BACHELOR

Sorting discrete samples

Colenbrander, D.G.J.

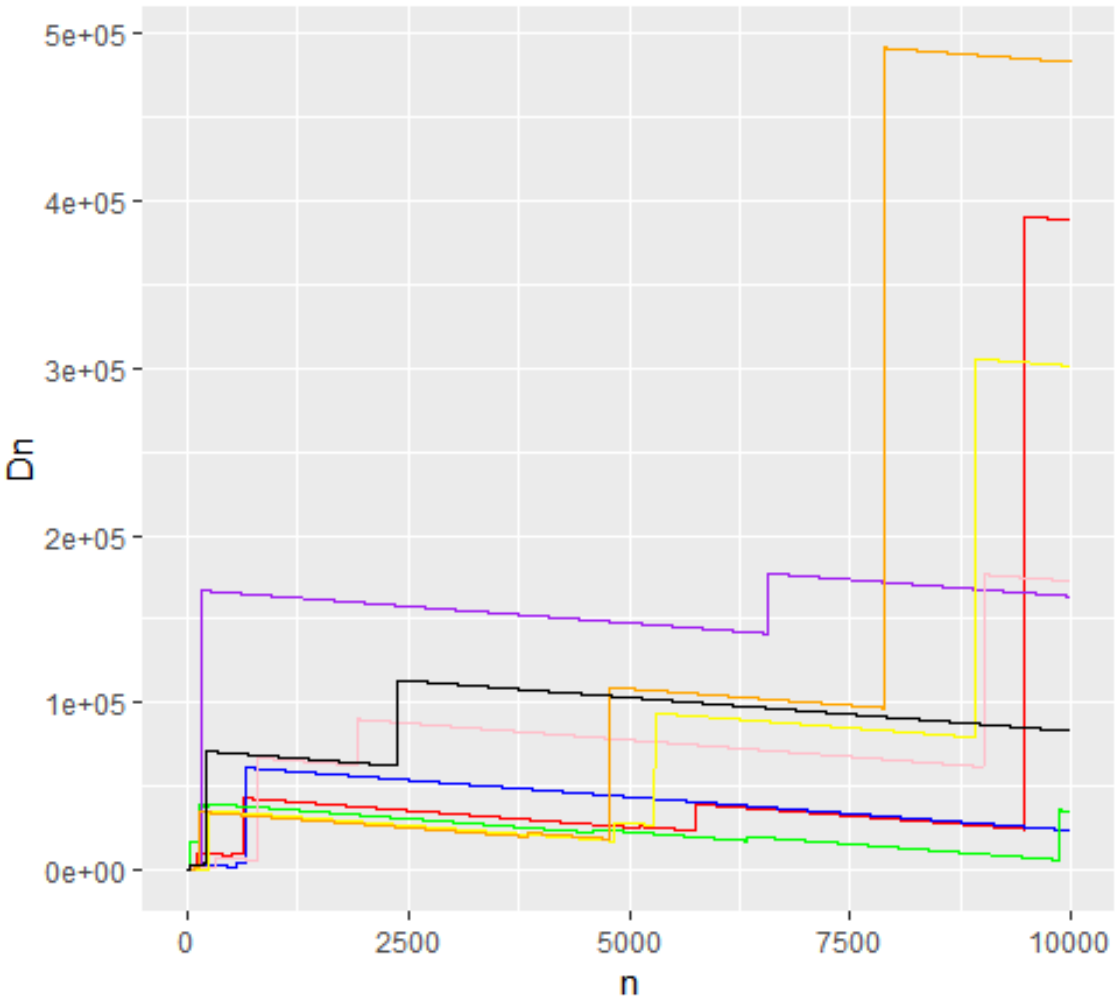
Award date:
2019

Bachelor Project

Sorting discrete samples

Author:

D.G.J. Colenbrander



November 21, 2019

1 Introduction

Let $(X_n)_{n\geq 0}$ be a sequence of independent and identically distributed discrete random variables. We wish to group these random variables in blocks of size m. One method to efficiently achieve this is the following. First, a random sample X_1 is generated which takes values from the set $S = \{1, ..., k\}, k \in \mathbb{N} \cup \{\infty\}$ with respective probability $\mathbb{P}(X_i = j) = p_j$ for $j \in S$. More random samples $X_2, X_3, ...$ are added to the sequence till there are exactly m samples with the same value as X_1. After the m'th sample is found these samples are marked as *used*, this does not change the position of the other samples in the sequence. We look at the first following sample which has not been marked as used, which we call *free*, and again try to find m samples with equal value. Chances are that some of these samples have already been generated. Only when there are less then m of these samples in the sequence, more samples will have to be generated till the required amount is found. This process is repeated a given number of times.

Let us look at an example where the random samples are represented by a fair die, given we have an infinite collection of dice and an infinitely long table to place them on. We have $S = \{1, ..., 6\}$ with $p_1 = ... = p_6 = \frac{1}{6}$ and take the size of our sorted groups $m = 3$. Before we start the process, we are at step 0 and the table is empty. Say the first die is to be the value 2. We now keep throwing new dice from our collection and place them next to our previously thrown die till we see three dice in our sequence with the value 2. Say our sequence of dice looks like:

2	1	3	4	2	6	1	2

We now take each die with value 2 and replace it with an asterisk, so after our first step the sequence looks like

*	1	3	4	*	6	1	*

Notice that the position nor value of the remaining dice has not changed. Now we look at the first die in our sequence, which has value 1. As we do not have three dice with this value, we throw some new dice till we do. We end up with the sequence

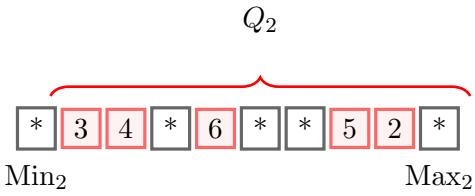
*	1	3	4	*	6	1	*	5	2	1

Therefore our second step of the process ends with the sequence

*	*	3	4	*	6	*	*	5	2	*

After this process has been repeated a number of steps we end up with a sequence of used samples before we see the first free sample. We are interested in the sequence which starts at the first free sample and ends with the last used sample. We define this sequence as $(Q_n)_{n\in\mathbb{N}}$,

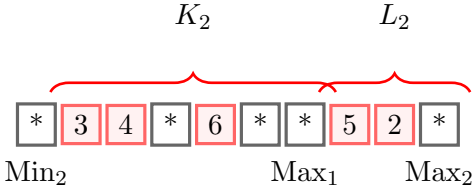
where n is the number of sorting steps of the process, and define the position of the last used sample as Max_n. As the first free sample can be after Max_n in the case all samples are used, we define the sample *before* the first free sample as Min_n. Going back to our example, we have



In this report we shall focus on the properties of Q_n, in particular, we are interested in the length of the sequence $D_n = |Q_n| = \mathrm{Max}_n - \mathrm{Min}_n$ and its distribution. In order to analyze the behaviour of D_n we study properties of Max_n and Min_n, such as the rate of growth and its fluctuations.
As some steps make use of previously generated samples, we find that the number of samples that need to be generated is conditionally dependent of the previous states. Because of these dependencies the state space of the sequence is rather complex, so we define it as $\Omega = \{(q_i)_{i=1}^{D_n} : \exists n\, \mathbb{P}(Q_n = (q_i)_{i=1}^{D_n}) > 0\}$ with $q_i \in \{S, *\}$. The sequence Q_n starts with the first free sample and ends with the last used sample such that it is not possible that $D_n = 1$, as this would mean that this one element is both used and free which is not possible by definition. We define the elements of our sequence $(Q_n(i))_{i=1}^{D_n}$ as

$$Q_n(i) = \begin{cases} X_{\mathrm{Min}_n+i} & \text{if } X_{\mathrm{Min}_n+i} \text{ is free,} \\ * & \text{if } X_{\mathrm{Min}_n+i} \text{ is used,} \end{cases} \tag{1.1}$$

where we start with the empty list Q_0, such that $D_0 = 0$. As shown in section three, the transition probabilities of this sequence are only dependent on its last state which allows us to make use of Markov properties (3.1). As every step of the process starts with a previously generated part and a part of samples that still has to be generated we can split our sequence in two parts $Q_n = (K_n, L_n)$.



The sequence K_n starts at the first free sample after the first up to m samples with the same value as $Q_{n-1}(1)$ are marked as used and ends at the sample at position Max_{n-1}. So this sequence has already been determined by the outcome of the previous sequence Q_{n-1}. In the case that there are m or less free samples in Q_{n-1} and they all have the same value, we find that on step n all these free samples are used and therefore $K_n = \emptyset$. Define the number of samples with value $Q_{n-1}(1)$ which we still need to find after grouping the previously generated

samples with this value as r, such that

$$r = \max\{0, m - \sum_{i=1}^{D_{n-1}} \mathbb{1}_{\{Q_{n-1}(i)=Q_{n-1}(1)\}}\}. \tag{1.2}$$

L_n is the random sequence conditioned on r and the respective probability of $Q_{n-1}(1)$. As all the samples are i.i.d. we find that the number of steps till we generate the next sample with the right value is Geometrically distributed with success probability $p_{Q_{n-1}(1)}$. Given that the sequence $K_n \neq \emptyset$, such that all the generated samples this step are part of the sequence Q_n, it follows that the length of the sequence $|L_n|$ is Negative Binomial distributed with success probability $p_{Q_{n-1}(1)}$ and the number of successful experiments r. Assume that $Q_{n-1}(1) = k$, then the p.m.f. of $|L_n|$ is

$$\mathbb{P}(|L_n| = l) = \binom{l-1}{r-1} p_k^r (1-p_k)^{l-r}. \tag{1.3}$$

All the samples in L_n with a different value as $Q_{n-1}(1)$ are free. Let S_i denote the number of occurrences of the value $i \in S$ in the sequence L_n, given $K_n \neq \emptyset$. We can approach the number of times every value appears in L_n with the Multinomial distribution with l trials, conditioned on that the value i appears r times. Say for ease of notation that $i = k$, such that for $l \geq r$

$$\begin{aligned}
\mathbb{P}(S_1 = s_1, ..., S_{k-1} = s_{k-1}, l | S_k = r) &= \frac{\mathbb{P}(S_1 = s_1, ..., S_{k-1} = s_{k-1}, S_k = r)}{\mathbb{P}(S_k = r)} \\
&= \frac{\frac{l!}{s_1!...s_{k-1}!r!} p_1^{s_1} ... p_{k-1}^{s_{k-1}} p_k^r}{\binom{l}{r} p_k^r (1-p_k)^{l-r}} \\
&= \frac{(l-r)!}{s_1!...s_{k-1}!} \prod_{i=1}^{k-1} \left(\frac{p_i}{1-p_k}\right)^{s_i} \\
&= \frac{(l-r)!}{s_1!...s_{k-1}!} \prod_{i=1}^{k-1} \pi_i^{s_i},
\end{aligned} \tag{1.4}$$

which is the Multinomial distribution with $l-r$ trials $k-1$ values with probabilities $\pi_i = \frac{p_i}{1-p_k}$. Now that we have analysed the structure of our sequence, we will define the first and last element of our sequence and find the bounds of its position.

As Max_n is the position of the last of the $m \cdot n$ used samples, we define it as the total number of samples generated to complete n steps, i.e.,

$$\begin{aligned}
\text{Max}_n &= \max_i \{i \in \mathbb{N} : X_i \text{ used during the first } n \text{ sorting steps}\} \\
&= m \cdot n + \sum_{j=1}^{D_n} \mathbb{1}_{\{Q_n(j) \neq *\}},
\end{aligned} \tag{1.5}$$

with $\text{Max}_0 = 0$. As there are $m \cdot n$ samples sorted after n steps but no limit to how many samples we need to generate before we find them, we have that

$$m \cdot n \leq \text{Max}_n < \infty. \tag{1.6}$$

The same way we can define Min_n, as all the samples up to sample Min_n are used we know the remaining $m \cdot n - \text{Min}_n$ used samples are elements of Q_n. From this we find the expression

$$\begin{aligned} \text{Min}_n &= \min_i \{i \in \mathbb{N} : X_{i+1} \text{ not used the first } n \text{ sorting steps}\} \\ &= m \cdot n - \sum_{j=1}^{D_n} \mathbb{1}_{\{Q_n(j)=*\}}, \end{aligned} \tag{1.7}$$

with $\text{Min}_0 = 0$.

Theorem 1.1: Bounds for Min_n

Let Min_n be defined as (1.7). Then $\max\{\text{Min}_n, n \in \mathbb{N}\} = m \cdot n$ and

$$\min\{\text{Min}_n, n \in \mathbb{N}\} = \begin{cases} n, & \text{for } n < k \\ m \cdot n - (m-1)(k-1), & \text{for } n \geq k. \end{cases} \tag{1.8}$$

Proof. *Upper bound* After n sorting steps we know that in total $m \cdot n$ samples have been used. If the first $m \cdot n$ samples have all been used, it follows that the first free sample is $X_{m \cdot n+1}$ such that $\text{Min}_n = m \cdot n$. If this is not the case it follows that there is some $X_i, i \leq m \cdot n$, that has not been sorted. Such that $\text{Min}_n = i - 1 \leq m \cdot n$. Therefore, it holds that $\text{Min}_n \leq m \cdot n$.

Lower bound $n < k$: As every sorting step the first free sample is always used, it follows that $\text{Min}_n > \text{Min}_{n-1}$. Therefore, as $\text{Min}_0 = 0$ we know that $\text{Min}_n \geq n$. For this to be the maximal lower bound, we want to show that it is possible that $\text{Min}_n = n$. Take the sequence $(X_i)_{i=1}^k$, for which $X_1 \neq X_2 \neq .... \neq X_k$, which is possible as we have k unique values. As every sorting step only one value is grouped, it follows that it takes k sorting steps till every sample is used, where for the first $k-1$ steps it holds that $\text{Min}_n = n$.

Lower bound $n \geq k$: To show why this lower bound makes sense, we look at the definition of Min_n in terms of the number of used samples. Namely, we know that the number of used values which are elements of Q_n is $m \cdot n - \text{Min}_n$. So for (1.8) to be the lower bound, it needs to hold that that it is not possible to have more then

$$m \cdot n - m \cdot n + (m-1)(k-1) = (m-1)(k-1) \tag{1.9}$$

used samples in Q_n. Therefore, we need to prove the following proposition:

Proposition 1.1: Upper bound used samples with the same value in Q_n

Define $G_{n,x} = \#\{i : Q_n(i) = *, X_{\text{Min}_n+i} = x\}$.

$$\max\{G_{n,x}\} \leq m - 1, \forall n \in \mathbb{N}, x \in S \tag{1.10}$$

Proof. Define the position of the i'th appearance of value x in the sequence as

$$x_i = \min\{y : \sum_{j=1}^{y} \mathbb{1}_{\{X_j=x\}} = i\}. \tag{1.11}$$

As every time a value is grouped the first m free samples with this value are used, it follows that if we sort value x for the $(s+1)$st time on step n that $\text{Min}_n = x_{s\cdot m+1}$. We will proof by induction by first showing that the upper bound holds for the first time the value is sorted. Say that we are sorting value x for the first time on step n_1 such that $\text{Min}_{n_1} = x_1$. As $X_{\text{Min}_{n_1}} \notin Q_{n_1}$, it follows that

$$G_{n_1,x} \leq |\{x_2, ..., x_m\}| = m - 1. \tag{1.12}$$

Now assume this upper bound also holds after sorting this value s times, such that for $n_1 < n_s$

$$G_{n_s,x} \leq |\{x_{(s-1)m+2}, ..., x_{s\cdot m}\}| = m - 1. \tag{1.13}$$

As we know that all samples up to $\text{Min}_{n_{s+1}} = x_{m\cdot s+1}$ are not element of $Q_{n_{s+1}}$, and $x_1 < x_2 < ... < x_{s\cdot m} < x_{s\cdot m+1}$, it follows that

$$G_{n_{s+1},x} \leq |\{x_{s\cdot m+2}, ..., x_{(s+1)\cdot m}\}| = m - 1. \tag{1.14}$$

□

From this proof we also find that if we are sorting a certain value on step $n+1$, then there are no used values in Q_n which held this value. Therefore, up to $(k-1)$ different values can have used samples in sequence Q_n, such that there can only be up to $(m-1)(k-1)$ used samples in Q_n. For this to be the maximum lower bound, Min_n needs to be able to hold this value. To show this, we construct a method to achieve this lower bound for every m and k. As achieving the lower bound for $n < k$ is trivial, which is always the case for $k = \infty$, we construct a method to achieve this lower bound for $n \geq k$ with $k < \infty$:

1. Let the first $k-1$ samples have unique values, so $X_1 \neq X_2 \neq .. \neq X_{k-1}$, such that after $k-2$ sorting steps we have $\text{Min}_{k-2} = k-2$.

2. To achieve the lower bound at step n let the value of X_{k-1} and the value that has been used yet alternately in groups of m for $n-(k-2)-1$ times. So we have for $X_{k+m} \neq X_1, ..., X_{k-1}$ and $X_{sm+k-1} \neq X_{(s+1)m+k}$ for $s = 0, ..., n-k+1$, that

   $$X_{sm+k-1} = X_{sm+k} = ... = X_{(s+1)m+k-1}. \tag{1.15}$$

   It follows that at step $n-1$ we have $\text{Min}_{n-1} = m \cdot n - (m-1)(k-1) - 1$.

3. At the end of the previous step the samples with the two values which appeared in groups of m positioned after Min_{n-1} are free. So take

   $$X_{m\cdot n-(m-1)(k-1)} = X_{k-1}, \quad X_{m\cdot n-(m-1)(k-1)+1} = X_{k+m}. \tag{1.16}$$

   Then on step n sample $X_{m\cdot n-(m-1)(k-1)}$ is used and sample $X_{m\cdot n-(m-1)(k-1)+1}$ is free. It follows that $\text{Min}_n = m \cdot -(m-1)(k-1)$, which is our lower bound.

As this can sound quite complicated we can look at table 1 for an example.

Table 1: *Sequence to achieve lower bound for* $k = 5, m = 3$ *for* $n = 5$ *and* $n = 6$

n	Sequence	Min_n	n	Sequence	Min_n
0	1,2,3,4,4,4,5,4	0	4	*,*,*,*,*,*,5,5,5,4,5	6
1	*,2,3,4,4,4,5,4	1	5	*,*,*,*,*,*,*,*,*,4,5	9
2	*,*,3,4,4,4,5,4	2	6	*,*,*,*,*,*,*,*,*,*,5	10
3	*,*,*,4,4,4,5,4	3			
4	*,*,*,*,*,*,5,4	6			
5	*,*,*,*,*,*,*,4	7			

Now we have shown that it is possible to reach the lower bound for every value of m and k and it is not possible to get any lower values we find

$$m \cdot n - (m-1)(k-1) \leq \text{Min}_n \leq m \cdot n, \text{ for } n \geq k. \tag{1.17}$$

□

Now that the basic properties have been analysed the rest of this thesis is organized as follows. In the next section we will analyse the properties for Q_n as the number of sorting steps grows. Namely, we will prove that the Min_n and Max_n for $k < \infty$ and Min_n for $k = \infty$ all have the same rate of growth. In section 3 we will prove that the sequence Q_n is a Markov Chain by looking at the transition probabilities and showing that for every state, the transition probability is only conditioned on composition of the previous state. The following sections will then show that Q_n only has a stationary distribution if and only if $k < \infty$ by proving the sequence is irreducible for all $k \in \mathbb{N} \cup \{\infty\}$, but only positive recurrent for $k < \infty$. We follow by further analysing the behaviour of Q_n for $k = \infty$ to get a better idea why these results hold. Finally, we will discuss the results shown in this thesis and suggest problems which will require further analysis.

2 Rate of growth

This section we will analyse the rate of growth for Min_n for both finite and infinite support and use this to prove that the distribution of the grouped values converges to the distribution of the samples. For Max_n we show that it has the same rate of growth as Min_n for the finite support case.

2.1 Finite support

Let us start by looking at the distribution $X_{\text{Min}_n+1} = Q_n(1)$, the value that is being grouped on step n. The first intuition could be that this distribution is simply the same as the distribution for X_i, but by looking at a simple example we can see that this does not hold for all n. Take $m = 2$, $k = 3$ and $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. For step $n = 1$ the probability that the current value being sorted is $i \in S$ is simply the probability that the first generated sample has value i, which has probability p_i. For $n = 2$ we see that the distribution is different. As every step uses two samples, we can already see which value will be sorted for $n = 2$ after a total of 3 samples are generated. We can simply add up all the sequences for which i will be sorted second and add up all their respective probabilities.

$$
\begin{aligned}
\mathbb{P}(Q_2(1) = 1) &= p_1p_1p_1 + p_2p_1p_1 + p_3p_1p_1 + p_2p_1p_2 + p_3p_1p_3 \\
&\quad + p_2p_1p_3 + p_3p_1p_2 + p_2p_2p_1 + p_3p_3p_1 \\
&= \frac{1}{8} + \frac{1}{16} + \frac{1}{16} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} \\
&= \frac{7}{16}, \\
\mathbb{P}(Q_2(1) = 2) &= p_2p_2p_2 + p_1p_2p_2 + p_3p_2p_2 + p_1p_2p_1 + p_3p_2p_3 \\
&\quad + p_1p_2p_3 + p_3p_2p_1 + p_1p_1p_2 + p_3p_2p_2 \\
&= \frac{1}{64} + \frac{1}{32} + \frac{1}{64} + \frac{1}{16} + \frac{1}{64} + \frac{1}{32} + \frac{1}{32} + \frac{1}{16} + \frac{1}{64} \\
&= \frac{9}{32} \\
&= \mathbb{P}(Q_2(1) = 3).
\end{aligned}
\tag{2.1}
$$

The distribution for the grouped values is different for low values of n, but in Figure 1 it seems that the distribution of the grouped values converges to the distribution of the samples as $n \to \infty$.

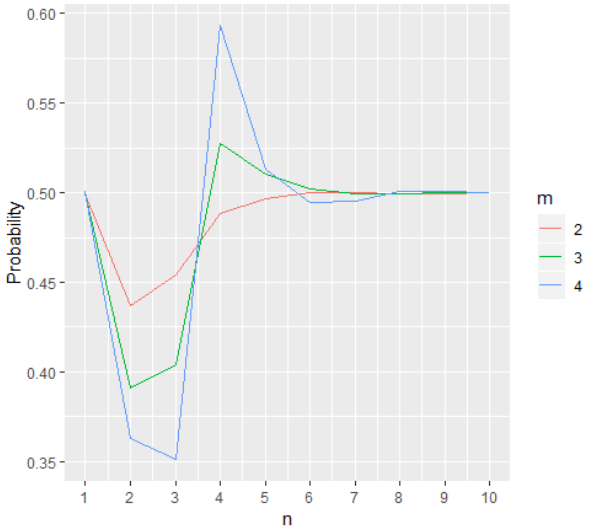


Figure 1: *Simulation with* 10^6 *runs for the probability that the grouped value* $X_{\text{Min}_n+1} = 1$ *after* n *steps with* $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$.

Let us look at the case that $k = 2$. As seen in the proof for the following proposition, there are some properties which only hold for $k = 2$ which enable us to construct an expression for the probabilities of the grouped values:

Proposition 2.1: Convergence distribution of grouped values for $k = 2$

Let $\hat{p}_{n,i}$ be the probability that value $i \in S$ was sorted on step n. If the number of values $k = 2$ it holds that $\hat{p}_{n,i} \xrightarrow{n\to\infty} p_i$.

Proof. Take $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 2) = (1 - p)$. If the value 1 is sorted on step $n + 1$, the value has appeared a multiple of m times before we find the first sample for which we start sorting step $n + 1$. Say we sorted value 1 l times before step $n + 1$ and value 2 $n - l$ times. This means that before we find the $(l \cdot m + 1)$st sample with value 1 we find this value $l \cdot m$ times and the value 2 between $m(n - 1) - l \cdot m + 1$ and $m(n - l)$ times. For $k = 2$ we find the expression

$$\begin{aligned}
\hat{p}_{i,n+1} &= p\left(p^{mn} + \sum_{j=1}^{m}\sum_{l=0}^{n-1}\binom{(n-1)m+j}{lm}(1-p)^{m(n-1)-lm+j}p^{lm}\right) \\
&= p\left(p^{mn} + \sum_{j=1}^{m}(1-p)^{m(n-1)+j}\sum_{l=0}^{n-1}\binom{(n-1)m+j}{lm}(\frac{p}{1-p})^{lm}\right) \qquad (2.2)\\
&= p\left(p^{mn} + \sum_{j=1}^{m}A_{n+1,j}\right).
\end{aligned}$$

To find the limit of this equation we need to simplify $A_{n+1,j}$. One way to do this is to rewrite

it in the form of the binomial formula $a_n \sum_{l=0}^{n} \binom{n}{l} x^l = a_n(1+x)^n$. One difference is that the series in $A_{n+1,j}$ only sums the multiples of m. This can be rewritten by introducing the primitive m-root of unity $\omega = e^{\frac{2\pi i}{m}}$.

Definition 2.1: Primitive m-root of unity

An mth root of unity, where m is a positive integer, is a number ω satisfying the equation

$$\omega^m = 1 \text{ and } \omega^k \neq 1 \text{ for } k = 1, ..., m-1. \tag{2.3}$$

The primitive m root has the property that

$$\frac{1}{m} \sum_{j=0}^{m-1} \omega^{jl} = \begin{cases} 1 & \text{if } l = 0 \mod (m), \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Proof. Let us start with the case that l is a multiple of m, such that $\exists z \in \mathbb{N} : l = z \cdot m$. From the property of the primitive m-root of unity it follows that

$$\sum_{j=0}^{m-1} \omega^{jl} = \sum_{j=0}^{m-1} (\omega^m)^{zj} = \sum_{j=0}^{m-1} 1^{zj} = m. \tag{2.5}$$

Now assume l is not a multiple of m. We find that the sum is the same as

$$1 + (\omega^l) + (\omega^l)^2 + ... + (\omega^l)^{m-1}, \tag{2.6}$$

which is the Geometric series with common factor (ω^l). Therefore, we find that

$$\sum_{j=0}^{m-1} \omega^{jl} = \frac{1 - (\omega^l)^m}{1 - \omega^l} = \frac{1 - (\omega^m)^l}{1 - \omega^l} = \frac{1 - 1^l}{1 - \omega^l} = 0, \tag{2.7}$$

as $\omega^l \neq 1$. □

If we sum over the terms $l = 0, ..., m(n-1) + i$ and multiply by (2.4) we find that all the extra terms that are not a multiple of m are multiplied by 0. In the case that $j = m$, the last term needs to be subtracted as this term is not multiplied by 0. Applying these changes, we find our equation in the form of the binomial formula. For $1 \leq j < m$ we have

$$\begin{aligned} A_{n+1,j} &= \frac{(1-p)^{m(n-1)+j}}{m} \left(\sum_{k=0}^{m(n-1)+j} \binom{m(n-1)+j}{k} \left(\frac{p}{1-p}\right)^k \sum_{l=0}^{m-1} \omega^{lk} \right) \\ &= \frac{(1-p)^{m(n-1)+j}}{m} \left(\sum_{l=0}^{m-1} (1 + \frac{\omega^l p}{1-p})^{m(n-1)+j} \right) \\ &= \frac{1}{m} \sum_{l=0}^{m-1} (1 - p(1-\omega^l))^{m(n-1)+j}. \end{aligned} \tag{2.8}$$

For $j = m$ we find

$$\begin{aligned} A_{n+1,m} &= \frac{(1-p)^{mn}}{m}\left(\sum_{k=0}^{mn}\binom{mn}{k}\left(\frac{p}{1-p}\right)^k\sum_{l=0}^{m-1}\omega^{lk} - m(\frac{p}{1-p})^{mn}\right) \\ &= \frac{(1-p)^{mn}}{m}\left(\sum_{l=0}^{m-1}(1+\frac{\omega^l p}{1-p})^{mn} - m(\frac{p}{1-p})^{mn}\right) \\ &= \frac{1}{m}\left(\sum_{l=0}^{m-1}(1-p(1-\omega^l))^{mn} - mp^{mn}\right). \end{aligned} \tag{2.9}$$

Substituting these results in (2.2) holds

$$\begin{aligned} \lim_{n\to\infty}\hat{p}_{i,n+1} &= \lim_{n\to\infty} p(p^{mn} + \sum_{j=1}^{m-1} A_{n+1,j} + A_{n+1,m}) \\ &= \lim_{n\to\infty}\frac{p}{m}\sum_{j=1}^{m}\sum_{l=0}^{m-1}(1-p(1-\omega^l))^{m(n-1)+j}. \end{aligned} \tag{2.10}$$

As $|(1-p(1-\omega^l))| < 1$ for $l \neq 0$ and is equal to 1 for $l = 0$, we find that all the terms, except for $l = 0$, converge to 0. Therefore,

$$\lim_{n\to\infty}\hat{p}_{i,n+1} = \frac{p}{m}\sum_{j=1}^{m} 1 = p. \tag{2.11}$$

□

Constructing a similar proof for $k > 3$ will be difficult, as we won't be able to use the same properties to find an expression for $\hat{p}_{i,n}$ which is solvable. Instead we will use the bounds for Min_n to prove the following theorem.

Theorem 2.1: Convergence of Min_n/n for $k < \infty$

Let the size of the the outcome space for X_i be $k < \infty$ and let $\hat{p}_{n,i}$ be the probability that value s_i was sorted on step n. Then

$$\text{Min}_n/n \xrightarrow{n\to\infty} m. \tag{2.12}$$

Consequently,

$$\hat{p}_{n,i} \xrightarrow{n\to\infty} p_i \tag{2.13}$$

Proof. Using the lower and upper bound (1.17) for Min_n, it follows that

$$\begin{aligned} &\lim_{n\to\infty}\frac{m\cdot n-(m-1)(k-1)}{n} \leq \lim_{n\to\infty}\frac{\text{Min}_n}{n} \leq \lim_{n\to\infty}\frac{m\cdot n}{n} \\ \Rightarrow &\lim_{n\to\infty} m - \frac{(m-1)(k-1)}{n} \leq \lim_{n\to\infty}\frac{\text{Min}_n}{n} \leq m. \end{aligned} \tag{2.14}$$

As the constant on the left side converges to 0 as $n \to \infty$ the limit is bounded on both sides by m. Therefore, as a result of the Sandwich theorem it holds that

$$\lim_{n\to\infty} \frac{\text{Min}_n}{n} = m. \tag{2.15}$$

This proves the first part of the theorem. Now denote the fraction of times value $i \in S$ has been sorted up till step n as

$$N_i(n) = \frac{1}{n}\#\{1 \le t \le n : Q_{t-1}(1) = i\}, \tag{2.16}$$

so that

$$\mathbb{E}[N_i(n)] = \hat{p}_{n,i}. \tag{2.17}$$

Note that the sample $Q_{n-1}(1)$ is the $(\text{Min}_n + 1)$st sample in the sequence. So we know that the position of $Q_{n-1}(1)$ is in $[mn - (m-1)(k-1) + 1, mn + 1]$. As the samples are i.i.d. it follows that

$$\mathbb{E}[\#\{t \le mn + 1 : X_t = i\}] = p_i(mn + 1). \tag{2.18}$$

Say value i appears s_i times in the first $\text{Min}_n + 1$ samples. Every time this value is sorted, m samples with this value are used. If there are less than m, but more than 0 samples with this value free after a number of sorting steps, the remaining samples with this value will be found after sample $\text{Min}_n + 1$. This means that the value i has been sorted up to $\frac{s_i}{m} \le \lceil \frac{s_i}{m} \rceil \le \frac{s_i}{m} + 1$ times, before we start sorting value $Q_{n-1}(1)$. We know that $\text{Min}_n + 1$ is in $[m \cdot n - (m-1)(k-1) + 1, m \cdot n + 1]$ and for n large it follows from the Law of large numbers that after $m \cdot n$ samples, we see the value i approximately $p_i \cdot m \cdot n$ times. Therefore, we find a lower and upper bound that the value i has been sorted is

$$p_i \frac{m \cdot n - (m-1)(k-1) + 1}{m} \le \#\{1 \le t \le n : Q_{t-1}(1) = i\} \le p_i \frac{m \cdot n + 1}{m} + 1. \tag{2.19}$$

Dividing by n to get the ratio of sorting steps with value i and taking the limit of $n \to \infty$ gives

$$\lim_{n\to\infty} \frac{p_i(m \cdot n - (m-1)(k-1) + 1)}{mn} \le \lim_{n\to\infty} \hat{p}_{n,i} \le \frac{p_i(m \cdot n + 1) + m}{nm}. \tag{2.20}$$

As both sides converge to p_i we find that by the Sandwich theorem that $\hat{p}_{n,i} \xrightarrow{n\to\infty} p_i$. □

Using this result we can now prove that for finite support Max_n has the same rate of growth as Min_n.

Theorem 2.2: Convergence of Max_n/n for $k < \infty$

$$\text{Max}_n/n \xrightarrow{\mathbb{P}} m. \tag{2.21}$$

Proof. We want to show that $\forall \varepsilon > 0$, it holds that

$$\mathbb{P}(\left|\frac{\text{Max}_n}{n} - m\right| > \varepsilon) \xrightarrow{n\to\infty} 0. \tag{2.22}$$

To do this, we will find an upper bound and show that this bound converges to 0.

$$\begin{aligned}\mathbb{P}(|\frac{\text{Max}_n}{n} - m| > \varepsilon) &= \mathbb{P}(\text{Max}_n - m \cdot n > \varepsilon n) + \mathbb{P}(\text{Max}_n < m \cdot n - \varepsilon n) \\ &= \mathbb{P}(\text{Max}_n > m \cdot n + \varepsilon n) \\ &= 1 - \mathbb{P}(\text{Max}_n - m \cdot n \leq \varepsilon n) \\ &\leq 1 - \mathbb{P}(\text{Max}_n - \text{Min}_n \leq \varepsilon n),\end{aligned} \tag{2.23}$$

where the second equality holds as $\text{Max}_n \geq m \cdot n$. The inequality holds as $\text{Min}_n \leq m \cdot n$, such that

$$\mathbb{P}(\text{Max}_n - m \cdot n \leq \varepsilon n) \geq \mathbb{P}(\text{Max}_n - \text{Min}_n \leq \varepsilon n). \tag{2.24}$$

As the probability that $\text{Max}_n - \text{Min}_n \leq \varepsilon n$ is greater than the probability that this holds for all $i = 1, 2, ..., n$ it follows that

$$\mathbb{P}(|\frac{\text{Max}_n}{n} - m| > \varepsilon) \leq 1 - \mathbb{P}(\bigcap_{i=1}^{n} \text{Max}_i - \text{Min}_i \leq \varepsilon n). \tag{2.25}$$

If it holds that the maximum number of samples between two consequent samples with equal value is $\frac{\varepsilon n}{m-1}$ then it follows that $\text{Max}_i - \text{Min}_i \leq \varepsilon n, i = 1, ..., n$, as starting from any X_i, the following $m - 1$ samples with equal values are always found within $(m-1) \cdot \frac{\varepsilon n}{m-1} = \varepsilon n$ samples. Define Y_j as the event that there does not exist $l \in [j+1, ..., j + \lfloor \frac{\varepsilon n}{m-1} \rfloor]$ such that $X_j = X_l$, which has probability

$$p := \mathbb{P}(Y_j = 1) = \sum_{i=1}^{k} p_i (1 - p_i)^{\lfloor \frac{\varepsilon n}{m-1} \rfloor}, \tag{2.26}$$

and expected value

$$\mathbb{E}[Y_j] = 0 \cdot (1 - p) + 1 \cdot p = p. \tag{2.27}$$

As $\text{Max}_n \leq n(m + \varepsilon)$ if the event holds, we take $\hat{n} = \lfloor n(m + \varepsilon) - \frac{\varepsilon n}{m-1} \rfloor$. Now define $Y = \sum_j Y_j, j = 1, ..., \hat{n}$, such that if $Y \geq 1$ it means that there is a X_j for which the first following sample with equal value is more than $\lfloor \frac{\varepsilon n}{m-1} \rfloor$ samples away. Therefore, it holds that

$$1 - \mathbb{P}(\bigcap_{i=1}^{n} \text{Max}_i - \text{Min}_i \leq \varepsilon n) = \mathbb{P}(Y \geq 1). \tag{2.28}$$

As a result of Markov's inequality, it holds that

$$\mathbb{P}(Y \geq 1) \leq \mathbb{E}[Y] = \sum_{j=1}^{\hat{n}} \mathbb{E}[Y_j] = \hat{n}p \xrightarrow{n \to \infty} 0. \tag{2.29}$$

As this is the upper bound for our probability, it follows that

$$\mathbb{P}(|\frac{\text{Max}_n}{n} - m| > \varepsilon) \leq \mathbb{P}(Y \geq 1) \xrightarrow{n \to \infty} 0. \tag{2.30}$$

□

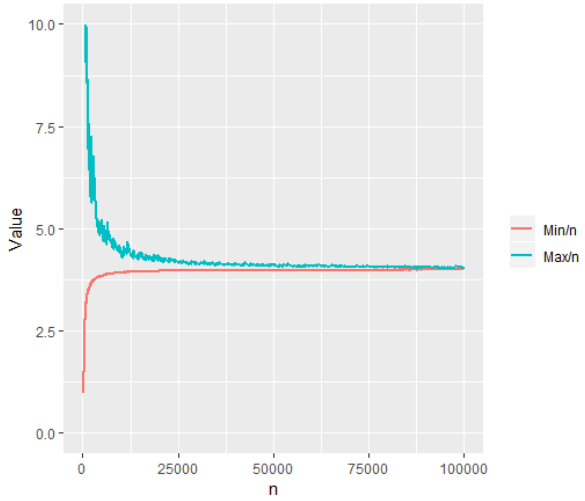


Figure 2: *Simulation of* Min_n/n *and* Max_n/n *for* $X_i \sim Unif[1, 500]$*, and for* $m = 4$

Now that we have shown that both Max_n/n and Min_n/n converge to the same constant, we can prove that $D_n = o(n)$ as seen in Figure 2.

Theorem 2.3: Convergence of $\text{Max}_n/n - \text{Min}_n/n$ for $k < \infty$

For $k < \infty$ it holds that

$$\text{Max}_n/n - \text{Min}_n/n \xrightarrow{\mathbb{P}} 0. \tag{2.31}$$

Proof. $\forall \varepsilon > 0$ it holds that

$$\begin{aligned} \mathbb{P}(|\text{Max}_n/n - \text{Min}_n/n| > \varepsilon) &= \mathbb{P}(|\text{Max}_n/n - m + m - \text{Min}_n/n| > \varepsilon) \\ &\leq \mathbb{P}(|\text{Max}_n/n - m| + |m - \text{Min}_n/n| > \varepsilon), \end{aligned} \tag{2.32}$$

where the inequality results from the triangle inequality. As $|m - \text{Min}_n/n| = |\text{Min}_n/n - m|$ we find

$$\begin{aligned} \mathbb{P}(|\text{Max}_n/n - \text{Min}_n/n| > \varepsilon) &\leq \mathbb{P}(|\text{Max}_n/n - m| + |\text{Min}_n/n - m| > \varepsilon) \\ &\leq \mathbb{P}\left((|\text{Max}_n/n - m| > \frac{\varepsilon}{2}) \cup (|\text{Min}_n/n - m| > \frac{\varepsilon}{2})\right) \\ &\leq \mathbb{P}(|\text{Max}_n/n - m| > \frac{\varepsilon}{2}) + \mathbb{P}(|\text{Min}_n/n - m| > \frac{\varepsilon}{2}), \end{aligned} \tag{2.33}$$

where the last inequality holds as a result of the Union bound. We have already shown that as $\text{Max}_n/n \xrightarrow{\mathbb{P}} m$ and $\text{Min}_n/n \xrightarrow{\mathbb{P}} m$, that both probabilities in the last upper bound converge to 0. It follows that

$$\mathbb{P}(|\text{Max}_n/n - \text{Min}_n/n| > \varepsilon) \xrightarrow{n \to \infty} 0. \tag{2.34}$$

□

2.2 Infinite support

We want to use similar methods as for $k < \infty$ to show that Min_n/n converges for $k = \infty$. To do this we want to show that the Min_n converges to some domain $[m \cdot n - C(n), m \cdot n]$, where $C(n)$ grows slower than n. For $k < \infty$ this is a constant dependent on the number of values. Denote the number of unique values found in the first n samples as U_n. We will show that for any discrete distribution, U_n will grow slower than $O(n)$.

Proposition 2.2: Order of unique values found in a discrete sequence

Let the r.v X_i be discreet. Then $U_n/n \xrightarrow{\mathbb{P}} 0$.

Proof. Take some $\varepsilon > 0$. It follows that

$$\begin{aligned} \mathbb{P}\, U_n \geq \varepsilon n) &\leq \mathbb{P}(\#\{i : X_i \geq \frac{\varepsilon n}{2}\} \geq \frac{\varepsilon n}{2}) \\ &\leq \frac{n\,\mathbb{P}(X_i \geq \frac{\varepsilon n}{2})}{\frac{\varepsilon n}{2}} \\ &= \frac{2}{\varepsilon}\,\mathbb{P}(X_i \geq \frac{\varepsilon n}{2}) \xrightarrow{n\to\infty} 0, \end{aligned} \tag{2.35}$$

where the second inequality is the result of Markov's inequality. □

We can use this result to prove using the similar methods that the $\text{Min}_n/n \xrightarrow{\mathbb{P}} m$ for all values of k.

Theorem 2.4: Convergence of Min_n/n

Let $k \in \mathbb{N} \cup \{\infty\}$, then

$$\text{Min}_n/n \xrightarrow{\mathbb{P}} m. \tag{2.36}$$

Consequently,

$$\hat{p}_{n,i} \xrightarrow{n\to\infty} p_i. \tag{2.37}$$

Proof. We have already showed it holds for $k < \infty$. As we have $U_{mn} < \infty$ unique values in the sequence $(X_i)_{i=1}^{mn}$, we find using the same method used to prove the bounds for Min_n for $k < \infty$ that

$$m \cdot n - (m-1)(U_{mn} - 1) \leq \text{Min}_n \leq m \cdot n, \tag{2.38}$$

holds for $k = \infty$. As we have shown that U_{mn} grows slower than n it follows that

$$\lim_{n\to\infty} \frac{(m-1)(U_{mn}-1)}{n} = 0. \tag{2.39}$$

Therefore, it follows that

$$m \leq \lim_{n\to\infty} \frac{\text{Min}_n}{n} \leq m, \tag{2.40}$$

such that by the Sandwich theorem we find that $\text{Min}_n/n \xrightarrow{\mathbb{P}} m$.
For the second part of the theorem we can use the same steps as for $k < \infty$. By the Law of large numbers, it follows that for n large the value i appears approximately $p_i \cdot m \cdot n$ times in the first $m \cdot n$ samples. Therefore, we find a lower and upper bound that the value i has been sorted is

$$p_i \frac{m \cdot n - (m-1)(U_{mn} - 1) + 1}{m} \leq \#\{1 \leq t \leq n : Q_{t-1}(1) = i\} \leq p_i \frac{m \cdot n + 1}{m} + 1. \quad (2.41)$$

Dividing by n to get the ratio of sorting steps with value i and taking the limit of $n \to \infty$ gives

$$\lim_{n\to\infty} \frac{p_i(m \cdot n - (m-1)(U_{mn} - 1) + 1)}{mn} \leq \lim_{n\to\infty} \hat{p}_{n,i} \leq \frac{p_i(m \cdot n + 1) + m}{nm}. \quad (2.42)$$

As U_{mn} grows slower than n it follows that both sides converge to p_i, such that that by the Sandwich theorem it results that $\hat{p}_{n,i} \xrightarrow{n\to\infty} p_i$. □

In Figure 3 we see the distribution of Min_n/n for $n = 10^4$. Notice that Min_n/n seems to converge to m more slowly for the Negative Binomial distribution. The reason for this can be seen by comparing the probability ratios for both distributions

$$\begin{aligned} \frac{p(1-p)^x}{p(1-p)^{x-1}} &= (1-p) \text{ for } X_i \sim Geo(p) \\ \frac{\binom{x}{r-1} p^r (1-p)^{x-r+1}}{\binom{x-1}{r-1} p^r (1-p)^{x-r}} &= \frac{x(1-p)}{x-r+1} \text{ for } X_i \sim NegBin(p, r). \end{aligned} \quad (2.43)$$

For the Geometric distribution the respective probability of the values decreases faster as the value increases. Therefore, U_n will increase slower such that the bounds for Min_n will also increase slower.

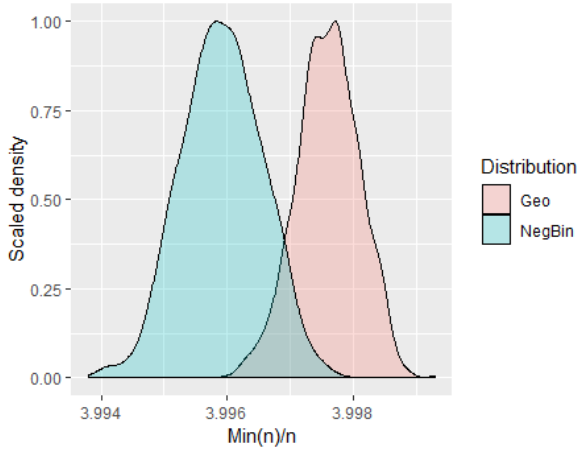


Figure 3: *Plot of the distribution of* Min_n/n *for* $n = 10^4$ *for* $X_i \sim Geo(\frac{1}{2})$ *and* $X_i \sim NegBin(\frac{1}{2}, 5)$. *Number of runs for each distribution is* 10^3.

3 The Markov Chain and its transition probabilities

In this section we will show that the sequence Q_n is a Markov Chain.

Definition 3.1: Markov chain

The process Q_n is a Markov chain if it satisfies the Markov condition:

$$\mathbb{P}(Q_n = q_n | Q_1 = q_1, ..., Q_{n-1} = q_{n-1}) = \mathbb{P}(Q_n = q_n | Q_{n-1} = q_{n-1}), \tag{3.1}$$

for all $n \geq 1$ and all $q_1, ..., q_n \in \Omega$.

This is done by showing that the states the sequence can transition to and its probabilities only depend on its previous state. We divide the outcome states of Q_{n-1} in two different events:

$$\begin{aligned} E_1 &= \{\text{There are } m \text{ or more samples with value } Q_{n-1}(1) \text{ in sequence } Q_{n-1}\} \\ E_2 &= \{\text{There are less than } m \text{ samples with value } Q_{n-1}(1) \text{ in sequence } Q_{n-1}\}. \end{aligned} \tag{3.2}$$

When the event E_1 occurs, there are enough samples in the sequence with equal value to $Q_{n-1}(1)$ so that no more samples have to be generated. In this case we have $L_n = \emptyset$ so the transition is deterministic.
When the event E_2 occurs, there are not enough samples with the correct value in the sequence so that more have to be generated. In this case we have that $L_n \neq \emptyset$ and the transition is random.

3.1 Deterministic case

We will first analyse the case that E_1 occurs. As we base the samples which are sorted on the value of $Q_{n-1}(1)$ we have that $\text{Min}_n > \text{Min}_{n-1}$ always holds, but as $L_n = \emptyset$, no samples are added to the sequence so that $\text{Max}_n = \text{Max}_{n-1}$. Therefore, we find that $d_n = \text{Max}_n - \text{Min}_n < \text{Max}_{n-1} - \text{Min}_n = d_{n-1}$. Namely, if we have that

$$K_n = (X_{\text{Min}_n+1}, ..., X_{\text{Max}_{n-1}}), \tag{3.3}$$

where for the first $m - (\text{Min}_n - \text{Min}_{n-1})$ samples with $k_n(i) = q_{n-1}(1)$ we have $k_n(i) = *$. As all the information for the transition is in the sequence Q_{n-1} and any additional information about the previous states will not effect which value will be sorted, we find that

$$\begin{aligned} &\mathbb{P}(K_n = k_n | Q_{n-1} = q_{n-1}, ..., Q_1 = q_1) \\ =&\mathbb{P}(K_n = k_n | Q_{n-1} = q_{n-1}) = 1. \end{aligned} \tag{3.4}$$

So this transition satisfies the conditions for the Markov Chain.

3.2 Random case

We will now analyse the case that event E_2 occurs, so new samples have to be generated and we have that $L_n \neq \emptyset$. We will divide this event in two sub events:

1. All free samples in q_{n-1} have the same value, but there are less than m.

2. Not all free samples in q_{n-1} have the same value and there are less then m with the same value as $q_{n-1}(1)$.

The reason for this division is because the conditions for these two events differ, which will become clear during the analysis.

Case(1): In this event all the free samples in q_{n-1} are used on step n, but more need to be generated to find a total of m with value $q_{n-1}(1)$. Say we have $m-r$ free samples q_{n-1}, so we have to find r samples with equal value to $q_{n-1}(1)$. This also means that the number of used samples in q_n has to be equal or less then r. So assume the number of used values in q_n is $\alpha \leq r$. This means that the first $r-\alpha$ generated samples have to be equal to $q_{n-1}(1)$, as they are not part of the new sequence which starts at the first sample with a different value. We get the following transition probability.

$$\mathbb{P}(Q_n = q_n | Q_{n-1} = q_{n-1}) = \begin{cases} p^r_{q_{n-1}(1)} & \text{if } d_n = 0 \\ p^{r-\alpha}_{q_{n-1}(1)} \cdot \prod_{i=1}^{d_n} p_{q_n(i)} & \text{if } d_n \geq 2. \end{cases} \tag{3.5}$$

Case(2): As $K_n \neq \emptyset$ we need that

$$K_n = \left(X_{\text{Min}_n+1}, ..., X_{\text{Max}_{n-1}}\right), \tag{3.6}$$

where for all the samples with $k_n(i) = q_{n-1}(1)$ we have $k_n(i) = *$. In this case all the generated samples are part of the sequence as we find sample Min_n in the determined part of the sequence, so the transition probability is

$$\mathbb{P}(Q_n = q_n | Q_{n-1} = q_{n-1}) = \prod_{i=|k_n|+1}^{d_n} p_{l_n(i)}. \tag{3.7}$$

We find that in either case the possible transitions and its probabilities are completely conditioned on just the previous state and have shown that our sequence Q_n is a Markov Chain.

4 Stationary distribution for finite support case

We have shown that for all possible states the conditions for a Markov Chain hold, we want to show this sequence has a stationary distribution if k is finite. To do this we first introduce the following definitions:

Definition 4.1: Irreducible

The Markov chain Q_n is irreducible if for all states $i, j \in \Omega$ it holds that $i \leftrightarrow j$ in a finite number of steps.

Definition 4.2: Positive recurrent

Let $T_i = \min\{n \geq 1 | Q_n = i\}$ be the number of steps till the first return to $i \in \Omega$. Then state i is positive recurrent if $\mathbb{E}[T_i | Q_0 = i] < \infty$.

With these definitions we can use the following theorem:

Theorem 4.1: Stationary distribution [2](227)

The Markov chain Q_n has a well-defined stationary distribution if and only if the chain is irreducible and all the states are positive recurrent.

We will first show that our Markov Chain is irreducible. This means that starting from every state in Ω we are able to transition to any other state in our state space in a finite number of steps. As we have defined our state space Ω as all the possible states reachable after a finite number of steps starting from the empty set, we just need to show it is possible to get from every state back to the empty set. Denote the number of times some value $i \in S$ appears in the sequence Q_l by s_i. To transition to the empty state, we need that all the generated samples are used. For this to happen we need to find every value in the sequence Q_l at least $\hat{s}_i = s_i \mod (m)$ additional times. There is a great number of ways this can happen, but we only need to show it is possible. If we look at the possibility it happens in one specific way, we can see that

$$\mathbb{P}(Q_n = \emptyset | Q_l = q_l) \geq \prod_{i=1}^{k} p^{\hat{s}_i} > 0, \tag{4.1}$$

where $n - l \geq \frac{1}{m} \sum_{i=1}^{k} s_i + \hat{s}_i$, as every step uses m free samples. Now we have shown that starting from any state, there is a non-zero possibility to return to the empty state in a finite number of steps, and so we can conclude that the sequence is irreducible. To show that all the states in Ω are positive recurrent, we start with the following properties.

1. State j is recurrent if $\sum_{n=1}^{\infty} \mathbb{P}(Q_n = j | Q_0 = j) = \infty$. [2](221)

2. A recurrent state j is positive if and only if $\lim_{n\to\infty} \mathbb{P}(Q_n = j) > 0$. [2](222)

3. If $i \leftrightarrow j$, then i is positive recurrent if and only if j is positive recurrent. [2](224)

We have already shown that Q_n is irreducible, such that we know that $i \leftrightarrow j$ for all $i, j \in \Omega$. Therefore, we only need to show for one state that it is positive recurrent. The most obvious choice will again be the empty state. As we know that if for some sequence a_n that

$$\lim_{n\to\infty} a_n > 0 \Rightarrow \sum_{n} a_n = \infty, \tag{4.2}$$

we only need to show that the following theorem holds:

Theorem 4.2: Q_n positive recurrent for $k = 2$

Let $k = 2$, then

$$\lim_{n\to\infty} \mathbb{P}(Q_n = \emptyset) = \frac{1}{m}. \tag{4.3}$$

Consequently Q_n is positive recurrent.

Proof. If $Q_n = \emptyset$ then exactly $m \cdot n$ samples have been generated, which have all been sorted. It follows that both values have been generated exactly a multiple of m times. We can approach this probability with the Negative Binomial distribution:

$$\mathbb{P}(Q_n = \emptyset) = \sum_{i=0}^{n} \binom{mn}{im} p^{im}(1-p)^{mn-im}. \tag{4.4}$$

By using the primitive m-root (2.3) we can rewrite this in the form of the Binomial formula.

$$\begin{aligned}
\mathbb{P}(Q_n = \emptyset) &= \frac{1}{m} \sum_{i=0}^{mn} \binom{mn}{i} p^i (1-p)^{mn-i} \left(\sum_{j=0}^{m-1} \omega^{ji} \right) \\
&= \frac{(1-p)^{mn}}{m} \sum_{j=0}^{m-1} \sum_{i=0}^{mn} \binom{mn}{i} \left(\frac{p\omega^j}{1-p} \right)^i \\
&= \frac{(1-p)^{mn}}{m} \sum_{j=0}^{m-1} \left(1 + \frac{p\omega^j}{1-p} \right)^{mn} \\
&= \frac{1}{m} \sum_{j=0}^{m-1} \left(1 - p(1-\omega^j) \right)^{mn} \xrightarrow{n\to\infty} \frac{1}{m}.
\end{aligned} \tag{4.5}$$

where the limit holds as $|1 - p(1 - \omega^j)| < 1$ for $j \neq 0$ and equal to 1 for $j + 0$. It follows that

$$\sum_{n=0}^{\infty} \mathbb{P}(Q_n = \emptyset) = \infty, \tag{4.6}$$

such that Q_n is recurrent. As $\lim_{n\to\infty} \mathbb{P}(Q_n = \emptyset) > 0$ it follows that it is positive recurrent. □

We have now shown that Q_n is both irreducible and positive recurrent for $k = 2$. Therefore, it follows that Q_n has a stationary distribution for $k = 2$.

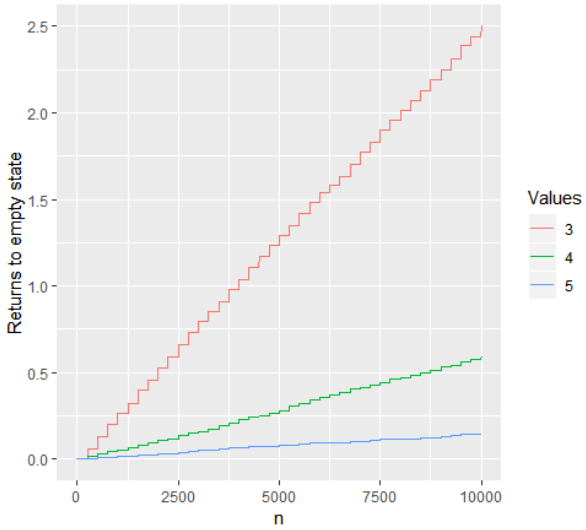


Figure 4: Return time for Uniform distribution with $m = 4$.

To prove that Q_n has a stationary distribution for $k < \infty$ we will extend the proof used for $k = 2$.

Theorem 4.3: Stationary distribution for Q_n

The Markov chain Q_n has a well-defined stationary distribution if and only if $k < \infty$.

Proof. We have already shown that for $k \in \mathbb{N} \cup \{\infty\}$ the Markov chain is irreducible. We shall use a proof by induction to show Q_n is positive recurrent for all $k < \infty$ and use this result to show that this is not the case for $k = \infty$.
Let us start by proving it holds for $k = 3$ by extending the prove used for $k = 2$. If Q_n is the empty list, it follows that all three values appear exactly a multiple of m times in the first $m \cdot n$ samples. Take $p_3 = 1 - p_1 - p_2$, such that

$$\begin{aligned} \mathbb{P}(Q_n = \emptyset) &= \sum_{i=0}^{n} \binom{m \cdot n}{i \cdot m} p_2^{i \cdot m} \sum_{j=0}^{n-i} \binom{m \cdot (n-i)}{j \cdot m} p_1^{j \cdot m} (1 - p_1 - p_2)^{m \cdot (n-i) - j \cdot m} \\ &= \sum_{i=0}^{n} \binom{m \cdot n}{i \cdot m} p_2^{i \cdot m} A_{2,n-i}. \end{aligned} \tag{4.7}$$

We can simplify $A_{2,n-i}$ by taking the sum over all $m \cdot (n - j)$ terms, instead of just multiples

of m, and multiplying with (2.4) such that all the extra terms are multiplied with 0.

$$\begin{aligned} A_{2,n-i} &= \frac{1}{m}\sum_{j=0}^{m\cdot(n-i)}\binom{m\cdot(n-i)}{j}p_1^j(1-p_1-p_2)^{m\cdot(n-i)-j}\left(\sum_{l=0}^{m-1}\omega^{jl}\right) \\ &= \frac{(1-p_1-p_2)^{m\cdot(n-i)}}{m}\sum_{l=0}^{m-1}\sum_{j=0}^{m\cdot(n-i)}\binom{m\cdot(n-i)}{j}\left(\frac{p_1\cdot\omega^l}{1-p_1-p_2}\right)^j \\ &= \frac{(1-p_1-p_2)^{m(n-i)}}{m}\sum_{l=0}^{m-1}\left(1+\frac{p_1\cdot\omega^l}{1-p_1-p_2}\right)^{m(n-i)} \\ &= \frac{1}{m}\sum_{l=0}^{m-1}\Big(1-p_1(1-w^l)-p_2\Big)^{m(n-i)} \end{aligned} \tag{4.8}$$

Substituting this in (4.7) we can simplify by using the primitive m-root once more.

$$\begin{aligned} \sum_{i=0}^{n}\binom{m\cdot n}{i\cdot m}p_2^{i\cdot m}A_{2,n-i} &= \frac{1}{m^2}\sum_{z=0}^{m-1}\sum_{l=0}^{m-1}\sum_{i=0}^{m\cdot n}\binom{m\cdot n}{i}(p_2\cdot\omega^z)^i\Big(1-p_1(1-w^l)-p_2\Big)^{mn-i} \\ &= \sum_{z=0}^{m-1}\sum_{l=0}^{m-1}\frac{\big(1-p_1(1-w^l)-p_2\big)^{mn}}{m^2}\sum_{i=0}^{m\cdot n}\binom{m\cdot n}{i}\left(\frac{p_2\cdot\omega^z}{1-p_1(1-w^l)-p_2}\right)^i \\ &= \sum_{z=0}^{m-1}\sum_{l=0}^{m-1}\frac{\big(1-p_1(1-w^l)-p_2\big)^{mn}}{m^2}\left(1+\frac{p_2\cdot\omega^z}{1-p_1(1-w^l)-p_2}\right)^{m\cdot n} \\ &= \frac{1}{m^2}\sum_{z=0}^{m-1}\sum_{l=0}^{m-1}\Big(1-p_1(1-\omega^l)-p_2(1-\omega^z)\Big)^{m\cdot n}. \end{aligned} \tag{4.9}$$

As $|1-p_1(1-\omega^l)-p_2(1-\omega^z)|<1$ for all values of z and l, except for $z=l=0$ for which the value is 1. Therefore, it follows that

$$\mathbb{P}(Q_n=\emptyset)\xrightarrow{n\to\infty}\frac{1}{m^2}, \tag{4.10}$$

for $k=3$. To simplify notation, denote

$$g(i_1,...,i_{k-2})=\big(1-p_1(1-\omega^{i_1})-...-p_{k-3}(1-\omega^{i_{k-3}})-p_{k-2}\big). \tag{4.11}$$

Assume that for $k-1$ values it holds that

$$A_{k-1,n}=\frac{1}{m^{k-2}}\sum_{i_1=0}^{m-1}...\sum_{i_{k-2}=0}^{m-1}g(i_1,...,i_{k-2})^{m\cdot n}. \tag{4.12}$$

We have shown that this holds for $k=3$, but now we want to show that it holds for all $k\in\mathbb{N}$.

Using the same method as for $k = 3$ we find

$$\begin{aligned}
\mathbb{P}(Q_n = \emptyset) &= \sum_{j=0}^{n} \binom{m \cdot n}{j \cdot m} p_{k-1}^{j \cdot m} A_{k-1,n-j} \\
&= \sum_{i_1=0}^{m-1} \dots \sum_{i_{k-1}=0}^{m-1} \frac{g(i_1, \dots, i_{k-2})^{m \cdot n}}{m^{k-1}} \sum_{j=0}^{m \cdot n} \binom{m \cdot n}{j} \left(\frac{p_{k-1} \cdot \omega^{i_{k-1}}}{g(i_1, \dots, i_{k-2})} \right)^j \\
&= \sum_{i_1=0}^{m-1} \dots \sum_{i_{k-1}=0}^{m-1} \frac{g(i_1, \dots, i_{k-2})^{m \cdot n}}{m^{k-1}} \left(1 + \frac{p_{k-1} \cdot \omega^{i_{k-1}}}{g(i_1, \dots, i_{k-2})} \right)^{m \cdot n} \\
&= \frac{1}{m^{k-1}} \sum_{i_1=0}^{m-1} \dots \sum_{i_{k-1}=0}^{m-1} \left(1 - p_1(1 - \omega^{i_1}) - \dots - p_{k-1}(1 - \omega^{i_{k-1}}) \right)^{m \cdot n} .
\end{aligned} \tag{4.13}$$

As $|1 - p_1(1 - \omega^{i_1}) - \dots - p_{k-1}(1 - \omega^{i_{k-1}})| < 1$ if $\exists j : i_j \neq 0$ and is 1 if $\forall j, i_j = 0$, it follows that

$$\mathbb{P}(Q_n = \emptyset) \xrightarrow{n \to \infty} \frac{1}{m^{k-1}} > 0, \forall k \in \mathbb{N} . \tag{4.14}$$

We have found that for all $k \in \mathbb{N}$ the sequence Q_n is irreducible and positive recurrent, such that it has a stationary distribution. Another result we have found is that the probability $\mathbb{P}(Q_n = \emptyset)$ only depends on m and k for large n. We can use this to prove that for $k = \infty$ the empty state is null recurrent by taking $k \to \infty$.

$$\lim_{k \to \infty} \lim_{n \to \infty} \mathbb{P}(Q_n = \emptyset) = \lim_{k \to \infty} \frac{1}{m^{k-1}} = 0, \tag{4.15}$$

from which we conclude the empty state is null recurrent. Therefore, as Q_n is irreducible it follows that every state is null recurrent, so there is no stationary distribution for $k = \infty$. □

As this proof does not clearly show why there is no stationary distribution for $k = \infty$, a different proof is introduced the following section.

4.1 Simulations

In Figure 5 we see the value of D_n for $n = 0, 1, ..., 10000, m = 5$ and $X_i \sim Unif[1, 500]$ for eight different simulations.

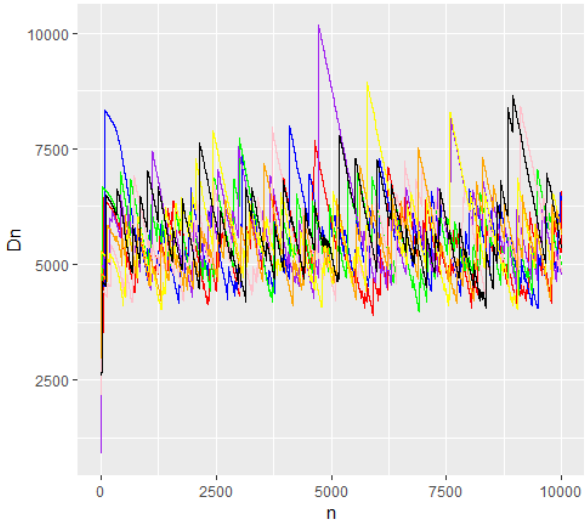


Figure 5: *Simulation of D_n for 500 Uniformly distributed values, and for $m = 5$*

In Figure 6 we see the empirical distribution for D_n with $m = 4$ and $X_i \sim Unif[1, 500]$ at $n = 10^4$, $n = 5 * 10^4$ and $n = 10^5$. Number of runs for every n is 10^3.

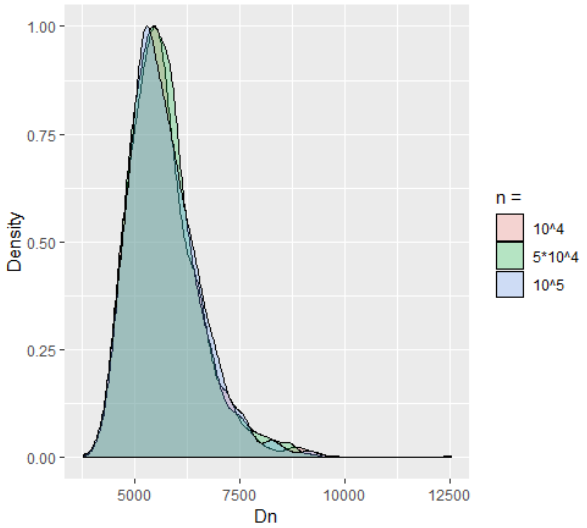


Figure 6: *Simulation of density for D_n for $X_i \sim Unif[1, 500]$ and $m = 4$. Number of runs 10^3.*

5 Divergence for infinite support case

Last section we have shown that Q_n has a stationary distribution if and only if $k < \infty$. This section we will analyse the behaviour of Q_n for $k = \infty$ by proving that $k = \infty$ has no stationary distribution in another way. In contrary to this case the outcome space S has no maximum value if $k = \infty$. This means that as more samples are being generated, the lowest respective probability that is found in the sequence will eventually decrease. As an increasingly number of samples will need to be generated to sort these values, the maximum value for D_n will increase with n. If this happens frequently enough, the length of the sequence will keep jumping in value before it has a chance to sort the added samples. An example of this for the Geometric distribution we can see in Figure 7. Although if these jumps happen infrequently, it could be the case that the process has the time to sort the added samples before the next jump in sequence length. If we look at some probability which decreases as n grows $p = \frac{a}{n^j}, a > 0$ its probability that we find the value with this respective probability in the first n samples converges to

$$\lim_{n\to\infty} 1 - (1 - \frac{a}{n^j})^n = \begin{cases} 0 & \text{for } j > 1, \\ 1 - e^{-a} & \text{for } j = 1, \\ 1 & \text{for } 0 < j < 1. \end{cases} \tag{5.1}$$

As we want these jumps to happen frequently, we define $m_n = \inf\{i \in S : \mathbb{P}(X > i) \leq \frac{1}{n}\}$ and assume it has probability $p_{m_n} = \frac{a}{n^j}, j \leq 1$. With these conditions we will prove that D_n will hold increasingly greater values as n grows and therefore has no stationary distribution as a consequence.

Proposition 5.1: Q_N has no limiting distribution for S infinite

If the size of the outcome space $k = \infty$ and for $m_n = \inf\{i \in S : \mathbb{P}(X > i) \leq \frac{1}{n}\}$ with $p_{m_n} = \frac{a}{n^j}, j \leq 1$ then $\forall \varepsilon > 0$ then

$$\limsup_{n\to\infty} \mathbb{P}(\text{Max}_n - \text{Min}_n \geq \varepsilon n) = c_\varepsilon > 0. \tag{5.2}$$

Consequently Q_n has no stationary distribution.

Proof. 1.) Let the r.v. G_n be the number of samples generated before we find a sample with value m_n. We find that $\forall l > 0$ that

$$\mathbb{P}(G_n > ln) \geq \left(1 - \frac{a}{n^j}\right)^{ln} \geq \left(1 - \frac{a}{n}\right)^{ln} \xrightarrow{n\to\infty} e^{-al} > 0. \tag{5.3}$$

As $\mathbb{P}(\text{Min}_n \leq m \cdot n) = 1$ we find a lower bound by taking the probability of just one event for which $D_n \geq \varepsilon n$, which is the event that there is a sample with value m_n in the first n

generated samples and the next is generated after more than $(\varepsilon + m)n$ samples. Therefore,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(\text{Max}_n - \text{Min}_n \geq \epsilon n) &\geq \lim_{n\to\infty} \left(1 - (1 - \frac{a}{n^j})^n\right) \left(1 - \frac{a}{n^j}\right)^{n(\varepsilon+m)} \\ &\geq \lim_{n\to\infty} \left(1 - (1 - \frac{a}{n})^n\right) \left(1 - \frac{a}{n}\right)^{n(\varepsilon+m)} \\ &= (1 - e^{-a})e^{-a(\varepsilon+m)} > 0.\end{aligned} \tag{5.4}$$

2.) Now we show shall use a proof by contradiction that Q_n has no stationary distribution. Denote $\pi_j^{(n)} = \mathbb{P}(Q_n = j), j \in \Omega$. Say (5.2) holds and Q_n has a limiting distribution such that $\pi^{(n)} \xrightarrow{n\to\infty} \boldsymbol{\pi}$. We have that

$$\mathbb{P}(\text{Max}_n - \text{Min}_n = l) = \sum_{j\in\Omega:|j|=l} \mathbb{P}(Q_n = j) \xrightarrow{n\to\infty} \sum_{j\in\Omega:|j|=l} \pi(j) = q_l, \tag{5.5}$$

with $\sum_{i=0}^{\infty} q_i = 1$. It should hold that $\forall\delta > 0\ \exists l(\delta) \in \mathbb{N} : \forall l \geq l(\delta)$ we have that

$$\sum_{i=l+1}^{\infty} q_i < \delta. \tag{5.6}$$

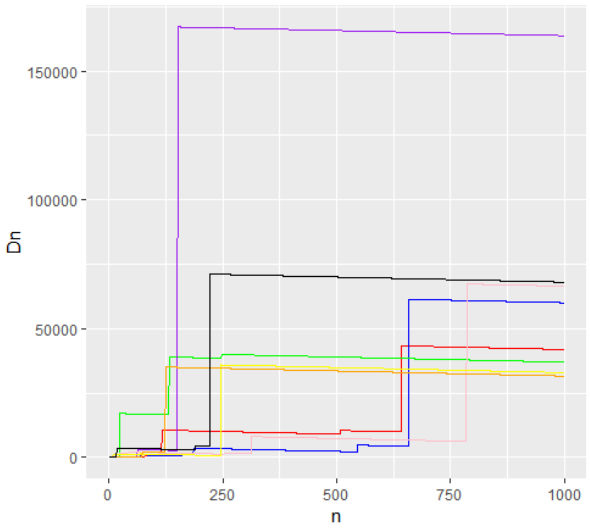
This would mean that

$$\lim_{n\to\infty} \mathbb{P}(\text{Max}_n - \text{Min}_n \geq l + 1) \xrightarrow{l\to\infty} 0. \tag{5.7}$$

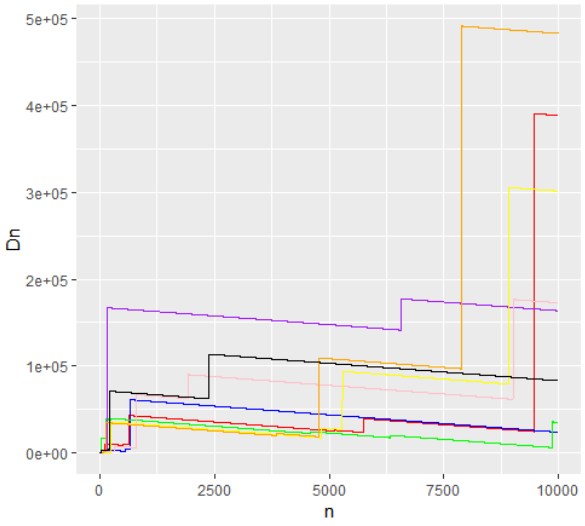
This contradicts (5.2) so we can conclude that there is no limiting distribution. □

5.1 Geometric distribution

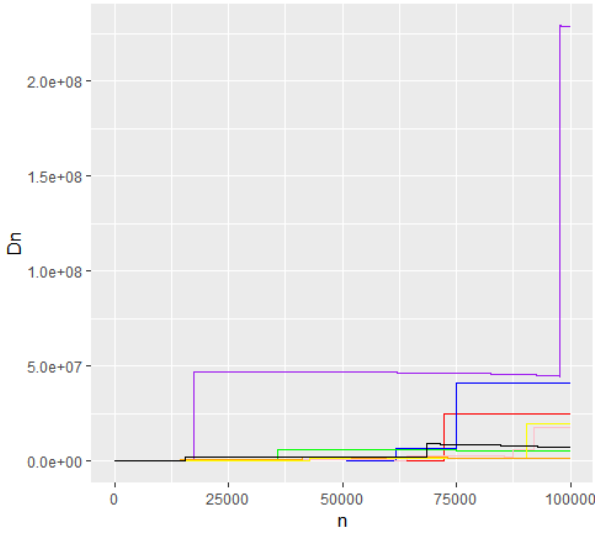
We will show that our theorem holds for the Geometric distribution. In Figure 7 we see sudden jumps in the value of D_n, which happen if a value with a respective low probability is sorted.



(a) 1000 sorting steps



(b) 10.000 sorting steps



(c) 100.000 sorting steps

Figure 7: *Value of D_n for $n = 1, ..., 100.000$ with $X_i \sim Geo(\frac{1}{2})$ for eight simulations.*

Take the Geometric distributionof the form

$$\mathbb{P}(X_i = x) = p(1-p)^{x-1}, \text{for } x = 1, 2, ..., \tag{5.8}$$

As the probability decreases as the value rises, we would like to find the distribution of the maximum, and so the least likely, value in the first n samples $M_n = \max\{X_1, ..., X_n\}$.

$$\begin{aligned} \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, ..., X_n \leq x) \\ &= \mathbb{P}(X_i \leq x)^n \\ &= (1 - (1-p)^x)^n. \end{aligned} \tag{5.9}$$

We can use this to prove the following theorem:

Theorem 5.1: Almost sure converge of maximum geometric sequence

Let $M_n = \max\{X_1, ..., X_n\}$ with $X_i \sim Geo(p)$. Then for every $\epsilon > 0$

$$\mathbb{P}(|\frac{M_n}{\log(n)} - \frac{1}{\log(\frac{1}{1-p})}| > \varepsilon) \xrightarrow{n\to\infty} 0, \tag{5.10}$$

so that $M_n/\log(n) \xrightarrow{\mathbb{P}} \frac{1}{\log(\frac{1}{1-p})}$.

Proof. Take $q = \frac{1}{1-p}$.

$$\begin{aligned}\mathbb{P}(|\frac{M_n}{\log(n)} - \frac{1}{\log(q)}| > \epsilon^*) &= \mathbb{P}(|\frac{M_n}{\log(n)} - \frac{1}{\log(q)}| > -\frac{\epsilon}{\log(q)}) \\ &= \mathbb{P}(\frac{M_n}{\log(n)} - \frac{1}{\log(q)} > \frac{\varepsilon}{\log(q)}) + \mathbb{P}(\frac{1}{\log(q)} - \frac{M_n}{\log(n)} > \frac{\varepsilon}{\log(q)}) \\ &= \mathbb{P}(M_n > (1+\varepsilon)\log_q(n)) + \mathbb{P}(M_n < (1-\varepsilon)\log_q(n)).\end{aligned} \tag{5.11}$$

As M_n only takes integer values, we find that

$$\mathbb{P}(M_n > (1+\varepsilon)\log_q(n)) = \mathbb{P}(M_n > \lfloor(1+\varepsilon)\log_q(n)\rfloor) \leq 1-(1-(1-p)^{(1+\varepsilon)\log_q(n)-1})^n, \tag{5.12}$$

and

$$\mathbb{P}(M_n < (1-\varepsilon)\log_q(n)) = \mathbb{P}(M_n \leq \lfloor(1-\varepsilon)\log_q(n)\rfloor) \leq (1-(1-p)^{(1-\varepsilon)\log_q(n)})^n. \tag{5.13}$$

We can simplify this upper bound by taking

$$(1-p)^{(1-\varepsilon)\log_q(n)} = (1-p)^{\frac{\log(n^{-(1-\varepsilon)})}{\log(1-p)}} = n^{-(1-\varepsilon)}. \tag{5.14}$$

If we do this for both upper bounds, it follows that

$$\begin{aligned}\mathbb{P}(|\frac{M_n}{\log(n)} - \frac{1}{\log(q)}| > \epsilon^*) &\leq 1 - (1 - \frac{q}{n^{1+\varepsilon}})^n + (1 - \frac{1}{n^{1-\varepsilon}})^n \\ &= 1 - \left((1 - \frac{q}{n^{1+\varepsilon}})^{n^{1+\varepsilon}}\right)^{\frac{1}{n^{\varepsilon}}} + \left((1 - \frac{1}{n^{1-\varepsilon}})^{n^{1-\varepsilon}}\right)^{\frac{1}{n^{-\varepsilon}}} \\ &\approx 1 - \left(\frac{1}{e^q}\right)^{\frac{1}{n^{\varepsilon}}} + \left(\frac{1}{e}\right)^{n^{\varepsilon}} \xrightarrow{n\to\infty} 0.\end{aligned} \tag{5.15}$$

□

This theorem shows us that for large values of n, M_n is close to $\frac{\log(n)}{\log(q)}$. So assume n has a large value and take $m_n = \lfloor\frac{\log(n)}{\log(q)}\rfloor$ and look at the probability $p_m(n)$ that we find m_n. Take the random value G_n as the number of steps it takes to find the next sample with value m_n. We want to show that for some $l \in \mathbb{R}$, we have that

$$\mathbb{P}(G_n > ln) = (1 - p_m(n))^{ln} \geq c_l > 0. \tag{5.16}$$

We need to compute

$$p_m(n) = \mathbb{P}(X_i = \lfloor \frac{\log(n)}{\log(q)} \rfloor) = p(1-p)^{\lfloor \frac{\log(n)}{\log(q)} \rfloor - 1}, \tag{5.17}$$

or an acceptable upper bound as the value of (5.16) is lower for higher values of $p_m(n)$. As $\lfloor \frac{\log(n)}{\log(q)} \rfloor > \frac{\log(n/q)}{\log(q)} = \frac{\log(n)}{\log(q)} - 1$ this can be used to find an upper bound,

$$\begin{aligned} p(1-p)^{\lfloor \frac{\log(n)}{\log(q)} \rfloor - 1} &\leq p(1-p)^{\frac{\log(n)}{\log(q)} - 2} \\ &= p(1-p)^{\log_{1-p}(n^{-1}) - 2} \\ &= \frac{p}{n(1-p)^2}. \end{aligned} \tag{5.18}$$

Substituting this in (5.16) we find

$$\mathbb{P}(G_n > ln) > (1 - \frac{p}{n(1-p)^2})^{ln} \xrightarrow{n\to\infty} e^{-\frac{pl}{(1-p)^2}} > 0. \tag{5.19}$$

Now as $\text{Min}_n \leq mn$, we can take $l = 2m$ to find

$$\mathbb{P}(\text{Max}_n - \text{Min}_n \geq mn) > e^{-\frac{2mp}{1-p}} > 0. \tag{5.20}$$

Therefore, by Proposition 5.1, we can conclude that Q_n has no stationary distribution if $X_i \sim Geo(p)$.

6 Conclusions and open problems

The main results this thesis is that there only exists a stationary distribution if and only if the random variables have finite support. In this case the position of the first free sample and the last used sample have the same rate of growth. An explanation for this behaviour is that as there are a finite number of values, these will eventually all have been generated. Therefore, the maximal number of samples needed to complete a sorting step will eventually stop increasing. For the infinite support case, the number of visits to the empty state, the state where all the generated samples are sorted, is finite. As the lowest respective probability that is found in the sequence decreases as the number of generated samples increases. Therefore, the maximal number of samples needed to be generated to complete a sorting step keeps increasing, such that the probability that every sample is generated decreases.

Due to time constraint there are still some open problems which require further analysis. For the finite support case we know there exists a stationary distribution, but we have not found what this distribution is. The only state for which we found the probability is the empty state for which the respective probability is $\frac{1}{m^{k-1}}$. Another point of analysis is the rate at which the probabilities converge to the stationary distribution.
For the infinite support case it only the rate of growth for the position of the first free sample was proven, namely the same rate as for the finite support case. In Figure 8 we see the results of simulations for $X_i \sim Geo(\frac{1}{2})$ any various sorting steps for the position of the last used samples divided by the number of sorting steps n. From the simulations it seems that the position converges in distribution $n \cdot Y$, where Y is some random variable.

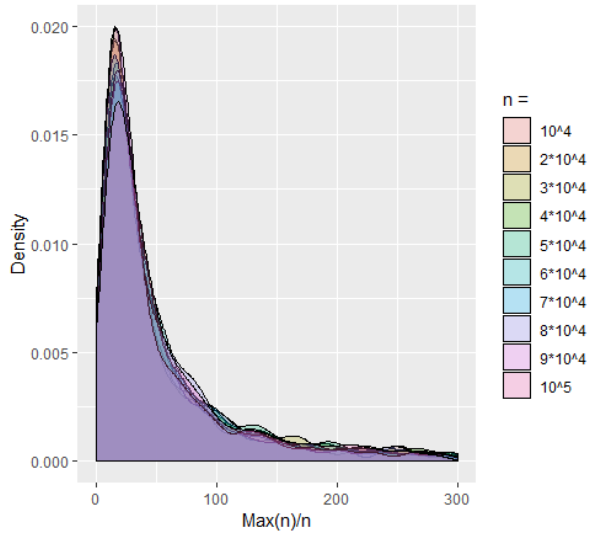


Figure 8: *Density of* Max_n/n *for different values of* n*, for* $X_i \sim Geo(\frac{1}{2})$*. Number of runs for every* n *is* 10^3*.*

References

[1] Pierre Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer, 1999

[2] Geoffrey Grimmett & David Stirzaker, *Probability and Random Processes*, Oxford, 2001