CS648 Project

Aayush Gupta (210020, aayushg21@iitk.ac.in)
Ridin Datta (210840, ridind21@iitk.ac.in)
Shrey Wadhawan (200948, shreyw20@iitk.ac.in)

1 Abstract

This study presents three proofs regarding the expected length of the smallest interval among $n+1$ intervals formed by randomly selecting n points from the $[0,1]$ line segment. The first proof establishes a Θ bound, the second employs geometric arguments, and the third utilizes probabilistic techniques.

2 Problem Statement

Finding the expected length of the smallest interval

n points are picked randomly uniformly and independently from the $[0,1]$ line segment. This will create $n+1$ intervals. What is the expected length of the smallest interval among these intervals?

3 Empirical Observations and Visualiser

For empirical observations we created a python script to simulate the problem statement and drew inferences from it.
We also created an interactive graphic visualiser for visualizing the problem statement. Please check it out at https://rhydberg.github.io/cs648-visualizer/

4 Claim

The expected length of the smallest interval is $\Theta(n^{-2})$

5 Proofs

5.1 Proof by Randomized Incremental Construction

Notation
δ_i := Length of the smallest interval after dropping i points randomly and uniformly on the $[0,1]$ line

5.1.1 Lemma

$$\alpha \cdot (i+1) \leq \mathbb{P}(\delta_{i+1} < \delta_i | \delta_i = \alpha) \leq 2\alpha \cdot (i+1) \tag{1}$$

Proof
Let us call the region where dropping the $(i+1)$th point causes $\delta_{i+1} < \delta_i$ the "favourable" region. If $\delta_i = \alpha$, for maximum possible favourable region all the i neighbouring points need

to be separated by a distance of atleast 2α. Moreover the distance between the leftmost point and 0 and the rightmost point and 1 individually need to be greater than 2α. Since the length of the favourable region directly translates to probability of the $(i+1)$th point falling in the favourable region, from these arguments, $\mathbb{P}(\delta_{i+1} < \delta_i | \delta_i = \alpha) \leq 2\alpha \cdot (i+1)$

The situation representing the minimum possible favourable region is the case where all neighbouring points are separated by a distance of α including the distance between 0 and the leftmost point and 1 and the rightmost point. Thus $\alpha \cdot (i+1) \leq \mathbb{P}(\delta_{i+1} < \delta_i | \delta_i = \alpha)$

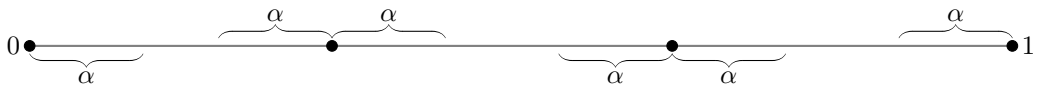

Figure 1: Illustration of maximum possible favourable region for 2 points

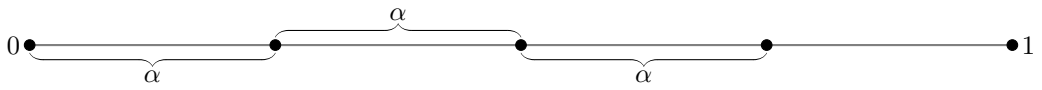

Figure 2: Illustration of minimum possible favourable region for 3 points

5.1.2 Lemma

$$\mathbb{P}(\delta_{i+1} < \delta_i) = \frac{2}{i+1} \tag{2}$$

Proof
In the randomized incremental construction framework let us consider that i points have been dropped. The smallest interval is defined by 2 points. Let us call them $\hat{p}$ and $\hat{q}$. If either $\hat{p}$ or $\hat{q}$ are chosen as the $(i+1)$th point they would define the smallest interval. Otherwise the smallest interval remains the same as δ_i. Since the points have been chosen random uniformly, every point has the same probability of being the last dropped point. Hence $\mathbb{P}(\delta_{i+1} < \delta_i) = \frac{2}{i+1}$

From (1)

$$\alpha \cdot (i+1) \leq \mathbb{P}(\delta_{i+1} < \delta_i | \delta_i = \alpha) \leq 2\alpha \cdot (i+1)$$

$$\implies \alpha \cdot (i+1) \cdot \mathbb{P}(\delta_i = \alpha) \leq \mathbb{P}(\delta_{i+1} < \delta_i | \delta_i = \alpha) \cdot \mathbb{P}(\delta_i = \alpha) \leq 2\alpha \cdot (i+1) \cdot \mathbb{P}(\delta_i = \alpha)$$

$$\implies \int \alpha \cdot (i+1) \cdot \mathbb{P}(\delta_i = \alpha) d\alpha \leq \int \mathbb{P}(\delta_{i+1} < \delta_i | \delta_i = \alpha) \cdot \mathbb{P}(\delta_i = \alpha) d\alpha \leq \int 2\alpha \cdot (i+1) \cdot \mathbb{P}(\delta_i = \alpha) d\alpha$$

$$\implies (i+1)\mathbb{E}[\delta_i] \leq \mathbb{P}(\delta_{i+1} < \delta_i) \leq 2(i+1)\mathbb{E}[\delta_i]$$

$$\implies \frac{\mathbb{P}(\delta_{i+1} < \delta_i)}{2(i+1)} \leq \mathbb{E}[\delta_i] \leq \frac{\mathbb{P}(\delta_{i+1} < \delta_i)}{i+1}$$

From (2)

$$\frac{1}{(i+1)^2} \leq \mathbb{E}[\delta_i] \leq \frac{2}{(i+1)^2}$$

Thus

$$\mathbb{E}[\delta_i] = \Theta(i^{-2})$$

$$\implies \mathbb{E}[\delta_n] = \Theta(n^{-2})$$

QED

5.2 Proof by geometric arguments

Claim

The kth largest interval's expected length is equal to

$$\frac{\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{n+1}}{n+1}$$

Hence expected length of the smallest interval is

$$\frac{1}{(n+1)^2}$$

Proof. Without loss of generality, assume the $[0, 1]$ segment is broken into segments of length $s_1 \geq s_2 \geq \cdots \geq s_n \geq s_{n+1}$, in that order.

Note that,

$$\sum_{i=1}^{n+1} s_i = 1 \tag{3}$$

Define:

$$x_i := s_i - s_{i+1}, \forall i \in [1, n] \quad \text{with} \quad x_{n+1} = s_{n+1}$$

This implies,

$$s_i = \sum_{k=i}^{n+1} x_k \tag{4}$$

Note that $x_i \geq 0$ for each i

Substituting (2) in (1)

$$\implies \sum_{i=1}^{n+1} \sum_{k=i}^{n+1} x_k = x_1 + 2x_2 + \cdots + (n+1)x_{n+1} = \sum_{i=1}^{n+1} ix_i = 1 \tag{5}$$

Define: $y_i = ix_i$, observe that $x_i \in [0, \frac{1}{i}]$ and $y_i \in [0, 1]$ for all i.

Substituting in (3)

$$\implies y_1 + \cdots + y_{n+1} = 1$$

Claim: $\mathbb{E}[y_i] = \frac{1}{n+1}$ for all i

Illustration: We drop $n = 3$ points forming intervals s_1, s_2, s_3, s_4 with s_1 being the largest and s_4 being the smallest interval. We denote $s_4 = x_4$ and proceed in an incremental manner to mark $x_i = s_i - s_{i+1}$.

(***Note:*** The intervals are drawn only in an illustrative manner and can be in any order, not necessarily in the order shown in Figure-3. Each of the s_i are separated by darker lines and different colors represent different x_i)

Consider the rearrangement in Figure-4 where each ix_i are separated by darker lines. Once these darker lines are placed, the lighter lines are *uniquely* determined by the dark lines as they equally partition each subrectangle formed by ix_i. In this case, these three random lines can be put anywhere randomly uniformly independently on the side with total length 1. Hence the expected value of each such ix_i = $\frac{1}{4} = \frac{1}{n+1}$.

Proof: There is a clear bijection between the representation of any configuration $(s_1; s_2; \ldots; s_{n+1})$ (in the general case) between Figure 1 and Figure 2.

We have already shown the construction of Figure 4 from Figure 3 via a simple rearrangement i.e. Figure 3 $\implies$ Figure 4

To show the reverse, Figure 4 $\implies$ Figure 3, follow the steps:

1. On the side with length 1, randomly draw n vertical lines. Call the $n+1$ regions so formed $y_1, y_2, \ldots, y_{n+1}$.
2. Divide y_i into i equal parts, such that $x_i = y_i/i$ is the length of one part of y_i.
3. Now, there are i copies of x_i. Rearrange them into groups such that $([x_1, x_2, \ldots, x_{n+1}]; [x_2, x_3, \ldots, x_{n+1}]; \ldots; [x_n, x_{n+1}]; [x_{n+1}]) = s_1; s_2; \ldots; s_n; s_{n+1})$.

Since the division of y_i into subparts of length x_i in Step 3 takes place after Step 1. So, step 3 does not affect the expected value of y_i in step 1. As the n lines were drawn randomly uniformly, we get $\mathbb{E}[y_i] = \mathbb{E}[y_j]$, $\forall i \neq j$.

Hence,

$$\sum_{i=1}^{n+1} y_i = 1 \implies \mathbb{E}[y_i] = \frac{1}{n+1}$$

Using the above claim,

$$\implies \mathbb{E}[x_i] = \frac{\mathbb{E}[y_i]}{i} = \frac{1}{i(n+1)}$$

for each $i \in [1, n+1]$

By linearity of expectation,

$$\implies \mathbb{E}[s_i] = \mathbb{E}[x_i] + \cdots + \mathbb{E}[x_{n+1}] = \frac{1}{n+1}\left(\frac{1}{i} + \cdots + \frac{1}{n+1}\right)$$

QED

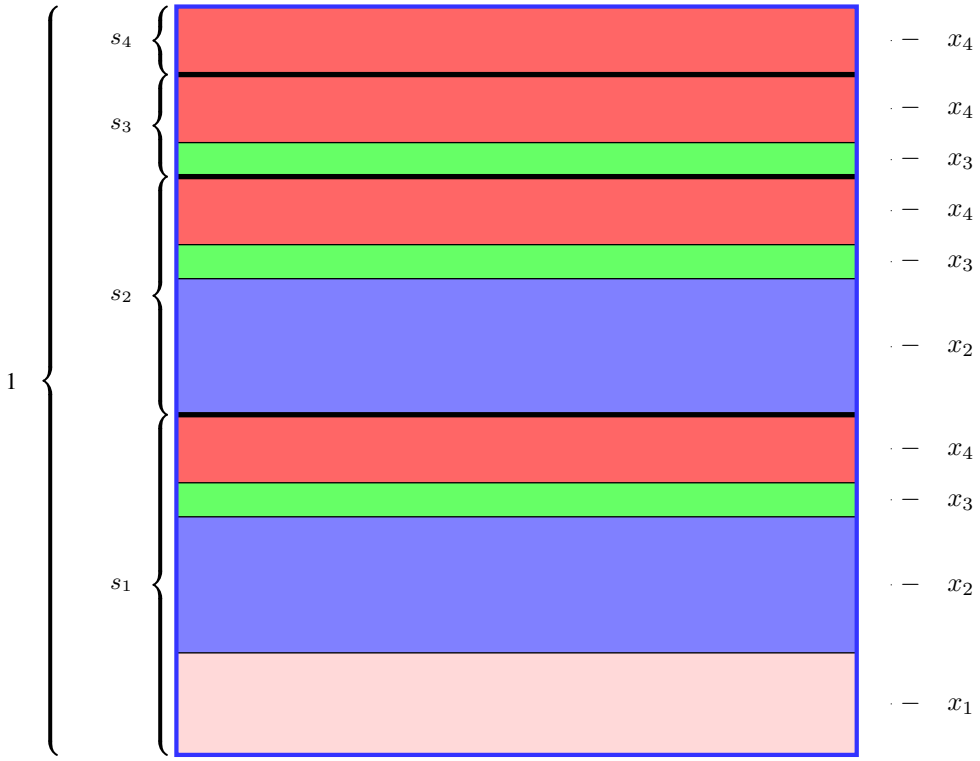

Figure 3: Rearrangement 1

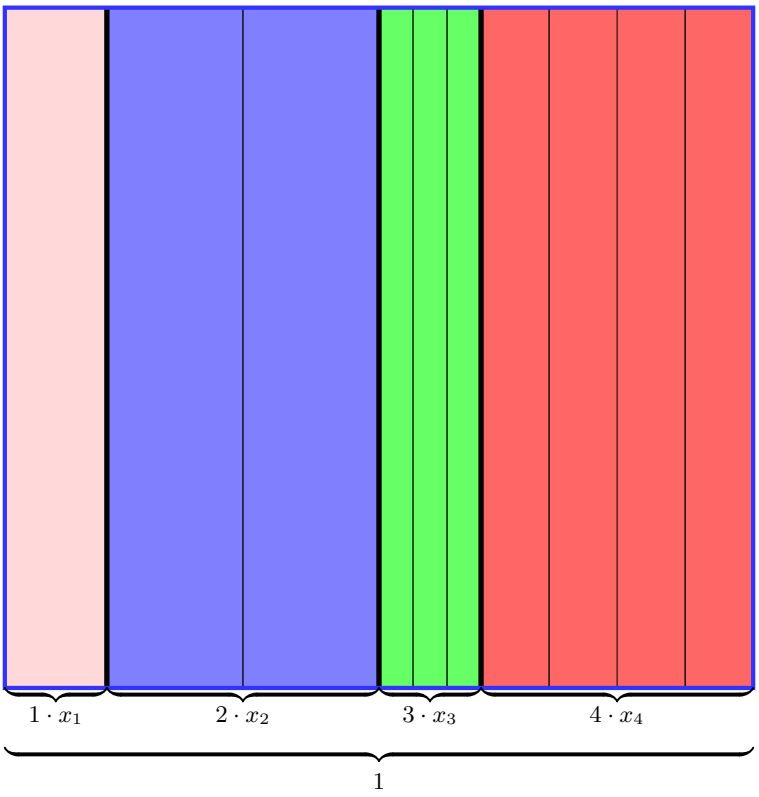

Figure 4: Rearrangement 2

5.3 Proof by probability methods

Let δ_n denote the shortest length when we drop n points randomly uniformly in the interval $[0, 1]$

Aim: Find $\mathbb{P}[\delta_n > x]$ = $\mathbb{P}$[all intervals have lengths greater than x]

Claim: $\mathbb{P}[\delta_n > x] = (1 - (n + 1)x)^n$

Proof: Note that $\mathbb{P}$[choosing a point] = [length of the interval]

For the above event to occur, all of the points need to be placed in the interval $[(n + 1)x, 1]$ i.e. we reserve a total amount of space $(n + 1)x$ at the start. This can be proved by showing a bijection.

If we drop n points in the region $[(n + 1)x, 1]$, we can extend all the $n + 1$ segments by length x which is possible because a total amount of $(n + 1)x$ was reserved. Conversely, any such valid instance where every segment has a length of at least x can be transformed to the above case where a length of x is removed from each segment to reserve a total length of $(n + 1)x$.

A generalized version of this is proved in [David and Nagaraja's Order Statistics, p. 135] which states that the probability that any particular k of the $n + 1$ segments simultaneously have lengths longer than $c_1, c_2, \ldots, c_k$, respectively (where $\sum_{i=1}^{k} c_i \leq 1$), is

$$(1 - c_1 - c_2 - \ldots - c_k)^n$$

Hence, for all the cuts to be in $[(n + 1)x, 1]$

$$\mathbb{P}[\delta_n > x] = \underbrace{(1 - (n + 1)x) \ldots (1 - (n + 1)x)}_{n \text{ times}} = (1 - (n + 1)x)^n$$

Tail Sum Theorem: The expectation for a non-negative discrete random variable X (similarly for continuous) is

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} \mathbb{P}(X > i)$$

Proof:

$$\sum_{i=0}^{\infty} \mathbb{P}[X > i] = \sum_{i=0}^{\infty} \sum_{r=i+1}^{\infty} f_X(r)$$

where $f_X(x)$ is the probability mass function of X i.e. $f_X(x) = \mathbb{P}(X = x)$. Exchanging the order of summation and changing the limits to appropriate values,

$$\implies \sum_{r=1}^{\infty} \sum_{i=0}^{r-1} f_X(r) = \sum_{r=1}^{\infty} r f_X(r) = \mathbb{E}[X]$$

Hence, by using the Tail-Sum Theorem and the claim (where X denotes the length of the smallest segment δ_n),

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X > x) dx$$

$$\implies \int_0^{\frac{1}{n+1}} \mathbb{P}[\delta_n > x] dx = \int_0^{\frac{1}{n+1}} (1 - (n+1)x)^n = \frac{1}{(n+1)^2}$$

QED

6 Conclusion

In conclusion, our study has provided three distinct proofs concerning the expected length of the smallest interval among $n+1$ intervals formed by randomly selecting n points from the $[0, 1]$ interval. Each proof contributes a unique perspective to our understanding of this fundamental problem and extends on it by finding the expected length of the kth smallest segment as well.

While more complex approaches exist, involving n-dimensional geometry or finding the probability distributions of the order statistics, our simple approaches offer succinct proofs without sacrificing rigor. We end by mentioning the problem, commonly referred to as the broken stick model, has applications in various areas such as the propagation of infectious diseases, traffic flow, finance, etc., and is a good match for a variety of real-world data sets. The model of the broken stick is also used in statistical science and in machine learning and can be used to develop new sampling methods.

References

[1] H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2003.

[2] William Verreault. On the probability of forming polygons from a broken stick, 2021.

[1] [2]