Pattern avoidance, entropy, and hitting time

Jacob Richey

June 28, 2024

Contents

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

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

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

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

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' σ, i.e. having no subsequence $I = (i_1, i_2, \ldots, i_k)$ with $i_1 < i_2 < \cdots < i_k$ 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 probabilitistic 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 $n \to \infty$, i.e. a measure on Ω_∞ (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, $S = \{w\}, w = w_1 w_2 \cdots w_l$ 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 λ_w denote the asymptotic growth rate of $|\Omega_n(w)|$, i.e.*

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

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 Ω_n. □

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

One can also compute combinatorially, which involves typical recurison/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

$$\begin{aligned}
\mathbb{E}\tau &= \sum_{t\geq 0} \mathbb{P}(\tau \geq t) && (2.2)\\
&= \sum_{t\geq 0} \frac{\#\text{ words of length } t \text{ avoiding } w}{2^t} && (2.3)\\
&\approx \sum_{t\geq 0} c_w \left(\frac{\lambda_w}{2}\right)^t && (2.4)\\
&= c_w \frac{1}{1-\lambda_w/2}. && (2.5)
\end{aligned}$$

Observe that this looks like it's increasing in λ (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.

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

(this is just re-wording the definition of entropy), but understanding how close τ is to being *exactly* exponentially distributed with parameter λ 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*

$$\mathcal{O} = \{i \in [l] : w_1 w_2 \cdots w_i = w_{l-i+1} w_{l-i+2} \cdots w_l\} \tag{2.7}$$

and its correlation polynomial *is the polynomial function*

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

A folklore probability result is:

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

See section 2.9 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 Odlyzdko (1980). We give an alternate proof:

Proof. For any word w, let τ_w denote the hitting time of w in an iid sequence uniform over $[q]$. From the martingale argument in Section 2.9, we will use the explicit formula 2.129 for the generating function $f(z) = \sum_{t \geq 1} z^t \mathbb{P}(\tau \geq t)$, namely

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

(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

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

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*

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

has a unique root $\alpha^*(A) \in (1, q)$. *If* $A \prec_{lex} B$ *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

$$\alpha^*(A) < \alpha^*(B) \iff A(\alpha^*(A)) < B(\alpha^*(A)). \tag{2.12}$$

This relies on the fact that α_B^* is the unique solution to (2.11) with polynomial B, and $\lim_{\alpha \to q^-} \frac{\alpha}{q-\alpha} = +\infty$, 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

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

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

$$\frac{c_k}{q - c_k} \le A(c_k), \tag{2.14}$$

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

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

where the equality uses the definition of c_k.

□

Here's a proof sketch for the uniqueness of the solution α^*, 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 $j \ge 2$, $h^j(x) = -(j-1)f^{j-1}(x) + (q-x)f^j(x)$ (where superscript denotes the jth 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 $j^* \in \{1, 2, \ldots, k\}$ such that $h^j(1) > 0$ for $i \le j^*$, 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 kth 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

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

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

In this section we study a specific graph construction for one dimensional SFTs with a single forbidden word, which allows easy comparisons between different such shift spaces.

Definition 2.7. *Fix $q \in \{2, 3, \ldots\}$ and $w \in [q]^k$, $w = w_1 w_2 \cdots w_k$. Let L_w denote the* ***follower set graph*** *with vertex set $\{\emptyset, w_1, w_1 w_2, \ldots, w_1 \cdots w_{k-1}\}$, the set of prefixes of w (which we identify with $\{0, 1, \ldots, k-1\}$ in the obvious way), and for each $i \leq k-1$ and $a \in [q]$ except for $i = k-1$ and $a = w_k$, the labeled directed edge $i \to d_i(a)$*

$$d_i(a) = \max\{j : w_1 w_2 \cdots w_j = w_{i-j+2} w_{i-j+3} \cdots w_i a\}, \tag{2.17}$$

which is assigned label a. When $q = 2$ we use the shorthand $d_i = d_i(1 - w_{i+1})$.

Infinite paths in the graph L_w are in bijection with the shift space where the word w is forbidden. The connection is the following: imagine a one-sided infinite word $x = (\ldots, x_{-2}, x_{-1}, x_0) \in [q]^{\mathbb{N}}$ that we modify one step at a time by adding a new letter on the right. The states of L correspond to the maximal frontier of x that agrees with a prefix of w, and the edges of L correspond to appending that edge's label to x to obtain $x' = (\ldots, x_{-1}, x_0, a)$ for some $a \in [q]$. Although there are infinitely many graphs that realize the same correspondence, we use this construction because it has many nice properties, including the following.

Proposition 2.8. *Fix $w \in [q]^k$ and let d_i be as in Definition 2.7. Then we have:*

a. *All states $i \leq k-2$ have out-degree q, and state $k-1$ has out-degree $q-1$.*

b. *All incoming edges to state $i > 0$ have label w_i, while incoming edges to state $i = 0$ can have any label except w_1.*

c. *$d_i(a) \leq i$ for all $i \in \{0, 1, \ldots, k-1\}$ and $a \neq w_{i+1}$, and $d_i(w_{i+1}) = i + 1$.*

d. *$i - d_i(a) = i' - d_{i'}(a')$ only if at least one of the following three conditions holds: $i = i'$; one of $d_i(a), d_{i'}(a) = 0$; or $a = w_{i+1}, a' = w_{i'+1}$.*

Proof. Parts a and b are clear from the definition. For b, it follows immediately from the definition that $d_{i+1}(a) \leq i + 1$ for all a, and if $a \neq w_{i+1}$, $d_{i+1}(a) = i + 1$ is impossible by part b. For part d, write $d_i = d_i(a)$ and $d_{i'}(a')$ for short, and suppose for the sake of contradiction that $i - d_i = i' - d_{i'}$ for some $i \neq i'$ with $i, i' \geq 1$. Then also $d_i \neq d_{i'}$ by re-arranging, so assume WLOG $d_i > d_{i'}$. If $a = w_{i+1}$, then by part c, $a' = w_{i'+1}$, and similarly with the roles of a and a' reversed, so assume $a \neq w_{i+1}$ and $a' \neq w_{i'+1}$. Unraveling the definition of d_i gives

$$w_{i-d_i+j+1} = w_j \text{ for } j = 1, 2, \ldots, d_i - 1, \text{ and } w_{i+1} \neq w_{d_i}. \tag{2.18}$$

If $d_{i'} \neq 0$, and thus also $d_i \neq 0$ by assumption, setting $j = d_{i'}$ gives

$$w_{d_{i'}} = w_{i-d_i+d_{i'}+1} = w_{i'-d_{i'}+d_{i'}+1} = w_{i'+1}, \tag{2.19}$$

contradicting the last part of Equation 2.18 for i'.

□

The properties in Proposition 2.8 are not sufficient to characterize the set of graphs that arise as L_w for some word w and integer alphabet size $q \geq 2$: we do not know such a set of sufficient conditions.

We now give a few applications of proposition 2.8, starting with a characterization of which of the L_w graphs are irreducible.

Proposition 2.9. *Except for the special cases $q = 2$ and $w \in \bigcup_{k\geq 1}\{10^{k-1}, 1^{k-1}0, 01^{k-1}, 0^{k-1}1\}$, the graph L_w is irreducible.*

Proof. We start with the easier case $q \geq 3$. Since the graph L_w always has the path $0 \to 1 \to \cdots \to k-1$, it suffices to show that for each $i > 0$, $d_i(a) < i$ for some letter $a \in [q]$. Proposition 2.8 parts a and c together imply that there is exactly one edge $i \to i+1$ and at most one edge $i \to i$, and thus at least one edge $i \to d_i(a) < i$.

So assume $q = 2$, and assume that there is no path $k-1 \to 0$ in L_w: we will show that w must be one of the exceptional words. For this to occur, there must be some $i \in \{1, 2, \ldots, k-1\}$ such that $d_j \geq i$ for all $j = i, i+1, \ldots, k-1$ – otherwise, it would be possible to escape the set $\{i, i+1, \ldots, k-1\}$ for every i, and thus to reach 0 from $k-1$. Pick the largest such i. If $i = k-1$, then $d_{k-1} = k-1$, and unraveling the definition of d_{k-1} gives $w_1 = w_2 = \cdots = w_{k-1} = \overline{w_k}$ (where $\overline{a} = 1 - a$), i.e. $w = 1^{k-1}0$ or $0^{k-1}1$.

So suppose $i < k-1$. By part d of Proposition 2.8, since $i > 0$, the values $j - d_j$ must all be distinct for $j \in \{i, i+1, \ldots, k-1\}$. So $j - d_j \in \{0, 1, \ldots, k-1-j\}$ for $j = i, i+1, \ldots, k-1$; and since there are exactly $k-i-1$ many j, each such value occurs exactly once. Since $d_j \leq j$, we must have $d_i = i$, and by induction over j, $d_j = i$ for all $j \in \{i, i+1, \ldots, k-1\}$. Unraveling the definition of d_j, we obtain that for $j \in \{i, i+1, \ldots, k-1\}$,

$$w_m = w_{j+1-i+m} \text{ for } m \in \{1, \ldots, i-1\}, \text{ and also } w_i = \overline{w_{j+1}}. \tag{2.20}$$

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

□

When $q = 2$, the graph L_w, viewed as a vertex-labeled (but not edge-labeled) graph on $\{0, 1, \ldots, k-1\}$, is enough information to determine the word w, up to permutations of the alphabet (i.e. bit flipping, for the binary alphabet):

Fact 2.10. *Let $w \in \{0,1\}^k$. Then for each $i \in \{1, \ldots, k\}$,*

$$w_i = \begin{cases} w_1, & \text{if } d_{i-1} = 0 \\ \overline{w_{d_{i-1}}}, & \text{otherwise} \end{cases}$$

Let $\mathcal{W}$ denote the family of equivalence classes of words in $\{0,1\}^k$ 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.

$$\hat{\mathcal{L}} = \{L_w : w \in \mathcal{W}\}, \tag{2.21}$$

where the graphs L_w have vertex labels $\{0, 1, \ldots, k-1\}$ corresponding to the frontier representation given by w. It turns out that even if we forget about the vertex labels, no two of the graphs $\hat{L}, \hat{L}' \in \mathcal{L}$ 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.

Proposition 2.11. *$|\mathcal{L}| = |\hat{\mathcal{L}}|$, i.e. for any $\hat{L}, \hat{L}' \in \hat{\mathcal{L}}$, L and L' are not isomorphic.*

Proof. Suppose $\varphi : \hat{L} \to \hat{L}'$ is an isomorphism of the underlying unlabeled graphs, viewed as a permutation on the labels $[k]$. Since $k \in \hat{L}'$ is the unique state with outdegree 0, we must have $\varphi(k) = k$. Assume by induction that $\varphi(i) = i$ for $i = k, k-1, \ldots, j$. By Proposition 2.8 part c, each incoming edge to state $j \in \hat{L}'$ has its other end at some state $m \geq j$, except one edge $j-1 \to j \in \hat{L}$. Since $\varphi(j-1)$ must be a state that has an edge $\varphi(j-1) \to j \in \hat{L}'$, and the other values $m \geq j$ are already taken, we must have $\varphi(j-1) = j-1$. □

The follower set graph L_w also admits a connection between entropy and hitting time, which is a strengthening of the result of GO under an additional assumption. To explain the connection, we make a short excursion into discrete probability.

Definition 2.12. *For a word $w \in [q]^k$, let τ_w denote the **hitting time** random variable of w by an iid sequence of Uniform($[q]$) random variables $(Y_i)_{i \in \mathbb{N}}$, i.e.*

$$\tau_w = \min\{t \in \mathbb{N} : Y_t Y_{t+1} \cdots Y_{t+k-1} = w\} \tag{2.22}$$

.

Also recall the stochastic dominance order for random variables:

Definition 2.13. *$A \prec_{st} B$ if $\mathbb{P}(A \geq t) \leq \mathbb{P}(B \geq t)$ for all t, or equivalently, if there is a coupling of A and B on the same probability space such that $A \leq B$.*

It follows easily from the results of GO that $h(X^w) \geq h(X^{w'})$ if and only if $\tau_w \prec_{\text{st}} \tau_{w'}$. In Propositions 2.14 and 2.16 we prove the same result under a natural assumption on the edges d_i:

Theorem 2.14. *Let $w, w' \in [q]^k$ be any words with follower set graphs $L = L_w, L' = L_{w'}$ given by edges d_i, d'_i, and assume that for all i there exists a permutation π_i of $[q]$ such that $d_i(\pi_i(a)) \leq d'_i(a)$ for all a. Then $\tau_w \succ_{st} \tau_{w'}$. Additionally, if for some i and a we have $d_i(a) < d'_i(\pi_i(a))$, then τ_w and $\tau_{w'}$ are not equal in distribution.*

For example, in the binary alphabet case $q = 2$, the assumption on the d_i is equivalent to $d_i \leq d'_i$ for all i.

Proof. Consider the markov chain on the graph $G_w = L_w \cup \{k\}$ where the (directed) edges have transition probabilities $1/2$, and the extra state k is added at the end of L_w, i.e. there is an edge $k-1 \to k$ corresponding to hitting the word w. (For completeness one can add outgoing edges from state k corresponding to appending any digit to w, so G_w is an edge shift representation for the full shift, and this markov chain is the maximum entropy such chain.) Let X, X' denote the random walk trajectories on the graphs $G_w, G_{w'}$ respectively, stopped on hitting state k, and by a slight abuse of notation, let τ and τ' denote the first hitting times of state k started from state 0 in each chain respectively.

To couple the pairs (X, τ) and (X', τ'), generate X' at random along with an iid sequence (Y_i) of Uniform($[q]$) random variables, then build X from X' and the Y_i in the following way. Given $\tau' = t'$ and $X' = (0 = x'_0, x'_1, \ldots, x'_{t'} = k)$, set $X_0 = x'_0 = 0$ and $\sigma_0 = 0$, and for $s \geq 1$ recursively define (X_s, σ_s) by

$$(X_{s+1}, \sigma_{s+1}) = \begin{cases} (X_s + 1, \sigma_s + 1) & \text{if } X_s = x'_{\sigma_s} \text{ and } x'_{\sigma_s+1} = x'_{\sigma_s} + 1 \\ (d_{X_s(\pi_{X_s}(a))}, \sigma_s + 1) & \text{if } X_s = x'_{\sigma_s} \text{ and } x'_{\sigma_s+1} = d'_{x'_{\sigma_s}}(a) \\ (d_{X_s(a)}, \sigma_s) & \text{if } X_s \neq x'_{\sigma_s} \text{ and } Y_s = a \end{cases} \tag{2.23}$$

In words, when X and X' are at the same level, they move together (after translating by the permutation π), and when they are at different levels, X moves independently while X' remains frozen. The (random) counter σ_s is the number of steps taken by X' when X has taken s steps. By induction, using the assumption $d_i(\pi_i(a)) \le d'_i(a)$ for all i, and because the only simple path in both graphs from $i \to j$ with $i < j$ is the path $i \to i+1 \to \cdots \to j-1 \to j$, for all $s \ge 0$ we have $X_s \le x'_{\sigma_s}$ and $\sigma_s \le s$. Stop the X process at time $\tau = \inf\{s : \sigma_s = t'\}$. It follows immediately from the construction that X has the distribution of simple random walk on G_w, and τ, the first hitting time by X of state k, has the distribution of τ_w. Since $\sigma_s \le s$ for all s, $\tau \le \tau'$ under this coupling. Additionally, $\tau < \tau'$ if $X_s < x'_{\sigma_s}$ for some $s < \tau$, and the latter occurs with positive probability if $d_i(\pi_i(a)) < d'_i(a)$ for some a and i. □

Theorem 2.14 says that the stochastic ordering of the hitting times τ_w can be read off from the graphs L_w. It turns out that entropies can also be compared easily in this setting.

Proposition 2.15. *Fix $w \in [q]^k$. Let $A = A_w$ denote the adjacency matrix of L_w, and write $r = r_w$ for the right eigenvector of A corresponding to the top (Perron-Frobenius) eigenvalue $\lambda = \lambda_w$ of A. Then*

a. $\det(A_w) = \pm 1$*;*

b. r_w has strictly decreasing entries. In particular, the entries of r_w decrease at least exponentially: for $i = 0, \dots, k-2$,

$$(r_w)_{i+1} \le (\lambda - q + 1)(r_w)_i. \tag{2.24}$$

Proof. For part b, we proceed by induction. Assume (2.24) holds for $i = 0, 1, \dots, j-1$. Then for any $a \ne w_{j+1}$,

$$\lambda r_j = \sum_{a \in [q]} r_{d_j(a)} \tag{2.25}$$

$$\ge r_{j+1} + (q-1) r_j. \tag{2.26}$$

Here the first line uses that r is a right eigenvector, and the second line follows from Proposition 2.8 part c, the induction hypothesis, and the fact that $\lambda < q$ (since A has 0-1 entries and column sums at most q). Re-arranging gives Equation 2.24 for $i = j$. In the base case $i = 0$, we have $d_0(a) = 0$ for all $a \ne w_1$ and $d_0(w_1) = 1$, so $r_1 = (\lambda - (q-1)) r_0$ is an equality. □

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

$$\frac{r_{i+1}}{r_i} \le \lambda - \sum_{a \in [q] \setminus \{w_{i+1}\}} (\lambda - q + 1)^{-i + d_i(a)}. \tag{2.27}$$

We can now describe the ordering of the entropies of the L_w graphs in the same context as Theorem 2.14:

Theorem 2.16. *Fix $w, w' \in [q]^k$, and let r be as in Proposition 2.15. Assume that for all $i \in \{0, 1, \dots, k-2\}$*

$$\sum_{a \in [q]} r_{d_i(a)} \le \sum_{a \in [q]} r_{d'_i(a)}, \tag{2.28}$$

and that

$$\sum_{a\in[q]\setminus\{w_{i+1}\}} r_{d_i(a)} \leq \sum_{a\in[q]\setminus\{w'_{i+1}\}} r_{d'_i(a)}. \tag{2.29}$$

Then $\lambda_w > \lambda_{w'}$.

Note that by Proposition 2.15, the assumption on w, w' in Theorem 2.14 implies the assumption in Theorem 2.16.

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 λ, λ'. For $i = 0, 1, \ldots, k-2$, by the assumption,

$$(A'r)_i = \sum_{a\in[q]} r_{d_i(a)} \leq \sum_{a\in[q]} r_{d'_i(a)} = (Ar)_i = \lambda r_i, \tag{2.30}$$

and similarly, $(A'r)_{k-1} \leq \lambda r_{k-1}$. Thus $A'r \leq \lambda r$. Take a left eigenvector ℓ' for eigenvalue λ' in A' and calculate:

$$\lambda'\ell' \cdot r = \ell' A' r \leq \lambda \ell' \cdot r. \tag{2.31}$$

(Here $\cdot$ is the dot product.) If at least one of w and w' are not among the exceptional words in Lemma 2.9, then by that lemma and the Perron Frobenius theorem, either L or L' is irreducible, and thus either ℓ' or r is strictly positive. Additionally, one can check directly that the assumptions of Theorem 2.14 never hold if both L and L' are reducible (unless $q = 2$ and $w = \overline{w'}$, but in this case the result is trivial). So $\ell' \cdot r > 0$, and thus $\lambda' \leq \lambda$. (Annoying, still to fix, want a strict inequality here) □

Unfortunately, the relation $w \prec w'$ given by $d_i(\pi_i(a)) \leq d'_i(a)$ for all i and a and some permutations π_i is too coarse to fully recover the GO theorem: one can check that there are words w, w' that are incomprable 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 $w = 1^k$, is the unique minimial element of the poset generated by $\prec$, i.e. $1^k \prec w'$ for all w' of length k (since all d_i are equal to 0 for the word 1^k.)

Question 2.17. *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:

$$a_n^0 = a_{n-1}^0 \tag{2.32}$$
$$a_n^1 = a_{n-1}^0 + a_{n-1}^1 + a_{n-1}^2 = a_{n-1} \tag{2.33}$$
$$a_n^2 = a_{n-1}^1 \tag{2.34}$$

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:

$$a = a^0 + a^1 + a^2 \tag{2.35}$$
$$= 2a_{-1} - a_{-1}^2 \tag{2.36}$$
$$= 2a_{-1} - a_{-2}^1 \tag{2.37}$$
$$= 2a_{-1} - a_{-3}. \tag{2.38}$$

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

$$a_n(100) \sim \left(1 + \frac{2}{\sqrt{5}}\right)\varphi^n, \varphi = \frac{1}{2}(1+\sqrt{5}). \tag{2.39}$$

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

$$f^0(z) = z + zf^0(z) \tag{2.40}$$
$$f^1(z) = z + z(f^0(z) + f^1(z) + f^2(z)) \tag{2.41}$$
$$f^2(z) = z + zf^1(z) \tag{2.42}$$

The solution is

$$f^0(z) = \frac{z}{1-z}, f^1(z) = \frac{z}{1-2z+z^3}, f^2(z) = \frac{z^2}{1-2z+z^3}. \tag{2.43}$$

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

$$a_n^1 \sim \left(\frac{3+\sqrt{5}}{2\sqrt{5}}\right)\varphi^n, a_n^2 \sim a_n^1\varphi^{-1}. \tag{2.44}$$

The proportions of paths that end at $0, 1, 2$, i.e. $\lim_{n\to\infty} \frac{a_n^j}{a_n}$, 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,

$$\mathbb{P}(\text{the last digit of } X_n = 1) = \frac{1}{a_n(w)} \sum_{k=1}^{l-1} a_n^k(w) 1\{\text{the } k^{th} \text{ digit of } w = 1\}. \tag{2.45}$$

In the example with $w = 100$, we computed $\frac{a_n^1(100)}{a_n(100)} \to \varphi - 1$, 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

$$\lim_{n\to\infty} \mathbb{P}(\text{the } j\text{th digit of } X_n = 1) \tag{2.46}$$

for any fixed $j \in \mathbb{N}$, by enumerating paths in the markov graph L_w 'backwards.' These values should converge, as $j \to \infty$, to the average density of 1's in X_n, γ_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 2.18. *For a fixed word w, let γ_w denote the limiting fraction of bits that are 1 over all strings in $\Omega_n(w)$:*

$$\gamma_w = \lim_{n\to\infty} \frac{1}{n|\Omega_n(w)|} \sum_{\omega\in\Omega_n(w)} \#1\text{'s in } \omega. \tag{2.47}$$

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

$$a_{n,k} = a_{n-1,k} + a_{n-1,k-1} - a_{n-3,k-1}. \tag{2.48}$$

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 2.20. 100 is a selfless string.)

Standard generating function technology yields

$$f(z,w) = \sum_{n,k\geq 0} a_{n,k} z^n w^k = \frac{1}{1 - z(1+w) + z^3 w}, \tag{2.49}$$

and by extracting coefficients and taking limits, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X_n(i) = 1) = \frac{1}{na_n(100)} [z^n] \frac{\partial}{\partial w}\Big|_{w=1} f(z,w) \to \frac{5+\sqrt{5}}{10} \approx .7236. \tag{2.50}$$

(As expected, the density of 1's increases as a result of conditioning on avoiding 100.) (Another aside: Mathematica is a bit tempermental 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:

$$\text{Var}(\text{number of 1's in } X_n) \sim \frac{1}{5\sqrt{5}} n. \tag{2.51}$$

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 straighforward. In fact, a large class of words w share common recurrences.

Definition 2.19. *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 $w_1 w_2 \cdots w_j = w_{l-j+1} w_{l-j+2} \cdots w_{l-1} w_l$, 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 2.20. *Let w be a selfless word of length l containing exactly j 1's. Then*

$$a_{n,k}(w) = a_{n-1,k}(w) + a_{n-1,k-1}(w) - a_{n-l,k-j}(w). \tag{2.52}$$

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 2.21. *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 2.22. *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, $\lambda_w = g_{w'}$, and if w and w' have the same number of 1's, then $\gamma_w = \gamma_{w'}$. The common recursion is*

$$a_n(w) = 2a_{n-1}(w) - a_{n-l}(w), \tag{2.53}$$

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

Note that, in contarast 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 2.20 yields the generating function

$$\sum_{n,k \geq 0} a_{n,k}(w) z^n w^k = \frac{1}{1 - z(1+w) + z^l w^j}, \tag{2.54}$$

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

$$\psi_w(t) = \sum_{i \in \mathcal{O}(w)} (\#\{\textit{1's in the length } i \textit{ self-overlap of } w\})t^i, \tag{2.55}$$

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

$$\psi_w(2) \leq \psi_{w'}(2) \iff \gamma_w \geq \gamma_{w'}, \tag{2.56}$$

but maybe something like it is true.

Definition 2.24. *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 2.25. *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.*

$$\sum_{k=0}^{n} k a_{n,k}(w) = \frac{1}{2} n a_n(w). \tag{2.57}$$

Proof. It would be nice to have a bijective proof. The above can be checked directly using the generating function formula 2.54. wetting

$$f(z,w) = \sum_{n,k \geq 0} a_{n,k}(w) z^n w^k = \frac{1}{1 - z(1+w) + z^l w^{l/2}}, \tag{2.58}$$

and a quick computation shows

$$\left.\frac{\partial}{\partial w}\right|_{w=1} f = \frac{1}{2} z g'(z) \tag{2.59}$$

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 $\mathrm{Rev}(w)$ 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. $\overline{\mathrm{Rev}(w)} = \mathrm{Rev}(\overline{w})$ and $\overline{\overline{w}} = \mathrm{Rev}(\mathrm{Rev}(w)) = w$.

Definition 2.26. *Call a word w **sweet** if $\overline{w} = Rev(w)$.*

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 2.27. *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.*

$$\sum_{k=0}^{n} k a_{n,k}(w) = \frac{1}{2} n a_n(w). \tag{2.60}$$

Proof. It suffices to find a bijection $\omega \mapsto \omega'$ from $\Omega_n(w)$ to itself such that the number of 0's in ω' is equal to the number of 1's in ω. 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 $\omega \mapsto \mathrm{Rev} \circ \overline{\omega}$. 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. □

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 ω of length $l/2$, then forming the word $\omega \oplus \mathrm{Rev}(\overline{\omega})$, 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 2.28. *The limiting density of 1's in a uniform random 1001 avoiding string is*

$$\gamma_{1001} = \frac{2\left(-3+2\sqrt{5}\right)^{5/2}\sqrt{\frac{1}{55}\left(3+2\sqrt{5}\right)}\left(110\sqrt{-3+2\sqrt{5}}+44\sqrt{5\left(-3+2\sqrt{5}\right)}+\sqrt{11}\left(35+17\sqrt{5}\right)\right)}{11\left(-35+27\sqrt{5}\right)\left(\sqrt{11}+3\sqrt{-3+2\sqrt{5}}\right)} \tag{2.61}$$

$$\approx .494161. \tag{2.62}$$

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)

$$a_n = 2a_{n-1} - a_{n-3} + a_{n-4} \sim \frac{(27\sqrt{5}-35)(\sqrt{11}+3\sqrt{2\sqrt{5}-3})}{20(2\sqrt{5}-3)^{5/2}\sqrt{3+2\sqrt{5}}}\frac{1}{2^n}(1+\sqrt{3+2\sqrt{5}})^n, \tag{2.63}$$

and

$$a_{n,k} = a_{n-1,k} + a_{n-1,k-1} - a_{n-3,k-1} + a_{n-4,k-1}, \tag{2.64}$$

which satisfies

$$\sum_{n,k\geq 0} a_{n,k} z^n w^k = \frac{1+z^3 w}{1-z(1+w)+z^3 w - z^4 w}. \tag{2.65}$$

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

Conjecture 2.29. *The density of 1's γ_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 2.30. *$\gamma_w = 1/2$ only if w is balanced.*

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

Conjecture 2.31. *$\gamma_w > 1/2$ if and only if w has at least as many 1's as 0's.*

Question 2.32. *Classify the set of balanced strings w with $\gamma_w < 1/2$.*

Conjecture 2.33. *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 ν 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 λ_w and the letter density γ_w. Namely:

Fact 2.34. *λ_w is the (exponential of the) topological entropy of ν, and γ_w is $\nu(C_0)$, the measure of the cylinder set of 0 under ν.*

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

Proposition 2.35. *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 2.36. *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 2.37. *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

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

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 2.38. *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

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

Thus the parry measure is

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

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 λ.

Question 2.39. *Write a computer program that finds an expression for the Perron eigenvectors as polynomials in the entropy λ.*

We continue with some general computations:

Example 2.40. *Consider $w = 111\cdots1$, a string of l 1s. The graph $L_{11\cdots1}$ has adjacency matrix*

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

The characteristic polynomial is $\lambda^n - \lambda^{n-1} - \cdots - \lambda - 1$, 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 $\gamma_{11\cdots1} = 1 - \nu(0) = \lambda^n(\lambda-1)^2[\lambda^{n+1} - \lambda(n+1) + n]^{-1}$.

2.6 Autocorrelation and spectrum

There is a somewhat explicit way to go between the autocorrelation polynomial ϕ_w and the spectrum of the graph L_w: namely, letting A denote the adjacency matrix of L_w, by a result of Doug Lind,

$$\det(A - tI) = (1 - qt^{-1})\phi_w(t) + 1. \tag{2.66}$$

The left hand side can be expressed in the usual way, as a polynomial in the eigenvalues of A (here w is a word of length k over alphabet $[q]$):

$$\det(A - tI) = \prod_{\lambda \in \mathrm{sp}(A)} t - \lambda = \sum_j (-1)^{k-j} t^j \det(A) \sum_{|S|=j} \prod_{\lambda \in S} \lambda^{-1}. \tag{2.67}$$

Here S ranges over all j element subsets of the spectrum. On the other hand, the right hand side captures the self-overlaps of w. Let r_j denote the coefficients of ϕ_w, i.e. r_j is 1 if w has a self-overlap of length j, and 0 otherwise, so $\phi_w(t) = \sum_j r_j t^j$. Note in particular that when $t = 0$ we obtain

$$\det(A) = -q\phi_w(0) + 1 = 1 - q \cdot r_1. \tag{2.68}$$

Generally, for $j = 1, 2, \ldots, k$, taking jth derivatives of both expressions with respect to t, then setting $t = 0$ yields

$$(-1)^{k-j} \det(A) \sum_{|S|=j} \prod_{\lambda \in S} \lambda^{-1} = r_j - q r_{j+1}, \tag{2.69}$$

where we interpret $r_{k+1} = 0$. Thus for every $j = 1, 2, \ldots, k$,

$$\sum_{|S|=j} \prod_{\lambda \in S} \lambda^{-1} = (-1)^{j-k} \frac{r_j - q r_{j+1}}{1 - q r_1}. \tag{2.70}$$

Question 2.41. *Can this system be massaged to yield more information about how the eigenvalues determine the polynomial ϕ_w or vice versa?*

If w is selected uniformly at random, the autocorrelation coefficients R_j, the indicator that w has a self-overlap of length j, are not independent, but they might be approximately so in some sense. R_j is simply Bernoulli with parameter 2^{-j}. The R_j might decorrelate in some sense for long words w. Can this be made precise? What kind of limit theorems could we try to establish for the R_j's?

Question 2.42. *Let R_j denote the autocorrelation variables of a uniformly chosen binary word w of length n. Can we describe a joint limit law for the sequence $(R_j)_{j \in [n]}$? Or, even partial results in this direction, like decay of correlations, or some kind of stationarity, as long as you zoom in on R_j for $j \leq \frac{1}{4}n$ or something like this?*

2.7 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 Ber(1/2) bits, as $n \to \infty$, $\frac{1}{n}\mathbb{E}N_w \approx 2^{-l}$. We can also compute the covariance directly:

Proposition 2.43. *Let v, w be any words of lengths k and l, respectively. Then*

$$\mathrm{Cov}(N_v, N_w) = (\Phi(v,w) - 2^{-k-l}(k+l-1))n + O(1), \tag{2.71}$$

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

$$\Phi(v,w) = \sum_{x \in \mathcal{O}(v,w)} 2^{-len(x)}, \tag{2.72}$$

where $\mathcal{O}(v,w)$ is the set of all minimal words where v and w both occur exactly once and share at least one letter (and minimal means that no subword has this property).

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

$$\Phi(101, 11) = 2^{-4} + 2^{-4} = 2^{-3}. \tag{2.73}$$

Plugging in shows $\mathrm{Cov}(N_{101}, N_{11}) = O(1)$. A similar example is our favorite word $w = 1001$. We have $\Phi(1001, 1) = 2 \cdot 2^{-4}$, so $\frac{1}{n}\mathrm{Cov}(N_{1001}, N_1) \to \frac{1}{n}(2^{-3} - 2^{-5} \cdot 4)n + O(n^{-1}) = 0$ (!) This heuristically hints that γ_{1001} is close to 1/2, but fails to capture the fact that $\gamma_{1001} \neq 1/2$. This makes sense: the covariance calculation doesn't live in the SFT, which has exponentially small measure, but in the complement, which has full measure.

Corollary 2.44. $\mathrm{Cov}(N_w, N_1) = 2^{-l}(\#1s \text{ in } w - 2l)$. *In particular,* $\mathrm{Cov}(N_w, N_1) = 0$ *if and only if w is balanced.*

Question 2.45. *Are there any pairs v, w such that N_w and N_v are asymptotically independent?*

Note the special case

$$\Phi(w,w) = 2\phi_w(2^{-1}) - 2^{-l}. \tag{2.74}$$

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 ϕ is the usual self-overlap polynomial defined in 2.3. So as a corollary we get

Corollary 2.46. *For any w of length l,*

$$\frac{1}{n}\mathrm{Var}(N_w) \sim 2\phi_w(2^{-1}) - 2^{-l} - (2l-1)2^{-2l}. \tag{2.75}$$

Question 2.47. *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

$$Q = \begin{bmatrix} 1 & 1 \\ y & 0 \end{bmatrix} \tag{2.76}$$

(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

$$[y^k] \sum_{i,j} (Q^n)_{ij}. \tag{2.77}$$

In the golden mean example, Q has eigenvalues

$$\lambda_\pm = \frac{1}{2}\left(1 \pm \sqrt{1+4y}\right), \tag{2.78}$$

and one easily diagonalizes:

$$Q = ADA^{-1}, \text{ where } A = \begin{bmatrix} 1 & 1 \\ \lambda_+ - 1 & \lambda_- - 1 \end{bmatrix}, \; D = \operatorname{diag}(\lambda_+, \lambda_-). \tag{2.79}$$

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

$$[y^k] \frac{1}{\sqrt{1+4y}} \left(\lambda_+^n (1+y) + \lambda_-^n (y - \lambda_-)\right), \tag{2.80}$$

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

$$\begin{bmatrix} 1 & y \\ y & yz \end{bmatrix} \tag{2.81}$$

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?**

2.8 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 β. Generally, for any set of words $\mathcal{F}$, and for any (fixed) $n \in \mathbb{N}$ we have a measure on the set of strings of length n given by

$$\mu^n_{\mathcal{F},\beta}(x) = Z^{-1}_{w,\beta} \exp(\sum_{w \in \mathcal{F}} \beta_w N_w), \tag{2.82}$$

where $N_w(x)$ is the number of occurrences of w in $x \in \{0,1\}^n$ and $\beta \in \mathbb{R}^{\mathcal{F}}$. Write $\mu^n_{w,\beta}$ for the measure where $\mathcal{F} = \{w\}$ consists of a single word. First we note that there is a limit in n:

Theorem 2.48 (Thermodynamic limit)**.** *For any $\beta \in \mathbb{R}^{\mathcal{F}}$, there exists a probability measure $\mu_{\mathcal{F},\beta}$ on $\{0,1\}^{\mathbb{Z}}$ that is the thermodynamic limit of the μ^n. The limit is the same if we consider μ^n with or without periodic boundary. Additionally, the pressure functions $p^n(\beta) = \frac{1}{n} \log Z^n(\beta)$ converge to an analytic (?) function $p(\beta)$.*

It remains to be seen if we can include $\beta = +\infty$ in this statement – probably $\beta = -\infty$ is fine. The former corresponds to the SFT where some subset of $\mathcal{F}$ is forbidden, and the latter should be an atomic measure, where we pack words w with $\beta_w = \infty$ as tightly as possible, though this isn't as clear.

Proof. The correct approach should be via 'transfer matrix,' i.e. form the finite matrix $A(\beta)$ with entries $0, 1$ or e^{β_w} corresponding to transitions which create a copy of w. Then $p(\beta)$ is the log of the Perron Frobenius eigenvalue of $A(\beta)$, and μ is determined by the Parry chain associated to $A(\beta)$, namely for any cylinder $C(x)$ for $x \in \{0,1\}^{[0,n)\cap\mathbb{Z}}$, $\mu(C(x)) = \lambda^{-n} l(x_0) r(x_{n-1})$, where l and r are the left and right eigenvectors of $A(\beta)$ and λ is its PF eigenvalue, all of which depend on β. □

Question 2.49. *Do we get any properties of p for free, e.g. convexity?*

Let $\mathbb{E}_{w,\beta}$ denote the expectation under $\mu_{w,\beta}$. Looking at observables $N_v = 1\{x_0 x_1 \cdots x_r = v\}$ for another fixed word v 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

$$\frac{\partial}{\partial \beta} \mathbb{E}_{w,\beta}[N_1] < 0 \text{ for } \beta < 0, \tag{2.83}$$

this would imply that $\gamma_w < 1/2$ (density of 1s for the SFT forbidding w) 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 2.50. *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 2.51. *Let f be any function on $\{0,1\}^n$. Then*

$$\frac{\partial \mathbb{E}_{w,\beta}[f]}{\partial \beta} = \mathrm{Cov}_{w,\beta}(N_w, f). \tag{2.84}$$

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

$$\frac{\partial Z}{\partial \beta} = Z\mathbb{E}[N_w]. \tag{2.85}$$

Thus

$$\frac{\partial}{\partial \beta}\mathbb{E}[f] = -Z^{-2}\mathbb{E}[N_w]\sum_x f(x)\mu(x) + Z^{-1}\sum_x f(x)N_w(x)\mu(x) \tag{2.86}$$

$$= \mathbb{E}[N_w \cdot f] - \mathbb{E}[N_w]\cdot\mathbb{E}[f] \tag{2.87}$$

$$= \mathrm{Cov}(N_w, f). \tag{2.88}$$

□

In particular:

Corollary 2.52. *At $\beta = 0$, $\frac{1}{\partial\beta}\partial\mathbb{E}_{w,\beta}[N_v] = \frac{1}{\partial\beta}\partial\mathbb{E}_{v,\beta}[N_w]$.*

2.8.1 Derivatives of the pressure

Here we evaluate explicitly some derivatives of the pressure function $p(\beta)$ in the general setting of $\mu_{\mathcal{F},\beta}$. Let $w, v \in \mathcal{F}$ be fixed words, and denote by ∂_w or ∂v the derivative with respect to β_w or β_v. We conjecture the following:

Conjecture 2.53. *$\partial_v\partial_w p$ is either identically zero or never zero for $\beta_{w'} = 0, w' \neq w$, and $\beta_w < 0$.*

In words, starting from iid and tuning down w either always increases or always decreases the density of v, unless the covariance is always zero. Because of the nice form of the pressure, we can compute these derivatives exactly in terms of some matrix products. Write $\ell(\beta)$ and $r(\beta)$ for the left and right eigenvectors of the edge shift matrix $A(\beta)$ normalized so that $\ell r = 1$ (ℓ is a row and r is a column), and $\lambda(\beta)$ for its top eigenvalue. (For example, A could be the adjacency matrix of the DeBruijn graph but with transitions that create a copy of $w \in \mathcal{F}$ replaced by e^{β_w} instead of 1.) Let $U = U(\beta) = \ell^T r^T$, which satisfies $\lim_{n\to\infty}\lambda^{-n}A^n = U$ by the Perron-Frobenius theorem. Finally, for $w \in \mathcal{F}$, let E_w denote the 0-1 matrix with 1s in any entry where a w is created when that edge is traversed (so for the DeBruijn graph, E is an elementary matrix, i.e. it has a single one and the rest of its entries are zero).

Proposition 2.54. $\partial_w p(\beta) = \exp(\beta_w)\lambda(\beta)^{-1}\mathrm{Tr}(UE_wU) = \exp(\beta_w)\lambda(\beta)^{-1}\ell(\beta)A(\beta)r(\beta)$.

Proof. There are two approaches here: one can write $p(\beta) = \lim_{n\to\infty}\frac{1}{n}\log\mathrm{Tr}(A(\beta)^n)$ and use the chain rule to differentiate through the matrix trace to obtain the trace formula above. A more elegant approach is to work with $p(\beta) = \log\lambda(\beta)$. Writing $\lambda = \ell A r$ and differentiating, we get

$$\partial_w \log\lambda(\beta) = \lambda^{-1}\partial_w \ell A r \tag{2.89}$$

$$= \lambda^{-1}(\partial_w \ell A r + \ell(\partial A)r + \ell A\partial_w r) \tag{2.90}$$

$$= \lambda^{-1}(\lambda\partial_w(\ell r) + e^{\beta_w}\ell E_w r) \tag{2.91}$$

$$= \lambda^{-1}e^{\beta_w}\ell E_w r, \tag{2.92}$$

where we used the facts that $\ell A = \lambda\ell$, $Ar = \lambda r$, and $\ell r = 1$ to get rid of the first and third terms in the product rule, and also that $\partial_w A(\beta) = e^{\beta_w}E_w$ (this is basically the definition

of E_w). Finally, to see that this agrees with the trace formula $\mathrm{Tr}(UE_wU)$, use the cyclic rule for the trace, along with the fact that U is idempotent ($U^2 = \ell^T r^T \ell^T r^T = \ell^T \cdot 1 \cdot r^T = U$), so $\mathrm{Tr}(UE_wU) = \mathrm{Tr}(E_wU^2) = \mathrm{Tr}(E_wU)$, which agrees with $\ell E_w r$ by direct computation. □

So far we haven't found a nice way to do the second derivative, but using the trace description one can get a semi-explicit formula:

Proposition 2.55. $\partial_v\partial_w p = e^{\beta_w}\lambda^{-1}\left(-e^{\beta_v}\lambda^{-1}\mathrm{Tr}(E_wU) + \mathrm{Tr}(E_w\partial_vU)\right)$.

Proof. We differentiate the expression in Proposition 2.54 directly. Since $\lambda = \exp(p)$, we compute $\partial_v\lambda^{-1}$ by Proposition 2.54, and for the other term, the derivative passes through since Tr is linear and E_w is a constant matrix. □

The issue now is how to get our hands on ∂_vU, which somehow is not clear. It's a rank 1 matrix so this shouldn't be hard. Here is a possibly useful calculation:

$$\partial_vU = \lim_n \partial_v(\lambda^{-1}A) \tag{2.93}$$

$$= \lim_n \left[-n\lambda^{-n-1}\partial_vA + \lambda^{-n}\sum_{k=0}^{n-1} A^k\partial_vAA^{n-k-1}\right] \tag{2.94}$$

$$= \lambda^{-1}e^{\beta_v}\lim_n \left[-n\ell E_vrU + \sum_{k=0}^{n-1}(\lambda^{-1}A)^k\partial_vA(\lambda^{-1}A)^{n-k-1}\right] \tag{2.95}$$

$$= \lambda^{-1}e^{\beta_v}\lim_{n,m\to\infty}\left[-n\ell E_vrU + (n-2m+1)UE_vU + \sum_{k=0}^{m-1}UE_v(\lambda^{-1}A)^k + \sum_{k=0}^{m-1}(\lambda^{-1}A)^kE_vU\right] \tag{2.96}$$

$$= \lambda^{-1}e^{\beta_v}\left(UE_vY + YE_vU - UE_vU\right), \tag{2.97}$$

where $Y = \sum_{k\geq 0}(\lambda^{-k}A^k - U)$. In the above m is something $\ll n$, and we have used the fact that $\ell E_vrU = UE_vU$, which is easily checked (and is probably a consequence of the existence of this derivative, since otherwise the limit would blow up at order n).

Note that Y measures the rate of convergence of powers of A to the matrix U. The rate of this convergence is controlled by the spectral gap of A. I wonder if it is possible to get an analytical expression for Y. If one tries to follow the proof of the Perron Frobenius theorem, that $\lambda^{-k}A^k$ converges to U, it goes by putting A in Jordan form. So probably this can be carried further, decomposing by the top two eigenspaces. But what if the second eigenvalue isn't simple? Maybe this is a real obstacle.

2.8.2 Generating function

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

$$g_k(\beta) = n^{-k}\mathbb{E}_{w,\beta}[N_w^k], \tag{2.98}$$

the kth moment of $n^{-1}N_w$ under $\mu_{w,\beta}$. By Proposition 2.51,

$$\frac{\partial g_k(\beta)}{\partial\beta} = g_{k+1}(\beta) - g_k(\beta)g_1(\beta). \tag{2.99}$$

Now it is natural to form the moment generating function:

$$G(\beta, z) = \sum_{k \geq 1} g_k(\beta) z^k. \tag{2.100}$$

Playing with this a bit gives a functional relation on G:

$$\begin{aligned} \frac{\partial}{\partial \beta} G(\beta, z) &= \sum_{k \geq 1} z^k (g_{k+1}(\beta) - g_k(\beta) g_1(\beta)) && (2.101) \\ &= G(\beta, z)(z^{-1} - g_1(\beta)) - g_1(\beta) && (2.102) \\ &= z^{-1} G(\beta, z) - g_1(\beta)(1 + G(\beta, z)). && (2.103) \end{aligned}$$

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

Question 2.56. *What can be extracted from this formula? Having g_1 along with G in the equation is annoying. If we can reduce to a functional equation involving only $g_1(\beta) = m_\beta$ and $g_2(\beta)$, which would work if there was a CLT for N for every β...*

If you pretend that $g_1 = f$ is a fixed function of β, Mathematica gives the solution (to the equation $G(\beta, z) = G(\beta, z)(z^{-1} - f(\beta)) - f(\beta)$, for fixed z)

$$G(\beta, z) = \exp(z^{-1}\beta - p(\beta) + p(0)) \left[C - \int_0^\beta f(\alpha) \exp(-z^{-1}\alpha - p(\alpha) + p(0))\, d\alpha \right], \tag{2.104}$$

where $p(\beta) = \lim_{n \to \infty} \frac{1}{n} \log Z^{(n)}(\beta)$ is the pressure. Of course, f is not fixed, it's the z^1 coefficient of G. But maybe this (pretty explicit!) formula can help.

Note: if you try to work instead with $F(\beta) = \sum_{k \geq 1} \frac{1}{k!} g_k(\beta) \beta^k = \exp(p(\beta))$, using 2.98 leads to

$$\frac{\partial F}{\partial \beta} = 1 - g_1(\beta) F(\beta) + 2 \sum_{k \geq 1} \frac{1}{k!} g_{k+1}(\beta) \beta^k. \tag{2.105}$$

I'm not immediately seeing how to write the last term in terms of F, maybe this is a dead end, or I am missing an easy trick...? Here is another similar attempt to fix this: let

$$H(\beta, y) = \sum_{j \geq 0} \sum_{k \geq 1} g_{j+k}(\beta) \frac{\beta^k}{k!} y^j. \tag{2.106}$$

Then applying the same tricks yields

$$\frac{\partial H}{\partial \beta} = \frac{1}{1 - y} + 2y^{-1}(H(\beta, y) - F(\beta)) - g_1(\beta) H(\beta, y). \tag{2.107}$$

Observing that the summation in the expression above for the β derivative of F is the coefficient of y^1 in H, we get

$$\frac{\partial F}{\partial \beta} = 1 - g_1 F + 2 \partial_y H(\beta, 0). \tag{2.108}$$

Remember: our goal is to decide whether some partial derivatives of p are positive or negative (or zero), so whatever exact formulas pop out here may be useful for that, even if we can't solve these equations explicitly.

2.8.3 CLT?

Write $m_\beta = \lim_{n\to\infty} g_1(\beta)$ for short, and $\sigma_\beta^2 = \lim_{n\to\infty} \operatorname{Var}(n^{-1}N) = \lim_{n\to\infty} n^{-1}\left(\mathbb{E}[N^2] - n^2 c_\beta^2\right)$. 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 $\mu_{\beta,w}^n$ for fixed $\beta = 0$ and $n \to \infty$,

$$\frac{N - g_1 n}{\sigma\sqrt{n}} \to_d N(0,1). \tag{2.109}$$

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

$$N \approx m_\beta n + \sigma_\beta \sqrt{n} Z, \tag{2.110}$$

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

$$\begin{aligned}
\mathbb{E}[N^k] &\approx \mathbb{E}[(m_\beta n + \sigma_\beta\sqrt{n}Z)^k)] && (2.111)\\
&= m_\beta^k n^k + k m_\beta^{k-1}\sigma_\beta n^{k-1/2}\mathbb{E}[Z] + \binom{k}{2} m_\beta^{k-2}\sigma_\beta^2 n^{k-1}\mathbb{E}[Z^2] + \cdots && (2.112)\\
&= m_\beta^k n^k + \binom{k}{2} m_\beta^{k-2}\sigma^2 n^{k-1} + O(n^{k-2}). && (2.113)
\end{aligned}$$

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

$$\mathbb{E}[N^2] \approx m_\beta^2 n^2 + \sigma_\beta^2 n, \tag{2.114}$$

and

$$\mathbb{E}[N^3] \approx m_\beta^3 n^3 + 3 m_\beta \sigma_\beta^2 n^2. \tag{2.115}$$

Now we plug into Proposition 2.51:

$$\frac{\partial m_\beta}{\partial \beta} \approx \lim_{n\to\infty} \frac{1}{n}\operatorname{Cov}(N,N) = \sigma_\beta^2, \tag{2.116}$$

and

$$\begin{aligned}
\frac{\partial \sigma_\beta^2}{\partial\beta} &= \lim_{n\to\infty} \frac{\partial}{\partial\beta} n^{-1}\operatorname{Var}(N) && (2.117)\\
&= \lim_{n\to\infty} n^{-1}\left(\frac{\partial}{\partial\beta}\mathbb{E}[N^2] - \frac{\partial}{\partial\beta}(\mathbb{E}[N]^2)\right) && (2.118)\\
&= \lim_{n\to\infty} n^{-1}\left(\mathbb{E}[N^3] - \mathbb{E}[N^2]\mathbb{E}[N] - 2\mathbb{E}[N]\operatorname{Var}(N)\right) && (2.119)\\
&\approx \lim_{n\to\infty} n^{-1}(m_\beta^3 n^3 + 3m_\beta\sigma_\beta^2 n^2 - (m_\beta^2 n^2 + n\sigma_\beta^2)(m_\beta n) - 2m_\beta\sigma_\beta^2 n^2) && (2.120)\\
&= 0 \quad (!!!) && (2.121)
\end{aligned}$$

(Here we used Proposition 2.51 twice, plus the product rule, to evaluate the derivatives.)

What happened? The expression for the derivative of σ_β^2 cancelled, so everything vanished! This is all fine, but useless because 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 2.57. *Maybe the combinatorial/Markovian structure of the measure μ 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.

Question 2.58. *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 $\beta \to -\infty$ exists and agrees with γ_w, and the limit $\beta \to \infty$ 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 2.59. *According to these simulations, $\mathbb{E}_{w,\beta}[N_v]$ is always a monotone function of β 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 β: 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?)

2.9 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.

2.9.1 IID case

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

$$\tau_w = \min\{t : (X_{t-l+1}, X_{t-l+2}, \ldots, X_t) = w\}. \tag{2.122}$$

We construct a martingale with respect to the X process as follows. At each time $t = 1, 2, \ldots$, 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

$$W_t = \text{ net profit of the casino up to all bets on } X_1, \ldots, X_t. \tag{2.123}$$

Then W_t is clearly a martingale, in this case with bounded increments. We now apply the optional stopping theorem at the stopping time τ_w. At time τ_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_{τ_w}, the casino has collected gross profit τ_w, minus the winnings of all the bettors still alive at that time:

$$W_{\tau_w} = \tau_w - \sum_{j \in \mathcal{O}(w)} q^j = \tau_w - \phi_w(q). \tag{2.124}$$

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

$$0 = \mathbb{E}W_{\tau_w} = \mathbb{E}\tau_w - \phi_w(q) \implies \mathbb{E}\tau_w = \phi_w(q). \tag{2.125}$$

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

(Probabilities for τ) 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

$$S_\tau^s = 1\{\tau \geq s\} - \sum_{j \in \mathcal{O}(w)} q^j 1\{\tau = s + j - 1\}. \tag{2.126}$$

So by the OST,

$$0 = \mathbb{E}S_t^s = \mathbb{P}(\tau \geq s) - \sum_{j \in \mathcal{O}(w)} q^j \mathbb{P}(\tau = s + j - 1). \tag{2.127}$$

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

$$p_s = \sum_{j \in \mathcal{O}} q^j (p_{s+j-1} - p_{s+j}). \tag{2.128}$$

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 $s = 1, 2, \ldots, l$), obtaining

$$\sum_{s \geq 1} p_s z^s = \frac{z \phi_w(qz^{-1})}{1 - (z-1)\phi_w(qz^{-1})}. \tag{2.129}$$

This recovers a result from Guibas-Odlyzko. Using the martingale is much more elegant than futzing 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 τ.

The generating function $\mathbb{E}[z^\tau]$ can also be obtained by having a bettor put a bet z^{s-1} at each time s that the word w will occur, and following the same rules as before (when you lose you're out, and when you win you ante all your winnings). Then stopping at time τ gives

$$\mathbb{E}[1 + z + \cdots + z^{\tau - 1}] = \sum_{j \in \mathcal{O}} z^{\tau - j}, \tag{2.130}$$

which after some algebra gives

$$\mathbb{E}[z^\tau] = \sum_{s \geq 1} \mathbb{P}(\tau = s) z^s = \frac{1}{1 - (z-1)\phi_w(qz^{-1})}. \tag{2.131}$$

This is equivalent to 2.129 by writing out p_s as a sum and using Fubini.

(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 $s = 1, 2, \ldots$. 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 τ_w and compute. First a quick definition that will be useful throughout:

Definition 2.60. *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:*

$$\vec{\phi}_{w,v}(t) = \sum_{j \in \vec{\mathcal{O}}(w,v)} t^j. \tag{2.132}$$

Then we have

$$U_\tau^i = \tau - q^{l+1} 1\{X_{\tau - l} = i\} - \sum_{j \in \vec{\mathcal{O}}(w, iw)} q^j, \tag{2.133}$$

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

$$\mathbb{P}(X_{\tau - l} = i) = q^{-l-1}(\phi_w(q) - \vec{\phi}_{w,iw}(q)), \tag{2.134}$$

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

$$\phi_w(q) = \vec{\phi}_{w,vw}(q) + \sum_{i=1}^{k} q^{l+i}\mathbb{P}(Y_iY_{i-1}\cdots Y_1 = v_1v_2\ldots v_i)1\{w_1w_2\cdots w_{k-i} = v_{i+1}\cdots v_k, l-k+i \in \mathcal{O}(w)\}. \tag{2.135}$$

Note that the events involving Y appearing in the above formula include the cases where τ is small so that $Y_i = \dagger$ for some i, in which case 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 2.61. *If $l - 1 \notin \mathcal{O}(w)$, i.e. if $w \neq i^l$ for any $i \in [q]$,*

$$\mathbb{P}(Y_2 = i|Y_1 = j) = q^{-1}\frac{\phi_w(q) - \vec{\phi}_{w,jw}(q)}{\phi_w(q) - \vec{\phi}_{w,ijw}(q)} \tag{2.136}$$

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

$$\mathbb{P}(Y_3 = i|Y_2 = j) = q^{-1}\frac{q\phi_w(q) - \sum_{k\in[q]}\vec{\phi}_{w,ijkw}(q)}{q\phi_w(q) - \sum_{k\in[q]}\vec{\phi}_{w,jkw}(q)} \tag{2.137}$$

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 2.62. *Decide whether the backwards markov chain Y_s converges quantitatively as $s \to \infty$ 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 τ via this method. We illustrate this by showing a simple method to compute the expected number of copies of a word v in $(X_1, \ldots, X_\tau)$. 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 $X_1, \ldots, X_t$: 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

$$Q_\tau = \tau - q^kN_v(\tau) - \vec{\phi}_{w,v}(q) + q^k1\{k \in \vec{\mathcal{O}}(w, v)\}, \tag{2.138}$$

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

$$\mathbb{E}N_v(\tau_w) = q^{-k}\left(\phi_w(q) - \vec{\phi}_{w,v}(q) + q^k1\{k \in \vec{\mathcal{O}}(w, v)\}\right). \tag{2.139}$$

In particular:

Fact 2.63. *Let $w \in [q]^l$ be any word. For any $a \in [q]$,*

$$\mathbb{E}N_a(\tau_w) = q^{-1}\mathbb{E}\tau_w. \tag{2.140}$$

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

Annoyingly, it seems difficult to compute the moment generating function of $N_a(\tau)$. One attempt is to put at each time s a better with starting bet $z^{N_a(s-1)}$, and bet on the word w. One obtains

$$\mathbb{E}\left[\sum_{s=1}^{\tau-1} z^{N_a(s-1)}\right] = \sum_{j\in\mathcal{O}} q^j \mathbb{E}[z^{N_a(\tau-j)}]. \tag{2.141}$$

The RHS is a deterministic polynomial times the MGF of $N_a(\tau)$, but the LHS does not seem to have a simple form: for example, I don't see how to apply Wald's lemma in a useful way, despite the fact that τ is a stopping time for the $N_a(s)$ sequence and the $N_a(s)$ are built using iid randomness. (Maybe I am missing something here?)

Since $N_a(\tau)$ doesn't immediately help measure the asymptotic letter density γ_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 2.65. *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 2.66. *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 γ_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 γ_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 $(X_{s-k}, X_{s-k+1}, \ldots, X_{s-1}) = v$. As usual, since the bets are always made on fresh randomness (and are mean zero), this yields a martingale Q_t. We obtain

$$Q_\tau = N_1(\tau) - 1\{(X_{\tau-k+1}, \ldots, X_\tau) = v\} - \sum_{j\in[k]} 1\{(X_{\tau-j+1}, \ldots, X_\tau) = v_1, \ldots, v_j\}. \tag{2.142}$$

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

Question 2.67. *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 2.81 for a conjecture along this line). We hope to recover a generating function formula like 2.129, 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

$$\mathbb{P}(\tau \geq t) = \sum_{j \in \mathcal{O}(v)} q^j \mathbb{P}(\tau = \tau_v = t + j - 1) + \sum_{j \in \vec{\mathcal{O}}(w,v} q^j \mathbb{P}(\tau = \tau^w = t + j - 1). \tag{2.143}$$

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

$$\sum_{t \geq 1} \mathbb{P}(\tau \geq t) z^t = \sum_{t \geq 1} \sum_{s \geq t} \mathbb{P}(\tau = s) z^t \tag{2.144}$$

$$= \sum_{s \geq 1} \sum_{t=1}^{s} \mathbb{P}(\tau = s) z^t \tag{2.145}$$

$$= \sum_{s \geq 1} \mathbb{P}(\tau = s) \frac{z^{s+1} - z}{z - 1} \tag{2.146}$$

$$= za(z) + zb(z) - \frac{z}{1 - z}, \tag{2.147}$$

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

$$a(z) = \frac{1}{1 - z} \frac{\phi_w(x) - \vec{\phi}_{w,v}(x)}{\phi_v(x) + \phi_w(x) - \phi_v(x)\phi_w(x) - \vec{\phi}_{v,w}(x) - \vec{\phi}_{w,v}(x) + \vec{\phi}_{v,w}(x)\vec{\phi}_{w,v}(x)}, \tag{2.148}$$

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.

2.9.2 Markov chain

Suppose we have a markov process $X_1, X_2, \ldots$ 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 $(x_1, \ldots, x_l) \to (x_2, \ldots, x_l, y)$ 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 $\tau_w = \min\{t : (X_{t-l+1}, X_{t-l+2}, \ldots, X_t) = w\}$. 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 occurrances of the string w, and apply the optional stopping theorem at time τ. For any $t \geq 0$, any finite string $x \in [q]^t$, and any $i \in \{1, 2, \ldots, k\}$, 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

$$Q(i, t, x) = \mathbb{P}(X_{t+1} = w_{i+1} | (X_1, \ldots, X_t) = x), \tag{2.149}$$

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

$$G(i, t, x) = \begin{cases} \prod_{j=0}^{i-1} Q(j, t-i+j, (x_1, \ldots, x_{t-i+j}))^{-1}, & (i, t, x) \text{ streaking} \\ 0, & (i, t, x) \text{ not streaking} \end{cases} \tag{2.150}$$

(else if $Q(i, t, x) \in \{0, 1\}$ then $G(i, t, x) = G(i-1, t-1, (x_1, \ldots, x_{t-1}))$. Also denote by $\mathrm{Bet}(i, t, x)$ 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 $\mathrm{Bet} = 0$ 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 ith digit, and $\mathrm{Bet}(i, t, X)$ 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

$$q \cdot (q^{-1}g - g) + (1 - q) \cdot (-g) = 0. \tag{2.151}$$

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

$$W_t = t - \sum_{i=1}^{l} G(i, t, (X_s)_{s \leq t}). \tag{2.152}$$

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

$$\sum_{i=0}^{l-1} \mathrm{Bet}(i, t, x), \tag{2.153}$$

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 τ_w is sub-geometrically distributed, i.e.

$$\mathbb{P}(T_w > t) \leq \exp(-ct) \tag{2.154}$$

for some $c > 0$, and thus τ 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 τ, so all bets on digits through $X_\tau = w_l$ have been settled, yields

$$\mathbb{E}W_{\tau_w} = \mathbb{E}\tau - \sum_{i=1}^{l} \mathbb{E}[G(i, \tau, (X_s)_{s\leq\tau})] = \mathbb{E}W_0 = 0. \tag{2.155}$$

Looking back at the definition of G, the ith 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 τ, the last l digits of X are w. Thus we can write

$$\mathbb{E}\tau = \sum_{i\in\mathcal{O}(w)} \mathbb{E}[G(i, \tau, (X_s)_{s\leq\tau})], \tag{2.156}$$

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 $\mathbb{P}(a \to b)$ 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

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

(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:

$$G(l, \tau, (X_s)_{s\leq\tau}) = \mathbb{P}(X_{\tau-l} \to w_1)^{-1} \times \left(\prod_{j=1}^{l} \mathbb{P}(w_{l-j} \to w_{l-j+1})\right)^{-1} \tag{2.157}$$

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

$$G(i, \tau, (X_s)_{s\leq\tau}) = \lambda^i r(w_{l-i}) r(w_l)^{-1} \tag{2.158}$$

and

$$G(l, \tau, (X_s)_{s\leq\tau}) = \lambda^l r(X_{\tau-l}) r(w_l)^{-1}, \tag{2.159}$$

where λ is the entropy of the MME and ℓ and r are the left and right eigenvectors (with eigenvalue λ) of the edge shift graph on $[q]$ representing the SFT X, scaled so that ℓ 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

$$\nu(a,b) = \frac{r(b)}{\lambda r(a)}, \text{ for } a,b \in [q], \tag{2.160}$$

which implies

$$\nu(a,x_1,x_2,\ldots,x_n,b) = \lambda^{-n}\frac{r(b)}{r(a)}. \tag{2.161}$$

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

Theorem 2.68. *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 τ_w of the word w in X satisfies*

$$\mathbb{E}[\tau_w] = r(w_l)^{-1}\left(\lambda^l \mathbb{E}[r(X_{\tau-l})] + \sum_{i\in\mathcal{O}(w)\setminus l}\lambda^i r(w_{l-i})\right) \tag{2.162}$$

Example 2.69. *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*

$$\mathbb{E}[\tau_w] = \sum_{i\in\mathcal{O}(w)} q^i = \phi_w(q) \tag{2.163}$$

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 2.68.

Example 2.70. *Suppose the edge shift of X has uniform in-degree, i.e. the edge shift matrix for X is doubly stochastic (and let λ denote the exponential of the entropy). Then $r = \vec{1}$ is the right eigenvector for eigenvalue λ, so we obtain the same formula:*

$$\mathbb{E}[\tau_w] = \phi_w(\lambda). \tag{2.164}$$

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 2.71. *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 $x = (x_0, x_1, \ldots, x_k)$,*

$$\mathbb{P}(X = x | X_0 = x_0) = \varphi^{-N_{00}(x)}\varphi^{-2N_{01}(x)}, \tag{2.165}$$

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,

$$N_{00}(x) + 2N_{01}(x) = len(x) - 1 - x_i + x_f \tag{2.166}$$

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 $w_{l-i+1} = w_1 = 1$)

$$\mathbb{E}[\tau_w] = \varphi^{w_l} \phi_w(\varphi) \tag{2.167}$$

Question 2.72. *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 λ times $\phi_w(\lambda)$?*

Question 2.73. *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.*

2.10 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

$$U_w = \{v \in [q]^n : \mathcal{O}(v, w) = \emptyset\}, \tag{2.168}$$

where $\mathcal{O}(v, w)$ is all the overlaps, see Proposition 2.43. What is the size of U_w? Simulations suggest $|U_w| \sim b_w q^n$ for large n. We can prove:

Proposition 2.74. *If w is not one of the four 'reducible' words (see 2.9), 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.

$$A_k(w) = \{v \in [q]^n : v_{n-k+1} \cdots v_n = w_1 \cdots w_k\}, \tag{2.169}$$

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

$$q^n - U_w = \left| \bigcup_{k=1}^{n} A_k(w) \cup B_k(w) \right| \leq 2 \sum_{k=1}^{n} q^{n-k} = \frac{2(q^n - 1)}{q - 1}, \tag{2.170}$$

which is stricly 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,

$$\mathbb{P}(|\mathcal{O}(V, w)| \geq 1) \leq \mathbb{E}|\mathcal{O}(V, w)| = 2 - 2^{-n}, \tag{2.171}$$

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

Conjecture 2.75. *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 2.76. *Give a condition on a sequence $w_n \in [q]^n$ for $n = \mathbb{N}$ such that*

$$q^{-n}|U_{w_n}| \to c \in (0, \infty) \tag{2.172}$$

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 $n \mod 3$: when $n \equiv 0 \mod 3$, $|U_{w_n}| \sim 2^{-4.415}2^n$; when $n \equiv 1 \mod 3$, i.e. $w = (100)^k 1$, $|U_{w_n}| \sim 2^{-6}2^n$; when $n \equiv 2 \mod 3$, i.e. $w = (100)^k 10$, $|U_{w_n}| \sim 2^{-5}2^n$.

2.11 Sampling failed successfully?

Consider the following simple way to generate sequences that avoid a pattern w: generate iid uniform bits from $[q]$ one at a time, and when a w occurs on the frontier, delete all the digits of that copy of w, and continue generating one bit at a time. This algorithm generates a growing random sequence Z with no subword w. How similar is the resulting sequence to a uniform random w-avoider? It appears to be a different distribution altogether. Here is some partial, indirect evidence:

Fact 2.77. *Suppose w is balanced, i.e. w has length kq for some positive integer k, and it has exactly k copies of each letter $a \in [q]$. Let Z_t denote the random word generated this way, let N_t^a denote the number of as in Z_t, and let L_t denote the length of Z_t (so $L_t \leq t$, since the length decreases by kq whenever a copy of w appears.) Then for any letter $a \in [q]$,*

$$M_t = N_t^a - q^{-1} L_t \tag{2.173}$$

is a martingale.

Proof. When adding the next digit generates a copy of w, the next letter has chance q^{-1} to be a, and $\Delta L = 1$. When a copy of w is created, $\Delta N = -k$ and $\Delta L = -qk$. In either case, $\mathbb{E}\Delta M = 0$. □

Take $q = 2$ and $w = 1001$ as an example. We get as a corollary (by optional stopping) that if τ is any sufficiently nice stopping time for the Z_t process, say the first time when $L_t = n$ for positive integer n, then Z_τ has on average half 0s and half 1s. But for n large, a uniformly random 1001 avoiding block does not have half 1s and half 0s on average. Thus Z_τ cannot have the same distribution, even approximately asymptotically, as a uniformly random 1001-avoiding block.

Question 2.78. *Prove or disprove: for any w, $\mathrm{TV}(Z_{\tau_n}, \omega_n)$ does not converge to 0, where ω_n is sampled uniformly at random from $\Omega_n(w)$, and* TV *is total variation distance.*

Probably this is true, we just need a better statistic to distinguish between the two distributions, one that works in all cases. Maybe one can give a direct entropy gap between the two.

2.12 Disconnected patterns

Consider the following more general notion of a forbidden patterns, where we allow 'disconnected' sets of letters in the following way. For any positive integer k, call a coloring $w : [k] \to [q] \cup \{?\}$ a pattern, and say a block $x = x_1 x_2 \cdots x_r$ avoids w if there is no i such that for $j = 1, 2, \ldots, k$, either $w_j =?$ or $x_{i+j} = w_j$. In other words, a copy of pattern w in x is any subword of length k where the labels in w that are in q match those in x, but if the label in w is a ? then it doesn't need to match. This is the same as enlarging the forbidden set to be a union of q^m many words. For example, over the binary alphabet, forbidding the pattern $w = 1?1$ is equivalent to forbidding both 101 and 111.

Which results carry over to this context? The martingale formula for the expected hitting time already runs into trouble. You can define the same betting game, stopping when pattern w occurs for the first time, but to compute the winnings you need to know exactly what length k word occurred at the end, so you know which bettors had winnings. Again the self-overlaps of w are involved, but in a more complex way.

Example 2.79. *Take for example $w = 1??1$. Let $\mathcal{F} = \{1001, 1011, 1101, 1111\}$, τ the hitting time of pattern w (i.e. the minimum of the hitting times τ_f for $f \in \mathcal{F}$), and p_f the corresponding probabilities $p_f = \mathbb{P}(\tau = \tau_f)$. Using the general formula of Li (which applies the martingale construction for any finite number of forbidden words), one can compute everything explicitly:*

$$p_{1001} = \frac{25}{71}, p_{1011} = \frac{16}{71}, p_{1101} = \frac{18}{71}, p_{1111} = \frac{12}{71}, \text{ and } \mathbb{E}\tau = 542/71. \tag{2.174}$$

Let $Q = X_{\tau-2}, R = X_{\tau-1}$ be the digits of the two ?'s when pattern w first appears. The distribution of the pair (Q, R) can be computed from the p_f values: $\mathbb{P}(Q = q, R = r) = p_{1qr1}$. One finds $Q \sim Ber(\frac{28}{71})$, $R \sim Ber(\frac{30}{71})$, and $Corr(Q, R) = \sqrt{\frac{6}{61705}} \sim 9.86 \times 10^{-3}$.

Applying the martingale idea directly to pattern w, where bettors arrive, and bet in the same way but only on non-? digits, one obtains

$$W_\tau = \tau - (4 + 2 \cdot \mathbb{1}\{Q = 1\} + 2 \cdot \mathbb{1}\{R = 1\} + 2), \tag{2.175}$$

where W_τ is the total casino winnings. Taking expectations, this agrees with the values above.

In general, the formula 2.175 will contain a term corresponding to each self-overlap of the word w, where ?'s may overlap with any other letter.

Question 2.80. *Already I think it would be interesting to look at some restricted classes of these guys. How about $w_n = 1?^n1$? Is the entropy monotone in n? I guess we expect the entropy to converge to q as $n \to \infty$...? What do the hitting time distributions look like? What is the distribution of the intermediate digits at the hitting time? Maybe it has a nice limit in n!*

2.13 Conjectures/questions

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

Question 2.81. *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 2.82. *Let X be an irreducible shift of finite type with entropy λ, and let w, w' be allowable words in X with the same extender sets. Let λ^w and λ^w denote the entropies of the further subshifts obtained by additionally forbidding the words w or w'. Then*

$$\lambda^w \leq \lambda^w \iff \phi_w(\lambda) \leq \phi_w(\lambda) \iff \mathbb{E}\tau_w \leq \mathbb{E}\tau_{w'} \tag{2.176}$$

where ϕ is the auto-correlation polynomial, τ_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 2.9: we could try to mimic the proof sketch 2.1, if we can compute the corresponding generating function for a general shift space.

Question 2.83. *Can we massage the formula in 2.68 so the autocorrelation polynomial pops out? Does the expected hitting time always have a simple polynomial formula involving the entropy, as in example 2.71?*

The martingale construction in section 2.9 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 2.84. *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 2.85. *Understand the 'word counting' random variables N_w better (see section 2.7). 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 2.8 be made to make sense in the limit $n \to \infty$?*

It seems this kind of thing might be answerable by LDP theory of Varadhan. Question 2.59 is particularly tantalizing.

Question 2.86. *For any shift space X, describe the allowable words w such that further forbidding w from X gives the maximal/minimal entropy loss.*

Question 2.87. *Carry out the same analysis with topological pressure, i.e. for underlying measure iid 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 $\mathrm{Ber}(p)$ sequence is $> 1/2$, is when the shorter string has auto-correlations of all lengths; and in the limit $p \to 0$, 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 $j = 1, 2, \ldots$. Let X_n^σ be the conditional measure of $(Z_1, Z_2, \ldots, Z_n)$ on avoiding σ as a sub-pattern, in the sense of pattern avoiding permutations (see the definition 1.3). Note that σ can have repeated elements. For example, if $\sigma = 112$, then $(Z) = 13222$ is σ 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 $Z_1, \ldots, Z_n$ are distinct. For an arbitrary distribution p, we have the formula

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

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 $p_1, p_2, \ldots$. We can also write inclusion probabilities in this way:

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

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 σ-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:

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

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 $N \to \infty$, we have the distributional convergence*

$$N - M_{\lfloor \lambda n \rfloor, N} \to Geo(1 + \lambda), \tag{3.4}$$

i.e.

$$\mathbb{P}(M_{\lfloor \lambda n \rfloor, N} = N - s) \to \lambda(1+\lambda)^{-1-s} \text{ for } s = 0, 1, 2, \ldots. \tag{3.5}$$

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 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/exponeitially 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?