibbs measures . . . . .	21
2.8	Martingale and hitting time . . . . .	25
2.8.1	IID case . . . . .	25
2.8.2	Markov chain . . . . .	29
2.9	Auto/cross correlations . . . . .	33
2.10	Conjectures/questions . . . . .	34
<b>3</b>	<b>Subsequence patterns in iid sequences</b>	<b>36</b>
3.1	$\sigma = 11$ , arbitrary distribution . . . . .	36
3.2	Uniform distribution . . . . .	36

## 1 Introduction

In this note we consider words over an alphabet  $\mathcal{A}$ , typically  $\mathcal{A} = \{0, 1\}$ , (and possibly  $\mathcal{A} = [q]$  for some positive integer  $q$  or  $\mathcal{A} = \mathbb{N}$ ), conditioned on avoiding some pattern set  $S$ . This can mean a few different things. So far we have focused on taking ‘pattern’ to mean ‘subword,’ i.e. take  $S \subset \Omega$ , where  $\Omega_n = \mathcal{A}^n$  is the set of sequences of length  $n$  and  $\Omega = \cup_n \Omega_n$ , and write

$$\Omega_n(S) = \{\omega \in \Omega_n : \omega \text{ does not contain } s \text{ as a subword for any } s \in S\}. \quad (1.1)$$

Here ‘subword’ means ‘consecutive subsequence:’  $11$  is a subword of  $1101$ , but  $111$  is not. (Disallowing arbitrary subsequences to match  $S$  seems quite restrictive, but could be interesting too.) For example, with  $\mathcal{A} = \{0, 1\}$  and  $S = \{11, 1001\}$ , we have

$$\Omega_4(S) = \{0000, 0001, 0010, 0100, 1000, 1010, 0101\} \quad (1.2)$$

Of course, these  $\Omega_n(S)$  generate all possible events if any sets  $S$  are allowed: we have in mind ‘small’ sets  $S$ . We want to study random words sampled from some measure on  $\Omega_n(S)$ , or if it makes sense,  $\Omega(S)$ , or  $\Omega_\infty(S)$ : two examples to keep in mind are an iid word of fixed length conditioned to have no subword in  $S$ , or a word generated with iid bits, one bit at a time, and stopped on containing some word in  $S$  as a subword. The natural limits for these objects are *shifts of finite type*. There is also work focusing on the expected hitting time of a given word (Feller has a few pages on it), and facts about a related ‘intransitive dice’ game (originally from Conway).

A more topical connection is with pattern avoiding permutations, where a finite ‘pattern’ permutation on  $k$  letters is chosen, and a uniform random permutation  $X$  is conditioned on ‘avoiding’  $\sigma$ , i.e. having no subsequence  $I = (i_1, i_2, \dots, i_k)$  with  $i_1 < i_2 < \dots < i_k$  such that

$$X_I = (X_{i_j})_{j \in [k]} \text{ is order-isomorphic to } \sigma, \tag{1.3}$$

i.e. for any  $a, b \in [k]$ ,  $X_{i_a} \leq X_{i_b} \iff \sigma(a) \leq \sigma(b)$ . The same question can be asked for any random sequence  $X$ , say iid from a discrete distribution. Do we recover phenomena similar to the permutation case? It seems there is some work on this, in the permutation-avoiding literature, where there are some recursive techniques that apply to general sequences  $X$  (not just permutations).

Let  $X$  denote a random instance of one of these processes. The over-arching questions we are interested in are:

1. How does the conditioning affect typical properties of  $X$ , like the density of each letter of  $\mathcal{A}$ , or ‘random walk’ properties of  $X$ ? **We can compute these kinds of things exactly with linear algebra/generating functions for the limiting SFT.**
2. Is there a simple probabilistic description of the conditional law  $X$ ? **The limiting measure is a Markov Chain in the case of SFTs. Gibbs measures give a somewhat nice way to interpolate. Generally speaking  $X$  has complex structure.**
3. Viewed as the underlying randomness of a random walk, does a scaled version of  $X$  converge to a diffusion? (e.g. if the alphabet is  $\{-1, 1\}$ , does it converge to BM?) **The book *Analytic Pattern Matching* has some possibly relevant CLTs for this? Probably because everything is local it will converge to BM with drift.**
4. For non-trivial sets  $S$ ,  $|\Omega_n(S)| \ll |\Omega_n| = |\mathcal{A}|^n$ , so  $X$  lives on a set of vanishing measure. Despite this, is there a natural limiting measure as  $n \rightarrow \infty$ , i.e. a measure on  $\Omega_\infty$  (infinite strings) supported on strings that avoid  $S$ ? **For ‘isomorphic’ pattern avoidance, there is a limit in permuton space; for subword avoidance, shift spaces and measures of maximal entropy give a full description.**
5. Is  $X$  Markovian, or approximately markovian? Is it possible to construct  $X$  one bit at a time by recording the output of a simple markov graph? (Simulations suggest this is possible, up to some ‘edge’ effects. This would be nice – sampling random pattern avoiding permutations is a hot topic.) **A: For subword avoidance, yes: the measure of maximal entropy on a shift space is a markov chain in some presentation.**

These models seem to have similar flavour to the maximal greedy independent set and the hard core model.

## 2 Substring patterns

### 2.1 Definitions

Already the case of excluding single binary words leads to interesting phenomena. Set  $\mathcal{A} = [q]$ , and suppose  $S$  is a single word of length  $l$ ,  $S = \{w\}$ ,  $w = w_1w_2 \cdots w_l$  for some  $w_i \in \mathcal{A}$ . The first order of business here is to compute  $|\Omega_n(w)|$ .

**Definition 2.1.** For a fixed word  $w$ , let  $\lambda_w$  denote the asymptotic growth rate of  $|\Omega_n(w)|$ , i.e.

$$\log \lambda_w = \lim_{n \rightarrow \infty} \frac{1}{n} \log |\Omega_n(w)|. \quad (2.1)$$

We have that:

**Lemma 2.2.** Except in the trivial cases where  $q = 2$  and  $w \in \{0, 1, 10, 01\}$ , the limit in 2.1 exists and  $\lambda_w \in (1, 2)$ .

*Proof.* An elementary proof is to use sub-additivity of the  $\Omega_n$ . □

$\lambda_w$  is the (exponential of the) topological entropy of the shift of finite type with forbidden word  $w$ . Alternatively,  $\lambda_w$  is the Perron-Frobenius eigenvalue of the corresponding edge-shift matrix.

One can also compute combinatorially, which involves typical recursion/generating function ideas, but with some novel elements. See section 2.2.

We have the following basic heuristic regarding entropies. Generate iid digits, and stop when you observe the word  $w$  for the first time. Let  $\tau = \tau_w$  be the number of digits generated. Then

$$\mathbb{E}\tau = \sum_{t \geq 0} \mathbb{P}(\tau \geq t) \quad (2.2)$$

$$= \sum_{t \geq 0} \frac{\# \text{ words of length } t \text{ avoiding } w}{2^t} \quad (2.3)$$

$$\approx \sum_{t \geq 0} c_w \left( \frac{\lambda_w}{2} \right)^t \quad (2.4)$$

$$= c_w \frac{1}{1 - \lambda_w/2}. \quad (2.5)$$

Observe that this looks like it's increasing in  $\lambda$  (ignoring the constant  $c$  – it's just a heuristic, after all!) This suggests that the expected hitting time is measuring the same thing as the entropy. The hitting time is roughly exponentially distributed, i.e.

$$\lim_{n \rightarrow \infty} \frac{-1}{n} \log \mathbb{P}(\tau > t) = \log(q^{-1}\lambda), \quad (2.6)$$

(this is just re-wording the definition of entropy), but understanding how close  $\tau$  is to being *exactly* exponentially distributed with parameter  $\lambda$  is the real challenge.

It turns out that this heuristic is correct for the full shift. First a definition:

**Definition 2.3.** Given a word  $w$  of length  $l$ , its overlap set  $\mathcal{O}$  is the set

$$\mathcal{O} = \{i \in [l] : w_1w_2 \cdots w_i = w_{l-i+1}w_{l-i+2} \cdots w_l\} \quad (2.7)$$

and its correlation polynomial is the polynomial function

$$\phi(x) = \sum_{i \in \mathcal{O}(w)} x^i. \quad (2.8)$$

A folklore probability result is:

**Fact 2.4.**  $\mathbb{E}\tau_w = \phi_w(q)$

See section 2.8 for a proof (and a general martingale method for computing with these hitting times). Now the punch line:

**Theorem 2.5.** For two words  $w, w'$ ,  $\phi_w(q) \leq \phi_{w'}(q) \iff \lambda_w \leq \lambda_{w'}$ .

This is originally due to Guibas and Odlyzko (1980). We give an alternate proof:

*Proof.* For any word  $w$ , let  $\tau_w$  denote the hitting time of  $w$  in an iid sequence uniform over  $[q]$ . From the martingale argument in Section 2.8, we will use the explicit formula 2.110 for the generating function  $f(z) = \sum_{t \geq 1} z^t \mathbb{P}(\tau \geq t)$ , namely

$$f(z) = \frac{z\phi_w(qz^{-1})}{1 - (z-1)\phi_w(qz^{-1})}. \quad (2.9)$$

(This can probably be proved via the formulas and methods of GO, but the martingale derivation is much more elegant, and doesn't require any futzing around with recursions.) By 2.6,  $f(z)$  has radius of convergence exactly  $q\lambda^{-1}$ . Comparing with the explicit formula above, we see that  $\lambda = qr^{-1}$ , where  $r$  is the smallest positive root of the polynomial  $1 - (z-1)\phi_w(qz^{-1})$ .

Observe that since  $q \geq 2$ ,  $\phi_w(q) \geq \phi_{w'}(q)$  exactly when  $\phi_w \prec_{\text{lex}} \phi_{w'}$ , where  $A \prec_{\text{lex}} B$  for polynomials  $A(\alpha) = \sum_{j=1}^m a_j \alpha^j$  and  $B(\alpha) = \sum_{j=1}^m b_j \alpha^j$  with  $\{0, 1\}$  coefficients means

$$a_j = b_j, j \in \{m, m-1, \dots, k+1\} \text{ and } a_k = 0, b_k = 1 \text{ for some } k \geq 1. \quad (2.10)$$

Combining these observations, making the change of variables  $\alpha = r^{-1}$  and doing some algebra, the theorem reduces to showing the following:

**Claim 2.6.** Fix  $q \in [2, \infty)$ . Let  $A$  be any polynomial with  $\{0, 1\}$  coefficients and zero constant term. The equation

$$A(\alpha) = \frac{\alpha}{q - \alpha} \quad (2.11)$$

has a unique root  $\alpha^*(A) \in (1, q)$ . If  $A \prec_{\text{lex}} B$  are two such polynomials,  $\alpha^*(A) < \alpha^*(B)$ .

The result follows from the claim by taking  $A = \phi_w$  and  $B = \phi_{w'}$ . The proof that the root is unique is tedious but easy, I postpone it for later. Then we have

$$\alpha^*(A) < \alpha^*(B) \iff A(\alpha^*(A)) < B(\alpha^*(A)). \quad (2.12)$$

This relies on the fact that  $\alpha_B^*$  is the unique solution to (2.11) with polynomial  $B$ , and  $\lim_{\alpha \rightarrow q^-} \frac{\alpha}{q-\alpha} = +\infty$ , which together imply  $B(\alpha) > \frac{\alpha}{q-\alpha}$  for  $\alpha \in (1, \alpha_B^*)$  and  $B(\alpha) < \frac{\alpha}{q-\alpha}$  for  $\alpha \in (\alpha_B^*, q)$ . Set  $\alpha^*(A) = u$ . By the assumption  $A \prec_{\text{lex}} B$ , for some  $k \leq \deg(A)$  we have

$$B(u) - A(u) \geq u^k - \sum_{1 \leq j < k} u^j. \quad (2.13)$$

Letting  $c_k$  denote the largest root of the polynomial  $x^k - \sum_{1 \leq j < k} x^j$ , the difference above is positive if  $u \geq c_k$ , i.e. if

$$\frac{c_k}{q - c_k} \leq A(c_k), \quad (2.14)$$

(note that this again relies on the uniqueness of the solution  $\alpha_A^*$ ). Finally, this inequality follows directly from the assumption  $q \geq 2$  and the fact that  $k \leq \deg(A)$ :

$$\frac{c_k}{q - c_k} \leq \frac{c_k}{2 - c_k} = c_k^k \leq A(c_k), \quad (2.15)$$

where the equality uses the definition of  $c_k$ . □

Here's a proof sketch for the uniqueness of the solution  $\alpha^*$ , I'll write down the important ideas and fill in the precise details later. It feels a bit convoluted but I couldn't find a better way. The idea is to do an induction over the derivatives of the function  $h(x) = (q - x)A(x) - x$ . Let  $A$  have degree  $k$ . Easy computations show that: for  $j \geq 2$ ,  $h^j(x) = -(j - 1)f^{j-1}(x) + (q - x)f^j(x)$  (where superscript denotes the  $j$ th derivative);  $h^j(q) < 0$  for all  $j$ ;  $h^k(x) < 0$  for all  $x$ ; and a slightly tricky computation shows that there is an integer  $j^* \in \{1, 2, \dots, k\}$  such that  $h^j(1) > 0$  for  $i \leq j^*$ , and  $h^j(1) < 0$  for  $i > j^*$ . Then the induction step is: if  $f$  is a polynomial such that  $f'(1) < 0$  and  $f'$  has no roots in  $(1, q)$ , then  $f$  has no roots in  $(1, q)$  (this is obvious); and if the same assumptions hold except for  $f(1)$  is nonnegative instead of negative, and  $f(q) < 0$ , then  $f$  must have at most one zero (again, obvious); and if  $f'(1) > 0$ ,  $f'(q) < 0$ ,  $f'$  has a unique zero in  $(1, q)$  and  $f(1) > 0$ , then  $f$  has a unique zero in  $(1, q)$  (again, obvious, nothing to prove). Putting all this together, we 'ride the induction chain' up from the  $k$ th derivative to the function  $h$ ; at index  $j^*$  point the induction 'flips' from the first case to the second, and then the second to the third.

The only tricky fact in this whole business is showing that the  $j^*$  exists, but it just boils down to a relatively simple bound, I think. The exact expression is

$$\sum_{\mathcal{O} \ni n \geq i} (n)_{i-1}((q-1)(n-i+1) - (i-1)) - (i-1)(i-1)! \mathbf{1}\{i-1 \in \mathcal{O}\}. \quad (2.16)$$

We want to show that as  $i$  increases from 1 to  $k$ , this expression is  $> 0$  and then  $< 0$ . The key point is that the whole thing is dominated by the term  $n = k$ , whatever sign it has will determine the sign of the whole expression, and when that term is 0 (it can happen), then it won't matter which sign the expression has, because that will be the 'switch' point  $j^*$ .

## 2.2 Follower set representation

Given a forbidden word  $w$ , we construct a directed graph  $L_w$  which has state space  $\{0, 1, \dots, |w| = l\}$ . This is called the ‘follower set edge shift’ in symbolic dynamics lingo: infinite paths in the graph  $L_w$  will correspond (be in bijection) with the shift space where the word  $w$  is forbidden. Each state, aside from state  $l$ , (which can, for our purposes, be thought of as a ‘graveyard’ state with no outgoing edges), has 2 outgoing edges, given by appending both possible letters 0 or 1 to the right end of  $p$ , then finding the longest prefix of  $S$  that matches a *suffix* of  $pa$  (for each  $a \in \mathcal{A}$ ). One of these edges always goes from  $k$  to  $k + 1$ . We denote the other edge by

$$d_k = \max\{r : w_1 w_2 \cdots w_r = w_{k-r+2} w_{k-r+3} \cdots w_k \overline{w_{k+1}}\} \quad (2.17)$$

where  $\overline{w} = 1 - w$ . More generally, if the alphabet is  $[q]$  for some positive integer  $q$ , state  $k$  has  $q$  outgoing edges, and exactly one such edge goes from state  $k$  to state  $k + 1$ . First, we observe that the set  $\mathcal{T}(w)$  of all finite words over  $\mathcal{A}$  ending with  $w$  is described by paths in the graph  $L_w$  in a nice way. To make this map precise, for each edge  $(k, k')$  in  $L_w$ , let  $a(k, k')$  denote its label, i.e. the digit of  $\mathcal{A}$  that was appended to the  $k$ -prefix of  $w$  to obtain the  $k'$ -prefix (as a suffix). Also, let  $\Gamma(L_S)$  denote the set of all finite paths in  $L_w$  starting at state 0 and ending at state  $l$ .

**Fact 2.7.** *The map  $f : \Gamma(L_w) \rightarrow \mathcal{T}(w)$  given by  $f(\gamma) = (a(\gamma_i, \gamma_{i+1}))_i$  (i.e. read the labels of the path) is a bijection.*

*Proof.* For any finite word  $v$  not containing  $w$  as a subword, let  $m(v)$  denote the length of the maximal suffix of  $v$  that is a prefix of  $w$ . It suffices to show that

$$\{m(v0), m(v1)\} = \{m(v) + 1, d_{m(v)}\}, \quad (2.18)$$

i.e. that to determine  $m(va)$  for some digit  $a$ , one only needs to look at the last  $m(v)$  digits of  $v$ , and that  $m(va)$  is given by the endpoint of the corresponding edge in the graph  $L_w$ . Observe that for any letter  $a$ ,  $m(va) \leq m(v) + 1$ , since if  $va$  ends with a prefix of  $w$  of length  $j$ , then  $v$  ends with a prefix of  $w$  of length  $j - 1$ . In the case where  $a = s_{m(v)+1}$ , we have  $m(va) = m(v) + 1$ . Otherwise,  $m(va) \leq m(v)$ , so only the last  $m(v)$  digits of  $v$  are necessary to determine  $m(va)$ .  $\square$

We also note the following. Let  $\text{Rev}(w)$  denote the reversal of the string  $w$ , i.e.  $\text{Rev}(w) = w_k w_{k-1} \cdots w_2 w_1$ .

**Lemma 2.8.**  $|\Omega_n(w)| = |\Omega_n(\text{Rev}(w))|$

*Proof.*  $\omega \in \Omega_n(w) \iff \text{Rev}(\omega) \in \Omega_n(\text{Rev}(w))$ .  $\square$

Despite this simple fact, it isn’t obvious what the relationship is between  $L_w$  and  $L_{\text{Rev}(w)}$ .

**Question 2.9.** *Describe a simple mapping  $L_w \rightarrow L_{\text{Rev}(w)}$ .*

Now we fix  $q = 2$ . These graphs are more special than simply requiring outdegree 2 everywhere except at state  $l$ :

**Proposition 2.10.** *Let  $q = 2$ . For any such  $L_w$ ,*

1.  $d_k \leq k$
2.  $k - d_k = k' - d_{k'}$  only if  $k = k'$  or one of  $d_k, d_{k'} = 0$ .

*Proof.* (1) is immediate from the definition. For (2), suppose by contradiction that  $k - d_k = k' - d_{k'}$  for some  $k \neq k'$  with  $k, k' \geq 1$ . Then also  $d_k \neq d_{k'}$ , so assume WLOG  $d_k > d_{k'}$ . The definition of  $d_k$  implies

$$w_{k-d_k+j+1} = w_j \text{ for } j = 1, 2, \dots, d_k - 1, \text{ and } \overline{w_{k+1}} = w_{d_k}, \quad (2.19)$$

where  $\overline{w} = 1 - w$ . Taking  $j = d_{k'}$  gives

$$w_{d_{k'}} = w_{k-d_k+d_{k'}+1} = w_{k'-d_{k'}+d_{k'}+1} = w_{k'+1}, \quad (2.20)$$

contradicting the last part of 2.19 for  $k'$  (as long as neither  $d_k, d_{k'} = 0$ , in which case one of the equalities is trivial). □

We now give a few applications of proposition 2.10, starting with a characterization of which of the  $L_w$  graphs are irreducible:

**Lemma 2.11.** *The graph  $L_w$  is irreducible except if  $w \in \bigcup_{l \geq 1} \{10^l, 1^l 0, 01^l, 0^l 1\}$ .*

Irreducible means there is a path  $i \rightarrow j$  for any states  $i, j$ . Irreducibility of the edge-shift graph  $L$  is equivalent to irreducibility of the corresponding shift of finite type. One can easily check that the exceptional words listed above give rise to reducible graphs/shifts: the words  $10^l$  and  $01^l$  give graphs that have a single transient ‘source’ state, while the words  $1^n 0, 0^n 1$  give graphs that have a single absorbing ‘sink’ state.

*Proof.* Since the graph  $L_w$  always has the path  $0 \rightarrow 1 \rightarrow \dots \rightarrow l - 1$ , it suffices to show that if there is no path from  $l - 1$  to 0, then  $w$  must be one of the exceptional words. For this to occur, there must be some  $k \in \{1, 2, \dots, l - 1\}$  such that  $d_i \geq k$  for all  $i = k, k + 1, \dots, l - 1$  – otherwise, it would be possible to escape the set  $\{k, k + 1, \dots, l - 1\}$  for every  $k$ , and thus to reach 0 from  $l - 1$ . Pick the largest such  $k$ . If  $k = l - 1$ , then  $d_{l-1} = l - 1$ , and unraveling the definition of  $d_{l-1}$  gives  $w_1 = w_2 = \dots = w_{l-1} = \overline{w_l}$ , i.e.  $w = 1^{l-1} 0$  or  $0^{l-1} 1$ .

So suppose  $k < l - 1$ . By claim 2 of Proposition 2.10, since  $k > 0$ , the values  $i - d_i$  must all be distinct for  $i \in \{k, k + 1, \dots, l - 1\}$ . So  $i - d_i \in \{0, 1, \dots, l - 1 - k\}$  for  $i = k, k + 1, \dots, l - 1$ ; and since there are exactly  $k - l - 1$  many  $i$ , each such value occurs exactly once. Since  $d_i \leq i$ , we must have  $d_k = k$ , and by induction over  $i$ ,  $d_i = k$  for all  $i \in \{k, k + 1, \dots, l - 1\}$ . Unraveling the definition of  $d_i$ , we obtain that for  $i \in \{k, k + 1, \dots, l - 1\}$ ,

$$w_j = w_{i+1-k+j} \text{ for } j \in \{1, \dots, k - 1\}, \text{ and also } w_k = \overline{w_{i+1}}. \quad (2.21)$$

In particular, if  $k > 1$  (so the first case above is non-empty), taking  $i = k$  or  $k + 1$  and  $j = k - 1$  immediately gives the contradiction  $w_{k-1} = w_k = w_{k+1}$  and  $w_k = \overline{w_{k+1}}$ . Thus  $k = 1$  and  $w_1 = \overline{w_2} = \overline{w_3} = \dots = \overline{w_l}$ , i.e.  $w = 10^{l-1}$  or  $01^{l-1}$ , as desired. □

The properties in Proposition 2.10 are not sufficient to classify all such graphs  $L_w$ , since the number of possible sequences of  $d_k$  values that satisfy (1) and (2) is strictly larger than  $2^{l-1}$ , while the total number of possible graphs is at most that many (there are  $2^l$  strings, and the bit flip operation  $w \rightarrow \overline{w}$  preserves the graph,  $L_w = L_{\overline{w}}$ ).

**Question 2.12.** *Find a full characterization of the sequences of  $d_k$  that occur for some string  $w$ .*

When  $q = 2$ , the graph  $L_w$ , viewed as a labeled graph on  $\{0, 1, \dots, l\}$ , is enough information to determine the word  $w$ , up to permutations of the alphabet (i.e. bit flipping, for the binary alphabet). Let  $\mathcal{W}$  denote the family of equivalence classes of words in  $\{0, 1\}^l$  up to bit flip, i.e.  $w \sim w'$  if  $w = \overline{w'}$ . Also, let  $\hat{\mathcal{L}}$  denote the family of *labeled* directed graphs we obtain from the  $L_w$ 's, i.e.

$$\hat{\mathcal{L}} = \{L_w : w \in \{0, 1\}^l\}, \quad (2.22)$$

where the graphs  $L_w$  have vertex labels  $\{0, 1, \dots, l\}$  corresponding to the frontier representation given by  $w$ .

**Lemma 2.13.** *There is a (simple, algorithmic) bijection  $f : \hat{\mathcal{L}} \rightarrow \mathcal{W}$ .*

*Proof.* By the definition of  $\hat{\mathcal{L}}$ , it suffices to construct an injective  $f : \hat{\mathcal{L}} \rightarrow \mathcal{W}$ . Fix  $\hat{L} \in \hat{\mathcal{L}}$ , and WLOG assume the image  $w = f(\hat{L})$  has  $w_1 = 1$ . Suppose we have determined the digits  $w_1, w_2, \dots, w_{k-1}$  of  $w$ . Use the same notation  $d_i$  for the outgoing edges in  $\hat{L}$  with  $d_i \leq i$ . The definition of  $d_k$  implies  $\overline{w_{k+1}} = w_{d_k}$  if  $d_i > 0$ , while if  $d_k = 0$  then we must have  $\overline{w_{k+1}} \neq w_1 = 1$ . So in the first case, since  $d_k \leq k$ , we set  $w_{k+1} = \overline{w_{d_k}}$ , which was already a determined digit, and in the second case  $w_{k+1} = \overline{w_1} = 0$ . By induction we can determine the full word  $w$  (note that this includes the final digit of  $w$ .) □

Note that Lemma 2.13 fails when  $q \geq 3$ : for example, the two words 0012 and 0011 give the same graphs, but are not equivalent by a permutation on  $[q]$ .

It turns out that even if we forget about the vertex labels, no two of the graphs  $\hat{L}, \hat{L}' \in \hat{\mathcal{L}}$  are isomorphic as graphs. Write  $\mathcal{L}$  for the same family of graphs as in  $\hat{\mathcal{L}}$ , but viewed as *unlabeled* directed graphs, and for  $\hat{L} \in \hat{\mathcal{L}}$ , let  $L$  denote the same graph but with labels removed.

**Lemma 2.14.**  $|\mathcal{L}| = |\hat{\mathcal{L}}|$ , i.e. for any  $\hat{L}, \hat{L}' \in \hat{\mathcal{L}}$ ,  $L$  and  $L'$  are not isomorphic.

*Proof.* Suppose  $\varphi : \hat{L} \rightarrow \hat{L}'$  is an isomorphism of the underlying unlabeled graphs, viewed as a permutation on the labels  $[l]$ . Since  $l \in \hat{L}'$  is the unique state with outdegree 0, we must have  $\varphi(l) = l$ . Assume by induction that  $\varphi(i) = i$  for  $i = l, l-1, \dots, j$ . By Proposition 2.10 (1), all of the incoming edges to state  $j \in \hat{L}'$  has its other end at some state  $k \geq j$ , except one edge  $j-1 \rightarrow j \in \hat{L}$ . Since  $\varphi(j-1)$  must be a state that has an edge  $\varphi(j-1) \rightarrow j \in \hat{L}'$ , and the other values  $k \geq j$  are already taken, we must have  $\varphi(j-1) = j-1$ . So  $\varphi = Id$ . □

### 2.2.1 Combinatorial interpretation of Theorem 2.5

The matrix representation  $L$  gives a way to explain Theorem 2.5, though it only succeeds in a limited setting. It is a helpful representation because of the following fact. Recall the stochastic dominance order for random variables:  $X \prec_{st} Y$  if  $\mathbb{P}(X \geq t) \leq \mathbb{P}(Y \geq t)$  for all  $t$ , or equivalently, there is a coupling of  $X$  and  $Y$  where  $X$  is at most  $Y$  pointwise.

**Lemma 2.15.** *For any words  $w, w'$  of the same length  $l$ , with follower set graphs  $L, L'$  given by edges  $d_i, d'_i$  satisfying  $d_i \leq d'_i$  for all  $i$ , we have  $\tau_w \succ_{st} \tau_{w'}$ .*

*Proof.* The hitting time  $\tau = \tau_w$  can be thought of as a hitting time in a markov chain on the graph  $G_w = L \cup \{l\}$ , where the (directed) edges have transition probabilities  $1/2$ , and the extra state  $w$  is added at the end of  $L$ , i.e. there is an edge  $l-1 \rightarrow l$  corresponding to hitting the word  $w$ .



(For completeness one can add an outgoing edge from state  $l$  to state  $k$ , where  $k = \max \mathcal{O} \setminus \{l\}$ , so  $G_w$  is an edge shift representation for the full shift, and the markov chain with all  $1/2$  transition probabilities is the measure of maximal entropy.) For  $w$  and  $w'$  satisfying  $d_i \leq d'_i$  for all  $i$ , let  $X, X'$  denote the random walk processes on the graphs  $G_w, G_{w'}$  respectively, coupled in the following way: when  $X = X' = i$  (i.e. they are at the same index  $i$  each in their own graph), they either both take the edge  $i \rightarrow i+1$ , both take the edge  $i \rightarrow d_i$ . When  $X \neq X'$ , because of the assumption on the  $d_i$  we must have  $X < X'$ : in this case  $X$  moves independently of  $X'$  until it hits position  $X'$ . Under this coupling it is clear that  $\tau \geq \tau'$ .  $\square$

The lemma says that the ordering of the hitting times  $\tau_w$  is easy to understand in terms of the  $L$  representation. We can also show that the entropies are ordered nicely in this case. First we need the following fact. For a word  $w$ , let  $r_w$  denote the right eigenvector of the adjacency matrix of  $L_w$ .

**Lemma 2.16.** *Fix  $q = 2$ . Then  $r_w$  has strictly decreasing entries. In particular, the entries of  $r$  decrease at least exponentially: for  $i = 0, \dots, l-2$ ,*

$$r_{i+1} \leq (\lambda_w - 1)r_i. \quad (2.23)$$

*Proof.* We proceed by induction. Assume (2.23) holds for  $i = 0, 1, \dots, j-1$ . Then

$$\lambda r_j = r_{j+1} + r_{d_j} \quad (2.24)$$

$$\geq r_{j+1} + r_j. \quad (2.25)$$

Here the first line follows from the fact that  $r$  is a right eigenvector, and the second line follows from Proposition 2.10 and the induction hypothesis, using the fact that  $\lambda_w - 1 < 1$ . Re-arranging gives 2.23 for  $i = j$ . In the base case  $i = 0$ , we have  $d_0 = 0$ , so  $r_1 = (\lambda - 1)r_0$  is an equality.  $\square$

The proof actually shows something a bit stronger: namely that for  $i = 0, 1, \dots, l-2$ ,

$$\frac{r_{i+1}}{r_i} \leq \lambda - (\lambda - 1)^{d_i - i}. \quad (2.26)$$

Using Lemmas 2.15 and 2.16, we can partially recover the entropy/hitting time connection in the context of the follower set graphs.

**Theorem 2.17.** *Let  $w, w'$  be any words satisfying the assumptions of Lemma 2.15. Then  $\lambda_w > \lambda_{w'}$ .*

*Proof.* Let  $A, A'$  denote the adjacency matrices of  $L, L'$ , and write  $r, r'$  for the right eigenvectors of  $A, A'$  corresponding to the top eigenvalues  $\lambda, \lambda'$ . For  $i = 0, 1, \dots, l-2$ ,

$$(A'v)_i = v_{i+1} + v_{d'_i} \leq v_{i+1} + v_{d_i} = (Av)_i = \lambda v_i. \quad (2.27)$$

Here we used Lemma 2.16 and the assumption on the  $d_i, d'_i$  for the inequality. Similarly, when  $i = l-1$ , we have

$$(A'v)_{l-1} = v_{d'_{l-1}} \leq v_{d_l} = \lambda v_{l-1}. \quad (2.28)$$

Taking a left eigenvector  $\ell'$  for eigenvalue  $\lambda'$  in  $A'$ ,

$$\lambda \ell' \cdot v = \ell' A' v \leq \lambda \ell' \cdot v. \quad (2.29)$$

(Here  $\cdot$  is the dot product.) Since  $\ell' \cdot v \neq 0$  (why?),  $\lambda' \leq \lambda$ .  $\square$

Unfortunately, the relation  $w \prec w'$  given by  $d_i \leq d'_i$  for all  $i$  is too coarse to fully recover Theorem 2.5: one can check that there are words  $w, w'$  that are incomparable under  $\prec$ , but  $\lambda_w \neq \lambda_{w'}$ . It is not even the case that for each word  $w$  whose entropy is larger than the minimum possible entropy of all words of fixed length, there exists a word  $w'$  with  $w \prec w'$ : a minimal counterexample is  $w = 1011$ . On the other hand, the word  $w$  with maximal entropy over words of length  $l$ , namely  $w = 1^l$ , is the unique minimal element of the poset generated by  $\prec$ , i.e.  $1^l \prec w'$  for all  $w'$  of length  $l$  (since all  $d_i$  are equal to 0 for the word  $1^l$ .)

**Question 2.18.** *Describe the structure of the poset generated by  $\prec$  in more detail. What are the maximal elements? How long is a typical chain?*

### 2.3 Recursions

As an example of the usefulness of the graphs  $L_w$ , we work through the necessary computation explicitly for  $w = 100$ . Here the graph is given by  $d_1 = d_2 = 1$ . We are trying to solve for  $\Omega_n(100)$ , which can be thought of as the number of paths in the graph  $L_{100}$  of length  $n$ , starting at either state 0 or state 1, that never hit state 3. To count these, write  $a_n(100)$  as the number of such paths, and partition  $a_n$  into three further counts  $a^0, a^1$ , and  $a^2$ , where  $a^j$  is the number of such paths ending at state  $j$ . These lead to the following system of recursions, obtained by collecting the incoming edges to each state:

$$a_n^0 = a_{n-1}^0 \tag{2.30}$$

$$a_n^1 = a_{n-1}^0 + a_{n-1}^1 + a_{n-1}^2 = a_{n-1} \tag{2.31}$$

$$a_n^2 = a_{n-1}^1 \tag{2.32}$$

There doesn't seem to be a systematic way to solve such a system, other than plugging in recursively repeatedly until a recursion for  $a_n$  appears. In this case, it doesn't take too long:

$$a = a^0 + a^1 + a^2 \tag{2.33}$$

$$= 2a_{-1} - a_{-1}^2 \tag{2.34}$$

$$= 2a_{-1} - a_{-2}^1 \tag{2.35}$$

$$= 2a_{-1} - a_{-3}. \tag{2.36}$$

Thus  $a_n(100) = 2a_{n-1}(100) - a_{n-3}(100)$ , which yields the asymptotic formula

$$a_n(100) \sim \left(1 + \frac{2}{\sqrt{5}}\right) \varphi^n, \varphi = \frac{1}{2}(1 + \sqrt{5}). \tag{2.37}$$

In general it seems easier to work with the corresponding generating functions  $f_{100}^j(z) = \sum_{n \geq 1} a_n^j(100)z^n$  and  $f_{100}(z) = \sum_{n \geq 1} a_n z^n$ . These functions satisfy  $f(z) = f^0(z) + f^1(z) + f^2(z)$  and

$$f^0(z) = z + z f^0(z) \tag{2.38}$$

$$f^1(z) = z + z(f^0(z) + f^1(z) + f^2(z)) \tag{2.39}$$

$$f^2(z) = z + z f^1(z) \tag{2.40}$$

The solution is

$$f^0(z) = \frac{z}{1-z}, f^1(z) = \frac{z}{1-2z+z^3}, f^2(z) = \frac{z^2}{1-2z+z^3}. \tag{2.41}$$

Note that  $a_n^0 = n$ , and asymptotically

$$a_n^1 \sim \left(\frac{3 + \sqrt{5}}{2\sqrt{5}}\right) \varphi^n, a_n^2 \sim a_n^1 \varphi^{-1}. \tag{2.42}$$

The proportions of paths that end at 0, 1, 2, i.e.  $\lim_{n \rightarrow \infty} \frac{a_n^j}{a_n}$ , are respectively 0,  $\varphi - 1$ ,  $2 - \varphi$ , or  $\approx 0, .618, .382$ .

**See section 2.5 for a linear algebra approach.**

## 2.4 Letter densities

Computing the average density of 1's is not as simple as the counts  $a_n(w)$ . Let  $X_n$  denote a uniformly random chosen element of  $\Omega_n(w)$ . In the notation of 2.3,

$$\mathbb{P}(\text{the last digit of } X_n = 1) = \frac{1}{a_n(w)} \sum_{k=1}^{l-1} a_n^k(w) 1\{\text{the } k^{\text{th}} \text{ digit of } w = 1\}. \quad (2.43)$$

In the example with  $w = 100$ , we computed  $\frac{a_n^1(100)}{a_n(100)} \rightarrow \varphi - 1$ , so this is the limiting probability of seeing a 1 in the final position. However, this isn't the same as the density of 1's in the whole word, as we will see shortly. The method from 2.3 can likely be extended to compute

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{the } j\text{th digit of } X_n = 1) \quad (2.44)$$

for any fixed  $j \in \mathbb{N}$ , by enumerating paths in the markov graph  $L_w$  'backwards.' These values should converge, as  $j \rightarrow \infty$ , to the average density of 1's in  $X_n$ ,  $\gamma_w$  (defined below).

A natural quantity is the density of 1's the string  $X_n$ . Consider the average fraction of 1's in a uniformly random  $w$ -avoiding string:

**Definition 2.19.** For a fixed word  $w$ , let  $\gamma_w$  denote the limiting fraction of bits that are 1 over all strings in  $\Omega_n(w)$ :

$$\gamma_w = \lim_{n \rightarrow \infty} \frac{1}{n|\Omega_n(w)|} \sum_{\omega \in \Omega_n(w)} \#1\text{'s in } \omega. \quad (2.45)$$

How can this density be computed? It seems necessary to further partition the strings  $\Omega_n(w)$  into sets  $\Omega_{n,k}(w)$ , i.e. strings of length  $n$  with exactly  $k$  1's. Let  $a_{n,k}(w) = |\Omega_{n,k}(w)|$ . As an example, we continue with the string  $w = 100$ . The  $a_{n,k}(100)$  satisfy a recursion similar to that for  $a_n(100)$ , namely

$$a_{n,k} = a_{n-1,k} + a_{n-1,k-1} - a_{n-3,k-1}. \quad (2.46)$$

This can be proved by observing that each  $\omega \in \Omega_{n,k}(100)$  can be built from a unique string in  $\Omega_{n-1,k}(100) \cup \Omega_{n-1,k-1}(100)$  by appending either a 1 or a 0, except for the ones (of length  $n-1$ ) ending in 10, since adding a 0 would result in a 100. (There is something slightly subtle here. See the definition of *selfless* words below, and proposition 2.21. 100 is a selfless string.)

Standard generating function technology yields

$$f(z, w) = \sum_{n,k \geq 0} a_{n,k} z^n w^k = \frac{1}{1 - z(1+w) + z^3 w}, \quad (2.47)$$

and by extracting coefficients and taking limits, we obtain

$$\frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_n(i) = 1) = \frac{1}{na_n(100)} [z^n] \frac{\partial}{\partial w} \Big|_{w=1} f(z, w) \rightarrow \frac{5 + \sqrt{5}}{10} \approx .7236. \quad (2.48)$$

(As expected, the density of 1's increases as a result of conditioning on avoiding 100.) (Another aside: Mathematica is a bit temperamental about evaluating these kinds of expressions. It seems to be happiest when the derivative in  $w$  is evaluated first, then the coefficient of  $z^n$  is extracted.) A variance calculation can be performed too:

$$\text{Var}(\text{number of 1's in } X_n) \sim \frac{1}{5\sqrt{5}}n. \quad (2.49)$$

A WLLN follows for the number of 1s, since the variance is  $o(n^2)$ . (**Note that the number of copies of any string  $w$  is asymptotically normal by the  $k$ -dependent CLT.**)

Finding these recursions is sometimes very straightforward. In fact, a large class of words  $w$  share common recurrences.

**Definition 2.20.** Call a word  $w$  *selfless* if no prefix of  $w$  matches any suffix of  $w$ , i.e. if there exists no  $j < l$  such that  $w_1w_2 \cdots w_j = w_{l-j+1}w_{l-j+2} \cdots w_{l-1}w_l$ , where  $w$  has length  $l$ .

The word  $w = 100$  is selfless, and it shares the recurrence above with all other selfless words of length 3 with a single 1, via the same construction.

**Proposition 2.21.** Let  $w$  be a selfless word of length  $l$  containing exactly  $j$  1's. Then

$$a_{n,k}(w) = a_{n-1,k}(w) + a_{n-1,k-1}(w) - a_{n-l,k-j}(w). \quad (2.50)$$

*Proof.* To generate an arbitrary string in  $\Omega_{n,k}(w)$ , we can start with an arbitrary string of length  $n-1$  and append a 0 or a 1. This overcounts things slightly, since adding this final digit may have created an instance of  $w$ . So we need to throw away all strings of length  $n-1$  ending with the first  $l-1$  digits of  $w$ . To complete the proof, it suffices to note the following lemma:

**Lemma 2.22.**  $w$  is selfless if and only if the map from the set of strings in  $\Omega_{n-1,k}(w)$  ending in the first  $l-1$  digits of  $w$  to  $\Omega_{n-l,k-j}$  that chops off the last  $l-1$  digits is a bijection. □

Since  $a_n(w) = \sum_{k=0}^n a_{n,k}(w)$ , and the ‘base case’ values  $a_{n,k} = \binom{n}{k}$  for  $n < l$  or  $n = l, k \neq j$  and  $a_{l,j} = \binom{l}{j} - 1$  only depend on  $l$  and  $j$ , we get a large family of stastical coincidences:

**Proposition 2.23.** Fix  $l$ . If  $w$  and  $w'$  are any two selfless words of length  $l$ , then  $a_{n,k}(w) = a_{n,k}(w')$  and  $a_n(w) = a_n(w')$  for all  $n$  and  $k$ . In particular,  $\lambda_w = \lambda_{w'}$ , and if  $w$  and  $w'$  have the same number of 1's, then  $\gamma_w = \gamma_{w'}$ . The common recursion is

$$a_n(w) = 2a_{n-1}(w) - a_{n-l}(w), \quad (2.51)$$

and  $\lambda_w$  is the unique solution  $z \in (1, 2)$  to  $z^{l-1} = 1 + z + z^2 + \cdots + z^{l-2}$ .

Note that, in contrast to the previous proposition, we don't require that  $w$  and  $w'$  have the same number of 1's. The only difference is in the base case  $n = l$ . Solving the recurrence in Proposition 2.21 yields the generating function

$$\sum_{n,k \geq 0} a_{n,k}(w) z^n w^k = \frac{1}{1 - z(1+w) + z^l w^j}, \quad (2.52)$$

where  $l$  is the length of  $w$  and  $j$  is the number of 1's.

Note that a word  $w$  is selfless exactly when  $\phi_w(t) = t^l$ . Theorem 2.5 says that the entropy is constant over all words with common correlation polynomial, not just the selfless ones.

**Question 2.24.** *Is there a similar theorem for letter densities? For example, one could look at a statistic like*

$$\psi_w(t) = \sum_{i \in \mathcal{O}(w)} (\#\{1\text{'s in the length } i \text{ self-overlap of } w\})t^i, \quad (2.53)$$

*and hope that letter densities are ordered in the same way as these values. I think it is probably false in general that*

$$\psi_w(2) \leq \psi_{w'}(2) \iff \gamma_w \geq \gamma_{w'}, \quad (2.54)$$

*but maybe something like it is true.*

**Definition 2.25.** *Call a word  $w$  **balanced** if the number of 1's in  $w$  is half the length of  $w$ .*

Recall  $a_n(w) = |\Omega_n(w)|$ , the number of strings of length  $n$  avoiding  $w$ , and  $a_{n,k}(w)$  is the number of those with exactly  $k$  1s. We have:

**Proposition 2.26.** *If  $w$  is selfless and balanced, then  $\gamma_w = 1/2$ . In fact, for all  $n$ , the average density of 1's in a uniform random string avoiding  $w$  is exactly  $1/2$ , i.e.*

$$\sum_{k=0}^n k a_{n,k}(w) = \frac{1}{2} n a_n(w). \quad (2.55)$$

*Proof.* It would be nice to have a bijective proof. The above can be checked directly using the generating function formula 2.52. wetting

$$f(z, w) = \sum_{n,k \geq 0} a_{n,k}(w) z^n w^k = \frac{1}{1 - z(1 + w) + z^l w^{l/2}}, \quad (2.56)$$

and a quick computation shows

$$\left. \frac{\partial}{\partial w} \right|_{w=1} f = \frac{1}{2} z g'(z) \quad (2.57)$$

which is equivalent to the claim. □

Also note: the family of selfless words is quite large! The probability of a word being selfless is bounded away from 0 for any  $n$  (perhaps an interesting computation of its own?), so a constant proportion of words are selfless. (Simulation suggests the probability of being selfless is approximately .266 for  $n$  large. The ‘mean field’ calculation – i.e. assuming matching each suffix to each prefix are independent events – gives an estimate of  $\prod_{j \geq 1} 1 - 2^{-j} \approx .289$ .) **There is a recursion for selfless words which can be solved to some extent. There’s an OEIS entry, for example.**

There is another class of words for which the density can be easily seen to be exactly  $1/2$ . Recall that  $\text{Rev}(w)$  is the reversal of  $w$ , and  $\overline{w_1 w_2 \cdots w_l} = \overline{w_1} \overline{w_2} \cdots \overline{w_l}$ , where  $\overline{s} = 1 - s$  is the ‘bit flipping’ operation. Note that these two operations are commuting involutions, i.e.  $\text{Rev}(w) = \text{Rev}(\overline{w})$  and  $\overline{\overline{w}} = \text{Rev}(\text{Rev}(w)) = w$ .

**Definition 2.27.** *Call a word  $w$  **sweet** if  $\overline{w} = \text{Rev}(w)$ .*

Note that sweet words must be balanced, so all sweet words have even length. Conditioning on avoiding a sweet word keeps the 0-1 count balanced:

**Proposition 2.28.** *If  $w$  is a sweet word, then  $\gamma_w = 1/2$ . In fact, for all  $n$ , the average density of 1's in a uniform random string avoiding  $w$  is exactly  $1/2$ , i.e.*

$$\sum_{k=0}^n k a_{n,k}(w) = \frac{1}{2} n a_n(w). \quad (2.58)$$

*Proof.* It suffices to find a bijection  $\omega \mapsto \omega'$  from  $\Omega_n(w)$  to itself such that the number of 0's in  $\omega'$  is equal to the number of 1's in  $\omega$ . Indeed, the existence of such a bijection implies that the total number of 1's over all strings in  $\Omega_n(w)$  is the same as the total number of 0's, which implies the result. The bijection that works has the simple formula  $\omega \mapsto \text{Rev} \circ \bar{\omega}$ . This map is an involution that swaps 0's and 1's, and that  $w$  is sweet implies that it maps  $\Omega_n(w)$  to itself.  $\square$

It is worth noting that the number of sweet strings words exponentially, but still makes up a vanishing fraction of all words. Indeed, the sweet words of length  $l$  can be exactly enumerated by choosing an arbitrary word  $\omega$  of length  $l/2$ , then forming the word  $\omega \oplus \text{Rev}(\bar{\omega})$ , where  $\oplus$  is concatenation. So there are exactly  $2^{l/2}$  sweet words of length  $l$ .

Being balanced is not enough to guarantee that the conditioned string is balanced. Already there is a counterexample when  $l = 4$ . Note that of the 6 words of length 4 with two 1's, up to reversal and bit-flipping only one is not sweet: 1010 and 1100 are sweet, while 1001 is not. And we have:

**Fact 2.29.** *The limiting density of 1's in a uniform random 1001 avoiding string is*

$$\gamma_{1001} = \frac{2(-3 + 2\sqrt{5})^{5/2} \sqrt{\frac{1}{55}(3 + 2\sqrt{5})} \left( 110\sqrt{-3 + 2\sqrt{5}} + 44\sqrt{5(-3 + 2\sqrt{5})} + \sqrt{11}(35 + 17\sqrt{5}) \right)}{11(-35 + 27\sqrt{5}) \left( \sqrt{11} + 3\sqrt{-3 + 2\sqrt{5}} \right)} \quad (2.59)$$

$$\approx .494161. \quad (2.60)$$

Amazingly, conditioning on avoiding 1001 very slightly decreases the density of 1s!

This ostentatious constant comes from computing with generating functions exactly. (**See section 2.5 for a nicer calculation.**) Via the graph  $L$ , one finds recursions (where all  $a_n$  are interpreted as  $a_n(1001)$  for ease of notation)

$$a_n = 2a_{n-1} - a_{n-3} + a_{n-4} \sim \frac{(27\sqrt{5} - 35)(\sqrt{11} + 3\sqrt{2\sqrt{5} - 3})}{20(2\sqrt{5} - 3)^{5/2}\sqrt{3 + 2\sqrt{5}}} \frac{1}{2^n} (1 + \sqrt{3 + 2\sqrt{5}})^n, \quad (2.61)$$

and

$$a_{n,k} = a_{n-1,k} + a_{n-1,k-1} - a_{n-3,k-1} + a_{n-4,k-1}, \quad (2.62)$$

which satisfies

$$\sum_{n,k \geq 0} a_{n,k} z^n w^k = \frac{1 + z^3 w}{1 - z(1 + w) + z^3 w - z^4 w}. \quad (2.63)$$

Note also that 1001 is not selfless – the recursion for  $a_{n,k}$  requires additional ‘correction’ terms.

**Conjecture 2.30.** *The density of 1's  $\gamma_w$  in a uniform random element of  $\Omega_n(w)$  is  $1/2$  if and only if  $w$  is sweet or balanced and selfless and satisfies ...*

Simulation found counterexamples of length 8 to just being sweet/balanced, namely 10011010 and 10100110.

**Conjecture 2.31.**  *$\gamma_w = 1/2$  only if  $w$  is balanced.*

This has been confirmed by (approximate) simulations up to words  $w$  of length 20.

**Conjecture 2.32.**  *$\gamma_w > 1/2$  if and only if  $w$  has at least as many 1's as 0's.*

**Question 2.33.** *Classify the set of balanced strings  $w$  with  $\gamma_w < 1/2$ .*

**Conjecture 2.34.** *For all  $l$  and all strings  $w$  of length  $l$ ,  $|\gamma_w - 1/2| \leq C \exp(-cl)$ . If this holds, what is the optimal rate  $c$ ?*



## 2.5 Letter densities via the MME for SFT

Let  $\nu$  denote the measure of maximal entropy for the shift of finite type with single forbidden word  $w$ . This measure can be computed explicitly via some matrix computations with the graph  $L_w$ , and gives an alternate way to calculate the entropy  $\lambda_w$  and the letter density  $\gamma_w$ . Namely:

**Fact 2.35.**  $\lambda_w$  is the (exponential of the) topological entropy of  $\nu$ , and  $\gamma_w$  is  $\nu(C_0)$ , the measure of the cylinder set of 0 under  $\nu$ .

These values can be computed exactly from any representation of the corresponding SFT.  $\lambda_w$  is the largest eigenvalue of any graph representation of the corresponding SFT:  $L_w$  is the ‘minimal’ such representation. As for  $\gamma_w$ , recalling the graph  $L$ , and using (by a slight abuse of notation)  $\nu$  to refer also to the stationary MME – the ‘parry measure’ – on the graph  $L$ , we have:

**Proposition 2.36.** Let  $w = w_1 w_2 \dots w_l$  with  $w_1 = 1$ . Then

$$\gamma_w = \sum_{k < l: w_k = 1} \nu(k).$$

Also, the characteristic polynomial of  $L_w$  matches the recursion satisfied by  $|\Omega_n|$ :

**Proposition 2.37.** Let  $A_w$  denote the adjacency matrix of  $L_w$ , and let  $p_w(\lambda)$  denote its characteristic polynomial, say  $p_w(\lambda) = \sum_{i=0}^l c_i \lambda^i$ . Then  $a_n = |\Omega_n|$  satisfies

$$c_0 a_n = \sum_{i=1}^l c_i a_{n-i}.$$

**Example 2.38.** To illustrate, we recover the example  $w = 100$  via this method. The graph  $L_{100}$  has adjacency matrix

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

with characteristic polynomial  $\lambda^3 - 2\lambda^2 + 1 = (\lambda - 1)(\lambda^2 - \lambda - 1)$ , top eigenvalue  $\varphi = \frac{1+\sqrt{5}}{2}$ , and right/left eigenvectors

$$r_{100} = \begin{bmatrix} 1 \\ \varphi - 1 \\ 1 - \varphi^{-1} \end{bmatrix} \ell_{100} = [0 \quad 1 \quad \varphi^{-1}].$$

The the parry measure is given by  $\nu_j = \frac{1}{Z_{100}} r_j \ell_j$ ,  $j = 0, 1, 2$ , with  $Z_{100} = r \cdot \ell$ . We have

$$\nu = \frac{1}{3\varphi - 4} (0, \varphi - 1, 2\varphi - 3),$$

and the density of 1s is  $\gamma_{100} = \nu(1) = \frac{\varphi-1}{3\varphi-4} = \frac{5+\sqrt{5}}{10}$ , since the state where we match the first digit is the only state ending in a 1. This matches the calculation from the previous section.

For completeness we also carry out the analysis this way for:

**Example 2.39.** Let  $w = 1001$ , which has  $L_{1001}$  with adjacency matrix

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

(irreducible over  $\mathbb{Q}$ ) characteristic polynomial  $\lambda^4 - 2\lambda^3 + \lambda - 1$ , top eigenvalue  $\lambda \approx 1.866760$ , and right/left eigenvectors

$$r_{1001} = \begin{bmatrix} 1 \\ \lambda - 1 \\ (\lambda - 1)^2 \\ \lambda^{-1} \end{bmatrix} \ell_{1001} = [1 \quad \lambda^2(\lambda - 1) \quad \lambda(\lambda - 1) \quad \lambda - 1].$$

Thus the parry measure is

$$\nu = \frac{1}{-2\lambda^3 + 6\lambda^2 - 3\lambda + 3} (1, \lambda^2(\lambda - 1)^2, \lambda(\lambda - 1)^3, 1 - \lambda^{-1}),$$

$$\text{and } \gamma_{1001} = \nu(1) = \frac{\lambda^2(\lambda - 1)^2}{-2\lambda^3 + 6\lambda^2 - 3\lambda + 3}.$$

These exact rational functions for the eigenvectors had to be obtained by hand – so far I don't know a systematic way of determining the exact rational expressions in terms of the top eigenvalue  $\lambda$ .

**Question 2.40.** Write a computer program that finds an expression for the Perron eigenvectors as polynomials in the entropy  $\lambda$ .

We continue with some general computations:

**Example 2.41.** Consider  $w = 111 \dots 1$ , a string of  $l$  1s. The graph  $L_{11\dots 1}$  has adjacency matrix

$$\begin{bmatrix} 1 & 1 & 0 & \dots & & \\ 1 & 0 & 1 & 0 & \dots & \\ 1 & 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & & \\ 1 & 0 & \dots & \dots & \dots & 0 \end{bmatrix}$$

The characteristic polynomial is  $\lambda^n - \lambda^{n-1} - \dots - \lambda - 1$ , with right/left eigenvectors

$$r_j = \sum_{i=1}^{n-j+1} \lambda^{-i}, \ell_j = \lambda^{1-j}, j \in [n].$$

The density of 1s is  $\gamma_{11\dots 1} = 1 - \nu(0) = \lambda^n(\lambda - 1)^2[\lambda^{n+1} - \lambda(n + 1) + n]^{-1}$ .

## 2.6 Word counts

Rather than jumping straight to the SFT by forbidding a word  $w$ , it is natural (and possibly helpful) to study the substring counts  $N_w(x) = \#$  of copies of  $w$  appearing in  $x \in \{0, 1\}^n$ . Observe that for iid  $\text{Ber}(1/2)$  bits, as  $n \rightarrow \infty$ ,  $\frac{1}{n}\mathbb{E}N_w \approx 2^{-l}$ . We can also compute the covariance directly:

**Proposition 2.42.** *Let  $v, w$  be any words of lengths  $k$  and  $l$ , respectively. Then*

$$\text{Cov}(N_v, N_w) = (\Phi(v, w) - 2^{-k-l}(k+l-1))n + O(1), \quad (2.64)$$

where  $\Phi(v, w)$  is the symmetric overlap polynomial

$$\Phi(v, w) = \sum_{x \in \mathcal{O}(v, w)} 2^{-\text{len}(x)}, \quad (2.65)$$

where  $\mathcal{O}(v, w) = \{(x_1, \dots, x_r) : r < k+l \text{ and } x_1 \cdots x_k = v, x_{r-l+1} \cdots x_r = w \text{ or vice versa}\}$ .

For example, if  $w = 101$  and  $v = 11$ , there are two overlaps, namely 1101 and 1011, each of which have length 4, so

$$\Phi(101, 11) = 2^{-4} + 2^{-4} = 2^{-3}. \quad (2.66)$$

Plugging in shows  $\text{Cov}(N_{101}, N_{11}) = O(1)$ . Note the special case

$$\Phi(w, w) = 2\phi_w(2^{-1}) - 2^{-l}. \quad (2.67)$$

since each overlap of  $w$  with itself occurs twice in  $\mathcal{O}(w, w)$ , and we subtract the full overlap to fix the over counting, and where  $\phi$  is the usual self-overlap polynomial defined in 2.3. So as a corollary we get

**Corollary 2.43.** *For any  $w$  of length  $l$ ,*

$$\frac{1}{n} \text{Var}(N_w) \sim 2\phi_w(2^{-1}) - 2^{-l} - (2l-1)2^{-2l}. \quad (2.68)$$

**Question 2.44.** *Observe that the variance is increasing in  $\phi_w(2^{-1})$ : does this have any significance? Can it be proved without the explicit calculation?*

It is also possible to write an explicit formula for the distribution of  $N_w$ , via a matrix calculation. Let  $X$  be any SFT where  $w$  is an allowable word, and let  $P$  be the  $\{0, 1\}$  valued matrix with 0s at the transitions forbidden by  $X$ . Construct matrix  $Q$  identical to  $P$ , except that we replace any transition  $P_{ij} = 1$  which ‘creates’ a copy of  $w$  by a variable  $y$ . For example, when  $X$  is the shift where 11 is forbidden and  $w = 1$ , we use

$$Q = \begin{bmatrix} 1 & 1 \\ y & 0 \end{bmatrix} \quad (2.69)$$

(Here the states are simply  $\{0, 1\}$ , since the parry measure on the golden mean shift is a markov chain with memory 1.) The number of words of length  $n$  containing exactly  $k$  copies of  $w$  is given by

$$[y^k] \sum_{i,j} (Q^n)_{ij}. \quad (2.70)$$

In the golden mean example,  $Q$  has eigenvalues

$$\lambda_{\pm} = \frac{1}{2} \left( 1 \pm \sqrt{1 + 4y} \right), \quad (2.71)$$

and one easily diagonalizes:

$$Q = ADA^{-1}, \text{ where } A = \begin{bmatrix} 1 & 1 \\ \lambda_+ - 1 & \lambda_- - 1 \end{bmatrix}, \quad D = \text{diag}(\lambda_+, \lambda_-). \quad (2.72)$$

some algebra yields that the number of words of length  $n$  containing exactly  $k$  1s is

$$[y^k] \frac{1}{\sqrt{1 + 4y}} \left( \lambda_+^n (1 + y) + \lambda_-^n (y - \lambda_-) \right), \quad (2.73)$$

which can be unraveled to get explicit formulas. Computing the joint distribution of  $(N_w, N_v)$  for a pair of strings  $w, v$  is already a challenge. A similar tool works to get explicit formulas: for example, to count  $w = 1$  and  $v = 11$  over the full shift (no forbidden words), one would use the matrix

$$\begin{bmatrix} 1 & y \\ y & yz \end{bmatrix} \quad (2.74)$$

with  $y$  counting the occurrences of  $w$  and  $z$  counting the occurrences of  $v$ . One could follow the same procedure – diagonalize, then extract coefficients – to obtain some explicit formulas for the  $N_w$  and  $N_v$  counts. **It's not clear how useful this is. Maybe an appeal to generating function technology can give us a CLT, even a joint CLT?**

## 2.7 Gibbs measures

There is a natural Gibbs measure that interpolates between iid bits and the SFT obtained by forbidding  $w$ : namely, fix  $\beta > 0$ , and weight occurrences of  $w$  by  $\beta$ . That is, for any (fixed)  $n \in \mathbb{N}$  we have a measure  $\mu_w$  on the set of strings of length  $n$  given by

$$\mu_{w,\beta}(x) = Z_{w,\beta}^{-1} \exp(\beta N_w(x)), \quad (2.75)$$

where  $N_w(x)$  is the number of occurrences of  $w$  in  $x$ . Then  $\beta = 0$  is iid measure,  $\beta = -\infty$  is the SFT where  $w$  is forbidden, and  $\beta = \infty$  is atomic, supported on periodic copies of  $w$ . Looking at the observables  $N_v$  for another fixed word  $w$  under the measure  $\mu_{w,\beta}$  leads to some interesting questions. For example,  $\mathbb{E}_{w,\beta}[N_1]$  is the letter density of 1s in a typical  $\mu_{w,\beta}$  sample: if we could show that

$$\frac{\partial}{\partial \beta} \mathbb{E}_{w,\beta}[N_1] < 0 \text{ for } \beta < 0, \quad (2.76)$$

this would imply that  $\gamma_w < 1/2$  by integrating over  $\beta < 0$ . This appears to be true for  $w = 1001$ , for example.

For many pairs  $w$  and  $v$ ,  $\mathbb{E}_{w,\beta}[N_v]$  is either globally minimized or maximized at  $\beta = 0$ , but this appears to not always be the case.

**Question 2.45.** *Are  $\mathbb{E}_{w,\beta}[N_v]$  and  $\mathbb{E}_{v,\beta}[N_w]$  related in a canonical way?*

In general, the derivative has the following nice form:

**Proposition 2.46.** *Let  $f$  be any function on  $\{0, 1\}^n$ . Then*

$$\frac{\partial \mathbb{E}_{w,\beta}[f]}{\partial \beta} = \text{Cov}_{w,\beta}(N_w, f). \quad (2.77)$$

*Proof.* Note that  $Z_{w,\beta} = Z = \sum_x \exp(\beta N_w)$ , so

$$\frac{\partial Z}{\partial \beta} = Z \mathbb{E}[N_w]. \quad (2.78)$$

Thus

$$\frac{\partial}{\partial \beta} \mathbb{E}[f] = -Z^{-2} \mathbb{E}[N_w] \sum_x f(x) \mu(x) + Z^{-1} \sum_x f(x) N_w(x) \mu(x) \quad (2.79)$$

$$= \mathbb{E}[N_w \cdot f] - \mathbb{E}[N_w] \cdot \mathbb{E}[f] \quad (2.80)$$

$$= \text{Cov}(N_w, f). \quad (2.81)$$

□

In particular:

**Corollary 2.47.** *At  $\beta = 0$ ,  $\frac{1}{\partial \beta} \mathbb{E}_{w,\beta}[N_v] = \frac{1}{\partial \beta} \mathbb{E}_{v,\beta}[N_w]$ .*

Here is an attempt to calculate  $\mathbb{E}_{w,\beta}[N_w]$  for all  $\beta$ , which seems to nearly work(!) Recall that  $\mu$  is a measure on  $\{0, 1\}^n$  for fixed (large)  $n$ . Fix  $w$ , and write

$$g_k(\beta) = n^{-k} \mathbb{E}_{w,\beta}[N_w^k], \quad (2.82)$$

the  $k$ th moment of  $n^{-1}N_w$  under  $\mu_{w,\beta}$ . By Proposition 2.46,

$$\frac{\partial g_k(\beta)}{\partial \beta} = n(g_{k+1}(\beta) - g_k(\beta)g_1(\beta)). \quad (2.83)$$

Write  $m_\beta = \lim_{n \rightarrow \infty} g_1(\beta)$  for short, and  $\sigma_\beta^2 = \lim_{n \rightarrow \infty} \text{Var}(n^{-1}N) = \lim_{n \rightarrow \infty} n^{-1} \left( \mathbb{E}[N^2] - n^2 c_\beta^2 \right)$ . At  $\beta = 0$ ,  $N$  is an  $l$ -dependent sum of random variables, so it satisfies a CLT, i.e. viewed as a random variable in the measure  $\mu_{\beta,w}$  for fixed  $\beta = 0$  and  $n \rightarrow \infty$ ,

$$\frac{N - g_1 n}{\sigma \sqrt{n}} \rightarrow_d N(0, 1). \quad (2.84)$$

This should also be true for any  $\beta \in \mathbb{R}$ , but it requires understanding the infinite-volume limit of the measures  $\mu$  better. (Proof coming soon?) Using this as an approximation, we have that for each fixed  $\beta$ ,

$$N \approx m_\beta n + \sigma_\beta \sqrt{n} Z, \quad (2.85)$$

where  $Z$  is a Normal(0, 1). This gives approximations to the moments of  $N$ :

$$\mathbb{E}[N^k] \approx \mathbb{E}[(m_\beta n + \sigma_\beta \sqrt{n} Z)^k] \quad (2.86)$$

$$= m_\beta^k n^k + k m_\beta^{k-1} \sigma_\beta n^{k-1/2} \mathbb{E}[Z] + \binom{k}{2} m_\beta^{k-2} \sigma_\beta^2 n^{k-1} \mathbb{E}[Z^2] + \dots \quad (2.87)$$

$$= m_\beta^k n^k + \binom{k}{2} m_\beta^{k-2} \sigma_\beta^2 n^{k-1} + O(n^{k-2}). \quad (2.88)$$

In particular,  $\mathbb{E}[N] \approx m_\beta n$ ,

$$\mathbb{E}[N^2] \approx m_\beta^2 n^2 + \sigma_\beta^2 n, \quad (2.89)$$

and

$$\mathbb{E}[N^3] \approx m_\beta^3 n^3 + 3m_\beta \sigma_\beta^2 n^2. \quad (2.90)$$

Now we plug into Proposition 2.46:

$$\frac{\partial m_\beta}{\partial \beta} \approx \lim_{n \rightarrow \infty} \frac{1}{n} \text{Cov}(N, N) = \sigma_\beta^2, \quad (2.91)$$

and

$$\frac{\partial \sigma_\beta^2}{\partial \beta} = \lim_{n \rightarrow \infty} \frac{\partial}{\partial \beta} n^{-1} \text{Var}(N) \quad (2.92)$$

$$= \lim_{n \rightarrow \infty} n^{-1} \left( \frac{\partial}{\partial \beta} \mathbb{E}[N^2] - \frac{\partial}{\partial \beta} (\mathbb{E}[N]^2) \right) \quad (2.93)$$

$$= \lim_{n \rightarrow \infty} n^{-1} (\mathbb{E}[N^3] - \mathbb{E}[N^2] \mathbb{E}[N] - 2 \mathbb{E}[N] \text{Var}(N)) \quad (2.94)$$

$$\approx \lim_{n \rightarrow \infty} n^{-1} (m_\beta^3 n^3 + 3m_\beta \sigma_\beta^2 n^2 - (m_\beta^2 n^2 + n \sigma_\beta^2)(m_\beta n) - 2m_\beta \sigma_\beta^2 n^2) \quad (2.95)$$

$$= 0 \quad (!!!) \quad (2.96)$$

(Here we used Proposition 2.46 twice, plus the product rule, to evaluate the derivatives.)

What happened? The expression for the derivative of  $\sigma_\beta^2$  cancelled to constant order, so everything vanished! I believe most of this sketchy calculation is correct, except that the normal approximation to  $N$  is not good enough: with a better approximation, we would see that this expression *does not cancel* to linear order  $\Theta(n)$ , so we would get some actual expression in the limit.

**Question 2.48.** *Maybe the combinatorial/Markovian structure of the measure  $\mu$  is enough to compute  $\mathbb{E}[N]$ ,  $\mathbb{E}[N^2]$  and  $\mathbb{E}[N^3]$  precisely enough so that this calculation can be carried out?*

If it works, this would be a kind of bootstrapping: we first get good enough *approximations* to the moments of  $N$  so that we can solve for the moments explicitly via a system of differential equations! It may be that further moments of  $N$  pop out, so it may be a system involving the first  $l$  moments of  $N$ , or perhaps the full moment generating function would be required.

One can also jump directly to the moment generating function:

$$g(\beta, w) = \sum_{k \geq 0} g_k(\beta) w^k. \quad (2.97)$$

Playing with this a bit gives a functional relation on  $g$ :

$$\frac{\partial}{\partial \beta} g(\beta, w) = 1 + n \sum_{k \geq 1} w^{-1} w^{k+1} g_{k+1}(\beta) - n g_1(\beta) w^{-1} \sum_{k \geq 1} g_k(\beta) w^k \quad (2.98)$$

$$= 1 + n w^{-1} (g(\beta, w) - 1 - w g_1(\beta)) - n g_1(\beta) (g(\beta) - 1) \quad (2.99)$$

$$= 1 - n w^{-1} + n (w^{-1} - g_1(\beta)) g(\beta, w). \quad (2.100)$$

We also know the ‘initial values’  $g(\beta, 0) = 1$ , and  $g(0, w)$  can be calculated directly (see the ‘Analytic Pattern Matching’ book for explicit formulas.  $\beta = 0$  is the iid case.) Finally,

$$g_1(\beta) = \left. \frac{\partial}{\partial w} g(\beta, w) \right|_{w=0}. \quad (2.101)$$

so  $g$  is described by the single PDE

$$g_\beta(\beta, w) = 1 - n w^{-1} + n g(\beta, w) (w^{-1} - g_w(\beta, 0)). \quad (2.102)$$

Again we have the same ‘normalization’ issue – if we understood  $g$  better, the factor of  $n$  would cancel out and we would be left with a proper, parameter-free PDE. But this requires getting better approximations for the moments of  $N$ .

**Question 2.49.** *Does the function  $\mathbb{E}_{w,\beta}[N_v]$  always have a single critical point? When is it a global max/min? When is the extreme point at  $\beta = 0$ ?*

Some computer simulations have been carried out for this. I used Glauber dynamics to approximate the function  $f_{w,v}(\beta) = \mathbb{E}_{w,\beta}[N_v]$  for large  $n$ , and  $\beta \in [-b, b]$  for  $b \approx 5$ . The function  $f$  always appears to be smooth (probably analytic), the limit  $\beta \rightarrow -\infty$  exists and agrees with  $\gamma_w$ , and the limit  $\beta \rightarrow \infty$  exists and agrees with the density of  $T$ ’s in the ‘periodic tiling of  $\mathbb{Z}$  by  $w$ ’s. Here are the results I recorded:

- $v = 1$

- $w = 1001$ :  $f$  has a maximum at  $\beta = 0$
- $w = 100$ :  $f$  strictly decreasing
- $w = 11$ :  $f$  strictly increasing
- $w = 100110$ :  $f$  always exactly  $1/2$  (clear by symmetry)
- $w = 110010$ :  $f$  always exactly  $1/2$  (not sure why??)
- $w = 10011100$ :  $f$  always exactly  $1/2$  (not sure why??)
- $v = 11$ 
  - $w = 00$ :  $f$  strictly decreasing
  - $w = 1001$ :  $f$  strictly decreasing
  - $w = 111$ :  $f$  strictly increasing
  - $w = 10011100$ :  $f$  strictly increasing
- $v = 00, w = 1001$ :  $f$  decreasing then increasing, minimum around  $\beta = 1.2$  (!!)
- $v = 10, w = 1001$ :  $f$  strictly increasing
- $v = 01, w = 1001$ :  $f$  strictly increasing
- $v = 0000, w = 1001$ :  $f$  strictly decreasing
- $v = 11, w = 101$ :  $f$  has a maximum at  $\beta = 0$

**Question 2.50.** *According to these simulations,  $\mathbb{E}_{w,\beta}[N_v]$  is always a monotone function of  $\beta$  for  $\beta \in (-\infty, 0)$ . Is this always the case? It would allow us to just compute at  $\beta = 0$  and answer the density question! But maybe it's too much to hope for.*

I also simulated  $\mathbb{E}_{w,\beta}[N_w]$  as a function of  $\beta$ : it is strictly increasing, and appears to depend only on the correlation polynomial of  $w$ , i.e. these functions are identical for  $w, w'$  with the same correlation polynomial. (Perhaps this can be shown directly?)



## 2.8 Martingale and hitting time

In this section we recall the martingale method, which seems to have first been spelled out by Li in full detail, though it's semi-folklore, Conway and Feller knew it. We start with the iid case, where the martingale is surprisingly robust and versatile, then explain how it can be generalized to an arbitrary markov chain, with an eye toward measures of maximal entropy for SFTs, in the next section.

### 2.8.1 IID case

Let  $w \in [q]^l$  be any finite word. Generate iid digits  $X_i \in [q]$  uniformly at random. Denote by  $\tau_w$  the hitting time

$$\tau_w = \min\{t : (X_{t-l+1}, X_{t-l+2}, \dots, X_t) = w\}. \quad (2.103)$$

We construct a martingale with respect to the  $X$  process as follows. At each time  $t = 1, 2, \dots$ , imagine a better arrives just before digit  $X_t$  arrives and places a 1 dollar bet on the event  $X_t = w_1$ . Then we – the casino – pay out with odds  $1 : q$  if the bettor is successful, otherwise she loses her 1 dollar investment. If she is successful, then she bets again on the next digit of  $w$ , i.e. on the event  $X_{t+1} = w_2$ , always betting her total gross winnings. Whenever she loses, she leaves and never places another bet. Thus, at each time  $t$ , there may be up to  $k$  bettors in the game. Set

$$W_t = \text{net profit of the casino up to all bets on } X_1, \dots, X_t. \quad (2.104)$$

Then  $W_t$  is clearly a martingale, in this case with bounded increments. We now apply the optional stopping theorem at the stopping time  $\tau_w$ . At time  $\tau_w$  almost all bettors are gone, except the ones who successfully bet on  $v$  just before or while the first  $w$  occurred. So to compute  $W_{\tau_w}$ , the casino has collected gross profit  $\tau_w$ , minus the winnings of all the bettors still alive at that time:

$$W_{\tau_w} = \tau_w - \sum_{j \in \mathcal{O}(w)} q^j = \tau_w - \phi_w(q). \quad (2.105)$$

Thus by the optional stopping theorem (which is valid here since  $\tau$  is sub-exponential),

$$0 = \mathbb{E}W_{\tau_w} = \mathbb{E}\tau_w - \phi_w(q) \implies \mathbb{E}\tau_w = \phi_w(q). \quad (2.106)$$

This martingale construction can be modified in many ways. We now work through a number of examples.

**(Probabilities for  $\tau$ )** Fix a word  $w$  and an integer  $s \geq 1$ . We build a martingale  $S_t^s$  by putting a single bettor who arrives at time  $s$  and bets 1 dollar on the word  $w$ . Again we will apply the optional stopping theorem to  $S$  at time  $\tau = \tau_w$ . Observe that

$$S_\tau^s = 1\{\tau \geq s\} - \sum_{j \in \mathcal{O}(w)} q^j 1\{\tau = s + j - 1\}. \quad (2.107)$$

So by the OST,

$$0 = \mathbb{E}S_t^s = \mathbb{P}(\tau \geq s) - \sum_{j \in \mathcal{O}(w)} q^j \mathbb{P}(\tau = s + j - 1). \quad (2.108)$$

This can be thought of as a recursive equation that determines the probabilities  $p_s = \mathbb{P}(\tau \geq s)$ , namely

$$p_s = \sum_{j \in \mathcal{O}} q^j (p_{s+j-1} - p_{s+j}). \quad (2.109)$$

One can solve for the generating function  $\sum_{s \geq 1} z^s p_s$  by the usual methods, (using the initial values  $p_s = 1$  for  $s = 1, 2, \dots, l$ ), obtaining

$$\sum_{s \geq 1} p_s z^s = \frac{z \phi_w(qz^{-1})}{1 - (z-1) \phi_w(qz^{-1})}. \quad (2.110)$$

This recovers a result from Guibas-Odlyzko. Using the martingale is much more elegant than futzng around with recursions! Another similar example is the generating function for  $\mathbb{P}(\tau = t)$ : to obtain this, one can form the martingale where bettors still arrive at each time and bet that  $w$  will occur, but the initial bet size of the bettor at time  $s$  is  $z^s$  for some  $z \in \mathbb{R}$ . If  $z$  is sufficiently small, the OST can be applied, and one will obtain (after some algebra) the MGF of  $\tau$ .

**(Backwards chain)** Fix  $w \in [q]^l$ . Suppose we look backwards from time  $\tau = \tau_w$ , i.e. at the random process  $Y_s = X_{\tau-l+1-s}$  for  $s = 1, 2, \dots$ . Is  $Y_s$  a markov chain up to the stopping time  $\tau - l + 1$ ? Here is a strategy to at least compute the distribution of  $X_{\tau-l}$ . For definiteness, set  $X_s = \dagger$  for  $s \leq 0$ , so if  $\tau = l$ , then  $X_{\tau-l} = \dagger$ . For each  $i \in [q]$ , consider the word  $iw$ , i.e. append  $i$  to the front of  $w$ . Consider the martingale  $U_t^i$  where bettors arrive at each time, and they bet on the word  $iw$  occurring. Then we will stop at time  $\tau_w$  and compute. First a quick definition that will be useful throughout:

**Definition 2.51.** For any words  $w \in [q]^l, v \in [q]^k$ , let  $\vec{\mathcal{O}}(w, v)$  denote the ‘directed’ overlap set, namely for  $j \in [k \wedge l]$ ,  $j \in \vec{\mathcal{O}}(w, v)$  if the first  $j$  digits of  $v$  match the last  $j$  digits of  $w$ . Define the ‘directed’ overlap polynomial accordingly:

$$\vec{\phi}_{w,v}(t) = \sum_{j \in \vec{\mathcal{O}}(w,v)} t^j. \quad (2.111)$$

Then we have

$$U_\tau^i = \tau - q^{l+1} 1\{X_{\tau-l} = i\} - \sum_{j \in \vec{\mathcal{O}}(w, iw)} q^j, \quad (2.112)$$

where Applying the OST and plugging in the value for  $\mathbb{E}\tau$  gives

$$\mathbb{P}(X_{\tau-l} = i) = q^{-l-1} (\phi_w(q) - \vec{\phi}_{w, iw}(q)), \quad (2.113)$$

where  $\vec{\phi}$  is the corresponding ‘directed’ overlap polynomial. As a check, with  $q = 2$  and  $w = 11$ , we have  $\vec{\phi}_{w, 1w} = \phi_w$ , so  $\mathbb{P}(X_{\tau-l} = 1) = 0$  in this case, which is correct because if the digit before the first 11 was a 1, then the first 11 would have been one digit earlier. A more interesting example is  $q = 2$  and  $w = 1010$ : then  $\vec{\phi}_{w, 0w}(2) = 2^3 + 2$ , since 01010 overlaps 1010 in a size 3 prefix and a size 1 prefix; so in this case  $\mathbb{P}(X_{\tau-4} = 0) = 2^{-5}(2^4 + 2^2 - 2^3 - 2) = \frac{10}{32}$ , which is confirmed by simulations. Note that here  $\mathbb{P}(X_{\tau-4} = \dagger) = \mathbb{P}(\tau = 4) = 2^{-4}$ , and  $\mathbb{P}(X_{\tau-4} = 1) = \frac{5}{8}$ .

Generally, if  $v$  is any word of length  $k$ , betting on  $vw$  and stopping at time  $\tau = \tau_w$  and applying the same martingale method as above yields

$$\phi_w(q) = \vec{\phi}_{w, vw}(q) + \sum_{i=1}^k q^{l+i} \mathbb{P}(Y_i Y_{i-1} \dots Y_1 = v_1 v_2 \dots v_i) 1\{w_1 w_2 \dots w_{k-i} = v_{i+1} \dots v_k, l-k+i \in \mathcal{O}(w)\}. \quad (2.114)$$

Note that the events involving  $Y$  appearing in the above formula include the cases where  $\tau$  is small so that  $Y_i = \dagger$  for some  $i$ , so an event like  $\{Y_i \cdots Y_1 = v_1 \cdots v_i\}$  cannot occur. This formula can be used to give explicit expressions for transition probabilities of the chain  $Y_s$ . For example:

**Fact 2.52.** *If  $l - 1 \notin \mathcal{O}(w)$ , i.e. if  $w \neq i^l$  for any  $i \in [q]$ ,*

$$\mathbb{P}(Y_2 = i | Y_1 = j) = q^{-1} \frac{\phi_w(q) - \vec{\phi}_{w,jw}(q)}{\phi_w(q) - \vec{\phi}_{w,ijw}(q)} \quad (2.115)$$

*If  $l - 1, l - 2 \notin \mathcal{O}(w)$ ,*

$$\mathbb{P}(Y_3 = i | Y_2 = j) = q^{-1} \frac{q\phi_w(q) - \sum_{k \in [q]} \vec{\phi}_{w,ijkw}(q)}{q\phi_w(q) - \sum_{k \in [q]} \vec{\phi}_{w,jkw}(q)} \quad (2.116)$$

The quantities  $\vec{\phi}_{w,ijw}(q)$  and  $\vec{\phi}_{w,jw}(q)$  are closely related, since  $u \in \vec{\mathcal{O}}(w, ijw) \implies u - 1 \in \vec{\mathcal{O}}(w, jw)$  for  $u \geq 2$ .

**Question 2.53.** *Decide whether the backwards markov chain  $Y_s$  converges quantitatively as  $s \rightarrow \infty$  to the maximal entropy markov chain for the shift space where  $w$  is forbidden. This should be clear if we condition on  $\tau \geq t$  for  $t$  large, but even unconditionally there should be some kind of convergence, e.g. for the transition probabilities for  $Y$ .*

**(Word counts)** We can compute word counts up to the stopping time  $\tau$  via this method. We illustrate this by showing a simple method to compute the expected number of copies of a word  $v$  in  $(X_1, \dots, X_\tau)$ . Generally, fix words  $w \in [q]^l, v \in [q]^k$  for  $k \leq l$ , and assume  $w \neq v$  (the case  $w = v$  is trivial). Let  $N_v(t)$  denote the number of copies of  $v$  in  $X_1, \dots, X_t$ : we will give an exact formula for  $\mathbb{E}N_v(\tau_w)$ . We build a martingale  $Q_t$  by having bettors arrive at each time, but instead of betting on  $w$ , they bet on  $v$ . If a bettor ever witnesses a copy of  $v$ , they take their  $q^k$  winnings and go home. We stop at time  $\tau = \tau_w$ . Then

$$Q_\tau = \tau - q^k N_v(\tau) - \vec{\phi}_{w,v}(q) + q^k \mathbf{1}\{k \in \vec{\mathcal{O}}(w, v)\}, \quad (2.117)$$

since we payout  $q^k$  for every  $v$  that occurs, plus possibly some extra prefixes of  $v$  that we see as suffixes of  $w$ , but excluding the case where  $v$  is itself a suffix of  $w$ . Thus

$$\mathbb{E}N_v(\tau_w) = q^{-k} \left( \phi_w(q) - \vec{\phi}_{w,v}(q) + q^k \mathbf{1}\{k \in \vec{\mathcal{O}}(w, v)\} \right). \quad (2.118)$$

In particular:

**Fact 2.54.** *Let  $w \in [q]^l$  be any word. For any  $a \in [q]$ ,*

$$\mathbb{E}N_a(\tau_w) = q^{-1} \mathbb{E}\tau_w. \quad (2.119)$$

*Generally, if  $v = v_1 \cdots v_k \in [q]^k$ , then  $\mathbb{E}N_v(\tau_w)$  does not depend on  $v_k$ .*

So, no matter what the word  $w$  is, the average number of times some letter  $a$  appears is always an equal share of the hitting time. (Maybe there is a simpler proof of this fact? It would follow from the fact that  $\mathbb{E}N_a(\tau_w)$  doesn't depend on  $a$ , since  $\sum_{a \in [q]} N_a(\tau) = \tau$ . Of course, it also follows by using the OST on the martingale  $N_1(t) - q^{-1}t$ , but maybe there is an even more elementary argument.)

**Question 2.55.** *Describe the distribution of  $N_a(\tau)$ . Does it depend on  $a$  and  $w$ , or just  $\mathbb{E}\tau_w$ ?*

Since  $N_a(\tau)$  doesn't immediately help measure the asymptotic letter density  $\gamma_w$  for the corresponding SFT, maybe we should look at a normalized version of  $N_a$ , say  $\rho(t) = \frac{N_a(t)}{t}$ , and  $\rho_w = \rho(\tau_w) = \frac{N_a(\tau_w)}{\tau_w}$ .

**Question 2.56.** *Can we actually compute  $\mathbb{E}\rho_w$ ? Is it rational?*

Here is a striking finding: in  $10^7$  empirical trials, I got  $\mathbb{E}\rho_w < 1/2$  for  $w = 0110$  and  $> 1/2$  for  $w = 1001$ , the **reverse** of how the letter densities go! Balazs M. came up with a proof that this is a general phenomenon:

**Theorem 2.57.** *Fix  $q = 2$ . Assume that  $\gamma_{w,n} < \frac{1}{2}$  for all  $n$  and that  $\gamma_w < \frac{1}{2}$ , where  $\gamma_{w,n}$  is the letter density of 1s over words of length  $n$ , and  $\gamma_w$  is the limiting density of 1s. Then  $\mathbb{E}\rho_w \geq \frac{1}{2}$ .*

The idea behind this is the following. In measuring  $\gamma_w$ , we only look at words  $w$  in  $\Omega_n(w)$ , whereas in the density of 1s up to the hitting time, it's like measuring the density of 1s when a  $w$  is guaranteed to appear, which is more like living in the complement of  $\Omega_n(w)$ .

**(Specialized betting)** Here is a variant of the martingale that works, but seems annoying to work out. The idea is to only allow bettors to arrive and play the game *after a copy of some word  $v$  has occurred in the  $X$  sequence*. In other words, a bettor arrives and bets on sequence  $w$  at time  $s$  if and only if  $(X_{s-k}, X_{s-k+1}, \dots, X_{s-1}) = v$ . As usual, since the bets are always made on fresh randomness (and are mean zero), this yields a martingale  $Q_t$ . We obtain

$$Q_\tau = N_1(\tau) - 1\{(X_{\tau-k+1}, \dots, X_\tau) = v\} - \sum_{j \in [k]} 1\{(X_{\tau-j+1}, \dots, X_\tau) = v_1, \dots, v_j\}. \quad (2.120)$$

Indeed, by time  $\tau$  exactly  $N_1(\tau)$  bettors have entered the game, except we may have overcounted the case where  $v$  has just occurred at time  $\tau$ , in which case we see an extra  $v$  but no extra bettor.

**Question 2.58.** *Can this formula be decomposed in a nice way to involve some correlation polynomials, and 'backwards' chain probabilities that we can actually compute?*

**(Multiple words)** Here is a first attempt at generalizing the proof of Theorem 2.5 to the setting where the underlying shift of finite type is a different shift space, namely when we first forbid a fixed word  $v$ , then further forbid some word  $w$  (see Question 2.68 for a conjecture along this line). We hope to recover a generating function formula like 2.110, and thus describe the entropy of the further subshift where  $w$  and  $v$  are both forbidden as a root of some explicit polynomial, which will involve the correlation polynomials. To achieve this, we fix a time  $t \geq 1$ , and run the betting scheme where a single bettor arrives at time  $t$ , and bets on  $v$ , and we stop at time  $\tau = \tau_v \wedge \tau_w$ . Assuming that neither  $v$  nor  $w$  is a subword of the other, so that  $\tau_v = \tau_w$  is impossible, I obtain

$$\mathbb{P}(\tau \geq t) = \sum_{j \in \mathcal{O}(v)} q^j \mathbb{P}(\tau = \tau_v = t + j - 1) + \sum_{j \in \mathcal{O}(w,v)} q^j \mathbb{P}(\tau = \tau_w = t + j - 1). \quad (2.121)$$

The same equality holds when  $v$  and  $w$  are reversed, by symmetry. We want to solve for the generating functions  $a(z) = \sum_{t \geq 1} \mathbb{P}(\tau = \tau_v = t)z^t$  and  $b(z) = \sum_{t \geq 1} \mathbb{P}(\tau = \tau_w = t)z^t$ , which together are

$$\sum_{t \geq 1} \mathbb{P}(\tau \geq t) z^t = \sum_{t \geq 1} \sum_{s \geq t} \mathbb{P}(\tau = s) z^t \quad (2.122)$$

$$= \sum_{s \geq 1} \sum_{t=1}^s \mathbb{P}(\tau = s) z^t \quad (2.123)$$

$$= \sum_{s \geq 1} \mathbb{P}(\tau = s) \frac{z^{s+1} - z}{z - 1} \quad (2.124)$$

$$= za(z) + zb(z) - \frac{z}{1-z}, \quad (2.125)$$

where we used the fact that  $\mathbb{P}(\tau = s) = \mathbb{P}(\tau = \tau^v = s) + \mathbb{P}(\tau = \tau^w = s)$ . Note that you need to be a bit careful when evaluating at  $z = 1$ , which is not actually a pole. Combining these equations together, one can solve for  $a$  and  $b$ , obtaining

$$a(z) = \frac{1}{1-z} \frac{\phi_w(x) - \vec{\phi}_{w,v}(x)}{\phi_v(x) + \phi_w(x) - \phi_v(x)\phi_w(x) - \vec{\phi}_{v,w}(x) - \vec{\phi}_{w,v}(x) + \vec{\phi}_{v,w}(x)\vec{\phi}_{w,v}(x)}, \quad (2.126)$$

And similarly for  $b$ . (This should be rechecked.) It might be possible to analyze this the same way as in our proof of Theorem 2.5. The formula looks a bit ugly, but for fixed  $v = 11$  say, i.e. when the ambient shift is the golden mean (GM) shift and  $w$  is any allowable word in the GM shift, it might be possible to make the same proof work.

## 2.8.2 Markov chain

Suppose we have a markov process  $X_1, X_2, \dots$  with the natural filtration  $\mathcal{F}_t = \sigma[(X_s)_{s \leq t}]$ , and  $X_t$  taking values in our (finite) alphabet  $[q]$ . We're thinking of the  $X_i$  as digits generated from the MME (measure of maximal entropy) of some SFT (shift of finite type), and although this construction works perfectly well for arbitrary markov processes, to get nicer formulas we further assume that  $X$  is a markov chain, i.e. has memory at most 1. For the case of markov chains arising from measures of maximal entropy for SFTs, this comes at no cost, since we can always lift an arbitrary SFT to a higher block representation where it has memory 1. (Also, this lift is completely explicit: if  $X$  is an SFT with forbidden set  $\mathcal{F}$  having words of length  $\leq l$ , then the MME on  $X$  will have memory  $l$ ; and we can build a conjugate, i.e. isomorphic, shift  $Y$  on the alphabet  $[q]^l$  with memory 1, namely: all transitions  $(x_1, \dots, x_l) \rightarrow (x_2, \dots, x_l, y)$  are allowed for any  $y \in [q]$  such that no element of  $\mathcal{F}$  occurs as a subword of the latter.)

Fix a word  $w$  of length  $l$  which is allowed in the language of  $X$ , and consider the stopping time  $\tau_w = \min\{t : (X_{t-l+1}, X_{t-l+2}, \dots, X_t) = w\}$ . In this section we derive a general formula for  $\mathbb{E}[\tau_w]$ , which reduces to a relatively simple formula closely related to the autocorrelation polynomial of the word  $w$ .

The plan is to mimic the ideas from the iid case, i.e. build a betting game where we bet on occurrences of the string  $w$ , and apply the optional stopping theorem at time  $\tau$ . For any  $t \geq 0$ , any finite string  $x \in [q]^t$ , and any  $i \in \{1, 2, \dots, k\}$ , say the triple  $(i, t, x)$  is *streaking* if the last  $i$  letters of  $x$  are the first  $i$  letters of  $w$  (any triple with  $i = 0$  is streaking), and define

$$Q(i, t, x) = \mathbb{P}(X_{t+1} = w_{i+1} | (X_1, \dots, X_t) = x), \quad (2.127)$$

and if  $Q(i, t, x) \in (0, 1)$  let

$$G(i, t, x) = \begin{cases} \prod_{j=0}^{i-1} Q(j, t-i+j, (x_1, \dots, x_{t-i+j}))^{-1}, & (i, t, x) \text{ streaking} \\ 0, & (i, t, x) \text{ not streaking} \end{cases} \quad (2.128)$$

(else if  $Q(i, t, x) \in \{0, 1\}$  then  $G(i, t, x) = G(i-1, t-1, (x_1, \dots, x_{t-1}))$ ). Also denote by  $\text{Bet}(i, t, x)$  the probability distribution taking value  $G(i, t, x)Q(i, t, x)^{-1}(1-Q(i, t, x))$  with probability  $Q(i, t, x)$  and value  $-G(i, t, x)$  with the complementary probability if  $Q(i, t, x) \notin \{0, 1\}$  and  $(i, t, x)$  is streaking, and  $\text{Bet} = 0$  deterministically otherwise. To explain the terms, imagine that a bettor arrives at each time  $t \geq 0$  and bets one dollar on the digits of  $w$  occurring in  $X$  in order, started from digit  $t$ , and reinvests all her winnings on the next digit if she wins, or goes home if she ever loses. Then a triple  $(i, t, X)$  is streaking if the bettor who arrived at time  $t-i+1$  bets on the correct digit at least  $i$  times,  $Q(i, t, X)$  is her probability of betting correctly on the  $i+1$ st digit, and  $G(i, t, X)$  is her total fortune (including the initial 1 dollar investment) up to the  $i$ th digit, and  $\text{Bet}(i, t, X)$  is her gamble on the  $i+1$ st digit. The bet amounts are arranged so that each bet is fair (mean zero): if some amount  $g$  is bet on a digit that has probability  $q$  to occur, then our net gain is  $gq^{-1} - g$  if we win (with probability  $q$ ) and  $-g$  if we lose (probability  $1-q$ ), which has expected value

$$q \cdot (q^{-1}g - g) + (1-q) \cdot (-g) = 0. \quad (2.129)$$

We use all these bets to define a martingale  $W_t$  which is the total *net* profits of the casino up to time  $t$ , given by

$$W_t = t - \sum_{i=1}^l G(i, t, (X_s)_{s \leq t}). \quad (2.130)$$

Observe that conditionally on  $(X_s)_{s \leq t} = x$ , the increment  $\Delta W_t = W_{t+1} - W_t$  is equal in distribution to

$$\sum_{i=0}^{l-1} \text{Bet}(i, t, x), \quad (2.131)$$

which has expectation zero (by linearity of expectation). It follows that  $\mathbb{E}[W_{t+1} | \mathcal{F}_t] = W_t$ , i.e.  $W$  is indeed a martingale with  $\mathbb{E}W_t = 0$  for all  $t \geq 0$ . If the underlying markov process is irreducible (and  $w$  has positive probability to occur), then  $\tau_w$  is sub-geometrically distributed, i.e.

$$\mathbb{P}(T_w > t) \leq \exp(-ct) \quad (2.132)$$

for some  $c > 0$ , and thus  $\tau$  and  $W_t$  satisfy the conditions of the optional stopping theorem (OST). (This is an easy general fact about irreducible markov chains: hitting times are always sub-geometric. So since  $X$  is markov in some higher block representation, we also get exponential decay in the lower block representation, with constant  $c$  scaled by the ratio between the block lengths.) Applying the OST at time  $\tau$ , so all bets on digits through  $X_\tau = w_l$  have been settled, yields

$$\mathbb{E}W_{\tau_w} = \mathbb{E}\tau - \sum_{i=1}^l \mathbb{E}[G(i, \tau, (X_s)_{s \leq \tau})] = \mathbb{E}W_0 = 0. \quad (2.133)$$

Looking back at the definition of  $G$ , the  $i$ th term in this sum is nonzero exactly when the last  $i$  digits of  $w$  are equal to the first  $i$  digits of  $w$ , since at time  $\tau$ , the last  $l$  digits of  $X$  are  $w$ . Thus we can write

$$\mathbb{E}\tau = \sum_{i \in \mathcal{O}(w)} \mathbb{E}[G(i, \tau, (X_s)_{s \leq \tau})], \quad (2.134)$$

where  $\mathcal{O}(w)$  is the usual overlap set of  $w$  (i.e. the set of  $i$  such that the first  $i$  digits of  $w$  match the last  $i$  digits of  $w$ .) When  $X$  is markov with memory 1, the values  $Q(i, t, x)$  appearing in the product  $G(i, t, x)$  depend only on the previous digit of  $x$ : denote these probabilities by  $\mathbb{P}(a \rightarrow b)$  for  $a, b \in [q]$ . For  $i < l$  and  $t = \tau$ , the digit that the  $i$ -streaking bettor saw when they arrived is deterministically  $w_{l-i}$ , so we get

$$G(i, \tau, (X_s)_{s \leq \tau}) = \left( \prod_{j=1}^i \mathbb{P}(w_{l-j} \rightarrow w_{l-j+1}) \right)^{-1} \quad (\text{for } i \in \mathcal{O}(w) \setminus \{l\})$$

(where we used the fact that  $i \in \mathcal{O}(w)$  to get  $w_i = w_l$ ), plus the one special bettor who won the (random!) jackpot:

$$G(l, \tau, (X_s)_{s \leq \tau}) = \mathbb{P}(X_{\tau-l} \rightarrow w_1)^{-1} \times \left( \prod_{j=1}^l \mathbb{P}(w_{l-j} \rightarrow w_{l-j+1}) \right)^{-1} \quad (2.135)$$

Note that the first term in this product is not deterministic – it depends on the digit  $X_{\tau-l}$ . When  $X$  is the MME markov chain of a 1-step SFT, these formulas become

$$G(i, \tau, (X_s)_{s \leq \tau}) = \lambda^i r(w_{l-i}) r(w_l)^{-1} \quad (2.136)$$

and

$$G(l, \tau, (X_s)_{s \leq \tau}) = \lambda^l r(X_{\tau-l}) r(w_l)^{-1}, \quad (2.137)$$

where  $\lambda$  is the entropy of the MME and  $\ell$  and  $r$  are the left and right eigenvectors (with eigenvalue  $\lambda$ ) of the edge shift graph on  $[q]$  representing the SFT  $X$ , scaled so that  $\ell$  is a probability vector and  $\ell^T r = 1$ . (This comes from a general formula for the measure of maximal entropy for a SFT of memory 1, sometimes called the ‘Parry measure:’ it is given by the matrix

$$\nu(a, b) = \frac{r(b)}{\lambda r(a)}, \quad \text{for } a, b \in [q], \quad (2.138)$$

which implies

$$\nu(a, x_1, x_2, \dots, x_n, b) = \lambda^{-n} \frac{r(b)}{r(a)}. \quad (2.139)$$

See section 2.5 for some explicit examples.) Putting this together, we get:

**Theorem 2.59.** *Let  $X$  be a markov chain that realizes a measure of maximal entropy for a 1-step shift of finite type, and let  $w$  be a finite word (of length  $l$ ) in the language of  $X$ . Then the hitting time  $\tau_w$  of the word  $w$  in  $X$  satisfies*

$$\mathbb{E}[\tau_w] = r(w_l)^{-1} \left( \lambda^l \mathbb{E}[r(X_{\tau-l})] + \sum_{i \in \mathcal{O}(w) \setminus l} \lambda^i r(w_{l-i}) \right) \quad (2.140)$$

**Example 2.60.** When  $X$  is the full shift over  $[q]$ , i.e.  $X$  is iid over  $[q]$ ,  $\lambda = q$ ,  $\ell = q^{-1}\vec{1}$  and  $r = \vec{1}$ , and we get

$$\mathbb{E}[\tau_w] = \sum_{i \in \mathcal{O}(w)} q^i = \phi_w(q) \quad (2.141)$$

where  $\phi_w(q)$  is the auto-correlation polynomial of  $w$ .

More generally, whenever  $r$  is a constant vector, we don't have to deal with the pesky expectation in Theorem 2.59.

**Example 2.61.** Suppose the edge shift of  $X$  has uniform in-degree, i.e. the edge shift matrix for  $X$  is doubly stochastic (and let  $\lambda$  denote the exponential of the entropy). Then  $r = \vec{1}$  is the right eigenvector for eigenvalue  $\lambda$ , so we obtain the same formula:

$$\mathbb{E}[\tau_w] = \phi_w(\lambda). \quad (2.142)$$

An example of such a shift is with  $q = 3$ , and the forbidden words  $\mathcal{F} = \{11, 22, 33\}$ . Then the 1-block representation is a markov chain with edge shift matrix  $J_3 - I_3$ , i.e. the matrix of all 1s except for 0's on the diagonal, and we have  $\lambda = 2$  and right eigenvector  $\vec{1}$ .

**Example 2.62.** Let  $X$  be the golden mean shift, i.e. with forbidden word  $\{11\}$  over alphabet  $\{0, 1\}$ , and assume  $w_1 = 1$ . One computes directly that, for the 2 by 2 matrix representation of  $X$ , with entropy  $\log \varphi = \log \frac{1+\sqrt{5}}{2}$ ,  $\nu(0, 0) = \varphi^{-1}$ ,  $\nu(0, 1) = \varphi^{-2}$ . Thus for  $x = (x_0, x_1, \dots, x_k)$ ,

$$\mathbb{P}(X = x | X_0 = x_0) = \varphi^{-N_{00}(x)} \varphi^{-2N_{01}(x)}, \quad (2.143)$$

where  $N_{00}(x)$  and  $N_{01}(x)$  count the number of 00 or 01 subwords of  $x$ , respectively. But it's an easy exercise (by induction, for example) that for any  $x$  with initial and final digits  $x_i$  and  $x_f$ ,

$$N_{00}(x) + 2N_{01}(x) = \text{len}(x) - 1 - x_i + x_f \quad (2.144)$$

so using the assumption  $w_1 = 1$ , which implies  $X_{T-l} = 0$  deterministically, (and also  $w_{l-i} = 0$  for  $i \in \mathcal{O}(w)$ , since for such  $i$   $w_{l-i+1} = w_1 = 1$ )

$$\mathbb{E}[\tau_w] = \varphi^{w_l} \phi_w(\varphi) \quad (2.145)$$

**Question 2.63.** Compute explicitly some other small example that doesn't fit into any of the above examples. Do we still get a similar formula, i.e. some polynomial in  $\lambda$  times  $\phi_w(\lambda)$ ?

**Question 2.64.** Can we generalize the martingale construction in this case, just like in the iid case? It seems like some tricky terms arise that aren't as easy to deal with.



## 2.9 Auto/cross correlations

This section is devoted to a purely combinatorial problem, which has implications for the general question: when are two SFT's with a single forbidden word conjugate? Fix a word  $w \in [q]^n$ , or more generally a pattern  $w$  on some subset of  $\mathbb{Z}^d$ , and consider the set

$$U_w = \{v \in [q]^n : \mathcal{O}(v, w) = \emptyset\}, \quad (2.146)$$

where  $\mathcal{O}(v, w)$  is all the overlaps, see Proposition 2.42. What is the size of  $U_w$ ? Simulations suggest  $|U_w| \sim b_w q^n$  for large  $n$ . We can prove:

**Proposition 2.65.** *If  $w$  is not one of the four ‘reducible’ words (see 2.11), then  $|U_w| > 0$ .*

Chengyu wrote up a constructive proof. This is trivial for  $q > 2$ , but when  $q = 2$  there is something to do. Indeed, for  $k \in [n]$  let  $A_k(w)$  be the set of words in  $[q]^n$  that overlap with the first  $k$  digits of  $w$ , i.e.

$$A_k(w) = \{v \in [q]^n : v_{n-k+1} \cdots v_n = w_1 \cdots w_k\}, \quad (2.147)$$

and similarly let  $B_k(w)$  denote the words that overlap  $w$  in the last  $k$  digits of  $w$ . Then since  $|A_k(w)| = |B_k(w)| = q^{n-k}$ , we have by the triangle inequality that

$$q^n - U_w = \left| \bigcup_{k=1}^n A_k(w) \cup B_k(w) \right| \leq 2 \sum_{k=1}^n q^{n-k} = \frac{2(q^n - 1)}{q - 1}, \quad (2.148)$$

which is strictly less than  $q^n$  when  $q \geq 3$ . For  $q = 2$ , using Markov's inequality (which, in the following form, is probably equivalent to the above calculation?) doesn't work: letting  $V$  be a random word,

$$\mathbb{P}(|\mathcal{O}(V, w)| \geq 1) \leq \mathbb{E}|\mathcal{O}(V, w)| = 2 - 2^{-n}, \quad (2.149)$$

which is a factor of 2 off. Somehow the ‘reducible’ words need to play a role in such a proof.

**Conjecture 2.66.** *For every  $\epsilon > 0$ , there exists  $\delta > 0$  such that for all  $n$  sufficiently large and for at least  $(1 - \epsilon)$  many  $w \in [q]^n$ ,  $|U_w| \geq \delta q^n$ .*

It would be nice if there was a ‘linear algebra’ proof, i.e. by describing  $U_w$  as the (approximate?) solution set of some linear system of equations.

**Question 2.67.** *Give a condition on a sequence  $w_n \in [q]^n$  for  $n = \mathbb{N}$  such that*

$$q^{-n}|U_{w_n}| \rightarrow c \in (0, \infty) \quad (2.150)$$

Some examples where we have this convergence:

- $w_n = 110^{n-2}, c = 2^{-3}$
- $w_n = 110^{n-4}11, c = 2^{-4}$
- $w_n = 1^n, c = 2^{-2}$  (this is easy to see directly, and is exact for every  $n$ )
- $w_n = (10)^{n/2}, c = 2^{n/4}$ , this is exact for all  $n$  (even and odd, i.e. 10101 when  $n = 5$ )
- $w_n = (100)^{n/3}$ : the limit doesn't exist, but oscillates depending on the value of  $n \pmod 3$ : when  $n \equiv 0 \pmod 3$ ,  $|U_{w_n}| \sim 2^{-4.415}2^n$ ; when  $n \equiv 1 \pmod 3$ , i.e.  $w = (100)^k 1$ ,  $|U_{w_n}| \sim 2^{-6}2^n$ ; when  $n \equiv 2 \pmod 3$ , i.e.  $w = (100)^k 10$ ,  $|U_{w_n}| \sim 2^{-5}2^n$ .

## 2.10 Conjectures/questions

Here we collect the big motivating questions/conjectures. There are many other questions and low hanging fruit scattered throughout the writeup.

**Question 2.68.** *To what extent does the Guibas, Odlyzko result 2.5 hold for arbitrary shift spaces?*

Here is our best guess as of October 2023:

**Conjecture 2.69.** *Let  $X$  be an irreducible shift of finite type with entropy  $\lambda$ , and let  $w, w'$  be allowable words in  $X$  with the same extender sets. Let  $\lambda^w$  and  $\lambda^{w'}$  denote the entropies of the further subshifts obtained by additionally forbidding the words  $w$  or  $w'$ . Then*

$$\lambda^w \leq \lambda^{w'} \iff \phi_w(\lambda) \leq \phi_{w'}(\lambda) \iff \mathbb{E}\tau_w \leq \mathbb{E}\tau_{w'} \quad (2.151)$$

where  $\phi$  is the auto-correlation polynomial,  $\tau_w$  is the hitting time of  $w$ , and  $\mathbb{E}$  denotes expectation with respect to the measure of maximal entropy markov chain on  $X$ .

This is verified by computer simulations of some shift spaces  $X$  with one or two forbidden words. The ‘extender set’ condition is a combinatorial condition that we guessed after looking at some examples. The closest we can get to proving this or something like it in general is the generalized martingale argument in section 2.8: we could try to mimic the proof sketch 2.1, if we can compute the corresponding generating function for a general shift space.

**Question 2.70.** *Can we massage the formula in 2.59 so the autocorrelation polynomial pops out? Does the expected hitting time always have a simple polynomial formula involving the entropy, as in example 2.62?*

The martingale construction in section 2.8 is very robust, I think the results in the IID case should port to general shift spaces, but I haven’t tried much yet.

**Question 2.71.** *For the full shift over the binary alphabet, characterize the set of words  $w$  such that forbidding  $w$  causes the density of 1s to increase/decrease/equal  $1/2$ . Give a simple condition to determine which of two forbidden words  $w, w'$  will give larger density of 1s.*

A partial answer and more detailed questions are given at the end of section 2.4. More generally:

**Question 2.72.** *Understand the ‘word counting’ random variables  $N_w$  better (see section 2.6). Prove detailed limit theorems for the joint distribution  $N_w, N_{w'}$ , e.g. joint CLT should be easy, but we want more information about the error terms. Can the Gibbs measure calculations in section 2.7 be made to make sense in the limit  $n \rightarrow \infty$ ?*

It seems this kind of thing might be answerable by LDP theory of Varadhan. Question 2.50 is particularly tantalizing.

**Question 2.73.** *For any shift space  $X$ , describe the allowable words  $w$  such that further forbidding  $w$  from  $X$  gives the maximal/minimal entropy loss.*

**Question 2.74.** *(Asked by Miklos Abert) Suppose you pick a forbidden set  $\mathcal{F}$  randomly, say by choosing each word  $w \in \{0, 1\}^n$  to go in  $\mathcal{F}$  with probability  $p_n$  all independently. What rate should  $p_n$  grow at so that  $X^{\mathcal{F}}$  has zero entropy? Perhaps  $p_n = \Theta(1)$  is necessary: is there a phase transition as  $n \rightarrow \infty$  at some critical value  $p_c$ ?*

**Question 2.75.** *Carry out the same analysis with topological pressure, i.e. for underlying measure iid  $\text{Ber}(p)$  for arbitrary  $p \in (0,1)$  instead of just  $p = 1/2$ , or perhaps any distribution  $F$  on a countable alphabet (e.g. Poisson-generated letters), or one of the gibbs measures (where some words are weighted more heavily), or perhaps any general markov chain. Can we mimic the same ideas in this setting? What is the entropy/MME?*

REX project (UBC undergrads) looked Penny's game with arbitrary  $p$ . We found a few things, including: when  $p = 1/2$ , the only time a longer string beats a shorter string, i.e. probability of appearing first in an iid  $\text{Ber}(p)$  sequence is  $> 1/2$ , is when the shorter string has auto-correlations of all lengths; and in the limit  $p \rightarrow 0$ , for fixed strings  $v, w$ ,  $v$  beats  $w$  with probability in  $\{0, 1/2, 1\}$ . **(This last fact is easy to prove using the explicit formula for the probability that one word occurs before another – one can check directly that there is never any ‘cancellation’ in the Conway formula.)**

### 3 Subsequence patterns in iid sequences

Fix a ‘pattern’ of length  $k$ , i.e. a  $\sigma \in [k]^k$ , and let  $Z_n$  be iid according to some fixed, discrete distribution  $p$  on  $\mathbb{N}$ , i.e.  $\mathbb{P}(Z = j) = p_j$  for  $j = 1, 2, \dots$ . Let  $X_n^\sigma$  be the conditional measure of  $(Z_1, Z_2, \dots, Z_n)$  on avoiding  $\sigma$  as a sub-pattern, in the sense of pattern avoiding permutations (see the definition 1.3). Note that  $\sigma$  can have repeated elements. For example, if  $\sigma = 112$ , then  $(Z) = 13222$  is  $\sigma$  avoiding, but  $(Z) = 13223$  is not. What can we say about  $X$ ? As a first example, consider:

#### 3.1 $\sigma = 11$ , arbitrary distribution

This is equivalent to conditioning that  $Z_1, \dots, Z_n$  are distinct. For an arbitrary distribution  $p$ , we have the formula

$$\mathbb{P}((Z)_n \text{ is } 11\text{-avoiding}) = n! \sum_{|A|=n} \prod_{a \in A} p_a = n! E^n(p), \quad (3.1)$$

where the sum is over all subsets of  $\mathbb{N}$  of size  $a$ . This is known to combinatorialists as  $(n!$  times) the ‘elementary homogeneous symmetric polynomial,’ over the variables  $p_1, p_2, \dots$ . We can also write inclusion probabilities in this way:

$$\mathbb{P}(j \in X_n) = \frac{1}{E^n(p)} \sum_{j \in A, |A|=n} \prod_{a \in A} p_a = p_j \cdot \frac{E^{n-1}(p_{\setminus j})}{E^n(p)}, \quad (3.2)$$

where  $p_{\setminus j}$  denotes the sequence of  $p_i$ ’s, but with  $p_j$  removed. **More formulas can be obtained like this, but it’s not clear what they’re useful for.**

#### 3.2 Uniform distribution

A natural setting is to take  $Z$  to be a uniform random variable on  $[N]$  for some large integer  $N$ , and take  $n$  to be some function of  $N$ . Note that if  $N$  is much larger than  $n$ , say  $n = \log N$ , then it’s nearly identical to the situation where  $Z$  is uniform on  $(0, 1)$ , which is *exactly* the case of pattern avoiding uniformly random permutations.

So think of  $n$  as being large enough compared to  $N$  that there is a non-vanishing probability of choosing the same element twice, i.e. when  $N = O(n^2)$ . Let  $\mathcal{A}(\sigma, n, N)$  denote the set of  $\sigma$ -avoiding strings of length  $n$  over the alphabet  $[N]$ . For example,  $X_{n,N}^{11}$  is simply a uniform random subset of  $[N]$  of size  $n$ , and  $|\mathcal{A}(11, n, N)| = \binom{N}{n}$ . More interesting is  $X^{12}$ , i.e. conditioning  $Z$  to be non-increasing. These are not too hard to count:

$$|\mathcal{A}(12, n, N)| = \binom{N+n-1}{n} \quad (3.3)$$

by a typical ‘stars and bars’ count. Note that  $X_{n,N}^{12}$  can be thought of as a uniformly random element of  $\mathcal{A}(12, n, N)$ , since there is a unique order of the elements of  $X$  making it non-increasing. An interesting quantity to study here is  $M_{n,N} = \max X_{n,N}$ . Some calculations with binomials yield that:

**Lemma 3.1.** *Fix  $\lambda > 0$ . As  $N \rightarrow \infty$ , we have the distributional convergence*

$$N - M_{\lfloor \lambda n \rfloor, N} \rightarrow \text{Geo}(1 + \lambda), \quad (3.4)$$

*i.e.*

$$\mathbb{P}(M_{\lfloor \lambda n \rfloor, N} = N - s) \rightarrow \lambda(1 + \lambda)^{-1-s} \text{ for } s = 0, 1, 2, \dots \quad (3.5)$$

Thus the maximum value of  $X$  is tight to  $N$  for  $n = O(N)$ , and the distance away from  $N$  is geometrically distributed, with parameter  $1 + \lambda = 1 + n/N$ . For example, when  $n = N$ ,  $\lambda = 2$ , so the maximum is  $\text{Geo}(1/2)$  away from  $N$ .

**Question 3.2.** *Come up with a simple combinatorial explanation for this phenomenon.*

*Todo: figure out how it works for  $n = \sqrt{N}$  or  $n = N^\beta$ . There should be a similar limit theorem with some geometric/exponentially distributed distance. For example, when  $n = \sqrt{N}$ , the distance should be on order  $\sqrt{N}$ , I think – after scaling properly, what do we get?*