CS543 Final Project Report

Can Erozer

8 May 2025

1 Introduction

Testing whether an unknown discrete distribution is uniform is a fundamental problem in theoretical computer science and statistics, with broad applications in randomness testing, data validation, and property testing. Given a distribution over a finite domain of size m, the task is to distinguish whether the distribution is uniform or ϵ-far from uniform in total variation distance, using as few independent samples as possible.

Two classical approaches to this problem are the collision-based uniformity test (the one we covered in class) and singleton-based uniformity test (as introduced by Paninski) [1]. The collision-based test relies on counting the number of collisions— pairs of identical samples—observed in the sample set. This test is motivated by the fact that the collision probability is minimized under the uniform distribution. On the other hand, Paninski's singleton-based test uses the number of elements that appear exactly once in the sample set (singletons), capitalizing on the idea that the frequency of such rare events varies significantly between uniform and non-uniform distributions.

While both tests are conceptually simple and practically appealing, they differ in their theoretical guarantees. Early analyses suggested that the collision-based test required $O(\frac{\sqrt{n}}{\epsilon^4})$ samples to reliably detect non-uniformity, whereas Paninski's singleton-based test achieved an improved sample complexity of $O(\frac{\sqrt{n}}{\epsilon^2})$, aligning with known lower bounds for uniformity testing. This led to a common belief that collision-based testers were suboptimal.

However, recent results challenge this view. In particular, the paper "Collision-based Testers are Optimal for Uniformity and Closeness" demonstrates that with careful tuning and a more refined analysis, the collision-based test can in fact achieve the optimal sample complexity of $O(\frac{\sqrt{n}}{\epsilon^2})$. This finding bridges a gap between the practical simplicity of the collision method and its theoretical performance, and invites a re-evaluation of its utility compared to the singleton-based approach.

In this project, I aim to explore both the theoretical derivations and empirical behavior of these two uniformity tests. Specifically, I will:

- Analyze the mathematical foundations that lead to the sample complexity bounds for both testers.
- Investigate the assumptions, variance bounds, and concentration inequalities that underpin these results.
- Empirically compare the sample efficiency of both tests across a variety of distributional shifts
- Evaluate whether the theoretical improvements suggested by recent work on collision-based testers manifest in practice.

2 Theoretical Derivations

High-Level Idea

All tests aim to solve the uniformity testing problem:
Given i.i.d. samples from an unknown distribution D over a domain of size n, distinguish:

- Null Hypothesis (H_0): $D = U_{[n]}$ (the uniform distribution)
- Alternative Hypothesis (H_1): $||D - U_{[n]}||_1 \geq \epsilon$

Each test uses a statistic computed from the sample, and decides YES (uniform) or NO (non-uniform) depending on whether the statistic deviates too much from what we'd expect under uniformity.

2.1 Collision-Based Uniformity Test

I used the derivations showed in lecture.

Let D be the unknown distribution that we are testing its uniformity.
Let $U_{[n]}$ be the uniform distribution over $[n] = \{1, 2, ..., n\}$.

Then, $||D||_2^2$ would be the probability of collision for the unknown distribution and $||U_{[n]}||_2^2$ would be the collision probability of the uniform distribution.

It is known that the uniform distribution's collision probability is:

$$||U_{[n]}||_2^2 = \frac{1}{n}$$

In class we found that,

$$||D||_2^2 \geq \frac{4\epsilon^2 + 1}{n}$$

Now, what we want is to find $||D||_2^2$ so that we can argue whether D looks more like uniform distribution or not.

We can argue that if $||D||_2^2$ is close to $\frac{1}{n}$ then it is a uniform distribution. And, we can argue the opposite if it is far from $\frac{1}{n}$.

To write this mathematically, we can find a threshold by using the two lower bounds of probability of collision in the two cases. The threshold is the middle of the two lower bounds:

$$t = \frac{1}{2}(\frac{1}{n} + \frac{4\epsilon^2 + 1}{n}) = \frac{1 + 2\epsilon^2}{n}$$

.

So the algorithm is: IF $\|\|D\|\|_2^2 \geq \frac{1+2\epsilon^2}{n}$, OUTPUT NO (D is not uniform) ELSE, OUTPUT YES (D is uniform)

But, we need to find $||D||_2^2$ in order to perform this test. The Straightforward way is to keep drawing pairs and see the number of collisions you get in them. But, this way requires $\Omega(n)$ samples.

The better way that is proposed in class is: Draw S samples, $X_1, ..., X_S$ and use the number of collisions on all pairs:

$$||D||_2^2 => \frac{number\ of\ collisions\ in\ S\ samples}{\binom{S}{2}}$$

So, our test statistics is

$$number\ of\ collisions\ in\ S\ samples = Y = \sum_{i<j} Y_{i,j}$$

$$where\ Y_{i,j} = \begin{cases} 1, & if\ X_i = X_j \\ 0, & otherwise \end{cases}$$

So the algorithm becomes: Draw S samples from D Set $t = \binom{S}{2}\frac{1+2\epsilon^2}{n}$ Find Y IF $Y \geq t$, OUTPUT NO (D is not uniform) ELSE, OUTPUT YES (D is uniform)

Now, we need to find S- the number of samples required.

Finding S:
Let, $\hat{c} = \frac{Y}{\binom{S}{2}}$ be the estimator for $||D||_2^2$.

We want our estimator $\hat{c}$ to be close to $||D||_2^2$ with high constant probability. Given that

$$E(\hat{c}) = ||D||_2^2$$

We can use the Chebyshev's Inequality to find a bound for S:

$$Pr(|\hat{c} - E(\hat{c})| \geq \alpha) \leq \frac{Var(\hat{c})}{\alpha^2}$$

We don't want the error to be bigger than the gap $\triangle = \frac{1+4\epsilon^2}{n} - \frac{1}{n} = \frac{4\epsilon^2}{n}$

Setting $\alpha = \frac{\triangle}{2} = \frac{2\epsilon^2}{n}$

The Chebyshev's Inequality becomes:

$$Pr(|\hat{c} - E(\hat{c})| \geq \frac{2\epsilon^2}{n}) \leq \frac{4Var(\hat{c})}{\frac{16\epsilon^4}{n^2}} = 4Var(\hat{c})\frac{n^2}{16\epsilon^4}$$

Under uniformity the upper bound becomes,

$$O(\frac{1}{S}\sqrt{||D||_2^2}) = O(\frac{1}{S\sqrt{n}})$$

Under Non-Uniformity, we infer that,

$$S \geq \frac{C\sqrt{n}}{\epsilon^4}$$

So, it suffices to get $S = \left\lceil \frac{C\sqrt{n}}{\epsilon^4} \right\rceil$
We were able to have this conclusion by using the following lemma,

$$Var(Y) \leq 7(\binom{S}{2}||D||_2^2)^{3/2}$$

2.2 Optimal Collision-Based Uniformity Test [1]

Note: the variables used in this paper is not the same with the ones we used in class. So, I am converting the variables in the paper to keep things consistent.

The test analyzed in the paper is the same: collision-based uniformity test. But, they differ in the analysis of the test statistic, threshold, and variance bound, which directly affects the sample complexity.

The estimate for $||D||_2^2$ used is the same: $\hat{c} = \frac{Y}{\binom{S}{2}}$. So, the test statistic is Y

Recall, that we used the the bounds of $||D||_2^2$ under completeness (near-uniform) and soundness ($\epsilon - far$).

In this paper they used a slightly different bounds for $||D||_2^2$:

Under completeness:

$$||D||_2^2 \leq \frac{1 + \epsilon^2/2}{n}$$

Under soundness:

$$||D||_2^2 \geq \frac{1 + \epsilon^2}{n}$$

As we did before, to set the threshold we will get the middle point:

$$t = \frac{1}{2}(\frac{1 + \epsilon^2}{n}) = \frac{1 + \epsilon^2/2}{n} = \frac{1 + 3\epsilon^2/4}{n}$$

.

Finally,

$$t = \binom{S}{2}\frac{1 + 3\epsilon^2/4}{n}$$

In an attempt to find a bound for S, we are going to use the Chebyshev's Inequality:

$$Pr(|Y - E(Y)| \geq k\sigma) \leq \frac{1}{k^2}$$

where σ is $\sqrt{Var(Y)}$

The paper says that we want S to be closer to its expected value then threshold is to its expected value. This implies:

$$|Y - E(Y)| \leq |t - E(Y)|$$

So, the error happens if:

$$|Y - E(Y)| \geq |t - E(Y)|$$

Using Chebyshev's Inequality:

$$Pr(||Y - E(Y)| \geq |t - E(Y)|) \leq \frac{Var(Y)}{(t - E(Y))^2}$$

The authors said that we want this error probability to be at most $\frac{1}{4}$

$$\frac{Var(Y)}{(t - E(Y))^2} \leq \frac{1}{4}$$

$$\sqrt{4Var(Y)} \leq \sqrt{(t - E(Y))^2}$$

$$|t - E(Y)| \geq 2\sigma$$

Using lemma 2: $E(Y) = \binom{S}{2}||D||_2^2$

$$|t - E(Y)| = |E(Y) - t| = |\binom{S}{2}||D||_2^2 - (1 + 3\epsilon^2/4)/4| = \binom{S}{2}|\alpha - 3\epsilon^2/4|/n$$

$$\binom{S}{2}|\alpha - 3\epsilon^2/4|/n \geq 2\sigma$$

It suffices for the number of samples S to satisfy the slightly stronger condition that

$$\sigma \leq S^2 \frac{|\alpha - 3\epsilon^2/4|}{5n}$$

After isolating S,

$$S \geq \sqrt{\frac{5\sigma n}{|\alpha - 3\epsilon^2/4|}}$$

In the paper this inequality is denoted as lemma 4.

But, in order to find the exact lower bound for S, we need to find what σ is. As in the previous case, the bounds for σ change depending on the uniformity of D.

Let's take a look at the lemma 3, to find the upper bound of $Var(Y)$:

$$Var(Y) \leq S^2||D||_2^2 + S^3(||D||_3^3 - ||D||_2^4)$$

This new upper bound opens the path to find the optimal solution.

In the completeness case, $||D||_2^2 = \frac{1}{n}$ and $||D||_3^3 = ||D||_2^4 = \frac{1}{n^2}$
By plugging these into lemma 3, we get:

$$\sigma \leq \frac{S}{\sqrt{n}}$$

We also know that $\alpha = 0$, when D is uniform. Substituting these two facts into lemma 4, we get:

$$S \leq \frac{6\sqrt{n}}{\epsilon^2}$$

Now, let's take a look at the soundness case. In this case, since we don't know what D is, we don't know which term dominates the the upper bound of the $Var(Y)$. Recall that we have a quadratic and a cubic term of S in the upper bound:

$$Var(Y) \leq S^2||D||_2^2 + S^3(||D||_3^3 - ||D||_2^4)$$

Let's see what happens when the quadratic term $(S^2||D||_2^2)$ dominates the upper bound. For the sake of brewity of the report, I will skip the proof and just mention the result. Lemma 6 takes care of this case:

We know that $||D||_2^2 = \frac{1+\alpha}{n}$, for $\alpha \geq \epsilon^2$. If $S^2||D||_2^2$ dominates the upper bound, it implies $S^2||D||_2^2 \geq S^3(||D||_3^3 - ||D||_2^4))$. Substituting these two facts into lemma 4, we get:

$$S \leq \frac{48\sqrt{n}}{\epsilon^2}$$

Now, let's see what happens when the cubic term $(S^3(||D||_3^3 - ||D||_2^4))$ dominates the upper bound. For the sake of brewity of the report, I will skip the proof and just mention the result. Lemma 7 takes care of this case:

Again, we know that $||D||_2^2 = \frac{1+\alpha}{n}$, for $\alpha \geq \epsilon^2$. If $S^3(||D||_3^3 - ||D||_2^4))$ dominates the upper bound, it implies $S^2||D||_2^2 \leq S^3(||D||_3^3 - ||D||_2^4))$. Substituting these two facts into lemma 4, this time we get:

$$S \leq \frac{3200\sqrt{n}}{\epsilon^2}$$

Conclusion: As can be seen in both cases, completeness and soundness, S is upper bounded similarly:

Let's take a look at them at the same time. In the completeness case:

$$S \leq \frac{6\sqrt{n}}{\epsilon^2}$$

In the soundness case, we have two different bounds depending of the domination of different terms:

$$S \leq \frac{48\sqrt{n}}{\epsilon^2}$$

and

$$S \leq \frac{3200\sqrt{n}}{\epsilon^2}$$

Looking at these three upper bounds, we can safely infer that

$$S = O(\frac{\sqrt{n}}{\epsilon^2})$$

Since we all know how to set S and t and we already know how to find Y ($\sum_{i<j} Y_{i,j}$) we can construct the algorithm:

Algorithm:
Draw S samples from D
Set $t = \binom{S}{2}\frac{1+3\epsilon^2/4}{n}$
Find Y
IF $Y \geq t$, OUTPUT NO (D is not uniform)
ELSE, OUTPUT YES (D is uniform)

2.3 Paninski's Singleton-Based Uniformity Test [2]

Before talking about the details of Paninski's singleton-based uniformity test, I want to highlight two main different approaches that Paninski accomplished compared to collision-based uniformity test:

Firstly, he used a different type of test statistic. Instead of relying on second-moment statistics (like $||D||_2^2$ from collisions), he used the number of elements seen exactly once. These elements are called as singletons.

Secondly, he analyzed the test statistic via Poisson approximation. He showed that under sparse sampling (i.e., $S << n$), the counts X_i can be approximated as Poisson random variables. This simplifies the variance analysis of the singleton count S.

I will show the details of these in this section.

In his paper, Paninski started by mentioning the relation that as we observe more collisions, the distribution, D, more tends to be non-uniform. Based on this fact, he defines

$$K_1 = number\ of\ X_i's\ that\ are\ observed\ exactly\ once$$

or, equivalently,

$$K_1 = number\ of\ singletons$$

In contrast, as we observe more collisions, the value of K_1 tends to drop.

He defines a test statistic based on the following difference under the sparse regime ($S << n$):

$$T = E_U(K_1) - K_1$$

where $E_U(K_1) = expected\,number\,of\,singletons\,under\,uniform\,distribution$

The key idea is that under null hypothesis K_1 is expected to be high. But, under the alternative hypothesis, K_1 is expected to be low. s

Now, let's talk about how we set the threshold, t.

Looking at lemma 1,

$$E_U(K_1) - E_D(K_1) \geq \frac{S^2\epsilon^2}{n}(1 + O(\frac{S}{n}))$$

where $E_U(K_1) = S(\frac{n-1}{n})^{S-1}$ and $E_D(K_1) = \sum_{i=1} \binom{S}{1} D_i(1 - D_i)^{S-1}$

In order to find the upper bound of the variance let's take a look the lemma 2:

$$Var_D(K_1) \leq E_U(K_1) - E_D(K_1) + O(\frac{S^2}{n})$$

due to Efron-Stein Inequality

To find the threshold, we can find the middle point of the gap (difference between $E_U(K_1)$ *and* $E_D(K_1)$), defined in lemma 1.

If $n >> S$, then $E_U(K_1) - E_D(K_1) \sim \frac{S^2\epsilon^2}{n}$

So,

$$t = \frac{1}{2}\frac{S^2\epsilon^2}{n} = \frac{S^2\epsilon^2}{2n}$$

As we found the threshold, now we can construct a test for the null hypothesis:

If,

$$T \equiv E_U(K_1) - K_1 = S(\frac{n-1}{n})^{S-1} - K_1 > t$$

then reject H_0

Let's define the error: $|T - E_U(T)| \geq t$. Then, by Chebyshev's Inequality,

$$Pr_U(|T - E_U(T)| \geq t) \leq \frac{Var(T)}{t^2}$$

where $E_U(T) = 0$ and $Var_U(T) = O(\frac{S^2}{n})$ by lemma 2.
Equivalently,

$$Pr_U(T \geq t) \leq \frac{Var(T)}{t^2}$$

If $n >> S$,

$$\frac{Var(T)}{t^2} \sim \frac{S^2/n}{(S^2\epsilon^2/n)^2} = \frac{n}{S^2\epsilon^4}$$

This is small if $S^2\epsilon^4 >> n$. In other words, $\frac{n}{S^2\epsilon^4} \to 0$ if $S^2\epsilon^4 >> n$.

Thus,

$$S^2\epsilon^4 >> n$$

$$\sqrt{S^2} >> \sqrt{\frac{n}{\epsilon^4}}$$

$$S >> \frac{\sqrt{n}}{\epsilon^2}$$

To reliably test uniformity vs. $\epsilon - far$ alternatives, it suffices to set

$$S = \frac{C\sqrt{n}}{\epsilon^2}$$

where C is a constant

Since we all know how to set S and t and we already know how to find K_1, we can construct the algorithm:

$$t = \frac{S^2\epsilon^2}{2n}$$

Algorithm:
Draw S samples from D
Set $t = \frac{S^2\epsilon^2}{2n}$
Compute $E_U(K_1) = S(\frac{n-1}{n})^{S-1}$
Compute test statistic: $T = E_U(K_1) - K_1$
Find Y

IF $Y \geq t$, OUTPUT NO (D is not uniform)
ELSE, OUTPUT YES (D is uniform)

3 Empirical Bounds