Non-Private Uniformity and Identity Testing

Ari

September 20, 2022

Abstract

Re-deriving well-known results in Uniformity Testing

1 Problem Statement

These notes are derived from the excellent survey by Clement Canonne [Can22]. We have n i.i.d samples from an unknown distribution $p \in \texttt{Simplex}(k)$. We want to guess whether $p = U_k$ or $d_{TV}(p, U_k) \geq \epsilon$, where U_k is the uniform distribution over $\texttt{Simplex}(k)$. Furthermore, we would like to guess correctly at least two-thirds of the time i.e.

$$\mathbb{P}\Big[\text{Guess } p = U_k | p = U_k\Big] \geq \frac{2}{3} \tag{1}$$

$$\mathbb{P}\Big[\text{Guess } d_{TV}(p, U_k) \geq \epsilon | d_{TV}(p, U_k) \geq \epsilon\Big] \geq \frac{2}{3} \tag{2}$$

To guess we need to design some kind of test, that takes as input the parameters k, ϵ and the i.i.d samples and spits out a guess. One very natural test is to count how many of the samples are the same. If the distribution was truly uniformly random, we do not expect to see a lot of overlap between the samples. Before describing the test in the next section we state some general facts about norms and distances.

$$d_{TV}(p, q) = \frac{1}{2}||p - q||_1 \leq \frac{\sqrt{k}}{2}||p - q||_2 \tag{3}$$

The proof for (3) comes from Cauchy-Schwartz and can be found in any linear algebra textbook under the title "equivalence of norms". Thus if we have $d_{TV}(p, q) \geq \epsilon$, then we have $||p - q||_2 \geq \frac{4\epsilon^2}{k}$. Furthermore, if $q = U_k$, then we have $||p - U_k||_2^2 = ||p||_2^2 - \frac{1}{k}$. Thus combining all these facts we have

- If $d_{TV}(p, U_k) = 0$, then $||p||_2^2 = \frac{1}{k}$
- Otherwise, $d_{TV}(p, U_k) \geq \epsilon$, then $||p||_2^2 \geq \frac{1+4\epsilon^2}{k}$

Note that $||p||_k^k$ is also called the collision probability of k samples drawn from p having the same value.

2 Tester Based on Collisions

Let $X_1, \ldots, X_n$ represent n i.i.d samples from p. Let Z denote the number of pairwise collisions i.e

$$Z = \frac{1}{\binom{n}{2}} \sum_{1 \leq s < t \leq n} \mathbb{1}(X_s = X_t) \tag{4}$$

Z represents the fraction of pairwise collisions as a sum of i.i.d indicators and $\mathbb{P}\left[\mathbb{1}(X_s = X_t)\right] = \mathbb{E}[\mathbb{1}(X_s = X_t)] = ||p||_2^2$. Therefore $\mathbb{E}[Z] = ||p||_2^2$. The obvious test is to pick some threshold τ such that if $Z \leq \tau$, declare that $p = U_k$ and $d_{TV}(p, U_k) \geq \epsilon$ otherwise. Say we pick a threshold between $\frac{1}{k}$

and $\frac{1+4\epsilon^2}{k}$, so say$\tau = \frac{1+2\epsilon^2}{k}$. Then we want to know how many samples we need for equations (1) and (2) to hold.

We need to bound Type I and Type II errors to upper-bound the probability of making a mistake. To upper bound Type I error (or completeness error) we have that the data comes from the uniform distribution but our test threshold still crossed τ, so we need to upper bound the following.

$$\begin{aligned}\mathbb{P}\Big[Z \geq \tau | p = U_k\Big] &= \mathbb{P}\Big[Z \geq \frac{1+2\epsilon^2}{k}\Big] && (5)\\ &= \mathbb{P}\Big[Z \geq (1+2\epsilon^2)\mathbb{E}[Z]\Big] && (6)\\ &\leq \mathbb{P}\Big[Z \geq (1+\epsilon^2)\mathbb{E}[Z]\Big] && (7)\end{aligned}$$

We drop the conditional from (5) onwards to make the notation clearer. (6) comes from the assumption that $p = U_k$. To upper bound Type II error, we need upper bound the following quantity

$$\begin{aligned}\mathbb{P}\Big[Z \leq \tau | d_{TV}(p, U_k) \geq \epsilon\Big] &= \mathbb{P}\Big[Z \leq \frac{1+2\epsilon^2}{k}\Big] && (8)\\ &\leq \mathbb{P}\Big[Z \leq \frac{(1-\epsilon^2)(1+4\epsilon^2)}{k}\Big] && (9)\\ &\leq \mathbb{P}\Big[Z \leq (1-\epsilon^2)\mathbb{E}[Z]\Big] && (10)\end{aligned}$$

(9) comes from the fact that when $\epsilon \leq \frac{1}{2}$, we have $(1-\epsilon^2)(1+4\epsilon^2) \geq (1+2\epsilon^2)$. (10) comes from the fact that $\mathbb{E}[Z] \geq \frac{(1+4\epsilon^2)}{k}$ from our assumption. Thus if we combined equation (10) and (7) and applied Chebychev we would get

$$\mathbb{P}\Big[|Z - \mathbb{E}[Z]| \leq \epsilon^2\mathbb{E}[Z]\Big] \leq \frac{\text{Var}(Z)}{\epsilon^4\mathbb{E}[Z]^2} \qquad (11)$$

If the variance of the estimator is much larger than the wiggle room we've given us, then we would have no hope i.e. we would like $\sqrt{\text{Var}(Z)} \ll (\frac{1+4\epsilon^2}{k} - \frac{1}{k})$.

2.1 Simple but non-optimal upper bound for variance

$$\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \qquad (12)$$

We already have $EZ = ||p||_2^2$, so $EZ^2 = ||p||_2^4$. We need a manageable expression for $\mathbb{E}[Z^2]$.

$$\begin{aligned}\mathbb{E}[Z^2] &= \frac{1}{\binom{n}{2}^2}\mathbb{E}\Big[\sum_{1\leq s<t\leq n}\sum_{1\leq s'<t'\leq n}\mathbb{1}(X_s = X_t)\mathbb{1}(X'_s = X'_t)\Big] && (13)\\ &= \frac{1}{\binom{n}{2}^2}\sum_{1\leq s<t\leq n}\sum_{1\leq s'<t'\leq n}\mathbb{E}[\mathbb{1}(X_s = X_t)\mathbb{1}(X'_s = X'_t)] && (14)\\ & && (15)\end{aligned}$$

(14) : By Linearity of expectation. Now looking at all those indices in that awkward-looking expression $\mathbb{E}[\mathbb{1}(X_s = X_t)\mathbb{1}(X'_s = X'_t)]$, we can group them into 3 categories based on the cardinality of the set $\{s, t, s', t'\}$ (1) When all the indices are distinct (2) When both pairs are the same (3) When there are 3 distinct indices. Additionally, we will only need to consider the events when the product of the two terms is 1.

1. s, t, s', t' are all distinct i.e. $|\{s, t, s', t'\}| = 4$. Therefore the two collisions are independent of each other. Therefore we get $\mathbb{E}[\mathbb{1}(X_s = X_t)\mathbb{1}(X'_s = X'_t)] = ||p||_2^4$. There are $\binom{n}{2}\binom{n-2}{2} = 6\binom{n}{4}$ ways to select pairs such they are distinct.

2. $s = s'$ and $t = t'$ i.e. both pairs are the same $|\{s, t, s', t'\}| = 2$, as the square of an indicator function is just the value of the indicator we have $\mathbb{E}[\mathbb{1}(X_s = X_t)\mathbb{1}(X'_s = X'_t)] = ||p||_2^2$. There are $\binom{n}{2}$ ways of selecting 2 unique indices.

3. Finally the mankiest event, where we have $|\{s, t, s', t'\}| = 3$. Three distinct indices will take up the same value and one of the pairs will have a repeat index. This is the same as saying 3 items will collide and the probability of having 3 collisions is $||p||_3^3$. There are $\binom{n}{3}$ ways of picking three distinct indices. Then for each of those three indices, there are 2 choices for which the remaining index is duplicated. Thus the total number of options is $(2 \times 3)\binom{n}{3}$. Also by the monotonicity of norms, we have $||p||_3^3 \le ||p||_2^3$ which we will use below.

Thus we have

$$
\begin{align}
\mathrm{Var}(Z) &= \left[\frac{1}{\binom{n}{2}^2} \sum_{1\le s<t\le n} \sum_{1\le s'<t'\le n} \mathbb{E}[\mathbb{1}(X_s = X_t)\mathbb{1}(X_s' = X_t')]\right] - \mathbb{E}[Z]^2 \tag{16} \\
&= \frac{1}{\binom{n}{2}^2}\left[6\binom{n}{4}||p||_2^4 + \binom{n}{2}||p||_2^2 + 6\binom{n}{3}||p||_3^3\right] - \frac{\binom{n}{2}^2}{\binom{n}{2}^2}||p||_2^4 \tag{17} \\
&= \frac{1}{\binom{n}{2}^2}\left[||p||_2^4\left(6\binom{n}{4} - \binom{n}{2}^2\right) + \binom{n}{2}||p||_2^2 + 6\binom{n}{3}||p||_3^3\right] \tag{18} \\
&\le \frac{1}{\binom{n}{2}^2}\left[||p||_2^4\left(6\binom{n}{4} - \binom{n}{2}^2\right) + \binom{n}{2}||p||_2^2 + 6\binom{n}{3}||p||_2^3\right] \tag{19} \\
&\le \frac{1}{\binom{n}{2}^2}\left[\binom{n}{2}||p||_2^2 + 6\binom{n}{3}||p||_2^3\right] \tag{20} \\
&= \frac{1}{\binom{n}{2}^2}\left[\binom{n}{2}\mathbb{E}[Z] + 6\binom{n}{3}\mathbb{E}[Z]^{3/2}\right] \tag{21} \\
&\le O(\frac{1}{n^2})\mathbb{E}[Z] + O(\frac{1}{n})\mathbb{E}[Z]^{3/2} \tag{22}
\end{align}
$$

(19): Montonicity, $||p||_3^3 \le ||p||_2^3$

(20) $\left(6\binom{n}{4} - \binom{n}{2}^2\right) \le 0$, for $n \ge 2$

(22): $\frac{4}{n^2} \ge \frac{1}{\binom{n}{2}}$ and

Plugging this into (11) we have

$$
\begin{align}
\mathbb{P}\left[|Z - \mathbb{E}[Z]| \le \epsilon^2\mathbb{E}[Z]\right] &\le \frac{O(\frac{1}{n^2})\mathbb{E}[Z] + O(\frac{1}{n})\mathbb{E}[Z]^{3/2}}{\epsilon^4\mathbb{E}[Z]^2} \tag{23} \\
&= \frac{1}{\epsilon^4 n^2\mathbb{E}[Z]} + \frac{1}{\epsilon^4 n\mathbb{E}[Z]^{1/2}} \tag{24} \\
&\le \frac{k}{\epsilon^4 n^2} + \frac{\sqrt{k}}{\epsilon^4 n} \tag{25}
\end{align}
$$

(25): $\mathbb{E}[Z] \ge \frac{1}{k}$ for both cases.
Setting (25) equal to 1/3 and solving for n we get

$$
n \ge O(\frac{\sqrt{k}}{\epsilon^4}) \tag{26}
$$

2.2 Optimal Upper Bound

Unfortunately, this bound can be improved to

$$
n \ge O(\frac{\sqrt{k}}{\epsilon^2}) \tag{27}
$$

Thus we must account for some slack in the analysis for variance. Thus going back to our analysis where we kicked out the negative term

$$\begin{aligned}
\binom{n}{2}^2 \mathrm{Var}(Z) &= \left[||p||_2^4 \left(6\binom{n}{4} - \binom{n}{2}^2 \right) + \binom{n}{2} ||p||_2^2 + 6\binom{n}{3} ||p||_3^3 \right] && (28) \\
&= \left[||p||_2^4 \left(-6\binom{n}{3} - \binom{n}{2} \right) + \binom{n}{2} ||p||_2^2 + 6\binom{n}{3} ||p||_3^3 \right] && (29) \\
&= \binom{n}{2} ||p||_2^2 (1 - ||p||_2^2) + 6\binom{n}{3} (||p||_3^3 - ||p||_4^2) && (30)
\end{aligned}$$

(29) : $\binom{n}{2}^2 = 6\binom{n}{4} + 6\binom{n}{3} + \binom{n}{2}$

The first summand is the variance of a binomial distribution with parameters $\mathrm{Bin}(\binom{n}{2}, ||p||_2^2)$ and the second summand is always non-negative (where the middle inequality comes from cauchy-schwartz)

$$||p||_2^4 = \left(\sum_i p(i)^{3/2} p(i)^{1/2} \right)^2 \leq \sum_i p(i)^3 \sum_i p(i) = \sum_i p(i)^3 = ||p||_3^3 \qquad (31)$$

Thus going back to our variance analysis and dividing both sides by $\binom{n}{2}^2$ we get

$$\begin{aligned}
\mathrm{Var}(Z) &= \frac{1}{\binom{n}{2}^2} \left[\binom{n}{2} ||p||_2^2 (1 - ||p||_2^2) + 6\binom{n}{3} (||p||_3^3 - ||p||_4^2) \right] && (32) \\
&\leq \frac{1}{\binom{n}{2}^2} \left[\binom{n}{2} ||p||_2^2 + 6\binom{n}{3} (||p||_3^3 - ||p||_4^2) \right] && (33) \\
&\leq O(\frac{1}{n^2}) ||p||_2^2 + O(\frac{1}{n}) \Big(||p||_3^3 - ||p||_4^2 \Big) && (34)
\end{aligned}$$

Now if $p = U_k$, then $\mathrm{Var}(Z) = O(\frac{1}{n^2})\mathbb{E}[Z]$ since $||p||_3^3 = ||p||_4^2$, thus plugging it back to (11), we get

$$\mathbb{P}\left[Z \geq \tau\right] \leq \mathbb{P}\Big[Z \geq (1+\epsilon^2)\mathbb{E}[Z] \Big] \leq \frac{k}{n^2 \epsilon^4} \qquad (35)$$

Setting the RHS of (35) equal to 1/3 we get

$$n \geq O(\frac{\sqrt{k}}{\epsilon^2}) \qquad (36)$$

Okay but this only solves the case for Type I error, we still need to deal with the case that $d_{TV}(p, U_k) \geq \epsilon$. Define $\alpha^2 := k||p - U_k||_2^2 \geq 4\epsilon^2$, then $\mathbb{E}[Z_1] = \frac{1+\alpha^2}{k}$.

$$\begin{aligned}
\mathbb{P}\left[Z \leq \tau\right] &= \mathbb{P}\left[Z \leq \frac{1+2\epsilon^2}{k}\right] && (37) \\
&= \mathbb{P}\left[Z \leq \frac{1+2\epsilon^2}{1+\alpha^2} \mathbb{E}[Z] \right] && (38) \\
&= \mathbb{P}\left[Z \leq \Big(1 - \frac{\alpha^2 - 2\epsilon^2}{1+\alpha^2} \Big) \mathbb{E}[Z] \right] && (39) \\
&\leq \mathbb{P}\left[Z \leq \Big(1 - \frac{\alpha^2}{2(1+\alpha^2)} \Big) \mathbb{E}[Z] \right] && (40) \\
&\leq \frac{\mathrm{Var}(Z)}{\alpha^4 \mathbb{E}[Z]^2} 4(1+\alpha^2)^2 && (41) \\
&\leq \frac{16(1+\alpha^2)^2}{\alpha^4 \mathbb{E}[Z]^2} \frac{\mathbb{E}[Z]}{n^2} + \frac{16(1+\alpha^2)^2}{\alpha^4 n \mathbb{E}[Z]^2} \Big(||p||_3^3 - ||p||_4^2 \Big) && (42)
\end{aligned}$$

(38): How we defined α, (39): By the assumption that p is far from U_k, (40): $\alpha^2 \geq 4\epsilon^2$
Dealing with the left summand first. We have $\mathbb{E}[Z] = \frac{1+\alpha^2}{k}$. Thus we get

$$\frac{16(1+\alpha^2)^2}{\alpha^4 \mathbb{E}[Z]^2} \frac{\mathbb{E}[Z]}{n^2} = \frac{16(1+\alpha^2)k}{\alpha^4 n^2} \leq \frac{5k}{\epsilon^4 n^2} \tag{43}$$

(43): For $x > 0$, $\frac{1+x}{x^2}$ is decreasing function. Furthermore $\alpha^2 \geq 4\epsilon^2$ and $\epsilon \leq 1$, thus $1 + 4\epsilon \leq 5$

Ok so the one last thing to show is the last summand $\frac{4}{n\mathbb{E}[Z]^2}\left(||p||_3^3 - ||p||_4^2\right)$ and then we would be done. As $||p||_2^2 \geq \frac{1}{k}$ we have

$$\begin{aligned}
\left(||p||_3^3 - ||p||_4^2\right) &\leq ||p - U_k + U_k||_3^3 - \frac{1}{k^2} \tag{44}\\
&= ||p - U_k||_3^3 + \frac{3}{k}||p - u||_2^2 \tag{45}\\
&\leq ||p - U_k||_2^3 + \frac{3}{k}||p - u||_2^2 \tag{46}\\
&= \frac{\alpha^3}{k^{3/2}} + \frac{3\alpha^2}{k^2} \tag{47}
\end{aligned}$$

Finally, multiplying it with $\frac{16(1+\alpha^2)^2}{n\alpha^4 ||p||_2^4}$

$$\begin{aligned}
\frac{16(1+\alpha^2)^2}{n\alpha^4 ||p||_2^4}\left(||p||_3^3 - ||p||_4^2\right) &\leq \frac{16(1+\alpha^2)^2}{n\alpha^4 ||p||_2^4} \frac{\alpha^3}{k^{3/2}} + \frac{3\alpha^2}{k^2} \tag{48}\\
&\leq \frac{8\sqrt{k}}{\epsilon n} + \frac{12}{n\epsilon^2} \tag{49}
\end{aligned}$$

Combining the two summands, we get

$$n \geq O(\frac{\sqrt{k}}{\epsilon^2}) \tag{50}$$

References

[Can22] Clément L. Canonne. Topics and Techniques in Distribution Testing: A Biased but Representative Sample. March 2022.