COMS E6998-9: Algorithms for Massive Data (Fall'23) Oct 23, 2023

Lecture 14: Distribution testing: Uniformity

Instructor: *Alex Andoni* Scribes: *Yiming Fang*

1 Uniformity Testing

Last time, we introduced the problem of *uniformity testing*: given m samples from an unknown distribution D over $[n]$, we want to distinguish between the following cases:

- D is uniform
- D is ε-far from uniform, i.e., $||D - U_n||_{TV} \geq \frac{\varepsilon}{2}$, or equivalently, $||D - U_n||_1 \geq \varepsilon$.

Note that there are many possible similarity metrics for distributions, and ε-far is only one of them. We want to achieve this goal with the smallest sample complexity, m.

1.1 Attempt 1

Algorithm 1: (Testing via Learning)

- Learn $\hat{D}$ such that $||D - \hat{D}||_1 \leq \varepsilon$ by computing the *Empirical Distribution* of D on samples $\{x_1, \ldots, x_m\}$, which is defined as follows:

$$\hat{D}_i = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}[x_j = i], \quad \forall i \in [n].$$

- Compute $||D - U_n||_1$ directly.

Goal: Determine how large m needs to be such that $||D - \hat{D}||_1 \leq \varepsilon$.

Claim 1. *$m = O(n/\varepsilon^2)$ samples are enough for learning $\hat{D}$ such that $||D - \hat{D}||_1 \leq \varepsilon$.*

Before proving the claim, we first assume that it is true and demonstrate how we can use it to solve the uniformity problem. We can learn $\hat{D}$ such that $||D - \hat{D}||_1 \leq \varepsilon/3$, then compute $||\hat{D} - U_n||_1$ and compare it to $\varepsilon/2$. To see why this procedure outputs the correct answer, consider both cases:

- If D is uniform, then $||\hat{D} - U_n||_1 \leq \varepsilon/3$.
- If D is ε-far from uniform, then by triangle inequality

$$||\hat{D} - U_n||_1 \geq ||D - U_n||_1 - ||D - \hat{D}||_1 \geq \varepsilon - \varepsilon/3 > \varepsilon/2$$

Proof. (of Claim 1)

$$\begin{aligned}
\mathbb{E}\left[\|D-\hat{D}\|_1\right] &= \sum_{i\in[n]} \mathbb{E}\left[|D_i-\hat{D}_i|\right] \\
&\le \sum_{i\in[n]} \left(\mathbb{E}\left[|D_i-\mathbb{E}[D_i]|^2\right]\right)^{\frac{1}{2}} \\
&= \sum_{i\in[n]} \left(\mathrm{Var}[\hat{D}_i]\right)^{\frac{1}{2}} \\
&= \sum_{i\in[n]} \left(\mathrm{Var}\left[\frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[x_j=i]\right]\right)^{\frac{1}{2}} \\
&= \sum_{i\in[n]} \left(\frac{1}{m^2}\sum_{j=1}^{m} \mathrm{Var}\left[\mathbf{1}[x_j=i]\right]\right)^{\frac{1}{2}} \\
&\le \sum_{i\in[n]} \left(\frac{1}{m}D_i\right)^{\frac{1}{2}} \\
&= \sum_{i\in[n]} \left(\frac{1}{\sqrt{m}}D_i^{\frac{1}{2}}\right) \\
&\le \frac{1}{\sqrt{m}}\left(\sum_{i\in[n]} D_i\right)^{\frac{1}{2}}\sqrt{n} \\
&= \sqrt{\frac{n}{m}}
\end{aligned}$$

where in the second to last step, we used Cauchy-Schwartz Inequality, and the fact that that D is a probability distribution, so $\sum_{i\in[n]} D_i = 1$.

Therefore, if we let $m = 100\frac{n}{\varepsilon^2}$, then $\mathbb{E}\left[\|D-\hat{D}\|_1\right] \le \frac{\varepsilon}{10}$. By Markov Inequality, we get

$$\Pr_D\left[\|D-\hat{D}\|_1 > \varepsilon\right] < \frac{1}{10}$$

□

Question: It is natural to ask: can we achieve the same goal with $m << n$?

It is impossible to use much less than n samples to compute the empirical distribution and achieve the same goal. However, it is possible that we can use the samples in a different and more efficient way.

1.2 Attempt 2

Intuition: If our distribution D is not uniform (or is uniform on a much smaller support than $[n]$), then we should see collisions much earlier than if we were drawing example from U_n, because $x_i \in D$ comes from a smaller range and collide with higher probability.
Algorithm 2: (Collision counting)

Let $C = \sum_{1 \le i < j \le m} \mathbf{1}[x_1 = x_j]$ be the collision count. We test the uniformity by

- If $C \le \frac{\alpha}{n}$, then D is uniform;
- If $C > \frac{\alpha}{n}$, D is ε-far from uniform.

for some constant $\alpha = \alpha(\varepsilon)$ that we will fix later.

Claim 2.

$$\|D - U_n\|_2^2 = \|D\|_2^2 - \frac{1}{n}$$

.

Note: $\|U_n\|^2 = (\frac{1}{n})^2 n = \frac{1}{n}$. Also, if D is such that $\|D - U_n\|_1 \ge \varepsilon$, then $\|D - U_n\|_2 \ge \frac{1}{\sqrt{n}} \|D - U_n\|_1 \ge \frac{\varepsilon}{\sqrt{n}}$.

Claim 3.

$$\mathbb{E}\left[\frac{C}{\binom{m}{2}}\right] = \|D\|_2^2$$

.

Analysis of Algorithm 2: Assuming the claims above are true. We analyze the l_2 distance:

- If $D = U_n$, we have $\|D - U_n\|_1 = 0$ and $\|D - U_n\|_2^2 = 0$. So $\|D\|^2 = \frac{1}{n}$, and we have

$$\mathbb{E}\left[\frac{C}{\binom{m}{2}}\right] = \frac{1}{n}$$

- If $\|D - U_n\|_1 \ge \varepsilon$, we have $\|D - U\|_2 \ge \frac{\|D - U_n\|_1}{\sqrt{n}} \ge \frac{\varepsilon}{\sqrt{n}}$, then $\|D\|_2^2 \ge \frac{1}{n} + \frac{\varepsilon^2}{n}$. Then,

$$\mathbb{E}\left[\frac{C}{\binom{m}{2}}\right] = \frac{1 + \varepsilon^2}{n}$$

We will thus set α so that the algorithm threshold is in the middle of these two expectations: namely, $\alpha = 1 + \epsilon^2/2$. In addition to proving the above claim, we also want to prove that the expectation concentrates around the expectation, without passing erroneously this threshold.

Proof. (of Claim 2)

$$\begin{aligned}\|D-U\|_2^2 &= \sum_{i=1}^{m}\left(D_i-\frac{1}{n}\right)^2\\ &= \sum_{i=1}^{m}\left(D_1^2+\frac{1}{n^2}-\frac{2D_i}{n}\right)\\ &= \|D\|_2^2+\frac{1}{n^2}-\frac{2}{n}\sum_{i=1}^{m}D_i\\ &= \|D\|_2^2-\frac{1}{n}.\end{aligned}$$

□

Proof. (of Claim 3)

$$\begin{aligned}\mathbb{E}(C) &= \mathbb{E}\left[\sum_{1\le i<j\le m}\mathbf{1}[x_i=x_j]\right]\\ &= \sum_{1\le i<j\le m} Pr[x_i=x_j]\\ &= \sum_{1\le i<j\le m}\sum_{k\in[n]} D_k^2\\ &= \binom{m}{2}\|D\|_2^2.\end{aligned}$$

Rearranging the terms gives $\mathbb{E}\left(\frac{C}{\binom{m}{2}}\right)=\|D\|_2^2$. □

Claim 4. *For $m=\Omega(\sqrt{n}/\epsilon^4)$, we have:*

$$\Pr_D\left[\left|\frac{C}{\binom{m}{2}}-\|D\|^2\right|\le\frac{\varepsilon^2}{2n}\right]\ge 90\%$$

Proof. (of Claim 4)

$$
\begin{aligned}
\mathrm{Var}[C] &= \mathbb{E}\left[\left(\sum_{i<j} \mathbf{1}[x_i = x_j]\right)^2\right] - \left(\|D\|^2 \binom{m}{2}\right)^2 \\
&= \left(\sum_{i<j}\sum_{i'<j'} \mathbb{E}\left[\mathbf{1}[x_i = x_j \wedge x_{i'} = x_{j'}]\right]\right) - \left(\|D\|^2 \binom{m}{2}\right)^2 \\
&\le \cancel{(\mathbb{E}[C])^2 - \left(\|D\|^2 \binom{m}{2}\right)^2} + \sum_{i<j} \Pr_D[x_i = x_j] + 2\sum_{i<j}\sum_{j' \neq i,j} \Pr_D[\mathbf{1}[x_i = x_j = x_{j'}]] \\
&= \binom{m}{2}\|D\|^2 + 2\sum_{k=1}^{m}(mD_k)^3 \\
&= \binom{m}{2}\|D\|^2 + 2m^3\|D\|_3^3 \\
&\le \binom{m}{2}\|D\|^2 + 2m^3\|D\|_2^3 \\
&\le \binom{m}{2} n\|D\|^4 + 2m^3\sqrt{n}\|D\|_2^4,
\end{aligned}
$$

since $\|D\|_2 \ge 1/n$.

For $m = 1200\sqrt{n}/\epsilon^4$, we obtain that $\mathrm{Var}[C] \le 3\frac{\sqrt{n}}{m}(\binom{m}{2}\|D\|_2^2)^2 \le 0.0025\epsilon^4 \cdot \mathbb{E}[C]$. By Chebyshev bound, we have that:

$$\Pr[C \in E[C] \cdot (1 \pm \epsilon^2/2)] \ge 0.9.$$

□