Some combinatorial processes

Jacob Richey

January 31, 2022

Contents

1 The sock process ........................................... 2

2 The coupon collector process ......................... 6

2.1 Generalized coupon collector .......................... 9

3 Non-adjacent uniform placements ................... 11

3.1 MGIS, September 2021 .................................. 14

3.2 Distribution of the configuration .................. 14

4 Adjacent occurrences in a random permutation ...... 17

5 Sum of uniform random variables .................... 20

6 The continuous coupon collector process ........... 22

7 Intersection of random sets ............................ 25

8 Asymptotics of hitting times and expected value .... 27

9 Number of maximums of iid random variables ...... 29

10 Ruler distribution ......................................... 32

11 Random decreasing sequences ....................... 33

12 Fisherman’s Dilemma .................................... 35

13 Bases for \( \mathbb{N} \) ........................................... 37

14 Limits of multiplicative functions on \( \mathbb{N} \) ....... 38
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 $k$th 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 auxiliary 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

$$X_k = \text{ number of un-grouped objects after the } k^{th} \text{ draw} \quad (1.1)$$
$$Y_k = \text{ number of groups removed after the } k^{th} \text{ draw} \quad (1.2)$$

We suppress the parameters $n$ and $l$. Then

$$X_k = k - lY_k, \quad (1.3)$$

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

$$\mathbb{E}Y_k = \sum_{j=1}^{n} \mathbb{P}(\text{every member of group } j \text{ drawn by time } k) \quad (1.4)$$
$$= n \cdot \mathbb{P}(\text{every member of group 1 drawn by time } k) \quad (1.5)$$
$$= n \cdot \left( \frac{\binom{n-1}{k-1}}{\binom{n}{k}} \right) \quad (1.6)$$
$$= n \cdot \left( \frac{\binom{k}{l}}{\binom{ln}{l}} \right), \quad (1.7)$$

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

$$\mathbb{E}X_k = k - (ln) \frac{\binom{k}{l}}{(ln)_l} \quad (1.8)$$
$$\approx k - (ln) \frac{k^l}{\ln^n l^l} \quad (1.9)$$
$$= k \left( 1 - \left( \frac{k}{ln} \right)^{l-1} \right), \quad (1.10)$$

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 $f(t) = t(1 - t^{l-1})$. 

2
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

$$\mathbb{E}Y_k^2 = \mathbb{E}Y_k + (n^2 - n)\left(\frac{\ln(k-2)}{\ln 2}\right)$$

$$= n \cdot \frac{k}{\ln n} + (n^2 - n)\left(\frac{k}{\ln 2}\right).$$

Thus

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

$$= k^2 - (2kl - l^2) \cdot n \cdot \frac{k}{\ln n} + l^2(n^2 - n)\left(\frac{k}{\ln 2}\right),$$

which leads to

$$\text{Var}(X_k) = l^2(n \cdot \frac{k}{\ln n} + (n^2 - n)\left(\frac{k}{\ln 2}\right) - n^2)\left(\frac{k}{\ln 2}\right)^2$$

$$= l^2n\left((k/\ln)^4 + (n - 1)(k/\ln)^2 - n(k/\ln)^2\right)$$

$$= l^2n(k/\ln)^4(1 - (k/\ln)^4).$$

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

$$Z_t^{(n)} = \frac{1}{\sqrt{n}}\left(X_{\lfloor nt \rfloor} - \mathbb{E}X_{\lfloor nt \rfloor}\right)$$
for $t \in (0, 1)$, and the limiting object

\[ "Z_t = \lim_{n \to \infty} Z_t^{(n)} \] (1.19)

which satisfies

\[ \mathbb{E}Z_t = 0, \ Var(Z_t) = t^l(1-t^l) \] (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

\[ \mathbb{E}[X_kX_m] = km - l(kY_m + mY_k) + l^2\mathbb{E}[Y_kY_m]. \] (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,

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

The crucial calculation is

\[ \mathbb{P}(G_2(m)|G_1(k)) = \frac{\binom{n-2l}{m-2l}}{\binom{n-l}{m-l}} = \frac{(m - l)_t}{(ln - l)_t}. \] (1.23)

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

\[ \text{Cov}(Z_s, Z_t) = (s \wedge t)^l - s^lt^l. \] (1.24)

Thus, to show that $Z_t^{(n)}$ 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 $Z^n_t$ converge to a (time scaled) Brownian bridge as $n \to \infty$?

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

$$
P(X_k = m) = 2^m \frac{n}{(k-m)/2} \frac{(n-(k-m)/2)}{m} / \binom{2n}{k}.
$$

(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

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

(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

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

(1.27)

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

(1.28)

Thus

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

(1.29)

One can use this to show that

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

(1.30)

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

(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 \(X^{(n)}_k\) to be the number of different types of coupons seen after the \(k\)th draw, and let \(\tau^{(n)} = \min\{t : X^{(n)}_t = n\}\). The starting point is to compute the expectation of \(\tau\), which is straightforward: we can write (suppressing notation)

\[
\tau = \sigma_1 + \sigma_2 + \cdots + \sigma_n, \quad (2.1)
\]

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

Because the \(\sigma_i\)'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 \(1 \leq k \leq Cn\log n\)? For example, what is \(EX_k\)? It should be an increasing function of \(k\) with \(EX_0 = 0\) and \(EX_{n\log n} \approx n\).

Note that

\[
X_k = \sum_{j=1}^{n} 1\{\text{coupon } j \text{ drawn by time } k\}, \quad (2.2)
\]

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

\[
EX_k = n \left( 1 - \left( 1 - \frac{1}{n} \right)^k \right). \quad (2.3)
\]

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

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

One can use this to check that

\[
M_k = -k + X_k + \frac{1}{n} \sum_{j=1}^{k-1} X_j \quad (2.5)
\]

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

\[
0 = EM_k \implies EX_k = k - \frac{1}{n} \sum_{j=1}^{k-1} EX_j. \quad (2.7)
\]

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

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

6
which matches the formula for $\mathbb{E}X_k$ given above.

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

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

Applying the optional stopping theorem to $M$ at $\tau$ yields the interesting identity

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

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

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

Thus

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

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

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

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

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

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

(Alternatively, using the second-order identity

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

one can show that

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

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

The variance is

\[ \text{Var}(X_k) \approx n(n^{-\alpha} - n^{-2\alpha}) \approx n^{1-\alpha}. \]  (2.19)

It is natural to consider the scaled process

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

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

8
**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

$$F = \sum_{n \in \mathbb{N}} p_n \delta_n. \quad (2.21)$$

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

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

**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 $W_{\tau_n}$, where $\tau_n$ is the first time that $n$ is collected?

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

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

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

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

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

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

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

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

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

$$Z_n^2 - \sum_{j=1}^{n-1} (2Z_j + 1)g_F(W_j) \quad (2.27)$$
is a martingale – does this yield anything useful?

An interesting identity:

\[
\mathbb{P}(\tau_n < \tau_m) = \mathbb{P}(V_1, V_2, \ldots, V_{\tau_n} \neq m) \\
= \sum_{t \in \mathbb{N}} \mathbb{P}(\tau_n = t)\mathbb{P}(V_1, \ldots, V_t \neq m) \\
= \sum_{t \in \mathbb{N}} \mathbb{P}(\tau_n = t)(1 - p_m)^t \\
= \mathbb{E}[(1 - p_m)^{\tau_n}].
\]

Thus

\[
\mathbb{E}[(1 - p_m)^{\tau_n} + (1 - p_n)^{\tau_n}] = 1.
\]
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 \( \mathbb{E}X_n \)? \( \text{Var}X_n \)?

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

\[
a_n = 1 + \frac{1}{n} \sum_{i=1}^{n} (a_{i-2} + a_{n-i-1}) = 1 + \frac{2}{n} \sum_{i=1}^{n-2} a_i,
\]

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

\[
f'(x) = \frac{x^2}{(1-x)^2} + \frac{2x(1+f(x))}{1-x}.
\]

This (non-linear) ODE can be solved to give

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

Using series expansions, we get

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

and equating coefficients yields

\[
a_n = \frac{1}{2} \sum_{j=1}^{n} (-1)^{j+1} \frac{2j}{j!} (n+1-j).
\]

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

\[
a_n = \frac{n}{2} \sum_{j=1}^{n} (-1)^{j+1} \frac{2j}{j!} - \frac{1}{2} \sum_{j=2}^{n} (-1)^j \frac{(j-1)2^j}{j!}
\]

\[
\approx \frac{1}{2} (1 - e^{-2}) n + \frac{1}{2} (1 - 3e^{-2})
\]

\[
\approx .432332 n + .296997.
\]

**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

\[ X_n \overset{d}{=} \sum_{i \in [n]} 1\{U = i\}(1 + X_{i-2} + X_{n-i-1}). \tag{3.9} \]

(This is truly an abuse of notation – each time we write $X_m$ in every term of the sum we’re thinking of an independent copy of the process with $m$ seats. But writing this out would be burdensome.) Squaring both sides and using the fact that the events $\{U = i\}$ are disjoint over $i$ to eliminate cross terms yields

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

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:

\[ b_n = \left(\begin{array}{c} n+1 \\ 2 \end{array}\right) + \frac{1}{8} \sum_{j=2}^{n} \frac{(-1)^j}{j!} \left(4^j(1-j/4) - 2^j(2+j^2)\right)(n-j+1)(n-j+2) \tag{3.14} \]

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

\[ b_n \approx \frac{n^2}{4}(1 - e^{-2})^2 + \frac{n}{2}(1 - 4e^{-2} + 5e^{-4}) + \frac{1}{4}(1 - 6e^{-2} + 21e^{-4}) \]

\[ \approx .18691n^2 + .275119n + .143154. \tag{3.15} \]

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 $\text{Var} X_n$ is order $n$, namely

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

Usually, the next step is:
Conjecture 3.2. $X_n$ has the following central limit theorem:

$$
\frac{X_n - a_n}{\sqrt{\text{Var} X_n}} \overset{d}{\to} N(0, 1). \quad (3.19)
$$

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

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

Exercise: use the distributional recursion 3.9 to obtain

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

Now define the two-variable moment generating function

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

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

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

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

$$h(x, z) = -x\sqrt{z}\cosh[x\sqrt{z}] + (xz + z - 1)\sinh[x\sqrt{z}] \overline{(x - 1)\sqrt{z}\cosh[x\sqrt{z}] + (xz - 1)\sinh[x\sqrt{z}]}$$

$$= zx + z^2x^2 + \frac{1}{3}(z + 2z^2)x^3 + z^2x^4 + \frac{1}{15}(8z^2 + 7z^3)x^5 + \cdots \quad (3.24)$$

(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

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

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

$$I_t = \{ z : W_z < t, \text{ and } z - 1, z + 1 \notin I_t \}, \quad (3.27)$$

where $I_{t-} = \bigcup_{s \leq t} I_s$. 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 $\{0, 1, 2, \ldots n\}$,

$$\mathbb{P}(0 \in I(\{0, 1, 2, \ldots n\})) = \int_0^1 e^{-x} - e^{-2x} \, dx = 1 - e^{-1}, \quad (3.33)$$

and

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

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:

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

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 occurring 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 renewal 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

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

(3.36)

Integrating over all four variables yields

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

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

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

(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

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

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

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

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:

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

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

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

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

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

and thus

$$P(C_k) = P(K = k) = \int_{x>y} \frac{(x - y)^{k-1}}{(k-1)!} (1 - y) \, dx \, dy = \frac{k + 1}{(k + 2)!}. \quad (3.47)$$
So $K$ is sub exponential, and for example,

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

The generating function of $K$ is

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

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

$$\mathbb{P}(0, k \in I) \quad (3.50)$$
4 Adjacent occurrences in a random permutation

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

\[
A_j = 1\{\sigma(j + 1) = \sigma(j) + 1\}, \quad (4.1)
\]

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

\[
A_1 = A_2 = 1, A_3 = 0 . \quad (4.2)
\]

Consider the statistic

\[
Y_n = \sum_{j=1}^{n-1} A_j \quad (4.3)
\]

\( Y \) counts the number of ‘adjacencies’ preserved by \( \sigma \), 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 \( E A_j = \frac{1}{n-1} \) for all \( j \), so \( E Y_n = 1 \) for all \( n \). Note that the \( A_j \) are all pairwise dependent. It is not too hard to show that

\[
E A_i A_j = \frac{1}{(n-1)(n-2)} , \quad (4.4)
\]

so

\[
\text{Var}(Y_n) = \sum_j E[A_j^2] + \sum_{i \neq j} E[A_i A_j] - \sum_j E[A_j] = \sum_{i \neq j \in [n-1]} \frac{1}{(n-1)(n-2)} = 1 . \quad (4.5)
\]

and

\[
\text{Cov}(A_i, A_j) = \frac{1}{(n-1)^2(n-2)} \sim n^{-3} , \quad (4.6)
\]

To see \( 4.4 \) note that the event \( A_i A_{i+1} \) 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 \( P(A_i A_j) \) 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

\[
E A_1 A_2 \cdots A_k = \frac{1}{(n-1)_k} = \frac{1}{(n-1)(n-2) \cdots (n-k)} . \quad (4.7)
\]

Is it true, for example, that

\[
E A_1 A_3 A_5 = \frac{1}{(n-1)_k} ? \quad (4.8)
\]

17
This seems 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

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

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 $\sigma_n$ in in-line notation, $\sigma_{n+1}$ is obtained by sticking $n + 1$ in at a random spot. Then $\sigma_n$ has the same distribution as a uniform random permutation for each $n$, and $Y_n$ is described by the transition probabilities

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

**Proposition 4.3.** With the initial conditions $Y_1 = 0$, this recursive definition $Y_n$ agrees in distribution with $\frac{1}{n+1}$. (We abuse notation slightly by writing $Y_n$ and $\sigma_n$ 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 $\sigma_n$. $Y_n$ 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, $Y_n = Y_{n+1}$. \hfill \Box

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

**Theorem 4.4.** $Y_\infty \sim \text{Poisson}(1)$.

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

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

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

$$\frac{1}{n + 1} \frac{1}{e \, n + 1} \frac{1}{k!} = \frac{1}{e \, n + 1} \frac{1}{k!} \, (n - k + 1 + k). \quad (4.12)$$

It follows that Poisson(1) is the unique stationary measure for $Y$, and thus by convergence of markov chains, $Y_\infty$ has Poisson(1) distribution. \hfill \Box

One simple variant on $Y$ is the ‘full’ sum

$$X_n = \sum_{j=1}^n A_j, \quad (4.13)$$
where $A_n = 1\{\sigma(n) + 1 = \sigma(1)\}$. $X$ satisfies the recursive relation

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

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

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

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

$$B_1 = B_2 = B_3 = B_4 = 1. \quad (4.16)$$

Set

$$Z_n = \sum_{j=1}^{n} B_j. \quad (4.17)$$

What is the corresponding limit for $Z_\infty$? 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 Sum of uniform random variables

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 \( E\tau_x \) for \( x \in (0, 1) \). First, to prove that \( E\tau_x \) is continuous, note that

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

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

\[
E\tau_x = \mathbb{P}(U_1 \geq x) + \int_{0}^{x} (1 + E\tau_{x-y}) \mathbb{P}(U_1 = y)dy = 1 + \int_{0}^{x} E\tau_ydy. \tag{5.3}
\]

Since \( E\tau_x \) is continuous, the above equation shows that it is differentiable, and taking the derivative yields

\[
\frac{\partial}{\partial x} E\tau_x = E\tau_x. \tag{5.4}
\]

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

\[
E\tau_x = \int_{0}^{1} (1 + E\tau_{x-y})dy = 1 + \int_{x-1}^{x} E\tau_ydy, \tag{5.5}
\]

and so differentiating yields

\[
\frac{\partial}{\partial x} E\tau_x = E\tau_x - E\tau_{x-1}. \tag{5.6}
\]

(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 \( E\tau_{x-1} = e^{x-1} \), 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 \( x \in (n, n+1) \),

\[
E\tau_x = e^x \sum_{k=0}^{n} e^{-k} \frac{(-1)^k}{k!} (x-k)^k. \tag{5.9}
\]
For example, \( \mathbb{E} \tau_x = (1 + \frac{1}{x})e^x - xe^{x-1} \) 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 \( \text{Var}(\tau_x) \sim \frac{2}{3}x \), but this is not at all clear from the explicit formula. For example:

**Question 5.1.** Can it be proved directly that

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

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 \( x \in (0, 1) \), \( \text{Var}(\tau_x) = (1 + 2x)e^x - e^{2x} \) is not monotone on \((0, 1)\). In fact, it has a maximum when \( x \approx .858 \).

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

The entropy exhibits a similar phenomenon (see picture).
6 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 $\theta$, centered at $z_j$, and consider the set of ‘collected’ points $T_n = \bigcup_{j \leq n} I_j$. Our main goal will be to understand the stopping time

$$\tau_\theta = \min\{t : T_t = \mathbb{T}\}.$$  \hspace{1cm} (6.1)

**Question 6.1.** What is

$$\lim_{\theta \to 0} -\frac{\theta \mathbb{E}\tau_\theta}{\log \theta},$$  \hspace{1cm} (6.2)

if it exists?

**Question 6.2.** 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 $\theta$. Set $\tau_A$ 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 6.3.** The minimum value occurs when $A$ is an arc of length $l$, and the maximum occurs when $A$ is a union of approximately $l/\theta$ ‘equally spaced’ arcs of length $\theta$.

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

$$\sigma_N = \min\{t : \text{each } A_k^N, 1 \leq k \leq N, \text{ contains at least one } z_j, 1 \leq j \leq t\}.$$  \hspace{1cm} (6.3)
Then $\sigma_N \leq \tau$ deterministically, since if some arc $A^N_k$ contains no point $z_j$, then the center of that arc cannot belong to $T$. Moreover, $\sigma$ is an instance of the classical coupon collector process on $N$ coupons. Thus

$$\mathbb{E}\tau \geq \mathbb{E}\sigma \approx N \log N = \frac{1}{\theta} \log \frac{1}{\theta} = -\frac{1}{\theta} \log \theta.$$  \hspace{1cm} (6.4)

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

$$\mathbb{E}\tau \leq \mathbb{E}\sigma_{2N} \approx -\frac{2}{\theta} \log \theta.$$  \hspace{1cm} (6.5)

This is enough to see that

$$-\theta \mathbb{E}\tau \theta \log \theta \in [1, 2], \quad (\theta \to 0) \quad (6.6)$$

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

$$\mathbb{E}L_n = \int \mathbb{P}(w \in T_n) \, dw = \int (1 - (1 - \theta)^n) \, dw = 1 - (1 - \theta)^n.$$  \hspace{1cm} (6.7)

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

$$\mathbb{E}L^2_n = \int \Omega L^2_n$$

$$= \int \left( \int \mathbb{1}\{w \in T_n\} \, dw \right)^2$$

$$= \int \int \int \mathbb{1}\{z, w \in T_n\} \, dz \, dw$$

$$= \int \int \mathbb{P}(z, w \in T_n) \, dz \, dw.$$  \hspace{1cm} (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

$$\mathbb{P}(w, z \in T_n) = 1 - \mathbb{P}(z \notin T_n) - \mathbb{P}(w \notin T_n) + \mathbb{P}(z, w \notin T_n)$$

$$= 1 - 2(1 - \theta)^n + \begin{cases} (1 - 2\theta)^n, & d_T(z, w) > \theta \\ 2(1 - \theta)^n (1 - \theta + d_T(z, w))^n, & d_T(z, w) \leq \theta \end{cases}$$

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

$$\mathbb{P}(z, w \notin T_n) = \mathbb{P}(z, w \notin A_j \forall j \text{ and either } z \notin A_j \forall j \text{ or } w \notin A_j \forall j)$$

$$= \mathbb{P}(z, w \notin A_j \forall j) \cdot 2\mathbb{P} \left( z \notin T_n \big| z, w \notin A_j \forall j \right)$$

23
Integrating directly, we get
\[
\mathbb{E}L_n^2 = 1 - 2(1 - \theta)^n + (1 - 2\theta)^{n+1} + \frac{2}{n+1} \left((1 - \theta)^{n+2} - (1 - \theta)(1 - 2\theta)^{n+1}\right)
\]
\[+ \theta - \frac{1}{n+2} (1 - (1 - \theta)^{n+2}) \]

Using the scaling \(\theta = \frac{\alpha \log n}{n}\) for \(\alpha > 0\), these formulas become
\[
\mathbb{E}L_n \approx 1 - n^{-\alpha}
\]
and
\[
\mathbb{E}L_n^2 \approx 1 - 2n^{-\alpha} + \left(1 - \frac{2\alpha}{n}\right) n^{-2\alpha} + \frac{2}{n+2} \left(\frac{n+1}{n+2} \left(1 - \frac{\alpha}{n}\right)^2 n^{-\alpha}\right)
\]
\[+ \left(1 - \frac{\alpha}{n}\right) \left(1 - \frac{2\alpha}{n}\right) n^{-2\alpha} + \frac{\alpha}{n} - \frac{1}{n+2}\]
\[= (1 - n^{-\alpha})^2 + \frac{2}{n} n^{-\alpha}(1 - (1 + \alpha)n^{-\alpha}) + o(n^{-1})\]

Thus, the variance is (to leading order)
\[
\text{Var}(L_n) = \mathbb{E}L_n^2 - (\mathbb{E}L_n)^2 \approx 2n^{-1-\alpha} (1 - (1 + \alpha)n^{-\alpha})
\]

Chebychev’s inequality gives the crude estimate
\[
P(|L_n - 1 + n^{-\alpha}| \geq n^{-1/2}) \leq 2n^{-\alpha}.
\]

**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?
7 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 \( \theta_j \) 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

\[ Q_n(\theta) = \sup\{ t > 0 : te^{i\theta} \in I_n \} \tag{7.1} \]

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 \( \psi_n \) subtended by \( I_n \) is a simple Markov process in discrete time on \([0, 1]\), which is almost surely decreasing: it has transition probabilities

\[ \psi_{n+1} = \begin{cases} 0 & \text{w.p. } \frac{1}{2} - \frac{\psi_n}{2\pi} \\ \psi_n & \text{w.p. } \frac{1}{2} - \frac{\psi_n}{2\pi} \\ x & \text{w.p. } \frac{1}{n} dx, 0 \leq x \leq \psi_n \end{cases} \tag{7.2} \]

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

\[ \mathbb{E}[\psi_{n+1} | \psi_n] = \psi_n \left( \frac{1}{2} - \frac{\psi_n}{2\pi} \right) + \int_0^{\psi_n} x \cdot \frac{1}{\pi} dx = \frac{1}{2} \psi_n. \tag{7.3} \]

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

Exercise 7.1. More generally, if the half spaces are replaced by cones of angle \( \theta \), 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 \( F(s) = s^2 \mathbb{1}_{\{s \in [0,1]\}} + 1_{\{s \geq 1\}} \). 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,

\[ Q_n(\theta) = \min_j \{ R_j \sec |\theta - \theta_j| : |\theta - \theta_j| < \pi/2 \}. \tag{7.4} \]

Thus we have

\[ \mathbb{P}(Q_j(\theta) \leq q) = \mathbb{P}(R_j \sec |\theta - \theta_j| \leq q)^n, \tag{7.5} \]

where \( \sec |\theta - \theta_j| \) is taken to be infinite if \( |\theta - \theta_j| \geq \pi/2 \). This can be evaluated directly:
\[ P(R_j \sec |\theta - \theta_j| \leq q) = \frac{1}{2} P \left( R_j \sec |\theta - \theta_j| \leq q \bigg| |\theta - \theta_j| < \pi/2 \right) \]  
(7.6)

\[ = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \int_0^1 \{s \sec \phi \leq q\} \cdot 2s \, ds \, d\phi \]  
(7.7)

\[ = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} \int_0^q s \cos \phi \, ds \, d\phi \]  
(7.8)

\[ = \frac{q^2}{4}. \]  
(7.9)

It follows that

\[ P(Q_n(\theta) \leq q) = 1 - (1 - q^2/4)^n, \]  
(7.10)

and thus

\[ \frac{n}{4} Q_n^2 \rightarrow_d \text{Exp}(1). \]  
(7.11)

Roughly speaking, this suggests

\[ \text{Area}(I_n) \approx \frac{4\pi}{n} \cdot \text{Exp}(1) \]  
(7.12)

for \( n \) large.
8 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

\[
\frac{1}{k} \tau_k \to \frac{1}{\mathbb{E} Y} \text{ almost surely and in } L^1. \tag{8.1}
\]

In general, \( X_n \) and \( \tau_k \) 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

\[
\mathbb{E} X_n = N \left(1 - \left(1 - \frac{1}{N}\right)^n\right), \quad \text{and } \mathbb{E} \tau_k = N (H_N - H_{N-k}),
\]

where \( H_j \) is the \( j \)th 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:

\[
\mathbb{E} X_{\tau_k} = k, \tag{8.2}
\]

since \( X_{\tau_k} \) is always equal to \( k \), and

\[
\mathbb{E} \tau_{X_n} \approx -N \log \left(1 - \left(1 - \left(1 - \frac{1}{N}\right)^n\right)\right) = -nN \log\left(1 - \frac{1}{N}\right) \approx n. \tag{8.3}
\]

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

\[
\mathbb{E} \tau_y = \int_{\mathbb{R}} \mathbb{P}(\tau_y \geq t) dt = \int_{\mathbb{R}} \mathbb{P}(X_t \leq y) dt, \tag{8.4}
\]

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

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

\[
\tau_{X_n} = \min\{m : X_m \geq X_n\} \leq n, \tag{8.5}
\]

which implies

\[
\limsup_{n} \frac{\tau_{X_n}}{n} \leq 1. \tag{8.6}
\]

**Proposition 8.1.** 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

\[
\mathbb{P}(X_n = X_{n+1} = \cdots = X_{n+\epsilon n}) \leq c^n. \tag{8.7}
\]

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

\[
X_n = X_{n+1} = \cdots = X_{n+\epsilon n} \tag{8.8}
\]

occurs finitely often almost surely. Thus
\[
\frac{\tau_{X_n}}{n} \geq 1 - \epsilon
\quad (8.9)
\]

for \(n\) large; now let \(\epsilon \to 0\).

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

\textbf{Question:} Can we weaken the assumptions to include the case where \(X\) is not necessarily non-decreasing, or where \(X\) is constant on large intervals (as with the coupon collector)?
9 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

\[
\mathbb{P}(X_n = M_n) = \mathbb{P}(X_n > M_{n-1}) = \frac{1}{n},
\]

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

\[
\mathbb{P}(X_n = M_n) = \frac{1}{n} \mathbb{E}[\text{Max}_n],
\]

where \(\text{Max}_n = \{i \in [n] : X_i = M_n\}\) and \(\mathbb{E}[\text{Max}_n] > 1\) 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

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

The distribution of Max looks like

\[
\mathbb{P}(\text{Max}_n = m) = \sum_{k \in \mathbb{Z}} \binom{n}{m} p^m_k \mathbb{P}_{<k}^{n-m}.
\]

As a check, the binomial theorem gives

\[
\sum_{m=1}^{n} \mathbb{P}(\text{Max}_n = m) = \sum_{m=1}^{n} \sum_{k} \binom{n}{m} p^m_k \mathbb{P}_{<k}^{n-m}
\]

\[
= \sum_{k} \sum_{m=1}^{n} \binom{n}{m} p^m_k \mathbb{P}_{<k}^{n-m}
\]

\[
= \sum_{k} p^m_{<k} - \mathbb{P}_{<k}^{n}
\]

\[
= 1,
\]

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

\[
\mathbb{E}[\text{Max}_n] = \sum_{k} \sum_{m=1}^{n} \binom{n}{m} p^m_k \mathbb{P}_{<k}^{n-m}
\]

\[
= n \sum_{k} p_k \sum_{j=0}^{n-1} \binom{n-1}{j} p^j_k \mathbb{P}_{<k}^{n-j-1}
\]

\[
= n \sum_{k} p_k \mathbb{P}_{<k}^{n-1}
\]

\[
= n \mathbb{E} \mathbb{P}_{<X}^{n-1}
\]
Figure 6: Approximate value of $E|\text{Max}_n|$ 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

$$E|\text{Max}_n| = n \sum_k p q^k (1 - q^{k+1})^{n-1}.$$  \hspace{1cm} (9.13)

For $p = \frac{1}{2}$ and $n$ large, Mathematica gives the value $E|\text{Max}| \approx 1.442689883$. (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 9.1.** For some nice class of distributions $X$, show that

$$\lim_{n \to \infty} E|\text{Max}_n| \text{ exists } \in (0, \infty),$$  \hspace{1cm} (9.14)

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

$$\mathbb{P}(X_n > M_{n-1}) = \sum_k p_k p_{<k}^{n-1}. \hspace{1cm} (9.15)$$

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 9.2.** Show that if $X \sim \text{Poisson}(\lambda)$, for any $\lambda > 0$,

$$\liminf_n n\mathbb{P}(X_n > M_{n-1}) > 0.$$  \hspace{1cm} (9.16)
If this fails, it should fail for all $\lambda$, since otherwise there would be a phase transition point in $\lambda$, which would be too interesting. I suspect it holds for all $\lambda$.

**Question 9.3.** *Is there some class of distributions, supported on $\mathbb{N}$, such that*

$$\limsup_n n \mathbb{P}(X_n > M_{n-1}) = 0? \quad (9.17)$$

I suspect not – can this be proved?

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

$$\mathbb{P}(X_n > M_{n-1}) = k^{-n} \sum_{i=1}^{k-1} i^{n-1} \approx n^{-1} (1 - k^{-1})^n, \quad (9.18)$$

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

$$\mathbb{P}(X_n > M_{n-1}) \approx e^{-\alpha^{-1} n^{-1}}, \quad (9.19)$$

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


10 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

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

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

$$
\mathbb{E}X = \int_{\mathbb{R}} (1 - F_f(x)) \, dx
$$

$$
= \int_{\mathbb{R}} \int_{\mathbb{R}} f(t)1\{f(t) > x\} \, dt \, dx
$$

$$
= \int_{\mathbb{R}} f(t) \int_{0}^{f(t)} dx \, dt
$$

$$
= \int_{\mathbb{R}} f(t)^2 \, dt
$$

Loosely speaking, the density of $f$ is given by

$$
\frac{d}{dx}\mathbb{P}(X \leq x) = \int_{\mathbb{R}} f(t)\delta_x(f(t)) \, dt
$$

$$
= x \cdot |f^{-1}(x)|
$$

**Question 10.1.** 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

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

Thus, in this case,

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

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

**Conjecture 10.2.** Does it hold in general that

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

This could make good exercise for a probability course. Does it have any other value?
11 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 $n > 1$,

$$Y_i = X_i Y_{i-1} = X_1 X_2 \cdots X_i. \quad (11.1)$$

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

$$\mathbb{E}(Y_n^k) = \mathbb{E}(X_1^k \cdots X_n^k) = (\mathbb{E}X^k)^n \quad (11.2)$$

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

**Proposition 11.1.** $S < \infty$ almost surely.

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

$$\mathbb{E}S = \sum_n \mathbb{E}Y_n = \sum n \mu^n = \frac{\mu}{1 - \mu} < \infty. \quad (11.3)$$

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

**Question 11.2.** 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 11.3.** 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 $n > 1$,

$$g_n(y) = \int_y^1 \frac{1}{x} f(y/x) g_{n-1}(x) \, dx. \quad (11.4)$$

We can do the usual thing with logs:

$$\log Y_n = \sum_{i=1}^n \log X_i. \quad (11.5)$$

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 $S$:

$$S \sim XS + X', \quad (11.6)$$

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 $s > 0$, (11.6) gives TBD
12 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 \( n \)th 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, regardless of whether we have seen a fish. Given such a strategy, for \( k \in \mathbb{N} \), let \( T_k \) denote the time between catching our \((k-1)\)th and \( k \)th 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

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

\[
= q \left( t \sum_{j \geq 0} j p^j + \sum_{j \geq 0} p^j \mathbb{E}[X|X \leq t] \right)
\]

\[
= q \left( \frac{tp}{q^2} + \frac{1}{q} \int_{\mathbb{R}} \frac{\mathbb{P}(s < X \leq t)}{q} ds \right)
\]

\[
= \frac{1}{q} \left( tp + \int_{0}^{t} (F(s) - p)ds \right)
\]

\[
= \frac{1}{1 - F(t)} \int_{0}^{t} F(s)ds.
\]

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,

\[
\lim_{t \to 0^+} \mu_t(F) = \lim_{t \to 0^+} \frac{\int_{0}^{t} F(s)ds}{1 - F(t)} = \lim_{t \to 0^+} \frac{F(t)}{-F'(t)} = \lim_{t \to 0^+} \frac{1}{h(t)} = -1/F'(0).
\]

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

\[
\frac{d\mu_t}{dt} = \frac{d}{dt} \left[ \frac{1}{1 - F(t)} \int_{0}^{t} F(s)ds \right] = \frac{F(t)}{1 - F(t)} + \frac{F'(t)}{(1 - F(t))^2} \int_{0}^{t} F(s)ds.
\]

Thus

\[
\frac{d\mu_t}{dt} = 0 \iff F(1 - F) = -F' \int_{0}^{t} F(s)ds.
\]

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, \( \mu_t \) is constant when \( F \) is exponential:

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

(12.9)

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

\[
-\frac{1}{\lambda} F'(t) = F(t),
\]  

(12.10)

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

\[
\mathbb{E}[z^{T_1}] = \sum_{j \geq 0} p^j q^j t^j \mathbb{E}[z^X | X \leq t]
\]  

(12.11)

\[
= q \mathbb{E}[z^X | X \leq t] \cdot \sum_{j \geq 0} (p \cdot z^t)^j
\]  

(12.12)

\[
= q \left( \int_R \frac{P(s < z^X \leq z^t)}{q} ds \right) \cdot \frac{1}{1 - pz^t}
\]  

(12.13)

\[
= \frac{\log z}{1 - pz^t} \int_0^t z^u F(u) du.
\]  

(12.14)

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

\[
P(T_1 = m) = \left. \left. \frac{d^m}{dz^m} \right|_{z=0} \mathbb{E}[z^{T_1}] \right. = \left. \frac{1}{m!} \mathbb{E}[z^{T_1}] \right|_{z=0}.
\]  

(12.15)

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

\[
\mathbb{E}[e^{rT_1}] = \left. \left. \frac{d^l}{dr^l} \right|_{r=0} \mathbb{E}[e^{rT_1}] \right.
\]  

(12.16)

What is the expected number of fish caught up to time \( s \)? It can be calculated for a Poisson process, by integrating over all possible paths: I wonder if a similar calculation can be done for a general process, or possibly for the fisherman process, where we re-cast upon seeing a fish a time \( \leq t \). The latter seems much more complicated!
13 Bases for $\mathbb{N}$

For $k \in \{2, 3, \ldots\}$, a subset $B \subset \mathbb{N}$ is called a $k$-basis if $kB = \mathbb{N}$, where $kB$ is the $k$-fold sumset of $B$, i.e.

$$kB = \{b_1 + b_2 + \cdots + b_k : b_i \in B, i = 1, 2, \ldots, k\}.$$  \hspace{1cm} (13.1)

There are many open problems related to bases. Many questions relate to the ‘representation’ function

$$r(n) = \#\{(b_1, \ldots, b_k) : \sum b_i = n\}.$$  \hspace{1cm} (13.2)

For simplicity, we consider:

**Conjecture 13.1.** (Erdős-Turan) For any 2-basis $B$, $r(n)$ is unbounded.

This is the weakest form of the conjecture: it is believed that $r(n)$ should be at least $O(\log n)$ infinitely often. An equivalent form of the problem is to consider the generating function

$$f_B(z) = \sum_k 1\{k \in B\} z^k,$$  \hspace{1cm} (13.3)

so we have

$$f_B(z)^2 = \sum_k r(k) z^k.$$  \hspace{1cm} (13.4)

Thus [13.1] is equivalent to:

$$f_B(z)^2 \geq \frac{1}{1 - z} \implies \frac{1}{n!} \frac{\partial^n}{\partial z^n} f_B(z)^2 \text{ is unbounded.}$$  \hspace{1cm} (13.5)

or alternatively,

$$\frac{\partial^n}{\partial z^n} f_B(z)^2 \leq Cn! \implies \frac{\partial^n}{\partial z^n} f_B(z)^2 = 0 \text{ for some } n.$$  \hspace{1cm} (13.6)
14 Limits of multiplicative functions on $\mathbb{N}$

Consider the sum of divisors function

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

(14.1)

where $p$ ranges over the primes, and $n = \prod_p p^{\alpha_p}$. There are many questions about the asymptotic behavior of $\sigma$. 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 $k = 0, 1, \ldots$,

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

(14.2)

for any prime $p$. Thus an element of $\Omega$ 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 $\Omega$, since the largest power of $p$ dividing $X_n$ is very close to $Y_p$ in distribution. More generally:

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

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

(14.3)

Define

$$\tilde{f} = \prod_p f(p)^{Y_p}.$$  

(14.4)

Is it true that

$$f(X_n) \rightarrow_d \tilde{f}.$$  

(14.5)

Alternatively, one could define $\tilde{f}$ as a limit, by truncating the sequence of primes in some way. For example, one could look at sequences of exponents $Y_p$ for $p \leq q$, and let $q \rightarrow \infty$; or, in addition, condition that $n = \prod_p p^{Y_p} \leq N$ for some $N$, and let $N, q \rightarrow \infty$.

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 $g(p) = \log f(p)$:

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

(14.6)

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

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

(14.7)
Returning to \( \sigma \), which we think of as a random variable on \( \Omega \) given by
\[
\sigma = \prod_p \frac{1 - p^{-\gamma_p - 1}}{1 - p^{-1}},
\]
(14.8)
note that
\[
\mathbb{E} p^{-j\gamma_p} = \frac{1 - p^{-1}}{1 - p^{-j - 1}}
\]
(14.9)
for each prime \( p \) and \( j \geq 1 \). Thus,
\[
\mathbb{E} \sigma = \prod_p (1 - p^{-1} - 1 + \mathbb{E} p^{-\gamma_p - 1})
\]
(14.10)
\[
= \prod_p (1 - p^{-1} - 1 + \sum_{k \geq 0} (1 - p^{-1})p^{-2k - 1})
\]
(14.11)
\[
= \prod_p (1 - p^{-1} - 1 + \frac{p^{-1} - 1}{1 - p^{-2}})
\]
(14.12)
\[
= \prod_p \frac{1}{1 - p^{-1} - 1 + \frac{p^{-1}}{1 - p^{-2}}}
\]
(14.13)
\[
= \prod_p \frac{1}{1 - p^{-2}}
\]
(14.14)
\[
= \zeta(2).
\]
(14.15)
In particular, this shows \( \sigma < \infty \) almost surely, so – assuming the answer to question 14.1 is positive for \( f(n) = \sigma(n) \) – the limit distribution of \( \sigma(n) \) exists. Similarly, one can obtain
\[
\mathbb{E} \sigma^2 = \prod_p \frac{p^{-4} - 2p^{-2} - p^{-1} - 1}{(p^{-1} - 1)(p^{-1} + 1)(p^{-2} + p^{-1} + 1)} = \prod_p \left( 1 + \frac{3p^2}{p^4 + p^3 - p - 1} \right), \approx 2.0999
\]
(14.16)
and for any \( m \in \mathbb{N} \),
\[
\mathbb{E} \sigma^m = \prod_p p(1 - p^{-1})^{1-m} \sum_{j=0}^m (-1)^j \binom{m}{j} (p^{j+1} - 1)^{-1}.
\]
(14.17)

**Question 14.3.** How can we compute or approximate properties the distribution function for \( \sigma \)?

One open question of interest is to determine \( \mathbb{P}(\sigma > 2) \), 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 \( \sigma \) to within arbitrary precision, can we compute such probabilities to high precision?

The log of \( \sigma \) is perhaps easier to understand. Using the expansion \( \log(1 - z) = -\sum_{n \geq 1} \frac{z^n}{n} \),
\[
\log \sigma = \sum_p \left[ \log(1 - p^{-\gamma_p - 1}) - \log(1 - p^{-1}) \right]
\]
(14.18)
\[
= \sum_p \sum_{n \geq 1} \frac{p^{-n}}{n} (1 - p^{-n\gamma_p})
\]
(14.19)
This is a sum of independent random variables, each of which is relatively simple. Some algebra yields

\[
\mathbb{E} \log \sigma = \sum_p \sum_{n \geq 1} \frac{1}{n} \frac{p^n - 1}{p^{2n+1} - p^n} \approx 0.4457.
\]

(14.20)

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