The Infinite Hat Paradox

Shreya Mogulothu

September 25, 2021

Introduction

In this paper, we will discuss Bacon's Puzzle and some responses to the paradox it seems to create.

Bacon's Puzzle

Game

First we outline Bacon's Puzzle. We are given an infinite sequence of prisoners $P_1, P_2, \dots$. They have been allowed to discuss their strategy for the following "game", but after that they have been forbidden to communicate.

We give P_k a white (w) hat or P_k gets a black (b) hat with equal probability. Then each P_k is taken to a separate room to guess their own hat color. They cannot hear the responses of any other P_l, $l \neq k$. If P_k guesses correctly, they get to live, and otherwise they get shot.

Strategy

Surprisingly, there is a strategy for the prisoners such that the majority of them survive. Let $S = \{\{w, b\}^*\}$ be the set of all possible inputs. Then S can be partitioned into [1] as follows: two sequences a and b are in the same partition iff $\{i \in \mathbb{N} \mid a_i \neq b_i\}$ is finite.[2]

With this setup, we can invoke the Axiom of Choice: there exists some choice of representative element for each partition P that the prisoners can fix during their strategy meeting. Then, the hat colors of the prisoner line constitute an infinite bit sequence $p = (p_1, p_2, \dots) \in S$. Therefore each P_k knows the infinite subsequence of p: $(p_{k+1}, p_{k+2}, \dots)$. Now P_k might assume (arbitrarily) that $\forall i \leq k$, $p_i = w$. Then their assumed sequence is some $p' = (w, \dots, w, p_{k+1}, p_{k+2}, \dots)$. So we can verify p' is in the same partition P of S as p. But P_k, like all prisoners, knows what the representative element of P is,

[1] uncountably many sets

[2] Note that we can verify that this partition is a valid set of equivalence classes, as we have done in the course.

and we will call it r. Then r is in the same partition as p', and by transitivity r is in the same partition as p, so r and p have only finitely many differing values.

Therefore, if each P_k responds "white" if r_k is 0 and "black" if r_k is 1, then there can be at most finitely many prisoners who get shot.

Probability

But we know that the hat color of P_k is independent to that of any other prisoner, so P_k gains no new information about *themselves* by looking at the prisoners in front of them. In fact, no matter their strategy, $\Pr[P_k \text{ is right}]$ should be 50%. But how is this compatible with our strategy in the last section that guarantees only a finite number of failures (i.e., infinite number of successes)?

Rayo's Approach

In the course, Professor Rayo responds to this by discussing that (1) we cannot assume that a finite number of failures means an individual success probability of more than 50%, and (2) an individual success probability of 50% does not mean the outcome for the whole group is random at best. I think these claims and their arguments in the course are reasonable, and I will give my own arguments to elaborate.

Finite numbers

As humans, we can maybe understand the size of 10, or the size of 20, or the size of 100. We can even recognize that a number deemed to be "infinite" is magnitudes larger than 10, 20, or 100. But maybe we do not have such a great understanding of what it means for something to be "finite" and also "unbounded".

For example, we guarantee that for any sequence of hats $p \in S$, the number of prisoners who fail must be finite. But what can we say about the average number of prisoners who fail over all sequences p?

We will argue that this average cannot be upper bounded by any number $u \in \mathbb{R}$.

Preliminaries

Let p belong to partition P with representative element r. Then let the "diff-sequence" $x(p) = (x(p)_1, x(p)_2, \dots)$ be the bit sequence where $x(p)_i$ is 0 iff $p_i = r_i$; that is, the sequence logs a 1 if prisoner i fails. Then let "magnitude" $|x(p)|$ be the number of 1s in the sequence; that is, the number of prisoners that fail.

Then we wish to find the average $|x(p)|$ over all p in all partitions. Call this $\overline{|x|}$.

First we observe that any sequence p can be uniquely defined by its partition P and its diff-sequence $x(p)$, but only its diff-sequence contributes change to the average $\overline{|x|}$. Let $\mathcal{P}$ be the set of partitions and D be the set of all infinite binary sequences with finitely many 1s (i.e. all possible diff-sequences). Now, we can say

$$\overline{|x|} = \frac{\sum_{P\in\mathcal{P}} \sum_{p\in P} |x(p)|}{\text{total number of hat-sequences}}$$

and we can fix some P:[3]

$$\frac{\sum_{p\in P} |x(p)|}{\text{number of hat-sequences in } P}.$$

Finally, we can easily verify that each p in a partition corresponds with exactly one element of D:

$$\frac{\sum_{x\in D} |x|}{|D|}.$$

Therefore we know $\overline{|x|}$ is just the average number of 1s across each possible diff-sequence.

Proof

Claim. $\overline{|x|}$ *cannot be bounded.*

Proof. We will prove the claim by contradiction. That is, we assume $\overline{|x|} \leq u$ for some number u and try to find something that goes wrong in this universe. Intuitively, we know that such an assumption feels weird—what is so special about u specifically? In fact, it seems to us that there are many more diff-sequences x with $|x| > u$ than there are sequences x with $|x| \leq u$.

It is hard to argue, however, the truth of this statement with infinitely many candidates for x, because infinite cardinalities are very flexible.[4] But we can accurately argue this for a set of sequences X_n containing each binary diff-sequence x with the last 1 being in at most the nth index. That is, more formally: $X_n = \{x \,|\, \forall i, x_i = 1 \implies i \leq n\}$. Now we wish to find the average $|x|$ where $x \in X_n$. Call this average $\overline{X_n}$.

To find the average value $\overline{X_n}$, we add $|x|$ $\forall x \in X_n$ and then divide by the total number of possible sequences x. But we know the total number of possible sequences in X_n is 2^n, because every index $i > n$ must be such that $x_i = 0$, and we can choose every x_i where $i \leq n$ in two ways.

Also, for any $k \in \mathbb{N}$ and $0 \leq k \leq n$, there must be $\binom{n}{k}$ sequences $x \in X_n$ such that x has k ones. Therefore, we can write

[3] Because the partition does not affect $|x(p)|$, and each partition has the same number of hat-sequences.

[4] We aren't looking for bijective equivalence here—for example, then the number of diff-sequences with one error is the same as the number of sequences with more than two errors, since $\mathbb{N}$ has the same cardinality as the set of finite sequences. However it would be absurd to claim, therefore, that the median number of shot prisoners is 2, since there are an "equal" number of sequences on either side of 2.

$$\overline{X_n} = \frac{\binom{n}{0} \cdot 0 + \binom{n}{1} \cdot 1 + \cdots + \binom{n}{n} \cdot n}{2^n}$$

So,

$$\overline{X_n} \cdot 2^n = \binom{n}{0} \cdot 0 + \binom{n}{1} \cdot 1 + \cdots + \binom{n}{n-1} \cdot (n-1) + \binom{n}{n} \cdot n$$

or

$$\overline{X_n} \cdot 2^n = \binom{n}{n} \cdot n + \binom{n}{n-1} \cdot (n-1) + \cdots + \binom{n}{1} \cdot 1 + \binom{n}{0} \cdot 0.$$

But we know $\binom{n}{k} = \binom{n}{n-k}$, so

$$\overline{X_n} \cdot 2^n = \binom{n}{0} \cdot n + \binom{n}{1} \cdot (n-1) + \cdots + \binom{n}{n-1} \cdot 1 + \binom{n}{n} \cdot 0$$

and adding $\overline{X_n} \cdot 2^n$ to itself gives

$$\begin{aligned} 2 \cdot \overline{X_n} \cdot 2^n &= \binom{n}{0} \cdot (0+n) + \binom{n}{1} \cdot (1+(n-1)) \\ &\quad + \cdots + \binom{n}{n-1} \cdot ((n-1)+1) + \binom{n}{n} \cdot (n+0) \\ &= \binom{n}{0} \cdot n + \binom{n}{1} \cdot n + \cdots + \binom{n}{n-1} \cdot n + \binom{n}{n} \cdot n \\ &= n \cdot \left(\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} \right). \end{aligned}$$

But we know $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$ [5] and therefore we write

$$2 \cdot \overline{X_n} \cdot 2^n = n \cdot 2^n$$

and therefore

$$\overline{X_n} = \frac{n}{2}.$$

Finally we can choose any $m > 2u$ and write

$$\overline{X_m} - u = \frac{m}{2} - u > \frac{2m}{2} - u = 0$$

so we know that for any large enough n, our average $\overline{X_n}$ is more than u, and therefore $\overline{X_n}$ cannot be upper bounded by u.

[5] We can see this intuitively because $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}$ is just the sum of the number of sequences x where x has k 1s for all possible values of k.

Now, what happens as n increases? For any $n' > n$, we have that $\overline{|X_{n'}|} = \frac{n'}{2} > \frac{n}{2}$, and therefore $\overline{X_n}$ is increasing on n. Therefore as we take n higher, $\overline{X_n}$ also gets increasingly higher than $\overline{X_m} > u$. Now note that

$$\overline{|x|} = \lim_{i \to \infty} \overline{X_i}$$

because there is no upper bound on where the last 1 of any diff-sequence x lies. But since $\overline{X_i}$ increases with i, we have that

$$\overline{|x|} - k = \lim_{i \to \infty} \overline{X_i} - u > \overline{X_m} - u > 0$$

which means $\overline{|x|} > u$, as desired.

Given our assumed upper bound u we have shown that it cannot be true that $\overline{|x|}$ is upper bounded by u, a contradiction. Therefore it cannot be true for any possible u that u is an upper bound for $\overline{|x|}$. And therefore $\overline{|x|}$ has no upper bound (bigger than any finite number), and so $\overline{|x|}$ is infinite.

□

But this means that the average number of people who get shot in our prisoner problem is infinite as well.

Another way to think about it is to note that the *expected* number of shot prisoners is the same as the *average* $|x|$, since we have a uniform distribution over infinite sequences in S. But then this means that the expected $|x|$ is infinite, which can change our perspective on the effectiveness of our strategy. This means the restriction that $|x|$ should be finite is *not as strong a restriction* as we might have thought, and we can better reconcile with ourselves the fact that the individual failure probabilities $\Pr[P_k \text{ fails}] = 50\%$.

Conclusion

Even if we have proven that the expected number of shot prisoners is infinite, there is a lingering queasiness—we haven't explained away the phenomenon that for any *particular* hat-sequence, the number of prisoners who is shot is *finite*, a trifling minority of them, as guaranteed by our strategy.

We know that we can claim no such strategy if the prisoners don't have the privilege of seeing each others' hats in the line (then the problem is reduced to a line of prisoners each guessing the outcome of a random coin flip), so what about seeing their sequence gives the prisoners eonugh knowledge to give them an edge?

I think it has to do with the fact that each prisoner sees a vast majority of the sequence. Finite beings cannot imagine the monstrous line prisoner P_k faces as they look ahead of them. I will say that when P_k scans over the infinite tail of prisoners in the line, they get a Feeling of what sequence they are in, because they are able to identify what partition P they belong in.

Consider a scenario when there are 10 prisoners. Consider prisoner P_2, who looks at the line of 8 prisoners in front of them. P_2 has a good Feeling of

what kind of sequence they are standing in, because this prisoner knows all but the first 2 elements of the sequence. However, in this finite version, we know prisoner P_{10} in the front of the line has absolutely no Feeling about what the sequence might look like, because they can see none of it.

This is where the infinite sequence in Bacon's Puzzle is useful—since there is no "last" prisoner, and each prisoner P_k has as many people in front of them as any other prisoner $P_{k'}$, each prisoner Feels the *same* about what the whole sequence might look like. This is why they are able to coordinate when they give their answers, and bring themselves some advantage in their predicament.

However, we must note that at the end of the day, a Feeling about the sequence that can be felt the same by P_1 and $P_{1,000,000,000}$ must be a pretty vague Feeling. This is why the stategy can't give any specific guarantees—we know it doesn't do much to help each individual prisoner, and the expected number of shot prisoners is infinite.