COMS E6998-9: Algorithms for Massive Data (Fall'25) Nov 17, 2025

Lecture 21: Uniformity Testing

Instructor: *Alex Andoni* Scribes: *Yuval Shemla*

1 Recap and Problem Setup

We consider a discrete distribution D over a finite universe $[n] = \{1, \ldots, n\}$. Our goal is to test whether D is uniform or far from uniform using as few samples as possible. Throughout, we have sample access to D: we obtain independent samples $x_1, \ldots, x_m \sim D$.

Uniformity testing problem. Distinguish between the two cases

- $D = U$, where U is the uniform distribution on $[n]$ (so $U_i = 1/n$ for all i);
- $\|D - U\|_1 \geq \varepsilon$, i.e. D is ε-far from uniform in ℓ_1 distance.

Recall that for distributions P, Q on $[n]$, the ℓ_1 distance is

$$\|P - Q\|_1 \; = \; \sum_{i=1}^{n} |P_i - Q_i|.$$

2 Algorithm 1: Learning the Distribution

The first approach is to *learn* the entire distribution D up to small ℓ_1 error, and then use this estimate for testing.

2.1 Empirical distribution

Given samples $x_1, \ldots, x_m \sim D$, we define the empirical distribution $\hat{D}$ by

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

Intuitively, $\hat{D}_i$ is the fraction of samples that fell on element i.

Theorem 1. *There exists a constant $C > 0$ such that for $m \geq C \cdot \frac{n}{\varepsilon^2}$ we have*

$$\Pr\left[\, \|D - \hat{D}\|_1 > \varepsilon/2 \,\right] \leq 0.1.$$

In particular, $m = \Theta\left(\frac{n}{\varepsilon^2}\right)$ samples are sufficient to learn D in ℓ_1 distance.

Proof. We first bound the expectation of $\|D - \hat{D}\|_1$.

$$\mathbb{E}\big[\|D - \hat{D}\|_1\big] = \mathbb{E}\left[\sum_{i\in[n]} |D_i - \hat{D}_i|\right] \;=\; \sum_{i\in[n]} \mathbb{E}\big[\,|D_i - \hat{D}_i|\,\big]$$

$$\leq \sum_{i\in[n]} \Big(\mathbb{E}\big[(D_i - \hat{D}_i)^2\big]\Big)^{1/2}.^{1} \qquad (\star)$$

Note that D_i is a constant and

$$\hat{D}_i \;=\; \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[x_j = i],$$

so

$$\mathbb{E}\big[(D_i - \hat{D}_i)^2\big] = \mathrm{Var}(\hat{D}_i).$$

Thus

$$\mathbb{E}\big[\|D - \hat{D}\|_1\big] \;\leq\; \sum_{i\in[n]} \sqrt{\mathrm{Var}(\hat{D}_i)}.$$

Next we compute $\mathrm{Var}(\hat{D}_i)$ using independence of the samples:

$$\mathrm{Var}(\hat{D}_i) = \mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[x_j = i]\right)$$

$$= \frac{1}{m^2}\sum_{j=1}^{m} \mathrm{Var}\,\big(\mathbf{1}[x_j = i]\big),^{2} \qquad (\dagger)$$

where

$$\mathrm{Var}\,\big(\mathbf{1}[x_j = i]\big) = D_i(1 - D_i) \;\leq\; D_i.$$

‡[3] Therefore

$$\mathrm{Var}(\hat{D}_i) \;\leq\; \frac{1}{m^2}\cdot m \cdot D_i \;=\; \frac{D_i}{m}.$$

Plugging this into our bound on the expectation,

$$\mathbb{E}\big[\|D - \hat{D}\|_1\big] \leq \sum_{i\in[n]} \sqrt{\frac{D_i}{m}} \;=\; \frac{1}{\sqrt{m}}\sum_{i\in[n]} \sqrt{D_i}.$$

We now upper bound the sum of square-roots using Cauchy–Schwarz again:

$$\sum_{i\in[n]} \sqrt{D_i} \;\leq\; \sqrt{\left(\sum_{i\in[n]} 1^2\right)\left(\sum_{i\in[n]} D_i\right)} \;=\; \sqrt{n\cdot 1} = \sqrt{n}.^{4}$$

[3] For a Bernoulli random variable with mean p, the variance is $p(1-p) \leq p$. Here $p = D_i$.

[4] Apply Cauchy–Schwarz with vectors $(1,\ldots,1)$ and $(\sqrt{D_1},\ldots,\sqrt{D_n})$. The second sum is 1 because D is a probability distribution.

Hence

$$\mathbb{E}\big[\|D-\hat{D}\|_1\big] \;\le\; \sqrt{\frac{n}{m}}.$$

Now choose m so that this expectation is much smaller than the error threshold $\varepsilon/2$. For example, if we take

$$m \;\ge\; 100\cdot\frac{n}{\varepsilon^2},$$

then

$$\mathbb{E}\big[\|D-\hat{D}\|_1\big] \;\le\; \sqrt{\frac{n}{m}} \;\le\; \sqrt{\frac{\varepsilon^2}{100}} \;=\; \frac{\varepsilon}{10}.$$

Finally we convert this expectation bound to a high-probability bound via Markov's inequality:

$$\Pr\big[\|D-\hat{D}\|_1 > \varepsilon/2\big] \;\le\; \frac{\mathbb{E}\big[\|D-\hat{D}\|_1\big]}{\varepsilon/2} \;\le\; \frac{\varepsilon/10}{\varepsilon/2} \;=\; 0.2.$$

By adjusting the constant 100 appropriately (e.g. to 200) we can make this probability at most 0.1, proving the theorem. □

2.2 Using Algorithm 1 for uniformity testing

Corollary 2. *With $m=\Theta\big(\frac{n}{\varepsilon^2}\big)$ samples, we can test whether D is uniform or ε-far from uniform in ℓ_1 distance with constant success probability.*

Proof. Draw m samples from D, form $\hat{D}$, and compute $\|\hat{D}-U\|_1$. If $\|\hat{D}-U\|_1 \le \varepsilon/2$, output "uniform"; otherwise output "ε-far".
If $D=U$, then $\|D-U\|_1=0$, and by Theorem 1 we have $\|\hat{D}-D\|_1 \le \varepsilon/2$ with probability at least 0.9, which implies $\|\hat{D}-U\|_1 \le \varepsilon/2$ and we accept.
If $\|D-U\|_1 \ge \varepsilon$, then by the triangle inequality

$$\|\hat{D}-U\|_1 \;\ge\; \|D-U\|_1 - \|D-\hat{D}\|_1 \;\ge\; \varepsilon-\varepsilon/2=\varepsilon/2,$$

so with probability at least 0.9 we will reject. □

Note that the sample complexity here is *linear* in n. This is because Algorithm 1 essentially learns all n probabilities of D; this is more than we need just to test uniformity. Next we see how to do better by *not* learning D completely.

3 Algorithm 2: Collision-Based Uniformity Tester

Algorithm 1 "sees" the entire support of D in principle: if m is on the order of n, we expect to observe most elements of $[n]$. With only $\sqrt{n}$ samples, this is no longer true: we cannot hope to reconstruct D accurately from so few samples. However, we can still test uniformity by exploiting information carried by *collisions*.
Intuitively, if we see many collisions (repeated sample values), this indicates that some elements have relatively large probability mass. Among all distributions on $[n]$, the uniform distribution has the *fewest* collisions in expectation.

3.1 ℓ_2 distance and uniformity

It is convenient for this algorithm to work in ℓ_2 distance. For a distribution D, define

$$\|D\|_2^2 = \sum_{i=1}^{n} D_i^2.$$

We write $d := \|D\|_2^2$ for brevity.

Claim 3. *If* $\|D - U\|_1 \geq \varepsilon$ *then*

$$\|D - U\|_2^2 \;\geq\; \frac{\varepsilon^2}{n}.$$

Proof. By Cauchy–Schwarz,

$$\|D - U\|_1 = \sum_{i=1}^{n} |D_i - U_i| \leq \sqrt{n} \cdot \|D - U\|_2.$$

Rearranging gives $\|D - U\|_2 \geq \|D - U\|_1/\sqrt{n} \geq \varepsilon/\sqrt{n}$, and squaring gives the claim. $\square$

Claim 4. $\|D - U\|_2^2 = \|D\|_2^2 - \frac{1}{n}$.

Proof.

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

$\square$

Combining Claims 3 and 4, if D is ε-far from uniform in ℓ_1, then

$$\|D\|_2^2 = \frac{1}{n} + \|D - U\|_2^2 \;\geq\; \frac{1}{n} + \frac{\varepsilon^2}{n}.$$

On the other hand, if D is exactly uniform, then $\|D\|_2^2 = 1/n$. Thus uniformity testing reduces to distinguishing between

$$\|D\|_2^2 = \frac{1}{n} \quad \text{and} \quad \|D\|_2^2 \geq \frac{1}{n} + \frac{\varepsilon^2}{n}.$$

3.2 Collision count and its expectation

Fix m samples $x_1, \ldots, x_m \sim D$. Define the collision count

$$C := \#\{(i,j) : 1 \leq i < j \leq m,\; x_i = x_j\},$$

i.e., the number of colliding pairs among the samples. Let

$$M := \binom{m}{2}$$

be the total number of unordered pairs.

Claim 5 (Collision expectation)**.**

$$\mathbb{E}\left[\frac{C}{M}\right] = \|D\|_2^2 = d.$$

Proof. For each pair $i < j$, define the indicator

$$\chi_{i,j} := \mathbf{1}[x_i = x_j].$$

Then $C = \sum_{i<j} \chi_{i,j}$ and hence

$$\mathbb{E}[C] = \sum_{i<j} \mathbb{E}[\chi_{i,j}] = \sum_{i<j} \Pr[x_i = x_j].$$

For a fixed pair (i, j),

$$\Pr[x_i = x_j] = \sum_{z\in[n]} \Pr[x_i = z, x_j = z] = \sum_{z\in[n]} D_z^2 = \|D\|_2^2.$$

Thus

$$\mathbb{E}[C] = M \cdot \|D\|_2^2,$$

and dividing by M gives the claim. □

So the normalized collision count C/M is an unbiased estimator of d.

3.3 Variance of the collision estimator

To argue that C/M concentrates around its mean we need to bound its variance. Unlike Algorithm 1, the relevant indicators are now *pairs* of samples and are not independent, so we must compute the variance more carefully.

Claim 6.

$$\mathrm{Var}\left(\frac{C}{M}\right) \;\le\; O\left(\frac{d^{3/2}}{m}\right).$$

Proof. We first compute $\mathbb{E}[(C/M)^2]$:

$$\mathbb{E}\left[\left(\frac{C}{M}\right)^2\right] = \frac{1}{M^2}\mathbb{E}\left[\left(\sum_{i<j} \chi_{i,j}\right)^2\right].$$

Expanding the square,

$$\mathbb{E}\left[\left(\frac{C}{M}\right)^2\right] = \frac{1}{M^2}\sum_{i<j}\sum_{k<\ell} \mathbb{E}[\chi_{i,j}\chi_{k,\ell}].$$

We split the sum into two parts depending on whether the index sets $\{i,j\}$ and $\{k,\ell\}$ intersect.

Disjoint pairs: If $\{i,j\} \cap \{k,\ell\} = \emptyset$, then the events "$x_i$ and x_j collide" and "x_k and x_ℓ collide" are independent, so

$$\mathbb{E}[\chi_{i,j}\chi_{k,\ell}] = \mathbb{E}[\chi_{i,j}]\,\mathbb{E}[\chi_{k,\ell}] = d^2.$$

There are M^2 such ordered pairs of pairs, so the total contribution of these terms is at most $M^2 d^2/M^2 = d^2$.

Overlapping pairs: If $\{i,j\} \cap \{k,\ell\} \neq \emptyset$, then (i,j) and (k,ℓ) share at least one index. There are only $O(m^3)$ such quadruples. In this case, $\chi_{i,j}\chi_{k,\ell}$ is the indicator that three samples collide (e.g. $x_i = x_j = x_k$). For any three distinct indices a, b, c,

$$\Pr[x_a = x_b = x_c] = \sum_{z \in [n]} D_z^3 = \|D\|_3^3.$$

Thus the total contribution from overlapping pairs is

$$O\left(\frac{1}{M^2}\right) \cdot O(m^3) \cdot \|D\|_3^3 = O\left(\frac{1}{m}\right) \|D\|_3^3.$$

Putting the two parts together,

$$\mathbb{E}\left[\left(\frac{C}{M}\right)^2\right] \leq d^2 + O\left(\frac{1}{m}\right) \|D\|_3^3.$$

For a probability distribution, higher ℓ_p norms are at most the lower ones: $\|D\|_3 \leq \|D\|_2$,[5] so

$$\|D\|_3^3 \leq \|D\|_2^3 = d^{3/2}.$$

Hence

$$\mathbb{E}\left[\left(\frac{C}{M}\right)^2\right] \leq d^2 + O\left(\frac{d^{3/2}}{m}\right).$$

Finally,

$$\mathrm{Var}\left(\frac{C}{M}\right) = \mathbb{E}\left[\left(\frac{C}{M}\right)^2\right] - \left(\mathbb{E}\left[\frac{C}{M}\right]\right)^2 \leq \left(d^2 + O\left(\frac{d^{3/2}}{m}\right)\right) - d^2 = O\left(\frac{d^{3/2}}{m}\right),$$

where we used Claim 5 to identify $\mathbb{E}[C/M] = d$. □

3.4 Analysis of Algorithm 2

The collision-based uniformity tester is:

- Draw m samples $x_1, \ldots, x_m$ from D.

[5] This follows from the monotonicity of ℓ_p norms on $\mathbb{R}^n$: for $1 \leq p \leq q$, $\|x\|_q \leq \|x\|_p$. Here $p = 2$, $q = 3$, and x is the vector of probabilities $(D_1, \ldots, D_n)$.

- Compute C, the number of colliding pairs, and let $M = \binom{m}{2}$.
- If $\frac{C}{M} < \frac{1}{n} + \frac{\varepsilon^2}{2n}$, output "uniform"; otherwise output "ε-far from uniform".

We now sketch why this works for $m = \tilde{O}(\sqrt{n}/\varepsilon^4)$.

Case 1: D is uniform. Then $d = \|D\|_2^2 = 1/n$. By Claim 6 and Chebyshev's inequality,

$$\Pr\left[\left|\frac{C}{M} - d\right| \geq \frac{\varepsilon^2}{2n}\right] \leq \frac{\mathrm{Var}(C/M)}{(\varepsilon^2/2n)^2} \leq \frac{O(d^{3/2}/m)}{\varepsilon^4/n^2} = O\left(\frac{1}{m} \cdot \frac{n^{3/2}}{\varepsilon^4}\right).$$

Since $d = 1/n$, choosing $m = \Theta(\sqrt{n}/\varepsilon^4)$ makes this probability at most 0.1, so with probability at least 0.9 we will have $C/M \leq 1/n + \varepsilon^2/(2n)$ and accept.

Case 2: D is ε-far from uniform. Then by Claims 3 and 4,

$$d = \|D\|_2^2 \geq \frac{1}{n} + \frac{\varepsilon^2}{n}.$$

Again using Chebyshev and Claim 6, with the same choice of m we have with probability at least 0.9 that

$$\frac{C}{M} \geq d - \frac{\varepsilon^2}{2n} \geq \frac{1}{n} + \frac{\varepsilon^2}{2n},$$

so the algorithm correctly outputs "ε-far from uniform".

Thus Algorithm 2 distinguishes the two cases with constant success probability using $m = \tilde{O}(\sqrt{n}/\varepsilon^4)$ samples.