Some combinatorial processes

Jacob Richey

July 30, 2023

Contents

1 The sock process

Start with $2n$ socks in a laundry basket. Socks come in pairs of 2 (mine do, anyways). Draw one sock at a time uniformly at random and without replacement, laying the drawn socks on the bed. When the second sock from a pair is drawn, fold them together and put that pair away. Set $X_k =$ number of socks on the bed after the kth draw. Then $X_0 = X_{2n} = 0$. What does the process X_k look like?

Actually, we will consider a generalization of this process. Fix an integer $l \geq 2$, and suppose we draw uniformly at random from ln objects, divided into n groups of l. Once all objects of a single group have been drawn, they are removed. So the case $l = 2$ is the sock process. Interestingly, the sock process – centered and scaled – converges to a Brownian bridge, which can be proved by looking at an auxilary martingale. But this approach doesn't seem to work for other values of l, and it is not obvious how to prove a similar limit result in that case.

Set

$$X_k = \text{ number of un-grouped objects after the } k^{th} \text{ draw} \tag{1.1}$$

$$Y_k = \text{ number of groups removed after the } k^{th} \text{ draw} \tag{1.2}$$

We supress the parameters n and l. Then

$$X_k = k - lY_k. \tag{1.3}$$

To determine the natural scaling of X_k, we need to understand the scale of $\mathbb{E}X_k$. Note that

$$\mathbb{E}Y_k = \sum_{j=1}^{n} \mathbb{P}(\text{every member of group } j \text{ drawn by time } k) \tag{1.4}$$

$$= n \cdot \mathbb{P}(\text{every member of group 1 drawn by time } k) \tag{1.5}$$

$$= n \cdot \frac{\binom{ln-l}{k-l}}{\binom{ln}{k}} \tag{1.6}$$

$$= n \cdot \frac{(k)_l}{(ln)_l}, \tag{1.7}$$

where $(m)_i = m(m-1)\cdots(m-1+i)$ is the usual falling factorial. Thus

$$\mathbb{E}X_k = k - (ln)\frac{(k)_l}{(ln)_l} \tag{1.8}$$

$$\approx k - (ln)\frac{k^l}{l^l n^l} \tag{1.9}$$

$$= k\Big(1 - \Big(\frac{k}{ln}\Big)^{l-1}\Big), \tag{1.10}$$

where the approximation is assuming $k, n \to \infty$ with $k/n \in (0,1)$. Thus, X_k approximately looks like a scaled version of $f(t) = t(1 - t^{l-1})$.

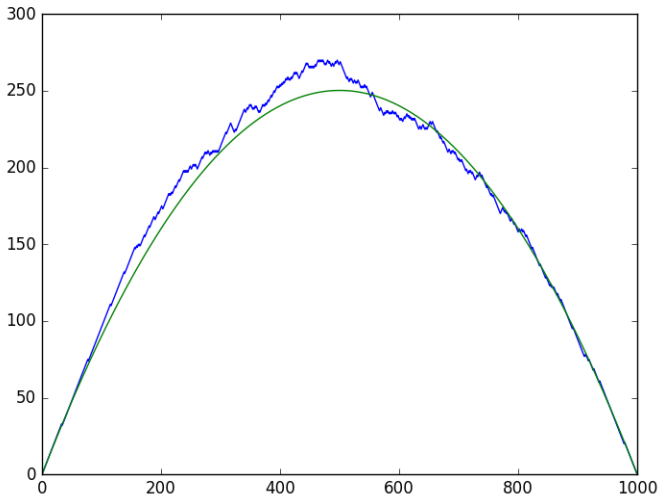

Figure 1: An instance of the sock process with $l = 2$ and $n = 500$, along with the theoretical expectation.

To determine the proper scaling, we also need the variance of X_k. By similar reasoning, we have

$$\mathbb{E}Y_k^2 = \mathbb{E}Y_k + (n^2 - n)\frac{\binom{ln-2l}{k-2l}}{\binom{ln}{k}} \tag{1.11}$$

$$= n \cdot \frac{(k)_l}{(ln)_l} + (n^2 - n)\frac{(k)_{2l}}{(ln)_{2l}}. \tag{1.12}$$

Thus

$$\mathbb{E}X_k^2 = k^2 - 2kl\mathbb{E}Y_k + l^2\mathbb{E}Y_k^2 \tag{1.13}$$

$$= k^2 - (2kl - l^2) \cdot n \cdot \frac{(k)_l}{(ln)_l} + l^2(n^2 - n)\frac{(k)_{2l}}{(ln)_{2l}}, \tag{1.14}$$

which leads to

$$\mathrm{Var}(X_k) = l^2\Big(n \cdot \frac{(k)_l}{(ln)_l} + (n^2 - n) \cdot \frac{(k)_{2l}}{(ln)_{2l}} - n^2\frac{(k)_l^2}{(ln)_l^2}\Big) \tag{1.15}$$

$$\approx l^2 n\Big((k/ln)^l + (n-1)(k/ln)^{2l} - n(k/ln)^{2l}\Big) \tag{1.16}$$

$$= l^2 n(k/ln)^l(1 - (k/ln)^l). \tag{1.17}$$

Thus, for k a fixed proportion of n, $\mathrm{Var}(X_k) \sim n$. This strongly suggests the scaled process

$$Z_t^{(n)} = \frac{1}{l\sqrt{n}}\Big(X_{\lfloor lnt \rfloor} - \mathbb{E}X_{\lfloor lnt \rfloor}\Big) \tag{1.18}$$

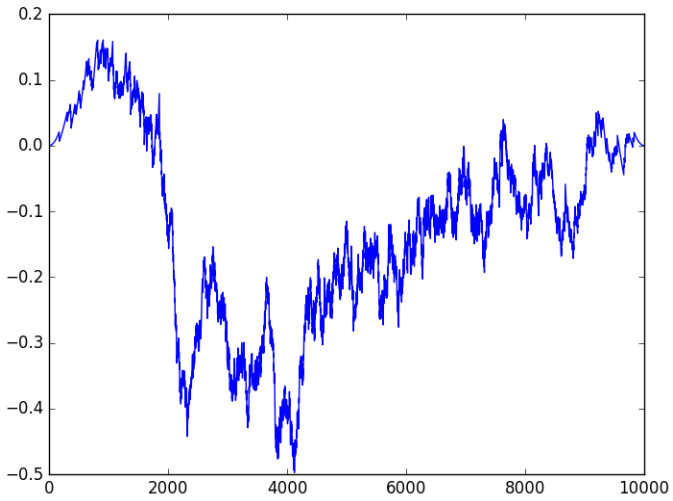

Figure 2: An instance of the process $Z_t^{(n)}$ with $l = 2$ and $n = 5000$.

for $t \in (0, 1)$, and the limiting object

$$\text{“}Z_t = \lim_{n\to\infty} Z_t^{(n)}\text{”} \tag{1.19}$$

which satisfies

$$\mathbb{E}Z_t = 0, \mathrm{Var}(Z_t) = t^l(1 - t^l) \tag{1.20}$$

looks like a time-scaled Brownian bridge. (The quotes are there because the limit Z is technically a limit of measures on $C[0, 1]$: this will take some extra work.) The covariance be computed directly via the same ideas as for the variance. We have

$$\mathbb{E}[X_k X_m] = km - l(kY_m + mY_k) + l^2\mathbb{E}[Y_k Y_m]. \tag{1.21}$$

Using $Y_k = \sum_i 1\{G_i(k)\}$, where $G_i(k)$ is the event that group i has been removed by the k^{th} draw,

$$\mathbb{E}Y_k Y_m = \mathbb{E}Y_{k\wedge m} + n(n-1)\mathbb{P}(G_1(k) \cap G_2(m)). \tag{1.22}$$

The crucial calculation is

$$\mathbb{P}(G_2(m)|G_1(k)) = \frac{\binom{ln-2l}{m-2l}}{\binom{ln-l}{m-l}} = \frac{(m-l)_l}{(ln-l)_l}. \tag{1.23}$$

Combining all of this, and approximating asymptotically with $k = lns, m = lnt$, one obtains

$$\mathrm{Cov}(Z_s, Z_t) = (s \wedge t)^l - s^l t^l. \tag{1.24}$$

Thus, to show that $Z_{t^{1/l}}$ is a Brownian bridge, it suffices to show that Z_t (exists and) is a continuous Gaussian process. It doesn't seem like it will be too difficult to show that Z_t exists and is continuous, but it is not clear why Z should be Gaussian. (One can relate the marginals of $Z^{(n)}$

to something that looks almost like a negative-binomial variable, but because the socks are drawn without replacement it isn't any typical discrete random variable. The idea would be to show that the distribution is close enough to something like a negative binomial, which is known to converge to normal when the parameters go to infinity.)

Question 1.1. *Does Z_t^n converge to a (time scaled) Brownian bridge as $n \to \infty$?*

One can show that the marginal distribution of X is given by

$$\mathbb{P}(X_k = m) = 2^m \binom{n}{(k-m)/2} \binom{n-(k-m)/2}{m} \Big/ \binom{2n}{k}. \tag{1.25}$$

To show that the limit is Brownian, these probabilities should converge to the Gaussian density when $n, m, k \to \infty$, as $k = 2nt, m = nx(?)$. To analyze the asymptotics, a useful formula is

$$\binom{N}{cN} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{c(1-c)}} c^{-c}(1-c)^{-(1-c)} \frac{1}{\sqrt{N}} \tag{1.26}$$

for $N \to \infty, c \in (0,1)$.

Note: The case $l = 2$ is special in that X_k is a Markov process with respect to itself, with transition probabilities

$$\mathbb{P}(X_{k+1} = X_k + 1 | X_k) = 1 - \frac{X_k}{2n-k}, \tag{1.27}$$

$$\mathbb{P}(X_{k+1} = X_k - 1 | X_k) = \frac{X_k}{2n-k}. \tag{1.28}$$

Thus

$$\mathbb{E}[X_{k+1} | X_k] = 1 + X_k(1 - \frac{2}{2n-k}). \tag{1.29}$$

One can use this to show that

$$M_k = \frac{2n(2n-1)}{(2n-k-1)(2n-k-2)} X_k - 2n \cdot \frac{k+1}{2n-k-2} \tag{1.30}$$

$$= \frac{2n}{2n-k-2} \left(\frac{(2n-1)X_k}{(2n-k-1)} - k - 1 \right) \tag{1.31}$$

is a martingale with respect to itself.

2 The coupon collector process

Consider the classical coupon collector setup: we draw repeatedly from $[n]$ with replacement, and wait for the first time when every element has been drawn at least once. Set $X_k^{(n)}$ to be the number of different types of coupons seen after the kth draw, and let $\tau^{(n)} = \min\{t : X_t^{(n)} = n\}$. The starting point is to compute the expectation of τ, which is straighforward: we can write (suppressing notation)

$$\tau = \sigma_1 + \sigma_2 + \cdots + \sigma_n, \tag{2.1}$$

where σ_i is the time it takes to get i coupons, after getting an $i-1$st one. The σ_i are exponentially distributed, with means $\mathbb{E}\sigma_i = \frac{n}{n-i+1}$. Thus $\mathbb{E}\tau = nH_n$, where $H_n = \sum_{i=1}^n \frac{1}{i} \approx \log n$.

Because the σ's are exponential, one can show that with high probability $\tau < Cn \log n$ for a universal constant C. So, what does the process X_k typically look like, for $1 \le k \le Cn \log n$? For example, what is $\mathbb{E}X_k$? It should be an increasing function of k with $\mathbb{E}X_0 = 0$ and $\mathbb{E}X_{n \log n} \approx n$.

Note that

$$X_k = \sum_{j=1}^n 1\{\text{coupon } j \text{ drawn by time } k\}, \tag{2.2}$$

and the probability of drawing any single coupon by time k is $1 - (1 - 1/n)^k$. Thus

$$\mathbb{E}X_k = n\left(1 - \left(1 - \frac{1}{n}\right)^k\right). \tag{2.3}$$

Alternatively, since X is a sub-martingale, it is natural to normalize X to be a martingale. Note that

$$\mathbb{E}[X_{k+1}|X_k] = \frac{X_k}{n} \cdot X_k + \left(1 - \frac{X_k}{n}\right)(X_k + 1) \tag{2.4}$$

$$= 1 + \frac{n-1}{n} \cdot X_k. \tag{2.5}$$

One can use this to check that

$$M_k = -k + X_k + \frac{1}{n}\sum_{j=1}^{k-1} X_j \tag{2.6}$$

is martingale (with respect to itself), with $M_0 = 0$. Thus

$$0 = \mathbb{E}M_k \implies \mathbb{E}X_k = k - \frac{1}{n}\sum_{j=1}^{k-1} \mathbb{E}X_j. \tag{2.7}$$

The recursion $a_k = k - \frac{1}{n} \cdot \sum_{j=1}^{k-1} a_j, a_0 = 0$ can be solved explicitly: the solution is

$$a_k = \sum_{j=0}^{k-1} \binom{k}{j+1} n^{-j}(-1)^j = \frac{n^k - (n-1)^k}{n^{k-1}}, \tag{2.8}$$

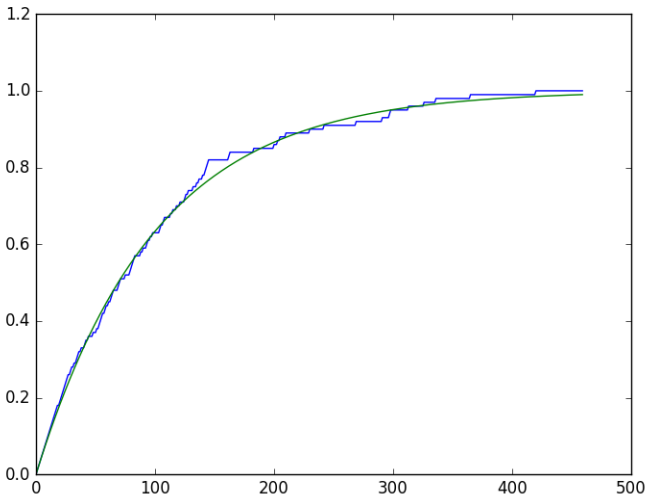

Figure 3: An instance of the process $\frac{1}{n}X_k^{(n)}$ with $n = 100$ and $1 \leq k \leq 100 \log 100 \approx 461$.

which matches the formula for $\mathbb{E}X_k$ given above.

Setting $k = \alpha n \log n$ yields the asymptotic formula

$$\mathbb{E}X_k = n\left(1 - \left(1 - \frac{1}{n}\right)^{\alpha n \log n}\right) \approx n(1 - n^{-\alpha}). \tag{2.9}$$

Applying the optional stopping theorem to M at τ yields the interesting identity

$$0 = \mathbb{E}M_\tau = -\mathbb{E}\tau + \mathbb{E}X_\tau + \frac{1}{n}\mathbb{E}\sum_{j=1}^{\tau-1} X_j \implies n^2(H_n - 1) = \mathbb{E}\sum_{j=1}^{\tau-1} X_j. \tag{2.10}$$

To compute the second moment, note that the probability that two distinct coupons are collected by time k is

$$\mathbb{P}(\text{coupons } i \text{ and } j \text{ drawn by time } k) = 1 - 2\left(1 - \frac{1}{n}\right)^k + \left(1 - \frac{2}{n}\right)^k. \tag{2.11}$$

Thus

$$\mathbb{E}X_k^2 = \mathbb{E}\left(\sum_{j=1}^{n} 1\{\text{coupon } j \text{ drawn by time } k\}\right)^2 \tag{2.12}$$

$$= \sum_{j=1}^{n} \mathbb{P}\left(\text{coupon } j \text{ drawn by time } k\right) + \sum_{i \neq j} \mathbb{P}\left(\text{coupons } i \text{ and } j \text{ drawn by time } k\right) \tag{2.13}$$

$$= \mathbb{E}X_k + (n^2 - n)\left(1 - 2\left(1 - \frac{1}{n}\right)^k + \left(1 - \frac{2}{n}\right)^k\right) \tag{2.14}$$

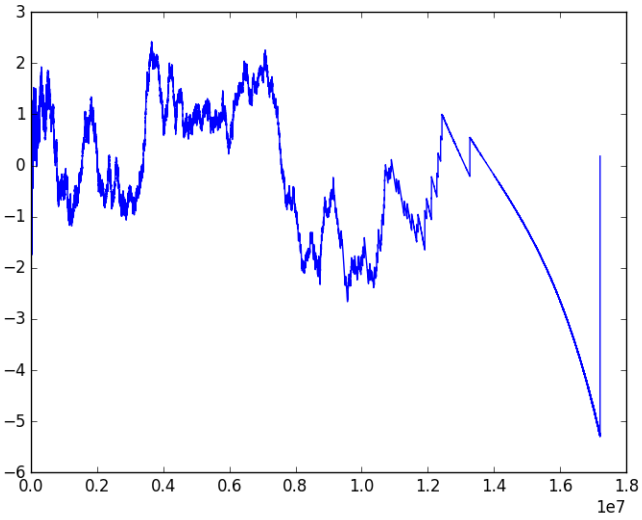

Figure 4: An instance of the process $Z_\alpha^{(n)}$ with $n = 10^6$. Note that $10^6 \log(10^6) \approx 1.4 \times 10^7$, which is around where the process stops looking like a Brownian bridge, as the variance is becoming extremely small.

$$= n^2 \left(1 - 2\left(1 - \frac{1}{n}\right)^k + \left(1 - \frac{2}{n}\right)^k\right) + n\left(\left(1 - \frac{1}{n}\right)^k - \left(1 - \frac{2}{n}\right)^k\right) \tag{2.15}$$

Setting $k = \alpha n \log n$ yields

$$\mathbb{E}X_k^2 \approx n^2(1 - n^{-\alpha})^2 + n^{1-\alpha}(1 - n^{-\alpha}). \tag{2.16}$$

(Alternatively, using the second-order identity

$$\mathbb{E}[X_{k+1}^2 | X_k] = X_k^2\left(1 - \frac{2}{n}\right) + X_k\left(2 - \frac{1}{n}\right) + 1, \tag{2.17}$$

one can show that

$$N_k = k + X_k^2 + \frac{2}{n}\sum_{j=1}^{k-1} X_j^2 - \left(2 - \frac{1}{n}\right)\sum_{j=0}^{k-1} X_j \tag{2.18}$$

is a martingale, and recover the second moment via a recursion.)

The variance is

$$\mathrm{Var}(X_k) \approx n(n^{-\alpha} - n^{-2\alpha}) \approx n^{1-\alpha}. \tag{2.19}$$

It is natural to consider the scaled process

$$Z_\alpha^n = \frac{X_{\lfloor \alpha n \log n \rfloor} - n(1 - n^{-\alpha})}{n^{(1-\alpha)/2}}, \tag{2.20}$$

where $\alpha \in (0, \infty)$. Of course, most of the action is taking place in $\alpha \in (0, 1)$.

Question 2.1. *Is it true that $Z^n_\alpha, \alpha \in (0,1)$ converges in distribution to a Brownian bridge process as $n \to \infty$?*

2.1 Generalized coupon collector

Consider a coupon collector process on $\mathbb{N}$: draw iid random variables V_i from a fixed distribution F on $\mathbb{N}$, say

$$F = \sum_{n \in \mathbb{N}} p_n \delta_n. \tag{2.21}$$

Then let $Z_n = \#$ of coupons collected by time n, i.e.

$$Z_n = \#\{V_i : i \in [n]\}. \tag{2.22}$$

Question 2.2. *What is the limiting behavior of Z_n?*

It should be tightly concentrated around $C_F \log n$ – how tightly? And what is C_F?

Question 2.3. *What is the distribution of $W_n = \{V_i : i \in [n]\}$?*

How close is it to Poisson process? (And what Poisson process? With values p_n, and some reduced intensity?)

Question 2.4. *What is the distribution of W_{τ_n}, where τ_n is the first time that n is collected?*

A similar martingale idea still works in this scenario: namely,

$$R_n = Z_n - \sum_{j=1}^{n-1} g_F(V_n) \tag{2.23}$$

is a martingale (with respect to the natural filtration), where for a finite subset $A \subset \mathbb{N}$,

$$g_F(A) = \sum_{n \notin A} p_n \tag{2.24}$$

is the probability of *not* selecting a new coupon at stage n, given that $V_{n-1} = A$. The OST yields

$$\mathbb{E}\left[\sum_{j=1}^{T_n - 1} g_F(W_j)\right] = n - 1, \tag{2.25}$$

where T_n is the first time t that $Z_t = n$; and also

$$\mathbb{E} Z_{S_m} = 1 + \mathbb{E}\left[\sum_{j=1}^{S_m - 1} g_F(W_j)\right] + 1, \tag{2.26}$$

where S_m is the first time t that $W_t \supset [m]$. One can also check that

$$Z_n^2 - \sum_{j=1}^{n-1} (2Z_j + 1) g_F(W_j) \tag{2.27}$$

is a martingale – does this yield anything useful?

An interesting identity:

$$\begin{aligned}
\mathbb{P}(\tau_n < \tau_m) &= \mathbb{P}(V_1, V_2, \ldots, V_{\tau_n} \neq m) && (2.28)\\
&= \sum_{t\in\mathbb{N}} \mathbb{P}(\tau_n = t)\mathbb{P}(V_1, \ldots, V_t \neq m) && (2.29)\\
&= \sum_{t\in\mathbb{N}} \mathbb{P}(\tau_n = t)(1-p_m)^t && (2.30)\\
&= \mathbb{E}[(1-p_m)^{\tau_n}]. && (2.31)
\end{aligned}$$

Thus

$$\mathbb{E}[(1-p_m)^{\tau_n} + (1-p_n)^{\tau_m}] = 1. \tag{2.32}$$

3 Non-adjacent uniform placements

My home movie theatre has n seats in a single row. One night I threw a big party, and at the end everyone wanted to watch the movie, but at a safe distance from everyone else (to avoid virus transmission). One by one people sat down in a (uniform) random seat that wasn't adjacent to anyone who is already seated. How many people got a seat? Can the distribution of occupied seats be described in a simple way?

Let X_n denote the number of people who have a seat when no more seats can be taken, i.e. when every seat is either occupied or adjacent to an occupied seat. Clearly $\lceil \frac{n}{3} \rceil \leq X_n \leq \lceil \frac{n}{2} \rceil$. What is $\mathbb{E}X_n$? $\operatorname{Var} X_n$?

The expectation can be calculated via a generating function idea. Let $a_n = \mathbb{E}X_n$. After the first person to sit down selects a location, this divides the seats into two smaller groups, and effectively 'deletes' the two seats adjacent to the first seat chosen. Further selections in each group are uniform over those groups. It follows that a_n satisfies the recursion

$$a_n = 1 + \frac{1}{n}\sum_{i=1}^{n}(a_{i-2} + a_{n-i-1}) = 1 + \frac{2}{n}\sum_{i=1}^{n-2} a_i, \tag{3.1}$$

where $a_{-1} = a_0 = 0$ by convention. Let $f(x) = \sum_{n\geq 1} a_n x^n$. Multiplying by x^{n-1} and summing over n in the above recursion yields

$$f'(x) = \frac{x^2}{(1-x)^2} + \frac{2x(1+f(x))}{1-x}. \tag{3.2}$$

This ODE can be solved to give

$$f(x) = \frac{1-e^{-2x}}{2(1-x)^2} = x + x^2 + \frac{5}{3}x^3 + 2x^4 + \frac{37}{15}x^5 + \cdots. \tag{3.3}$$

Using series expansions, we get

$$f(x) = \frac{1}{2}\left(\sum_{n\geq 0}(n+1)x^n\right)\left(\sum_{n\geq 1}(-1)^{n+1}\frac{2^n}{n!}x^n\right), \tag{3.4}$$

and equating coefficients yields

$$a_n = \frac{1}{2}\sum_{j=1}^{n}(-1)^{j+1}\frac{2^j}{j!}(n+1-j). \tag{3.5}$$

Mathematica gives $\frac{a_n}{n} = \frac{1}{n}\mathbb{E}X_n \to \frac{1}{2}\left(1-e^{-2}\right)$ as $n \to \infty$. It's not too hard to see this explicitly:

$$\begin{aligned} a_n &= \frac{n}{2}\sum_{j=1}^{n}(-1)^{j+1}\frac{2^j}{j!} - \frac{1}{2}\sum_{j=2}^{n}(-1)^j\frac{(j-1)2^j}{j!} && (3.6)\\ &\approx \frac{1}{2}(1-e^{-2})n + \frac{1}{2}(1-3e^{-2}) && (3.7)\\ &\approx .432332n + .296997. && (3.8) \end{aligned}$$

Question 3.1. *Is there a direct combinatorial/probabilistic explanation for this limiting constant?*

Other functions of X_n can also be computed this way. Let U denote the position of the first person to choose a seat, so U is uniform on $[n]$. Implicit in our proof of the recursion for a_n is the distributional identity

$$X_n \stackrel{d}{=} \sum_{i\in[n]} 1\{U=i\}(1+X'_{i-2}+X''_{n-i-1}). \tag{3.9}$$

(Here X' and X'' represent independent copies of X.) Squaring both sides and using the fact that the events $\{U=i\}$ are disjoint over i to eliminate cross terms yields

$$X_n^2 \stackrel{d}{=} \sum_{i\in[n]} 1\{U=i\}(1+X_{i-2}+X_{n-i-1})^2. \tag{3.10}$$

Setting $b_n = \mathbb{E}X_n^2$ and $g(x) = \sum_{n\geq 1} b_n x^n$, the same methods as above give the recursion

$$nb_n = 2na_n - n + 2\sum_{i=1}^{n-4} a_i a_{n-3-i} + 2\sum_{i=1}^{n-2} b_i, \tag{3.11}$$

the functional equation

$$g'(x) = 2f'(x) - \frac{1}{(1-x)^2} + 2x^2 f(x)^2 + \frac{2x}{1-x}g(x), \tag{3.12}$$

and the solution

$$g(x) = \frac{(1+x)(1+e^{-4x}) - 2(1-x+2x^2)e^{-2x}}{4(1-x)^3} = x + x^2 + 3x^3 + 4x^4 + \frac{19}{3}x^5 + \cdots. \tag{3.13}$$

With some additional effort, an explicit formula for b_n follows:

$$b_n = \binom{n+1}{2} + \frac{1}{8}\sum_{j=2}^{n} \frac{(-1)^j}{j!}\left(4^j(1-j/4) - 2^j(2+j^2)\right)(n-j+1)(n-j+2) \tag{3.14}$$

Expanding the polynomial $(n-j+1)(n-j+2)$ and summing terms separately yields

$$b_n \approx \frac{n^2}{4}(1-e^{-2})^2 + \frac{n}{2}(1-4e^{-2}+5e^{-4}) + \frac{1}{4}(1-6e^{-2}+21e^{-4}) \tag{3.15}$$

$$\approx .186911n^2 + .275119n + .143154. \tag{3.16}$$

Note that

$$\lim_{n\to\infty}\frac{b_n}{n^2} = \left(\lim_{n\to\infty}\frac{a_n}{n}\right)^2, \tag{3.17}$$

so that $\operatorname{Var} X_n$ is order n, namely

$$\operatorname{Var} X_n = e^{-4}(n+3) + O(\exp(-n)). \tag{3.18}$$

.

Usually, the next step is:

Conjecture 3.2. *X_n has the following central limit theorem:*

$$\frac{X_n - a_n}{\sqrt{\operatorname{Var} X_n}} \xrightarrow{d} N(0,1). \tag{3.19}$$

It is not yet clear how to prove this CLT. Given our nice recursive formula, it is natural to try a moment generating function approach. Set

$$c_n(z) = \mathbb{E}[z^{X_n}]. \tag{3.20}$$

Exercise: use the distributional recursion 3.9 to obtain

$$z^{-1} n c_n(z) = \sum_{i=1}^{n} c_{i-2}(z) c_{n-i-1}(z). \tag{3.21}$$

Now define the two-variable moment generating function

$$h(x,z) = \sum_{n \geq 1} c_n(z) x^n. \tag{3.22}$$

The above recursion, after some manipulations (similar to the variance calculation) leads to

$$z^{-1} \frac{\partial h}{\partial x}(x,z) = (1 + x + x h(x,z))^2. \tag{3.23}$$

If you ask Mathematica nicely, it provides the following solution:

$$h(x,z) = -\frac{x\sqrt{z}\text{Cosh}[x\sqrt{z}] + (xz + z - 1)\text{Sinh}[x\sqrt{z}]}{(x-1)\sqrt{z}\text{Cosh}[x\sqrt{z}] + (xz - 1)\text{Sinh}[x\sqrt{z}]} \tag{3.24}$$

$$= zx + zx^2 + \frac{1}{3}(z + 2z^2)x^3 + z^2 x^4 + \frac{1}{15}(8z^2 + 7z^3)x^5 + \cdots \tag{3.25}$$

Ideally, one could extract enough information about the coefficient of x^n in $h(x,z)$ to understand the limiting distribution of X_n. To prove the conjecture, it is enough to show that

$$\lim_{n \to \infty} \exp\left(-z a_n / \sqrt{\operatorname{Var} X_n}\right) \mathbb{E}\left[\exp\left(\frac{z}{\sqrt{\operatorname{Var} X_n}}\right)\right] = \mathbb{E}[\exp(zN(0,1))] = \exp(z^2/2). \tag{3.26}$$

Numerical approximations suggest that this holds.

Question 3.3. *Can the explicit formula for $h(x,z)$ be used to prove the MGF convergence 3.26?*

Other possible distributions: uniform random configuration (see 'Padovan' sequence), or left-to-right arrivals with some distribution (inter-distances are either 2 or 3 with any probability).

Question 3.4. *Which distribution, among those that are sufficiently 'random' and 'local', optimizes the average number of seats taken?*

How to make this question make sense? If seats are taken left to right, with all gaps of size one, then clearly the number of seats is optimized. Perhaps 'local' means that arrivals can only ask about a constant number of seats when deciding on availability... or something like this.

3.1 MGIS, September 2021

There is existing work on computing asymptotic value $\lim_{n\to\infty} X_n/n$ for a large class of (possibly random) graphs. The 'non-adjacent placements' process can be viewed as a 'maximal independent greedy set' on $\mathbb{Z}$ (or $\mathbb{N}$ or an interval) as follows. Assign to the sites $z \in \mathbb{Z}$ iid continuous random variables W_z, and define a sequence $(I_t(\mathbb{Z}) : t \in \mathbb{R})$ of increasing subsets of $\mathbb{Z}$ via

$$I_t = \{z : W_z < t, \text{ and } z-1, z+1 \notin I_{t^-}\}, \tag{3.27}$$

where $I_{t^-} = \bigcup_{s<t} I_s$. Then $I = I_\infty$ is a maximal independent set, in the sense that every point in $\mathbb{Z}$ is adjacent to some point of I, and this construction is equivalent to choosing 'seats' one by one for a finite interval.

With this construction, we can directly compute the expected density of I, i.e. $\mathbb{P}(0 \in I)$. Choose arrival times W_z that are iid with uniform $(0,1)$ distribution. First, observe that on a finite interval, say $\{0, 1, 2, \dots n\}$,

$$\begin{aligned}
\mathbb{P}(0 \in I | W_0 = x) &= 1 - \mathbb{P}(1 \text{ arrives before } 0 | W_0 = x) && (3.28)\\
&= 1 - \mathbb{P}(W_1 < W_0 | W_0 = x) + \mathbb{P}(2,1,0 \text{ arrive in that order} | W_0 = x) && (3.29)\\
&= 1 - \mathbb{P}(W_1 < x) + \mathbb{P}(W_2 < W_1 < x) && (3.30)\\
&\qquad - \mathbb{P}(3,2,1,0 \text{ arrive in that order} | W_0 = x) && (3.31)\\
&= \sum_{k=0}^{n} (-1)^k \frac{x^k}{k!}, && (3.32)
\end{aligned}$$

which converges to e^{-x} as $n \to \infty$. It follows that for MGIS on $\mathbb{N}$ or $\mathbb{Z}$,

$$\mathbb{P}(0 \in I(\mathbb{N})) = \int_0^1 e^{-x} = 1 - e^{-1}, \tag{3.33}$$

and

$$\mathbb{P}(0 \in I(\mathbb{Z})) = \int_0^1 e^{-2x} = \frac{1}{2}(1 - e^{-2}), \tag{3.34}$$

since conditionally on $W_0 = x$, the event that -1 arrives before 0 is independent of the event that 1 arrives before 0, so the above calculation splits into a product. Note that this recovers the same density as before. It also gives an explicit expression for finite intervals:

$$\mathbb{P}(0 \in I([-a,b])) = \int_0^1 \left(\sum_{j=0}^{a} (-1)^j \frac{x^j}{j!} \right) \left(\sum_{k=0}^{b} (-1)^k \frac{x^k}{k!} \right) dx. \tag{3.35}$$

Of course, this converges to $\mathbb{P}(0 \in I(\mathbb{Z}))$ quickly as a or $b \to \infty$.

3.2 Distribution of the configuration

A natural candidate for the distribution of I on $\mathbb{Z}$ is a renewal process, namely the ergodic process that has gaps of size 1 or 2 between points of the configuration, each gap occuring independently with the appropriate probability so that the configuration has density $\frac{1}{2}(1 - e^{-2})$.

Fact 3.5. *I does not have the distribution of this renewal process.*

To prove this, we can look at a probability for a possible configuration on a subinterval in $\mathbb{Z}$ of size 6. Let $R \subset \mathbb{Z}$ denote the renwal process configuration. Assuming sites 0 and 6 are occupied, there are two possible ways to fill the interval $[0,6] \cap \mathbb{Z}$: with two gaps of size 2, or three of size 1. Observe that

$$\mathbb{P}(0,2,4,6 \in I | W_0 = x, W_2 = y, W_4 = z, W_6 = w) = \tag{3.36}$$
$$e^{-x}e^{-w}(1-\min(x,y))(1-\min(y,z))(1-\min(z,w)). \tag{3.37}$$

Integrating over all four variables yields

$$\mathbb{P}(0,2,4,6 \in I) = \frac{1}{2} - \frac{79}{30e^2} \approx 0.1436170875 \tag{3.38}$$

For the renewal process R, we must first compute the probabilities of having gaps of size 1 or 2, say p_1 and p_2. The densities d_1, d_2 of gaps of size 1 and 2 satisfy

$$d_1 + d_2 = \frac{1}{2}(1-e^{-2}) \text{ and } 2d_1 + 3d_2 = 1. \tag{3.39}$$

(The second condition comes from counting the total length: gaps of size 1 correspond to length 2 strings 10, and gaps of size 2 to length 3 strings 100.) The probabilities are then given by

$$p_1 = \frac{d_1}{d_1+d_2}, \text{ and } p_2 = \frac{d_2}{d_1+d_2}. \tag{3.40}$$

The solution is $p_1 = \frac{e^2-3}{e^2-1}, p_2 = \frac{2}{e^2-1}$, and we can obtain

$$\mathbb{P}(0,2,4,6 \in R) = \mathbb{P}(0 \in R)\mathbb{P}(2,4,6 \in R | 0 \in R) \tag{3.41}$$
$$= \frac{1}{2}(1-e^{-2}) \cdot p_1^3 \tag{3.42}$$
$$\approx 0.1401590138 \tag{3.43}$$

Since these probabilities don't match, the distributions are different.

One way to get a handle on the configuration is via the following 'renewal' event:

$$C_k = \{W_0 > W_1 > \cdots > W_k < W_{k+1}\} \tag{3.44}$$

In words, C_k is the event where, searching right from site 0, the first site where there is a guaranteed element of I (because the weight at that site is smaller than both neighbors) occurs at k. Along with C_k we have the random variable

$$K = \min\{z \geq 0 : W_z < \min(W_{z-1}, W_{z+1})\} \tag{3.45}$$

Conditioning on W_0 and W_k allows us to compute the conditional probabilities of the C_k:

$$\mathbb{P}(C_k | W_0 = x, W_k = y) = \frac{(x-y)^{k-1}}{(k-1)!}(1-y), \tag{3.46}$$

and thus

$$\mathbb{P}(C_k) = \mathbb{P}(K=k) = \int_{x>y} \frac{(x-y)^{k-1}}{(k-1)!}(1-y)\,dx\,dy = \frac{k+1}{(k+2)!}. \tag{3.47}$$

So K is sub exponential, and for example,

$$\mathbb{E}[K] = \sum_{k\geq 0} \frac{k(k+1)}{(k+2)!} = e - 2, \mathrm{Var}(K) = e(3-e). \tag{3.48}$$

The generating function of K is

$$\mathbb{E}[w^K] = \frac{1 + (w-1)e^w}{w^2}. \tag{3.49}$$

What does this have to do with 'renewal'? Conditionally on W_0, the configuration to the right of 0 is independent of the configuration to the left, so the events C_k and C_{-k} (same event but left instead of right) are conditionally independent; and if we also condition on W_k, the configuration in $[0, k]$ is independent of the configuration in $[k+1, \infty)$. So we can build the configuration I by first generating two independent copies of K, one for the right of 0 and one for the left – note that these values completely determine the configuration in the (random) interval $[-K' - 1, K + 1]$ – then generating further copies starting at $K + 2$ and $-K' - 2$, and so on.

In principle, this should allow a calculation of the covariance between sites in I, i.e. of the probabilities

$$\mathbb{P}(0, k \in I) \tag{3.50}$$

4 Adjacent occurrences in a random permutation

Let σ be a uniformly random permutation on $[n]$, i.e. $\sigma : [n] \to [n]$ a uniformly random bijection. Define the indicator variables

$$A_j = 1\{\sigma(j+1) = \sigma(j) + 1\}, \tag{4.1}$$

for $j \in [n-1]$. For example, when $\sigma = 2341$, i.e. $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 4, \sigma(4) = 1$,

$$A_1 = A_2 = 1, A_3 = 0. \tag{4.2}$$

Consider the statistic

$$Y_n = \sum_{j=1}^{n-1} A_j \tag{4.3}$$

Y counts the number of 'adjacencies' preserved by σ, but Y ignores the final possible 'adjacent pair,' namely $\sigma(n)$ and $\sigma(1)$. Note also that Y does not count the numbers n and 1 as 'adjacent,' even though they are 1 apart mod n. This makes it easy to establish a limit theorem for Y, and small perturbations of Y will follow the same limit law.

Observe that $\mathbb{E}A_j = \frac{1}{n-1}$ for all j, so $\mathbb{E}Y_n = 1$ for all n. Note that the A_j are all pairwise dependent. It is not too hard to show that

$$\mathbb{E}A_iA_j = \frac{1}{(n-1)(n-2)}, \tag{4.4}$$

so

$$\mathrm{Var}(Y_n) = \sum_j \mathbb{E}[A_j^2] + \sum_{i \neq j} \mathbb{E}A_iA_j - \sum_j \mathbb{E}[A_j] = \sum_{i \neq j \in [n-1]} \frac{1}{(n-1)(n-2)} = 1. \tag{4.5}$$

and

$$\mathrm{Cov}(A_i, A_j) = \frac{1}{(n-1)^2(n-2)} \sim n^{-3}, \tag{4.6}$$

To see 4.4, note that the event A_iA_{i+1} is equivalent to having an ascending sequence of length 3 starting at position i. Whatever goes in position i, there is only one choice out of the possibilities for the next two numbers for which the sequence $\sigma(i), \sigma(i)+1, \sigma(i)+2$ occurs, and there are exactly $(n-1)(n-2)$ possibilities for these two slots. (This doesn't quite work unless we consider n and 1 an adjacent pair!) Similar reasoning takes care of the case $|i-j| > 1$ separately, but interestingly, the probabilities come out the same.

Question 4.1. *Is it a coincidence that the probability $\mathbb{P}(A_iA_j)$ does not depend on i and j, as long as $i \neq j$, or is there a simple explanation that does not rely on splitting into cases when $j = i \pm 1$ and otherwise?*

Question 4.2. *Does this happen for triples of the A_j's? For $k \in \mathbb{N}$, it always holds that*

$$\mathbb{E}A_1A_2 \cdots A_k = \frac{1}{(n-1)_k} = \frac{1}{(n-1)(n-2)\cdots(n-k)}. \tag{4.7}$$

Is it true, for example, that

$$\mathbb{E}A_1A_3A_5 = \frac{1}{(n-1)_k}? \tag{4.8}$$

This seeems likely to be false, because there can be 'collisions' of possible adjacent pairs when three pairs are considered.

One can think of the Y_n as a process in n. Start with the trivial permutation $\sigma_1 : [1] \to [1]$, and for $n \geq 1$, recursively define

$$\sigma_{n+1}(j) = \begin{cases} \sigma_n(j), & j = 1, 2, \ldots, U_{n+1} - 1 \\ n+1, & j = U_{n+1} \\ \sigma_n(j-1), & j = U_{n+1}+1, \ldots, n+1 \end{cases} \tag{4.9}$$

where $\{U_n\}_n$ is a sequence of independent uniform random variables, with U_n is uniform on $[n]$ for each n. In words, writing σ_n in in-line notation, σ_{n+1} is obtained by sticking $n+1$ in at a random spot. Then σ_n has the same distribution as a uniform random permutation for each n, and Y_n is described by the transition probabilities

$$Y_{n+1} = \begin{cases} Y_n - 1, & w.p. \ \frac{Y_n}{n+1} \\ Y_n, & w.p. \ \frac{n-Y_n}{n+1} \\ Y_n + 1, & w.p. \ \frac{1}{n+1} \end{cases} \tag{4.10}$$

Proposition 4.3. *With the initial conditions $Y_1 = 0$, this recursive definition Y_n agrees in distribution with 4.3.*

(We abuse notation slightly by writing Y_n and σ_n for both this process in n and the distribution functions.)

Proof. Y_n increases by 1 exactly when $n+1$ is placed just after n in σ_n. Y_n decreases by 1 exactly when $n+1$ is placed between two adjacent elements $\sigma(i)$ and $\sigma(i+1) = \sigma(i)+1$. There are exactly Y_n such positions i. In any other case, $Y_n = Y_{n+1}$. □

Let X_∞ and Y_∞ denote the limiting (stationary) distributions for the X and Y processes.

Theorem 4.4. $Y_\infty \sim$ Poisson(1).

Proof. One can show that the poisson distribution exactly satisfies the recursion implied by 4.10, namely the equation

$$\mathbb{P}(Y_{n+1} = k) = \frac{n-k}{n+1}\mathbb{P}(Y_n = k) + \frac{k+1}{n+1}\mathbb{P}(Y_n = k+1) + \frac{1}{n+1}\mathbb{P}(Y_n = k-1). \tag{4.11}$$

Indeed, substituting $\mathbb{P}(Y_n = k) = \frac{1}{e}\frac{1}{k!}$ yields the equality

$$\frac{1}{e}\frac{1}{k!} = \frac{1}{e}\frac{1}{n+1}\frac{1}{k!}\left(n - k + 1 + k\right). \tag{4.12}$$

It follows that Poisson(1) is the unique stationary measure for Y, and thus by convergence of markov chains, Y_∞ has Poisson(1) distribution. □

One simple variant on Y is the 'full' sum

$$X_n = \sum_{j=1}^{n} A_j, \tag{4.13}$$

where $A_n = 1\{\sigma(n) + 1 = \sigma(1)\}$. X satisfies the recursive relation

$$\mathbb{P}(X_{n+1} = k) = \frac{n-k-1}{n+1}\mathbb{P}(X_n = k) + \frac{k+1}{n+1}\mathbb{P}(X_n = k+1) + \frac{2}{n+1}\mathbb{P}(X_n = k-1). \tag{4.14}$$

This isn't satisfied by the poisson density. However, X is close enough to Y that it has the same limit: indeed, $X_n - Y_n = A_n$ converges to 0 in L^1, since the indicator variable A_n has $\mathbb{E}A_n = \frac{1}{n-1} \to 0$, which implies that $X_n - Y_n$ converges to 0 in distribution.

Another natural variant is to allow equality mod n. Define

$$B_j = 1\{\sigma(j+1) = \sigma(j) + 1\}, \tag{4.15}$$

where the equality and the addition are performed mod n. For example, with $\sigma = 2341$ as above,

$$B_1 = B_2 = B_3 = B_4 = 1. \tag{4.16}$$

Set

$$Z_n = \sum_{j=1}^{n} B_j. \tag{4.17}$$

What is the corresponding limit for Z_∞? The same idea does not apply. Thinking of Z_n as a process in n doesn't work as nicely as with Y, because the adjacent pair 1 and n mod n is no longer adjacent when $n+1$ is added.

5 Hitting times of sums

5.1 IID Uniform$(0,1)$

Let $S_n = \sum_{i=1}^n U_i$ be a sum of iid uniform $(0,1)$ random variables U_i, and for $x > 0$, consider the hitting time

$$\tau_x = \inf\{n > 0 : S_n \geq x\}. \tag{5.1}$$

There is an easy way to compute $\mathbb{E}\tau_x$ for $x \in (0,1)$. First, to prove that $\mathbb{E}\tau_x$ is continuous, note that

$$\mathbb{E}\tau_x = \sum_{k\geq 1} \mathbb{P}(\tau_x \geq k) \cdot k = \sum_{k\geq 1} \mathbb{P}(S_k < x) \cdot k. \tag{5.2}$$

Each of the probabilities $\mathbb{P}(S_k < x)$ is continuous in x, so it suffices to show that the infinite sum above is uniformly convergent. But this is clear: the probabilities $\mathbb{P}(S_k < x)$ decay exponentially if x is confined to a compact interval by the usual CLT, and so the tails of the sum above converge to zero uniformly (on any compact interval).

Also, by conditioning on the value of U_1, we have

$$\mathbb{E}\tau_x = \mathbb{P}(U_1 \geq x) + \int_0^x (1 + \mathbb{E}\tau_{x-y})\mathbb{P}(U_1 = y)dy = 1 + \int_0^x \mathbb{E}\tau_y dy. \tag{5.3}$$

Since $\mathbb{E}\tau_x$ is continuous, the above equation shows that it is differentiable, and taking the derivative yields

$$\frac{\partial}{\partial x}\mathbb{E}\tau_x = \mathbb{E}\tau_x. \tag{5.4}$$

Thus $\mathbb{E}\tau_x = e^x$ for $x \in (0,1)$. Now for $x \geq 1$,

$$\mathbb{E}\tau_x = \int_0^1 (1 + \mathbb{E}\tau_{x-y})dy = 1 + \int_{x-1}^x \mathbb{E}\tau_y dy, \tag{5.5}$$

and so differentiating yields

$$\frac{\partial}{\partial x}\mathbb{E}\tau_x = \mathbb{E}\tau_x - \mathbb{E}\tau_{x-1}. \tag{5.6}$$

(This equation actually holds for all x, if one defines $\tau_x = 0$ for $x \leq 0$.) This equation can be solved recursively, i.e. by solving it on $x \in (1,2)$, using the explicit formula $\mathbb{E}\tau_{x-1} = e^{x-1}$, and the fact that the differential equation

$$y' = y + f \tag{5.7}$$

has solution

$$y = Ce^x + e^x \int_0^x e^{-s} f(s)ds. \tag{5.8}$$

One obtains that for $x \in (n, n+1)$,

$$\mathbb{E}\tau_x = e^x \sum_{k=0}^n e^{-k} \frac{(-1)^k}{k!}(x-k)^k. \tag{5.9}$$

For example, $\mathbb{E}\tau_x = (1+\frac{1}{e})e^x - xe^{x-1}$ for $x \in (1,2)$. Interestingly, $\mathbb{E}\tau_x$ is smooth for $x \notin \mathbb{N}$, and it is C^n at $x = n$, i.e. it has exactly $n-1$ derivatives at n.

The LLN and CLT for renewal processes tell us that $\mathbb{E}\tau_x \sim 2x$ and $\mathrm{Var}(\tau_x) \sim \frac{2}{9}x$, but this is not at all clear from the explicit formula. For example:

Question 5.1. *Can it be proved directly that*

$$\lim_{x\to\infty} \frac{1}{x} e^x \sum_{k=0}^{\lceil x \rceil} e^{-k} \frac{(-1)^k}{k!} (x-k)^k = 2? \tag{5.10}$$

Additionally:

Question 5.2. *Numerical computations suggest that $\mathbb{E}\tau_x - 2x$ converges to $\frac{2}{3}$ as $x \to \infty$, at an exponential rate. Can this be proved?*

Higher moments can also be computed explicitly. For any $k \in \mathbb{N}$ and $x \in (0,1)$,

$$\mathbb{E}\tau_x^k = (1-x) + \int_0^x \mathbb{E}\left[(1+\tau_{x-y})^k\right] dy = 1 + \sum_{j=1}^{k} \binom{k}{j} \int_0^x \mathbb{E}\tau_y^j dy. \tag{5.11}$$

This is equivalent to a differential equation of the same form as in the case $k = 1$: solving yields the recursion

$$\mathbb{E}\tau_x^k = e^x \left(1 + \sum_{j=1}^{k-1} \binom{k}{j} \int_0^x e^{-s} \mathbb{E}\tau_y^j dy\right). \tag{5.12}$$

Solving this recursion gives

$$\mathbb{E}\tau_x^k = e^x \left(\sum_{j=1}^{k} j \cdot s_2(k,j) \cdot x^{j-1}\right), \tag{5.13}$$

where $s_2(k,j)$ are the Stirling numbers of the second kind, i.e. the number of partitions of $[k]$ into j non-empty subsets. (So $j s_2(k,j)$ is the number of such partitions with a distinguished subset.) The first few are

$$\mathbb{E}\tau_x = e^x, \mathbb{E}\tau_x^2 = e^x(1+2x), \mathbb{E}\tau_x^3 = e^x(1+6x+3x^2), \ldots \tag{5.14}$$

We note the following surprising fact: for $x \in (0,1)$, $Var(\tau_x) = (1+2x)e^x - e^{2x}$ is not monotone on $(0,1)$. In fact, it has a maximum when $x \approx .858$.

Find explicit formulae for $\mathrm{Var}(\tau_x)$ for any $x > 0$.

The entropy exhibits a similar phenomenon (see picture).

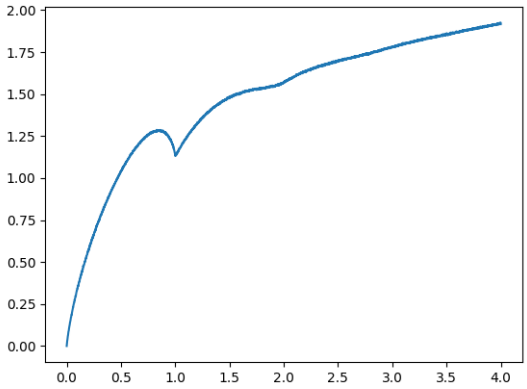

Figure 5: An empirical plot of the entropy of τ_x for $x \in [0, 4]$.

5.2 IID Geometric(p)

Now let $S_n = \sum_{i=1}^n X_i$ where X_i are iid Geometric(p) for $p \in (0, 1)$. We have hitting times

$$\sigma_k = \inf\{s \geq 0 : S_s \geq k\} \tag{5.15}$$

for any $k \in \mathbb{R}$. The the renewal equation gives

$$\mathbb{E}\sigma_k = 1 + \sum_{i=1}^{k-1} pq^{k-i-1}\mathbb{E}\sigma_i, \tag{5.16}$$

and one readily finds the generating function

$$\sum_{k\geq 1} z^k \mathbb{E}\sigma_k = \frac{z(1-qz)}{(1-z)^2}, \tag{5.17}$$

where $q = 1 - p$, which yields the exact (!) formula

$$\mathbb{E}\sigma_k = 1 + (k-1)p = pk + q \tag{5.18}$$

for $k \geq 1$. The following questions make sense in the context of the 'glass panes' game from Squid Game:

Question 5.3. *What is the distribution of σ_k? What is its median? For a given $L > 0$, which value j maximizes $\mathbb{P}(\sigma_k \in \{j, j-1, \ldots, j-L\})$?*

Given k and L, if we can solve for the maximizer j, that's where we should want to stand in line in the glass panes game. Probably all of these are explicitly answerable, since S_s has the negative binomial distribution, and S_s and σ_k are related in the usual way: $S_s \leq k \iff \sigma_k \geq s$.

6 Heatseekers

Start k heatseeking missiles in $\mathbb{R}^d$ or the torus $\mathbb{T}^d$, labelled $1, 2, \ldots, k$. Missiles move at unit speed, and missile i moves directly towards missle $(i+1) \mod k$ for each i. When missle i catches its target, one can imagine that it 'coalesces' with missile $i+1$, so we are left with one fewer missiles performing the same dynamics. Eventually, we are left with some stable configuration, where no more coalescences will ever occur. It seems natural to look at the limit $k \to \infty$, and start with iid initial points, or with a Poisson process of points with intensity k.

In dimension $d \geq 2$, stable configurations should be measure zero, so the process should always end with a single coalesced particle. Is there a quick proof of this? Something along the lines of: the only possible stable configurations are where all the particles are 'colinear,' and those have measure zero.

Question 6.1. *How many coalescences occur before a stable configuration is reached?*

Question 6.2. *Start from iid random points or a Poisson process. After j coalescences have occurred, what is the distribution of the remaining points?*

Question 6.3. *When a stable configuration is reached, what are the 'cluster sizes,' i.e. the number of coalesced particles 'contained' in each surviving particle?*

It's easy to calculate the measure of the set of stable configurations on the circle:

Fact 6.4. *Suppose the initial heatseeker positions are sampled iid uniformly over the circle $\mathbb{T}^1$. Then the probability that the configuration is stable, i.e. all missiles are traveling in the same direction, is 2^{k-1}.*

Proof. All particles have to be traveling in the same direction around the circle for the configuration to be stable. Each particle is initially traveling clockwise or counter clockwise, independent of the directions of the other particles. □

Observe that the only particle whose trajectory is affected by a collision is the one that found its target – the other trajectories are unchanged from time τ^- to time τ^+. So given an initial configuration of initial positions, we can write down a sequence of orientations for each particle, i.e an element of $\{\pm 1\}^k$, which evolves by: at each time, find a pair of arrows that disagree, $a_i \neq a_{i+1}$, and delete one of them from the sequence. Exactly how we choose the next arrow to delete seems complicated.

6.1 Mean field arrows

Suppose we start with an iid field of arrows in $\{\pm 1\}^k$, and choose *uniformly* from all disagreements, and delete the 'left' arrow from that disagreement. If we just want to keep track of the number of arrows of each type during the process, it is equivalent to the following 'random walk' formulation: let $X_0 = B \sim \text{Binomial}(k, 1/2)$ be the number of $+1$'s initially, and $Y_0 = k - B$ is the number of -1's. At each step we flip a fair coin and decrease either X or Y by 1. The process ends at the first time when either $X = 0$ or $Y = 0$. So we have a simple random walk (X_t, Y_t) taking steps $(-1, 0)$ or $(0, -1)$, each with probability $1/2$, started from $(B, k - B)$ and ended on hitting one of the axes $x = 0$ or $y = 0$. We can answer some questions for this model. For example, the total number of coalescences is the hitting time

$$T = \min\{t \geq 0 : X_t = 0 \text{ or } Y_t = 0\}. \tag{6.1}$$

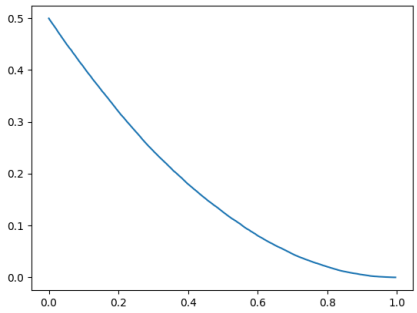

Figure 6: A plot of $\frac{1}{n}Z_{nt}$ for $t \in [0,1]$, where $n = 10^5$ and Z_i is the number of clusters at step i in the heatseeker meanfield process initialized with n missiles. Note that $Z_0 = n/2 + O(\sqrt{n})$, since initially the number of disagreements is exactly Binomial$(n, 1/2)$. The number of steps taken until the process ends (i.e. $Z_t = 0$) is $n + O(\sqrt{n})$, though this is not completely obvious a priori.

The asymptotics of T can be worked out explicitly: for example, it should be easy to see that the total number of steps needed is $n+O(\sqrt{n})$. But it won't help us understand Z_t or the cluster size distribution.

Trying to track the sizes of the 'clusters' of arrows all pointing the same direction in this model is an interesting question. At each time t the cycle of length $k-t$ is divided into connected clusters of arrows all pointing the same direction. The number of disagreements is the same as the number of clusters: associate each disagreement to the connected cluster preceding it. This number, call it Z_t, decreases by 0 or 1 at each step: it decreases by 1 if a disagreement associated to a cluster of size 1 is chosen for deletion, in which case the two neighboring clusters, say of sizes c and c' merge into a single cluster of size $c + c'$; it decreases by 0 if a disagreement associated to a cluster of size ≥ 1 is chosen for deletion, in which case that cluster decreases its size by 1, but the disagreement is still there (and nothing else is affected).

We are now entering the territory of 'fragmentation/coagulation' processes, for which there may be a nice hydrodynamic limit description (see figure 6 for a plot of Z_t, which appears to have a deterministic scaling limit).

7 The *continuous* coupon collector process

Consider the following 'continuous' version of the coupon collector problem. Start with a circle of unit circumference $\mathbb{T}$. Fix a number $\theta \in (0,1)$, and successively choose random points $z_j \in \mathbb{T}$ according to uniform measure on $\mathbb{T}$ for $j = 1, 2, \ldots$. Form random arcs $I_j \subset \mathbb{T}$ of length θ, centered at z_j, and consider the set of 'collected' points $T_n = \bigcup_{j \le n} I_j$. Our main goal will be to understand the stopping time

$$\tau_\theta = \min\{t : T_t = \mathbb{T}\}. \tag{7.1}$$

Question 7.1. *What is*

$$\lim_{\theta \to 0} -\frac{\theta \mathbb{E}\tau_\theta}{\log \theta}, \tag{7.2}$$

if it exists?

Question 7.2. *Suppose the process has run for some time, and we are given a set $A \subset \mathbb{T}$ which is a union of arcs of length θ. Set τ_A to be the first time t such that $A \cup \bigcup_{s=1}^{t} I_s = \mathbb{T}$. Which sets A of fixed length l maximize/minimize $\mathbb{E}\tau_A$?*

(Note: the maximization problem would be silly if we would allow A to be any Lebesgue measurable set (or even any union of arcs of arbitrarily small length), since then the maximum would occur when A is a dense subset of $\mathbb{T}$ of Lebesgue measure L, and we would have $\mathbb{E}\tau_A = \mathbb{E}\tau_\theta$ for that A.)

Conjecture 7.3. *The minimum value occurs when A is an arc of length l, and the maximum occurs when A is a union of approximately l/θ 'equally spaced' arcs of length θ.*

For convenience, suppose $\theta = \frac{1}{N}$ for some positive integer N. Partition S into N fixed, disjoint, connected, open arcs $\{A_k^N : 1 \le k \le N\}$ of angle measure $\theta = \frac{1}{N}$, and consider the stopping time

$$\sigma_N = \min\{t : \text{ each } A_k^N, 1 \le k \le N, \text{ contains at least one } z_j, 1 \le j \le t\}. \tag{7.3}$$

Then $\sigma_N \le \tau$ deterministically, since if some arc A_k^N contains no point z_j, then the center of that arc cannot belong to T. Moreover, σ is an instance of the classical coupon collector process on N coupons. Thus

$$\mathbb{E}\tau \ge \mathbb{E}\sigma \approx N \log N = \frac{1}{\theta} \log \frac{1}{\theta} = -\frac{1}{\theta} \log \theta. \tag{7.4}$$

Similarly, $\tau \le \sigma_{2N}$ deterministically, since if every arc A_k^{2N} contains a point, then every such arc is covered by T, and thus T covers $\mathbb{T}$. So

$$\mathbb{E}\tau \le \mathbb{E}\sigma_{2N} \approx -\frac{2}{\theta} \log \theta. \tag{7.5}$$

This is enough to see that

$$-\frac{\theta \mathbb{E}\tau_\theta}{\log \theta} \in [1, 2], \quad (\theta \to 0) \tag{7.6}$$

One can also consider the length covered by time n: set $L_n(\theta) = L_n = |T_n|$. In perfect analogy with the discrete coupon collector process, we have

$$\mathbb{E}L_n = \int_{\mathbb{T}} \mathbb{P}(w \in T_n)\, dw = \int_{\mathbb{T}} \left(1 - (1-\theta)^n\right)\, dw = 1 - (1-\theta)^n\,. \tag{7.7}$$

A more complicated quantity is $\mathbb{P}(w, z \in T_n)$ for two points $w \neq z$ on the circle: as usual, this is related to the second moment of L, via

$$\begin{aligned}
\mathbb{E}L_n^2 &= \int_\Omega L_n^2 \\
&= \int_\Omega \left(\int_{\mathbb{T}} 1\{w \in T_n\} dw \right)^2 \\
&= \int_\Omega \int_{\mathbb{T}} \int_{\mathbb{T}} 1\{z, w \in T_n\}\, dz\, dw \\
&= \int_{\mathbb{T}^2} \mathbb{P}(z, w \in T_n)\, dz\, dw.
\end{aligned}$$

(Here $\Omega = \Omega_n$ denotes the underlying probability space, and we have suppressed the implied probability measure for T_n.)

It is straightforward to check that

$$\begin{aligned}
\mathbb{P}(w, z \in T_n) &= 1 - \mathbb{P}(z \notin T_n) - \mathbb{P}(w \notin T_n) + \mathbb{P}(z, w \notin T_n) \\
&= 1 - 2\,(1-\theta)^n + \begin{cases} (1-2\theta)^n\,, & d_{\mathbb{T}}(z,w) > \theta \\ 2(1-\theta)^n\,(1 - \theta + d_{\mathbb{T}}(z,w))^n\,, & d_{\mathbb{T}}(z,w) \leq \theta \end{cases}
\end{aligned}$$

where $d_{\mathbb{T}}$ denotes the angle distance on $\mathbb{T}$, i.e. $d_{\mathbb{T}}(e^{i\alpha}, e^{i\beta}) = |\alpha - \beta| \in [0, \pi]$. The formula for the case when $d_{\mathbb{T}}(z,w) < \theta$ is the hard part: it boils down to the observation

$$\begin{aligned}
\mathbb{P}(z, w \notin T_n) &= \mathbb{P}(z, w \notin A_j \forall j \text{ and either } z \notin A_j \forall j \text{ or } w \notin A_j \forall j) \\
&= \mathbb{P}(z, w \notin A_j \forall j) \cdot 2\mathbb{P}\left(z \notin T_n \middle| z, w \notin A_j \forall j \right)
\end{aligned}$$

Integrating directly, we get

$$\begin{aligned}
\mathbb{E}L_n^2 = 1 - 2(1-\theta)^n + (1-2\theta)^{n+1} + \frac{2}{n+1}\Big(&(1-\theta)^{n+2} - (1-\theta)(1-2\theta)^{n+1} \\
&+ \theta - \frac{1}{n+2}\left(1 - (1-\theta)^{n+2}\right) \Big)
\end{aligned}$$

Using the scaling $\theta = \frac{\alpha \log n}{n}$ for $\alpha > 0$, these formulas become

$$\mathbb{E}L_n \approx 1 - n^{-\alpha} \tag{7.8}$$

and

$$\begin{aligned}\mathbb{E}L_n^2 &\approx 1 - 2n^{-\alpha} + \left(1 - \frac{2\alpha}{n}\right) n^{-2\alpha} + \frac{2}{n+2}\left(\frac{n+1}{n+2}\left(1 - \frac{\alpha}{n}\right)^2 n^{-\alpha}\right. \\ &\quad \left. - \left(1 - \frac{\alpha}{n}\right)\left(1 - \frac{2\alpha}{n}\right) n^{-2\alpha} + \frac{\alpha}{n} - \frac{1}{n+2}\right) \\ &= (1 - n^{-\alpha})^2 + \frac{2}{n} n^{-\alpha}(1 - (1+\alpha)n^{-\alpha}) + o(n^{-1})\end{aligned}$$

Thus, the variance is (to leading order)

$$\mathrm{Var}(L_n) = \mathbb{E}L_n^2 - (\mathbb{E}L_n)^2 \approx 2n^{-1-\alpha}\left(1 - (1+\alpha)n^{-\alpha}\right) \tag{7.9}$$

Chebychev's inequality gives the crude estimate

$$\mathbb{P}(|L_n - 1 + n^{-\alpha}| \geq n^{-1/2}) \leq 2n^{-\alpha}. \tag{7.10}$$

Idea: It would be nice to show that if L_n is 'close enough' to 1, then it equals 1 with high probability. (It seems unlikely that there is a (very) small gap missing in the union of arcs.) Perhaps there is a sub-martingale approach?

8 Intersection of random sets

(Half planes) Consider n iid half planes $H_j, j = 1, \ldots, n$, given by $H_j = \{p \in \mathbb{R}^2 : \langle p, e^{i\theta_j} \rangle \leq R_j\}$, where $\theta_j, j = 1, \ldots, n$ are iid uniform on $[0, 2\pi)$, and R_j are iid according to some distribution F supported on $(0, \infty)$. Set $I_n = \bigcap_{j=1}^n H_j$. What does I_n look like? Since all the H_j are convex, I_n is convex for all n. Also, except in the case where R_n is identically 0, I_n is compact for n sufficiently large. Because the θ_j are uniform, I_n has rotational symmetry in distribution, in the sense that any rotation of I_n about the origin is equal in distribution to I_n; also, the 'radial' random variables

$$Q_n(\theta) = \sup\{t > 0 : te^{i\theta} \in I_n\} \tag{8.1}$$

are identically distributed (though, of course not independent!) for $\theta \in [0, 2\pi)$.

Example 1: Suppose that $F = \delta_0$, i.e. all the R_j are 0. Then we are sampling random half planes through the origin in $\mathbb{R}^2$, and taking the intersection. The intersection will be a single point for some finite n almost surely, since each time an additional half plane is added there is a positive probability that it intersects I_n only in the origin (for $n \geq 2$). Since all the planes pass through the origin, I_n is a cone through the origin. In fact, the angle ψ_n subtended by I_n is a simple Markov process in discrete time on $[0, 1]$, which is almost surely decreasing: it has transition probabilities

$$\psi_{n+1} = \begin{cases} 0 & \text{w.p. } \frac{1}{2} - \frac{\psi_n}{2\pi} \\ \psi_n & \text{w.p. } \frac{1}{2} - \frac{\psi_n}{2\pi} \\ x & \text{w.p. } \frac{1}{\pi} dx, 0 \leq x \leq \psi_n \end{cases} \tag{8.2}$$

These three events correspond to H_{n+1} missing, containing, or partly intersecting I_n. One can easily compute the conditional expectation of ψ_{n+1} given ψ_n:

$$\mathbb{E}\left[\psi_{n+1} | \psi_n\right] = \psi_n \left(\frac{1}{2} - \frac{\psi_n}{2\pi}\right) + \int_0^{\psi_n} x \cdot \frac{1}{\pi}\, dx = \frac{1}{2}\psi_n. \tag{8.3}$$

Thus, setting $\psi_0 = 2\pi$, $\frac{1}{2\pi}\mathbb{E}\psi_n = 2^{-n}$.

Exercise 8.1. *More generally, if the half spaces are replaced by cones of angle θ, this should be equivalent to covering the circle by arcs of length $2\pi - \theta$... check this...*

What is the expected first time that $\psi_n = 0$?

Example 2: Suppose that $F(s) = s^2 1\{s \in [0, 1]\} + 1\{s \geq 1\}$. This is equivalent to choosing random points from area measure on the unit disk, and drawing half planes perpendicular to the radial lines to those points. In this case,

$$Q_n(\theta) = \min_j \{R_j \sec|\theta - \theta_j| : |\theta - \theta_j| < \pi/2\}. \tag{8.4}$$

Thus we have

$$\mathbb{P}(Q_j(\theta) \leq q) = \mathbb{P}(R_j \sec|\theta - \theta_j| \leq q)^n, \tag{8.5}$$

where $\sec|\theta - \theta_j|$ is taken to be infinite if $|\theta - \theta_j| \geq \pi/2$. This can be evaluated directly:

$$\mathbb{P}(R_j \sec|\theta-\theta_j| \le q) = \frac{1}{2}\mathbb{P}\left(R_j \sec|\theta-\theta_j| \le q \Big| |\theta-\theta_j| < \pi/2\right) \tag{8.6}$$

$$= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2}\int_0^1 1\{s\sec\phi \le q\}\cdot 2s\,ds\,d\phi \tag{8.7}$$

$$= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2}\int_0^{q\cos\phi} s\,ds\,d\phi \tag{8.8}$$

$$= \frac{q^2}{4}. \tag{8.9}$$

It follows that

$$\mathbb{P}(Q_n(\theta) \le q) = 1-(1-q^2/4)^n, \tag{8.10}$$

and thus

$$\frac{n}{4}Q_n^2 \to_d Exp(1). \tag{8.11}$$

Roughly speaking, this suggests

$$\text{Area}(I_n) \approx \frac{4\pi}{n}\cdot Exp(1) \tag{8.12}$$

for n large.

9 Asymptotics of hitting times and expected value

Consider any process $X_n, n \in \mathbb{N}$ (or more generally, X_t for $t \geq 0$). Consider also the hitting times $\tau_k = \min\{n \in \mathbb{N} : X_n \geq k\}$, for $k \in \mathbb{N}$ associated to X. If X is a partial sum process, e.g. $X_n = \sum_{k=1}^{n} Y_k$ where $Y_k \sim Y$ are iid, $\mathbb{E}Y < \infty$ and $Y \in [0, \infty)$, the LLN for renewal processes implies that

$$\frac{1}{k}\tau_k \to \frac{1}{\mathbb{E}Y} \text{ almost surely and in } L^1. \tag{9.1}$$

In general, X_n and τ_k should be inverse functions, roughly speaking. We explore under what conditions such a statement is true.

Example 1: X = discrete coupon collector process, with N coupons (see section 2). Then $\mathbb{E}X_n = N\left(1 - \left(1 - \frac{1}{N}\right)^n\right)$, and $\mathbb{E}\tau_k = N(H_N - H_{N-k})$, where H_j is the jth harmonic number, $H_j = 1 + \frac{1}{2} + \cdots + \frac{1}{j}$. Note that $\mathbb{E}\tau_k \approx -N \log\left(1 - \frac{k}{N}\right)$. These functions are 'near' inverses for N large:

$$\mathbb{E}X_{\tau_k} = k, \tag{9.2}$$

since X_{τ_k} is always equal to k, and

$$\mathbb{E}\tau_{X_n} \approx -N \log\left(1 - \left(1 - \left(1 - \frac{1}{N}\right)^n\right)\right) = -nN \log\left(1 - \frac{1}{N}\right) \approx n. \tag{9.3}$$

Example 2: X = Poisson process on $\mathbb{R}$ with rate λ. Then $\mathbb{E}X_t = \lambda t$, while

$$\mathbb{E}\tau_y = \int_{\mathbb{R}} \mathbb{P}(\tau_y \geq t)dt = \int_{\mathbb{R}} \mathbb{P}(X_t \leq y)dt, \tag{9.4}$$

which yields $\mathbb{E}\tau_y \approx \frac{y}{\lambda}$, by using the fact that X_t is Poisson distributed with mean λt.

We can prove the following. First, note that deterministically, $X_{\tau_k} = k$. Also, if X is non-decreasing,

$$\tau_{X_n} = \min\{m : X_m \geq X_n\} \leq n, \tag{9.5}$$

which implies

$$\limsup_n \frac{\tau_{X_n}}{n} \leq 1. \tag{9.6}$$

Proposition 9.1. *If X is non-decreasing and $\limsup_n \mathbb{P}(X_n = X_{n+1}) < 1$, then $\lim_{n\to\infty} \frac{\tau_{X_n}}{n} = 1$.*

Proof. The assumptions imply that for any $\epsilon > 0$, there exists $c < 1$ such that

$$\mathbb{P}(X_n = X_{n+1} = \cdots = X_{n+\epsilon n}) \leq c^{\epsilon n}. \tag{9.7}$$

By the Borel Cantelli lemma, since the above probabilities are summable over n (for fixed ϵ), the event

$$X_n = X_{n+1} = \cdots = X_{n+\epsilon n} \tag{9.8}$$

occurs finitely often almost surely. Thus

$$\frac{\tau_{X_n}}{n} \geq 1 - \epsilon \tag{9.9}$$

for n large; now let $\epsilon \to 0$. □

Thus, under the assumptions of the proposition, $\tau_{X_n} \sim n$ almost surely as $n \to \infty$.

Question: Can we weaken the assumptions to include the case where X is not necessarily non-decreasing, or where X is constant on large invervals (as with the coupon collector)?

10 Number of maximums of iid random variables

Let (X_i) be any iid sequence with common distribution X. Let $M_n = \max\{X_1, \ldots, X_n\}$. If X is continuous, then

$$\mathbb{P}(X_n = M_n) = \mathbb{P}(X_n > M_{n-1}) = \frac{1}{n}, \tag{10.1}$$

since there is a unique maximum among the n values. The situation is more interesting if X is has atoms. The event $\{X_n = M_n)$ becomes more likely, since conditioning on the number of maxes yields

$$\mathbb{P}(X_n = M_n) = \frac{1}{n}\mathbb{E}|\mathrm{Max}_n|, \tag{10.2}$$

where $\mathrm{Max}_n = \{i \in [n] : X_i = M_n\}$ and $\mathbb{E}|\mathrm{Max}_n| > 1$ if X is atomic. On the other hand, $\mathbb{P}(X_n > M_{n-1})$ can be as small as exponential in n (for example, when X is Bernoulli). Assume X takes values in $\mathbb{Z}$, and denote

$$p_k = \mathbb{P}(X = k), \text{ and } \mathbb{P}_{<k} = \sum_{j<k} p_j. \tag{10.3}$$

The distribution of Max looks like

$$\mathbb{P}(\mathrm{Max}_n = m) = \sum_{k\in\mathbb{Z}} \binom{n}{m} p_k^m p_{<k}^{n-m}. \tag{10.4}$$

As a check, the binomial theorem gives

$$\begin{aligned}
\sum_{m=1}^{n} \mathbb{P}(\mathrm{Max}_n = m) &= \sum_{m=1}^{n}\sum_{k} \binom{n}{m} p_k^m p_{<k}^{n-m} && (10.5)\\
&= \sum_{k}\sum_{m=1}^{n} \binom{n}{m} p_k^m p_{<k}^{n-m} && (10.6)\\
&= \sum_{k} p_{\le k}^n - p_{<k}^n && (10.7)\\
&= 1, && (10.8)
\end{aligned}$$

since the last sum telescopes. A similar calculation gives a formula for the expected size of Max:

$$\begin{aligned}
\mathbb{E}|\mathrm{Max}_n| &= \sum_{k}\sum_{m=1}^{n} m\binom{n}{m} p_k^m p_{<k}^{n-m} && (10.9)\\
&= n\sum_{k} p_k \sum_{j=0}^{n-1} \binom{n-1}{j} p_k^j p_{<k}^{n-j-1} && (10.10)\\
&= n\sum_{k} p_k p_{\le k}^{n-1} && (10.11)\\
&= n\mathbb{E}p_{\le X}^{n-1} && (10.12)
\end{aligned}$$

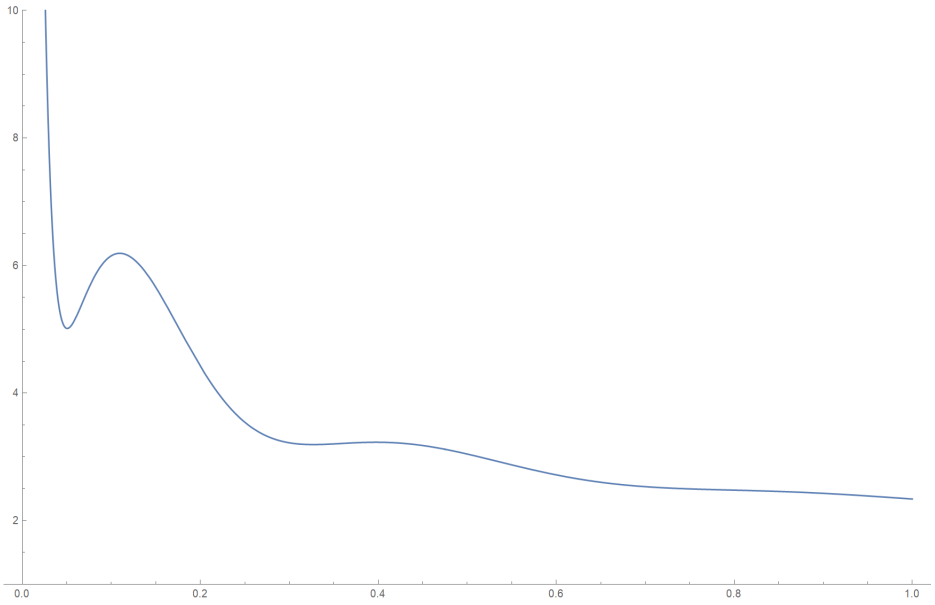

Figure 7: Approximate value of $\mathbb{E}|\text{Max}_n|$ for $X \sim \text{Poisson}(\lambda)$, with $n = 10^2$ and $\lambda \in (0, 1)$. This plot is way more interesting than it should be. The value appeared to smoothly decrease for $\lambda > 1$, though to what limit I'm not sure. What is happening in $(0, 1)$?! Could be a rounding issue...

This value is surprisingly difficult to calculate, even for nice distributions. For example, when X is Geometric(p), i.e. $p_k = pq^k$ for $k \geq 0$, we get the expression

$$\mathbb{E}|\text{Max}_n| = n \sum_k pq^k (1 - q^{k+1})^{n-1}. \tag{10.13}$$

For $p = \frac{1}{2}$ and n large, Mathematica gives the value $\mathbb{E}|\text{Max}| \approx 1.442689883$. (Attempts to evaluate these expressions have been in vain. Differentiating with respect to p goes nowhere fast, for example. Mathematica also seems stumped. The expression is valid for all real values of n, perhaps something can be done with that?)

Question 10.1. *For some nice class of distributions X, show that*

$$\lim_{n \to \infty} \mathbb{E}|Max_n| \text{ exists } \in (0, \infty), \tag{10.14}$$

and compute the limit.

The limit seems surprisingly difficult to compute, even for Poisson or Geometric.

Alternatively, we can work with the probability that X_n is the unique maximum. We have

$$\mathbb{P}(X_n > M_{n-1}) = \sum_k p_k p_{<k}^{n-1}. \tag{10.15}$$

It seems that this probability should be on the same order as $\mathbb{P}(X_n = M_n)$, i.e. roughly n^{-1}, under some mild conditions, but again the formulas are difficult to work with. Perhaps having all of $\mathbb{N}$ as support is sufficient?

Conjecture 10.2. *Show that if $X \sim Poisson(\lambda)$, for any $\lambda > 0$,*

$$\liminf_n n\mathbb{P}(X_n > M_{n-1}) > 0. \tag{10.16}$$

If this fails, it should fail for all λ, since otherwise there would be a phase transition point in λ, which would be too interesting. I suspect it holds for all λ.

Question 10.3. *Is there some class of distributions, supported on $\mathbb{N}$, such that*

$$\limsup_n n\mathbb{P}(X_n > M_{n-1}) = 0? \tag{10.17}$$

I suspect not – can this be proved?

One concrete example is when X_n are iid Uniform on $(0, k)$ for fixed k. Then

$$\mathbb{P}(X_n > M_{n-1}) = k^{-n} \sum_{i=1}^{k-1} i^{n-1} \approx n^{-1}(1 - k^{-1})^n, \tag{10.18}$$

so this probability decays exponentially. Alternatively, if we let $k = k_n$ grow with n, a natural choice being $k = \alpha n$ for $\alpha \in \mathbb{R}$, we obtain

$$\mathbb{P}(X_n > M_{n-1}) \approx e^{-\alpha^{-1}} n^{-1}, \tag{10.19}$$

i.e. the probability that a density α^{-1} Poisson process on $\mathbb{Z} \cap [1, n]$ has a unique maximum value.

11 Ruler distribution

Consider the following general class of distribution functions. Fix a function $f : \mathbb{R} \to \mathbb{R}^+$ with $\int f = 1$. (f may have atoms.) Then define the 'ruler' random variable X_f by

$$\mathbb{P}(X_f \leq x) = F_f(x) = \int_{-\infty}^{\infty} f(t) 1\{f(t) \leq x\}\, dt. \tag{11.1}$$

That is, the probability $\mathbb{P}(X_f \leq x)$ is obtained by putting a ruler at height x over the function f, and integrating below it. Observe that

$$\begin{aligned}
\mathbb{E}X &= \int_{\mathbb{R}} (1 - F_f(x))\, dx && (11.2)\\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} f(t) 1\{f(t) > x\} dt\, dx && (11.3)\\
&= \int_{\mathbb{R}} f(t) \int_0^{f(t)} dx\, dt && (11.4)\\
&= \int_{\mathbb{R}} f(t)^2\, dt && (11.5)
\end{aligned}$$

Loosely speaking, the density of f is given by

$$\begin{aligned}
\frac{d}{dx}\mathbb{P}(X \leq x) &= \int_{\mathbb{R}} f(t) \delta_x(f(t))\, dt && (11.6)\\
&= x \cdot |f^{-1}(x)| && (11.7)
\end{aligned}$$

Question 11.1. *Under what conditions on f does this actually hold? (When is it kosher to pass the derivative through the integral?)*

When f is a sum of atoms, i.e. $f = \sum_{n \geq 0} \delta_n p_n$, we obtain

$$\mathbb{P}(X_f = p_n) = \sum_{m : p_m = p_n} p_m = \#\{m : p_m = p_n\} p_n. \tag{11.8}$$

Thus, in this case,

$$\mathbb{E}X_f^\alpha = \sum_n p_n^{\alpha+1} \tag{11.9}$$

for any $\alpha \in \mathbb{R}$.

Conjecture 11.2. *Does it hold in general that*

$$\mathbb{E}X_f^\alpha = \int f^{\alpha+1}? \tag{11.10}$$

This could make good exercise for a probability course. Does it have any other value?

12 Random decreasing sequences

This note describes a natural model for a random sequence of probabilities $p_i \in (0,1)$ with $\sum_i p_i = 1$, which is surprisingly difficult to work with. Let F be any distribution function on $[0,1]$ other than $F = \delta_1$, the unit mass at 1 (up to zero measure changes), and $\{X_n\}_{n\in\mathbb{N}} \sim F$ i.i.d. Set $Y_1 = X_1$, and for $n > 1$,

$$Y_i = X_i Y_{i-1} = X_1 X_2 \cdots X_i. \tag{12.1}$$

First note that for any $n \in \mathbb{N}$ and $k \in \mathbb{R}$,

$$\mathbb{E}(Y_n^k) = \mathbb{E}(X_1^k \cdots X_n^k) = (\mathbb{E}X^k)^n \tag{12.2}$$

Let $S = \sum_{n\in\mathbb{N}} Y_n$. We have:

Proposition 12.1. *$S < \infty$ almost surely.*

Proof. Let μ denote the mean of F. Since F is not the atom at $1, \mu < 1$, so

$$\mathbb{E}S = \sum_n \mathbb{E}Y_n = \sum_n \mu^n = \frac{\mu}{1-\mu} < \infty. \tag{12.3}$$

□

If we want to use the Y_i as probabilities, it is natural to normalize by S. This leads us to:

Question 12.2. *What is the distribution of S?*

It seems like this question does not have a satisfying answer. The distribution of Y_n can be readily computed, using the following recursive formula (which is just convolution for products of independent random variables):

Proposition 12.3. *Suppose F has a continuous density f, and for $n \in \mathbb{N}$ let g_n denote the density of Y_n. Then g_n exists, with $g_1 = f$, and for $n > 1$,*

$$g_n(y) = \int_y^1 \frac{1}{x} f(y/x) g_{n-1}(x)\, dx. \tag{12.4}$$

We can do the usual thing with logs:

$$\log Y_n = \sum_{i=1}^n \log X_i. \tag{12.5}$$

If F is well behaved, so that the $\log X_i$ satisfy a CLT, then we'll have that $\log Y_n$ is approximately normal, on order n, i.e. Y_n is approximately log-normally distributed with parameters on order n. This at least gives us an approximation for the distribution of S: it is roughly an infinite sum of log normals, namely $L_n \sim \text{LogNormal}(n\mathbb{E}\log X, n\text{Var}(X))$ for $n \geq 1$.

Another approach is to write a distributional equation for S:

$$S \sim XS + X', \tag{12.6}$$

where $\sim$ denotes equality in distribution (and, as usual, we have mildly abused notation: the two copies of S are independent). Let H denote the CDF of S, i.e. $H(s) = \mathbb{P}(S \leq s)$. For any $s > 0$, 12.6 gives TBD

13 Fisherman's Dilemma

Suppose a fisherman is sitting in his boat on the lake, waiting for fish. He can re-cast his line at any time. What is the optimal strategy to catch the most fish? Assume that each time the fisherman casts his line, he has to wait a random amount of time before he gets a bite. Specifically, fix a distribution function F and a random variable X distributed like F, and let $\{X_n\}_{n\in\mathbb{N}}$ be iid copies of X. Then the nth time the fisherman casts his line, a fish will bite after time X_n. The fisherman chooses a strategy for maximizing his haul over time: he may choose to re-cast his line at any time, even if he hasn't seen a fish. We can ask different questions with this setup: if the fisherman knows F, what is his optimal strategy? What if he does not know F? What if it takes a fixed amount of time to reel in and recast the line?

It is easiest to understand the case where F is known to the fisherman. Since the waiting times between bites are independent, any optimal strategy should be of the form: pick a time $t \in \mathbb{R}$, and always re-cast at that time (if a fish hasn't yet arrived). Given such a strategy, for $k \in \mathbb{N}$, let T_k denote the time between catching our $(k-1)$st and kth fish, with $T_0 = 0$ by convention. Also, use $F(x) = \mathbb{P}(X > x)$ for the distribution of the X's, $p = F(t)$ and $q = 1 - p$. The T_k are iid, and

$$\mathbb{E}T_1 = \sum_{j\geq 0} \mathbb{P}(X > t)^j \mathbb{P}(X \leq t)\Big(jt + \mathbb{E}[X|X \leq t]\Big) \tag{13.1}$$

$$= q\Big(t\sum_{j\geq 0} jp^j + \sum_{j\geq 0} p^j \mathbb{E}[X|X \leq t]\Big) \tag{13.2}$$

$$= q\Big(\frac{tp}{q^2} + \frac{1}{q}\int_{\mathbb{R}} \frac{\mathbb{P}(s < X \leq t)}{q} ds\Big) \tag{13.3}$$

$$= \frac{1}{q}\Big(tp + \int_0^t (F(s) - p) ds\Big) \tag{13.4}$$

$$= \frac{1}{1 - F(t)} \int_0^t F(s) ds. \tag{13.5}$$

Set $\mu_t(F) = \mathbb{E}T_1(t, F)$ to be the mean time to catch one fish using the always-re-cast-at-t strategy, if the distribution of the X's is F. $\mu_t(F)$ is closely related to the hazard function of F, $h(t) = -F'(t)/F(t)$. By L'hopital's rule,

$$\lim_{t\to 0^+} \mu_t(F) = \lim_{t\to 0^+} \frac{\int_0^t F(s) ds}{1 - F(t)} = \lim_{t\to 0^+} \frac{F(t)}{-F'(t)} = \lim_{t\to 0^+} \frac{1}{h(t)} = -1/F'(0). \tag{13.6}$$

Given F, our job is simply to minimize the function μ_t over $t \in \mathbb{R}$: any such minimum value yields an optimal strategy. If F is differentiable, then

$$\frac{d\mu_t}{dt} = \frac{d}{dt}\Big[\frac{1}{1 - F(t)} \int_0^t F(s) ds\Big] = \frac{F(t)}{1 - F(t)} + \frac{F'(t)}{(1 - F(t))^2} \int_0^t F(s) ds. \tag{13.7}$$

Thus

$$\frac{d\mu_t}{dt} = 0 \iff F(1 - F) = -F' \int_0^t F(s) ds. \tag{13.8}$$

So finding an optimal strategy boils down to solving this equation.

There is one special distribution for this process: any exponential distribution $F(x) = e^{-\lambda x}$, $\lambda > 0$. Because of the memory-less property of the exponential distribution, μ_t is constant when F is exponential:

$$\mu_t(\text{Exponential}(\lambda)) = \frac{1}{1-e^{-\lambda t}} \int_0^t e^{-\lambda s} ds = \frac{1}{\lambda} \frac{-e^{-\lambda t}+1}{1-e^{-\lambda t}} = \frac{1}{\lambda}. \tag{13.9}$$

In fact, one can easily prove that the exponential distribution is the *only* distribution for which μ_t is always constant. Indeed, if $\mu_t \equiv 1/\lambda$, then cross-multiplying and differentiating (using the FTC) yields

$$-\frac{1}{\lambda} F'(t) = F(t), \tag{13.10}$$

and this ODE, subject to $F(0) = 1$ and $F(\infty) = 0$ only has the solution $F(x) = e^{-\lambda x}$.

One can actually do better than just the expectation of T_1: it is straightforward to compute the moment generating function of T_1. Using the same notation as above, for $z \in (0, p^{-1/t})$ we have

$$\begin{aligned}
\mathbb{E}z^{T_1} &= \sum_{j \geq 0} p^j q z^{jt} \mathbb{E}[z^X | X \leq t] && (13.11)\\
&= q\mathbb{E}[z^X | X \leq t] \cdot \sum_{j \geq 0} (p \cdot z^t)^j && (13.12)\\
&= q\Big(\int_{\mathbb{R}} \frac{\mathbb{P}(s < z^X \leq z^t)}{q} ds \Big) \cdot \frac{1}{1-pz^t} && (13.13)\\
&= \frac{\log z}{1-pz^t} \int_0^t z^u F(u) du. && (13.14)
\end{aligned}$$

This has all the information about the distribution of T_1: for example,

$$\mathbb{P}(T_1 = m) = \frac{1}{m!} \frac{d^m}{dz^m}\Big|_{z=0} \mathbb{E}z^{T_1}. \tag{13.15}$$

The moments can also be recovered directly, though it is easier if one substitutes $z = e^r$: then

$$\mathbb{E}T_1^l = \frac{d^l}{dr^l}\Big|_{r=0} \mathbb{E}e^{rT_1}. \tag{13.16}$$

14 Limits of multiplicative functions on $\mathbb{N}$

Consider the average sum of divisors function

$$\sigma(n) = \frac{1}{n}\sum_{d|n} d = \prod_p \frac{1-p^{-\alpha_p-1}}{1-p^{-1}}, \tag{14.1}$$

where p ranges over the primes, and $n = \prod_p p^{\alpha_p}$. There are many questions about the asymptotic behavior of σ. Much work has gone into studying the random variable $\sigma(X_n)$, where X_n is uniformly distributed on $[1, n]$. It can be shown that $\sigma(X_n)$ converges in distribution using the Erdős-Wintner theorem. An alternative model is to consider the probability space $\Omega = \mathbb{N}^{\mathbb{N}}$, equipped with the product measure of geometric random variables Y_p, namely for $k = 0, 1, \ldots$,

$$\mathbb{P}(Y_p = k) = (1-p^{-1})p^{-k} \tag{14.2}$$

for any prime p. Thus an element of Ω looks like $(Y_2, Y_3, Y_5, \ldots)$, which corresponds to the 'number' $\prod_p p^{Y_p}$. This extends the random variables X_n in a natural way to the limit space Ω, since the largest power of p dividing X_n is very close to Y_p in distribution (and jointly over p). More generally:

Question 14.1. *Suppose f is a multiplicative function on $\mathbb{N}$, so we can write*

$$f(n) = \prod_p f(p)^{\alpha_p}. \tag{14.3}$$

Define

$$\widetilde{f} = \prod_p f(p)^{Y_i} \tag{14.4}$$

Show that

$$f(X_n) \to_d \widetilde{f} \tag{14.5}$$

and quantify the convergence rate. There are many such functions of interest on $\mathbb{N}$: does this construction offer any insight?

(Here we are thinking that $f(n) = \Theta(1)$ so the limit is non-trivial, which can usually be achieved by scaling appropriately.) Alternatively, one could define $\widetilde{f}$ as a limit, by truncating the sequence of primes in some way. For example, one could look at sequences of exponents Y_p for $p \le q$, and let $q \to \infty$; or, in addition, condition that $n = \prod_p p^{Y_p} \le N$ for some N, and let $N, q \to \infty$.

A related question would be to possibly simplify the proof of the following theorem, due to Erdős and Wintner (1939):

Theorem 14.2. *A multiplicative function f has a limiting distribution if and only if the following three series converge, where $g(p) = \log f(p)$:*

$$\sum_{|g(p)|>1} \frac{1}{p}, \sum_{|g(p)|\le 1} \frac{g(p)^2}{p}, \sum_{|g(p)|\le 1} \frac{g(p)}{p}. \tag{14.6}$$

Moreover, the characteristic function of the limit law is given by

$$\phi(\tau) = \prod_p \left(1-\frac{1}{p}\right)\sum_{\nu\ge 0} \frac{p^{i\nu\tau}}{p^\nu}. \tag{14.7}$$

I took a look at the proof which seems quite complicated, though maybe this idea is already hidden inside there somewhere.

Returning to σ, which we think of as a random variable on Ω given by

$$\sigma = \prod_p \frac{1 - p^{-Y_p - 1}}{1 - p^{-1}}, \tag{14.8}$$

note that

$$\mathbb{E}p^{-jY_p} = \frac{1 - p^{-1}}{1 - p^{-j-1}} \tag{14.9}$$

for each prime p and $j \geq 1$. Thus,

$$\begin{aligned}
\mathbb{E}\sigma &= \prod_p (1 - p^{-1})^{-1} \left(1 - \mathbb{E}p^{-Y_p - 1}\right) && (14.10)\\
&= \prod_p (1 - p^{-1})^{-1} \left(1 - \sum_{k \geq 0} (1 - p^{-1}) p^{-2k-1}\right) && (14.11)\\
&= \prod_p (1 - p^{-1})^{-1} \left(1 - \frac{p^{-1}(1 - p^{-1})}{1 - p^{-2}}\right) && (14.12)\\
&= \prod_p \frac{1}{1 - p^{-1}} - \frac{p^{-1}}{1 - p^{-2}} && (14.13)\\
&= \prod_p \frac{1}{1 - p^{-2}} && (14.14)\\
&= \zeta(2). && (14.15)
\end{aligned}$$

In particular, this shows $\sigma < \infty$ almost surely. Similarly, one can obtain

$$\mathbb{E}\sigma^2 = \prod_p \frac{p^{-4} - 2p^{-2} - p^{-1} - 1}{(p^{-1} - 1)(p^{-1} + 1)(p^{-2} + p^{-1} + 1)} = \prod_p \left(1 + \frac{3p^2}{p^4 + p^3 - p - 1}\right), \approx 2.0999 \tag{14.16}$$

and for any $m \in \mathbb{N}$,

$$\mathbb{E}\sigma^m = \prod_p p(1 - p^{-1})^{1-m} \sum_{j=0}^{m} (-1)^j \binom{m}{j} (p^{j+1} - 1)^{-1}. \tag{14.17}$$

Question 14.3. *How can we compute or approximate properties the distribution function for σ? One open question, of Pomerance, is to determine to high precision $\mathbb{P}(\sigma > 2)$, i.e. the probability of being abundant, or more generally $\mathbb{P}(\sigma > u)$ for some fixed u. If we can get all the moments of σ to within arbitrary precision, can we compute such probabilities to high precision?*

(I think typically the moments aren't the right thing to work with when trying to compute probabilities. It is true that the moments determine the distribution if they grow slowly enough, which these probably do, but my sense is that this will lead to difficult questions about the Laplace transform. Maybe we want the Fourier transform ϕ from the EW theorem instead?)

The log of σ is perhaps easier to understand. Using the expansion $\log(1-z) = -\sum_{n\geq 1} \frac{z^n}{n}$,

$$\log \sigma = \sum_p \left[\log(1 - p^{-Y_p - 1}) - \log(1 - p^{-1})\right] \tag{14.18}$$

$$= \sum_p \sum_{n\geq 1} \frac{p^{-n}}{n} \left(1 - p^{-nY_p}\right) \tag{14.19}$$

This is a sum of independent random variables, each of which is relatively simple. Some algebra yields

$$\mathbb{E} \log \sigma = \sum_p \sum_{n\geq 1} \frac{1}{n} \frac{p^n - 1}{p^{2n+1} - p^n} \approx .4457. \tag{14.20}$$

[Note: Mathematica sims suggest that the inner sum is $p^{-2} + O(\exp(-p))$.]