Pattern avoiding strings

Jacob Richey

September 14, 2022

1 Introduction

In this note we consider strings over an alphabet $\mathcal{A}$, typically $\mathcal{A} = \{0, 1\}$, (and possibly $\mathcal{A} = [r]$ for some positive integer r or $\mathcal{A} = \mathbb{N}$), conditioned on avoiding some pattern set S. This can mean a few different things. As a warm-up, we can take 'pattern' to mean 'substring,' 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

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

Here 'substring' means 'consecutive subsequence:' 11 is a substring 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 strings sampled from some measure on $\Omega_n(S)$, or if it makes sense, $\Omega(S)$, or $\Omega_\infty(S)$: a two examples to keep in mind are an iid string of fixed length conditioned to have no substring in S, or a string generated with iid bits, one bit at a time, and stopped on containing some string in S as a substring. There is a body of work related to this model, focusing on the expected hitting time of a given string (Feller has a few pages on it), and facts about the 'intransitive dice' game (see below) from Conway and others.

A more topical connection is with pattern avoiding permutations, where a finite 'pattern' permutation $\sigma \in \Sigma_k$ is chosen, and a uniform random permutation X is conditioned on 'avoiding' σ, i.e. having no subsequence $i_1 < i_2 < \cdots < i_k$ such that

$$(X(i_1), X(i_2), \ldots, X(i_k)) \text{ is order-isomorphic to } \sigma. \tag{1.3}$$

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?
2. Is there a simple probabilitistic description of the conditional law X?

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?)

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 permutations, there is a limit in permuton space.)

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.)

This model is morally in the same family as the MGIS model: when $\mathcal{A} = \{0, 1\}$, and $S = \{11\}$, i.e. we condition on having no two 1's in a row, we are choosing from the set of 'independent' configurations (but they need not be maximal).

2 Substring patterns in binary strings

Already the case of excluding single binary strings leads to interesting phenomena. Set $\mathcal{A} = \{0, 1\}$, and suppose S is a single string of length l, $S = s_1 s_2 \cdots s_l$ for some $s_i \in \mathcal{A}$. The first order of business here is to compute $|\Omega_n(S)|$.

Definition 2.1. *For a fixed string S, let λ_S denote the asymptotic growth rate of $|\Omega_n(S)|$, i.e.*

$$\lambda_S = \lim_{n\to\infty} |\Omega_n(s)|^{1/n}. \tag{2.1}$$

We have that:

Lemma 2.2. *Except in the trivial cases $S = 0, 1, 10,$ or 01, the limit in 2.1 exists and $\lambda_S \in (1, 2)$.*

Proof. λ_S is the topological entropy of the shift of finite type with forbidden word S. □

The combinatorics of $\Omega_n(S)$ involves typical recurison/generating function ideas, but with some novel elements. To compute the count, and represent the corresponding process X_n, it helps to construct a corresponding graph. We build a string one letter at a time, and keep track of how many digits at the front of the string match S.

At this point it is natural to recall a connection with intransitive dice. For any two strings S, S', we can play the following game: generate a string Z with iid digits in $\{0, 1\}$, and stop when either S or S' appears as a substring. Whichever string occurs first 'wins.' Then for any S, there exists an S' (of the same length) such that S' will win with probability at least $1/2$. This game is naturally connected to the distribution of our Ω_n and X: the probability that the game is not over by time n is $\Omega_n(\{S\} \cup \{S'\})$, and the distribution of the string at that time is that of X_n.

Generate a string Z with iid digits from $\{0, 1\}$. The corresponding graph L_S has state space $\{0, 1, \ldots, |S| = l\}$. We set $L_S(n) = k$ where k is the largest integer such that $Z_{n-k+1} Z_{n-k+2} \cdots Z_n = s_1 s_2 \cdots s_k$, i.e. the frontier of Z matches k digits of S. Each state, aside from state l, (which can, for our purposes, be thought of as a 'graveyard' state with no outgoing edges), has two outgoing edges: every edge $k \to k + 1$ is present for $0 \le k \le l - 1$. Additionally, for each k, there is an edge $k \to d_k$ for some $d_k \le k$ (take this as the definition of d_k). It turns out that these graphs are characterized with one additional property:

Proposition 2.3. *Any such graph L_S, given by edges $(d_k : k = 0, 1, 2, \ldots, l - 1)$, necessarily satisfies*

(1) $d_k \le k$

(2) $k - d_k = k' - d_{k'}$ if and only if $k = k'$ or one of $d_k, d_{k'} = 0$.

Proof. To prove (1), note that by the definition of $L_S(n)$, the last $L_S(n) - 1$ digits of Z_{n-1} match the first $L_S(n) - 1$ digits of S. This immediately implies that $L_S(n - 1) \ge L_S(n) - 1$, or $L_S(n) \le L_S(n - 1) + 1$, so $d_k \le k + 1$. Clearly $d_k \ne k + 1$ (the value of L_S only increases when we match the next digit of Z to the next digit of S), so $d_k \le k$.

For (2), suppose by contradiction that $k - d_k = k' - d_{k'}$ for some $k \ne k'$ with $k, k' \ge 1$. Then also $d_k \ne d_{k'}$, so suppose $d_k > d_{k'}$. The definition of d_k implies

$$s_{k-d_k+j+1} = s_j \text{ for } j = 1, 2, \ldots, d_k - 1, \text{ and } \overline{s_{k+1}} = s_{d_k}, \tag{2.2}$$

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

$$s_{d_{k'}} = s_{k-d_k+d_{k'}+1} = s_{k'-d_{k'}+d_{k'}+1} = s_{k'+1}, \tag{2.3}$$

contradicting the last part of 2.2 for k' (as long as neither $d_k, d_{k'} = 0$, in which case one of the equalities is trivial).

□

The above properties are not necessary, simply because the number of possible sequences of d_k values that satisfy (1) and (2) is strictly larger than 2^{l-1}, while the total number of possible graphs is at most that many (there are 2^l strings, and the bit flip operation $S \to \overline{S}$ preserves the graph, $L_S = L_{\overline{S}}$).

Question 2.4. *Find a full characterization of the sequences of d_k that occur for some string S, and prove (or disprove): if $L_S = L_{S'}$, then $S' = \overline{S}$, i.e. there are exactly 2^{l-1} distinct such graphs.*

2.1 Binary example

As an example of the usefulness of the graphs L_S, we work through the necessary computation explicitly for $S = 100$. 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:

$$a_n^0 = a_{n-1}^0 \tag{2.4}$$
$$a_n^1 = a_{n-1}^0 + a_{n-1}^1 + a_{n-1}^2 = a_{n-1} \tag{2.5}$$
$$a_n^2 = a_{n-1}^1 \tag{2.6}$$

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:

$$a = a^0 + a^1 + a^2 \tag{2.7}$$
$$= 2a_{-1} - a_{-1}^2 \tag{2.8}$$
$$= 2a_{-1} - a_{-2}^1 \tag{2.9}$$
$$= 2a_{-1} - a_{-3}. \tag{2.10}$$

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

$$a_n(100) \sim \left(1 + \frac{2}{\sqrt{5}}\right)\varphi^n, \varphi = \frac{1}{2}(1+\sqrt{5}). \tag{2.11}$$

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

$$f^0(z) = z + zf^0(z) \tag{2.12}$$
$$f^1(z) = z + z(f^0(z) + f^1(z) + f^2(z)) \tag{2.13}$$
$$f^2(z) = z + zf^1(z) \tag{2.14}$$

The solution is

$$f^0(z) = \frac{z}{1-z}, f^1(z) = \frac{z}{1-2z+z^3}, f^2(z) = \frac{z^2}{1-2z+z^3}. \tag{2.15}$$

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

$$a_n^1 \sim \left(\frac{3+\sqrt{5}}{2\sqrt{5}}\right)\varphi^n, a_n^2 \sim a_n^1\varphi^{-1}. \tag{2.16}$$

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 $\approx 0, .618, .382$.

2.2 String reversal

We begin with another natural observation about the sets $\Omega_n(S)$. Let $\mathrm{Rev}(S)$ denote the reversal of the string S, i.e. $\mathrm{Rev}(S) = s_k s_{k-1} \cdots s_2 s_1$.

Lemma 2.5. $|\Omega_n(S)| = |\Omega_n(Rev(S))|$

Proof. $\omega \in \Omega_n(S) \iff \mathrm{Rev}(\omega) \in \Omega_n(\mathrm{Rev}(S))$. □

Despite this simple fact, it isn't obvious what the relationship is between L_S and $L_{\mathrm{Rev}(S)}$.

Question 2.6. *Describe a simple mapping $L_S \to L_{Rev(S)}$.*

2.3 Letter densities

Computing the average density of 1's is not as simple as the counts $a_n(S)$. Let X_n denote a uniformly random chosen element of $\Omega_n(S)$. In the notation of 2.1,

$$\mathbb{P}(\text{the last digit of } X_n = 1) = \frac{1}{a_n(S)} \sum_{k=1}^{l-1} a_n^k(S) 1\{\text{the } k^{th} \text{ digit of } S = 1\}. \tag{2.17}$$

In the example with $S = 100$, 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 string, as we will see shortly. The method from 2.1 can likely be extended to compute

$$\lim_{n\to\infty} \mathbb{P}(\text{the } j\text{th digit of } X_n = 1) \tag{2.18}$$

for any fixed $j \in \mathbb{N}$, by enumerating paths in the markov graph L_S 'backwards.' These values should converge, as $j \to \infty$, to the average density of 1's in X_n, γ_S (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 S-avoiding string:

Definition 2.7. *For a fixed string S, let γ_S denote the limiting fraction of bits that are 1 over all strings in $\Omega_n(S)$:*

$$\gamma_S = \lim_{n\to\infty} \frac{1}{n|\Omega_n(S)|} \sum_{\omega\in\Omega_n(S)} \#1\text{'s in } \omega. \tag{2.19}$$

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

$$a_{n,k} = a_{n-1,k} + a_{n-1,k-1} - a_{n-3,k-1}. \tag{2.20}$$

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* strings below, and proposition 2.9. 100 is a selfless string.)

Standard generating function technology yields

$$f(z,w) = \sum_{n,k\geq 0} a_{n,k} z^n w^k = \frac{1}{1 - z(1+w) + z^3 w}, \tag{2.21}$$

and by extracting coefficients and taking limits, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X_n(i) = 1) = \frac{1}{na_n(100)}[z^n]\frac{\partial}{\partial w}\Big|_{w=1} f(z,w) \to \frac{5+\sqrt{5}}{10} \approx .7236. \tag{2.22}$$

(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:

$$\text{Var(number of 1's in } X_n) \sim \frac{1}{5\sqrt{5}} n. \tag{2.23}$$

A WLLN follows for the number of 1s, since the variance is $o(n^2)$.

Finding these recursions is sometimes very straighforward. In fact, a large class of strings S share common recurrences.

Definition 2.8. *Call a string S* ***selfless*** *if no prefix of S matches any suffix of S, i.e. if there exists no $j < l$ such that $s_1 s_2 \cdots s_j = s_{l-j+1} s_{l-j+2} \cdots s_{l-1} s_l$, where S has length l.*

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

Proposition 2.9. *Let S be a selfless string of length l containing exactly j 1's. Then*

$$a_{n,k}(S) = a_{n-1,k}(S) + a_{n-1,k-1}(S) - a_{n-l,k-j}(S). \tag{2.24}$$

Proof. To generate an arbitrary string in $\Omega_{n,k}(S)$, 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 S. So we need to throw away all strings of length $n-1$ ending with the first $l-1$ digits of S. To complete the proof, it suffices to note the following lemma:

Lemma 2.10. *S is selfless if and only if the map from the set of strings in $\Omega_{n-1,k}(S)$ ending in the first $l-1$ digits of S to $\Omega_{n-l,k-j}$ that chops off the last $l-1$ digits is a bijection.*

□

Since $a_n(S) = \sum_{k=0}^{n} a_{n,k}(S)$, 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 2.11. *Fix l. If S and S' are any two selfless strings of length l, then $a_{n,k}(S) = a_{n,k}(S')$ and $a_n(S) = a_n S'$ for all n and k. In particular, $\lambda_S = g_{S'}$ and $\gamma_S = \gamma_{S'}$. The common recursion is*

$$a_n(S) = 2a_{n-1}(S) - a_{n-l}(S), \tag{2.25}$$

and λ_S 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 S and S' have the same number of 1's. The only difference is in the base case $n = l$. Solving the recurrence in Proposition 2.9 yields the generating function

$$\sum_{n,k \geq 0} a_{n,k}(S) z^n w^k = \frac{1}{1 - z(1+w) + z^l w^j}, \tag{2.26}$$

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

Definition 2.12. *Call a string S **balanced** if the number of 1's in S is half the length of S.*

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

Proposition 2.13. *If S is selfless and balanced, then $\gamma_S = 1/2$. In fact, for all n, the average density of 1's in a uniform random string avoiding S is exactly 1/2, i.e.*

$$\sum_{k=0}^{n} k a_{n,k}(S) = \frac{1}{2} n a_n(S). \tag{2.27}$$

Proof. It would be nice to have a bijective proof. The above can be checked directly using the generating function formula 2.26. Setting

$$f(z,w) = \sum_{n,k \geq 0} a_{n,k}(S) z^n w^k = \frac{1}{1 - z(1+w) + z^l w^{l/2}}, \tag{2.28}$$

and

$$g(z) = \sum_{n \geq 0} a_n(S) z^n = \frac{1}{1 - 2z + z^l}, \tag{2.29}$$

a quick computation shows

$$\left. \frac{\partial}{\partial w} \right|_{w=1} f = \frac{1}{2} z g'(z) \tag{2.30}$$

which is equivalent to the claim. □

Also note: the family of selfless strings is quite large! The probability of a string being selfless is bounded away from 0 for any n (perhaps an interesting computation of its own?), so a constant proportion of strings 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 another class of strings for which the density can be easily seen to be exactly 1/2. Recall that $\mathrm{Rev}(S)$ is the reversal of S, and $\overline{s_1 s_2 \cdots s_l} = \overline{s_1}\ \overline{s_2} \cdots\ \overline{s_l}$, where $\overline{s} = 1 - s$ is the 'bit flipping' operation. Note that these two operations are commuting involutions, i.e. $\overline{\mathrm{Rev}(S)} = \mathrm{Rev}(\overline{S})$ and $\overline{\overline{S}} = \mathrm{Rev}(\mathrm{Rev}(S)) = S$.

Definition 2.14. *Call a string S **sweet** if $\overline{S} = Rev(S)$.*

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

Proposition 2.15. *If S is a sweet string, then $\gamma_S = 1/2$. In fact, for all n, the average density of 1's in a uniform random string avoiding S is exactly 1/2, i.e.*

$$\sum_{k=0}^{n} k a_{n,k}(S) = \frac{1}{2} n a_n(S). \tag{2.31}$$

Proof. It suffices to find a bijection $\omega \mapsto \omega'$ from $\Omega_n(S)$ 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 $\Omega_n(S)$ 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 S is sweet implies that it maps $\Omega_n(S)$ to itself. □

It is worth noting that the number of sweet strings grows exponentially, but still makes up a vanishing fraction of all strings. Indeed, the sweet strings of length l can be exactly enumerated by choosing an arbitrary string ω of length $l/2$, then forming the string $\omega \oplus \mathrm{Rev}(\overline{\omega})$, where $\oplus$ is concatenation. So there are exactly $2^{l/2}$ sweet strings 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 strings 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 2.16. *The limiting density of 1's in a uniform random 1001 avoiding string is*

$$\gamma_{1001} = \frac{2\left(-3+2\sqrt{5}\right)^{5/2}\sqrt{\frac{1}{55}\left(3+2\sqrt{5}\right)}\left(110\sqrt{-3+2\sqrt{5}}+44\sqrt{5\left(-3+2\sqrt{5}\right)}+\sqrt{11}\left(35+17\sqrt{5}\right)\right)}{11\left(-35+27\sqrt{5}\right)\left(\sqrt{11}+3\sqrt{-3+2\sqrt{5}}\right)} \tag{2.32}$$

$$\approx .494161. \tag{2.33}$$

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

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

$$a_n = 2a_{n-1} - a_{n-3} + a_{n-4} \sim \frac{(27\sqrt{5}-35)(\sqrt{11}+3\sqrt{2\sqrt{5}-3})}{20(2\sqrt{5}-3)^{5/2}\sqrt{3+2\sqrt{5}}}\frac{1}{2^n}(1+\sqrt{3+2\sqrt{5}})^n, \tag{2.34}$$

and

$$a_{n,k} = a_{n-1,k} + a_{n-1,k-1} - a_{n-3,k-1} + a_{n-4,k-1}, \tag{2.35}$$

which satisfies

$$\sum_{n,k\geq 0} a_{n,k} z^n w^k = \frac{1+z^3 w}{1-z(1+w)+z^3 w - z^4 w}. \tag{2.36}$$

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

2.4 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 S. This measure can be computed explicitly via some matrix computations with the graph L_S, and gives an alternate way to calculate the entropy λ_S and the letter density γ_S. Namely:

Fact 2.17. *λ_S is the (exponential of the) topological entropy of ν, and γ_S 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_S is the largest eigenvalue of any graph representation of the corresponding SFT: L_S is the 'minimal' such representation. As for γ_S, 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 2.18. *Let $S = s_1 s_2 \dots s_l$ with $s_1 = 1$. Then*

$$\gamma_S = \sum_{k<l : s_k = 1} \nu(k).$$

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

Proposition 2.19. *Let A_S denote the adjacency matrix of L_S, and let $p_S(\lambda)$ denote its characteristic polynomial, say $p_S(\lambda) = \sum_{i=0}^{l} c_i \lambda^i$. Then $a_n = |\Omega_n|$ satisfies*

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

Example 2.20. *To illustrate, we recover the example $S = 100$ via this method. The graph L_{100} has adjacency matrix*

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

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

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

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 2.21. *Let* $S = 1001$, *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 $\gamma_{1001} = \nu(1) = \frac{\lambda^2 - \lambda + 1}{-2\lambda^3 + 6\lambda^2 - 3\lambda + 3}$.

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. There should be a way to have the computer do it!

We continue with some general computations:

Example 2.22. *Consider* $S = 111 \cdots 1$, *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*

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

The density of $1s$ *is* $\gamma_{11\cdots1} = 1 - \nu(0) = \lambda^l(\lambda - 1)^2[\lambda^{l+1} - \lambda(l - 1) + l]^{-1}$.

Note the following general fact about the right eigenvector r:

Proposition 2.23. *For any string S, the right eigenvector r_S has strictly decreasing entries. In particular, for any $i = 0, \ldots, |S| - 2$,*

$$(\lambda_S - 1)(r_S)_i > (r_S)_{i+1}. \tag{2.37}$$

Note that since $\lambda_S < 2$, the factor $\lambda - 1 < 1$, so the above is strictly stronger than having decreasing entries: the entries of r decrease at least geometrically at rate $\lambda - 1$.

Proof. We proceed by induction. Since r is a right eigenvector, using $d_i \leq i$ (so $r_{d_i} \geq r_i$ by induction),

$$\lambda r_i = r_{i+1} + r_{d_i} \tag{2.38}$$
$$\geq r_{i+1} + r_i. \tag{2.39}$$

Now subtract. In the base case $i = 0$, we have $d_0 = 0$, so $r_1 = (\lambda - 1)r_0$ is an equality. □

The proof actually shows something a bit stronger: namely that

$$\frac{v_{i+1}}{v_i} \leq \lambda - (\lambda - 1)^{i - d_i}. \tag{2.40}$$

2.5 Extremal entropy

The string of length k with the largest entropy is the all 1's string (Brian's argument). We conjecture that the selfless strings give the minimal entropy:

Conjecture 2.24. *Of the strings S of length k, λ_S is minimized when S is selfless.*

We have the following basic heuristic regarding entropies. Turn the graph L into a markov chain with uniform transition probabilities – so probability $\frac{1}{2}$ on all edges, except $l - 1 \to d_{l-1}$ which has probability 1 – and let $\tau_S(i)$ be the hitting time of string S started from state i, with $\mu_S(i) = \mathbb{E}[\tau_S(i)]$, and $\mu = \mu_S = \mu_S(0)$ for short. Then

$$\mu = \sum_{t \geq 0} \mathbb{P}(\tau \geq t) \tag{2.41}$$
$$= \sum_{t \geq 0} \frac{\#\text{ paths in } L \text{ started from 0 of length } t}{2^t} \tag{2.42}$$
$$\approx \sum_{t \geq 0} r_S(0)(\lambda_S/2)^t \tag{2.43}$$
$$= r_S(0) \frac{1}{1 - \lambda_S/2}. \tag{2.44}$$

This suggests the following conjecture:

Conjecture 2.25. *$\lambda_S < \lambda_{S'}$ if and only if $\mu_S < \mu_{S'}$.*

The conjecture holds for all strings up to length 25 ish according to explicit computer computations. It is true that:

Fact 2.26. *μ_S is maximized when $S = 11\cdots1$, and minimized when S is selfless.*

Proof. The martingale argument shows that

$$\mu_S = \sum_{j \in SF(S)} 2^j, \tag{2.45}$$

where $SF(S)$ is the 'selfish' set of indices in S, where $j \in [l]$ belongs to $SF(S)$ if the first j letters of S exactly match the last j letters of S. When $S = 11\cdots1$, $SF(S) = [l]$, and when S is selfless, $SF(S) = \{l\}$. □

One approach to getting the same result for entropy is to work with some poset of all string S (of the same length), with a relation that is easy to work with, and always agrees with $\lambda_S < \lambda_{S'}$. Brian's argument, along with Proposition 2.23, gives

Proposition 2.27. *If $d_i \leq d'_i$ for all i, then $\lambda_S > \lambda_{S'}$.*

(Note also that if $d_i \leq d'_i$ for all i, then $\mu_S > \mu_{S'}$.) So to prove the conjecture, it would be enough to show:

Conjecture 2.28. *S is selfless if and only if there exists no S' such that $d'_i \geq d_i$ for all i.*

It would suffice to show the 'if' part of this statement, by 2.11.

2.6 Conjectures/questions

Some conjectures:

Question 2.29. *How common are 'coincidences' among the λ_S values? There to seem to be some for small k values. How many different λ_S values are there among the strings of length k? Is it possible for multiple strings to have the same λ_S but different exact formulas $a_n(S)$?*

See Proposition 2.11 for a partial answer.

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

Simulation found counterexamples of length 8, namely 10011010 and 10100110.

Conjecture 2.31. *$\gamma_S = 1/2$ only if S is balanced.*

This has been confirmed by (approximate) simulations up to strings S of length 20.

Conjecture 2.32. *For all l and all strings S of length l, $|\gamma_S - 1/2| \leq C\exp(-cl)$. If this holds, what is the optimal rate c? Or does the convergence go even faster?*

Guiding questions moving forward:

Question 2.33. *Describe the strings / forbidden sets that have maximum/minimum entropy.*

Question 2.34. *Describe the class of strings S or families of strings with $\gamma_S = 1/2$.*

Question 2.35. *Prove a general LLN/CLT for the total number of 1's in X_n as $n \to \infty$, with mean γ_S.*

Question 2.36. *Carry out the same analysis with topological pressure, i.e. for underlying measure Ber(p) for arbitrary $p \in (0,1)$ instead of just $p = 1/2$. How do the densities/entropies depend on p?*

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 usual sense (see the introduction). 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?