COMS E6998-3: Algorithms for Massive Data (Fall'25) Nov 19, 2025

Lecture 22: Testing uniformity and beyond, Start on Learning-Augmented Algos

Instructor: *Alex Andoni* Scribes: *John Mulholland*

1 Announcements

Approaching the end of the semester. Project presentations are coming soon; some groups will present in-person and others online. Continue attending office hours for help with your final project.

2 Uniformity Testing

We consider a distribution D over $[n]$ and the problem of distinguishing:

$$D = U_n \quad \text{vs.} \quad \|D - U_n\|_1 > \varepsilon.$$

We draw m i.i.d. samples $X_1, \ldots, X_m \sim D$ and define the collision statistic:

$$c = \frac{1}{\binom{m}{2}} \sum_{i<j} \mathbf{1}[X_i = X_j].$$

2.1 Collision Statistic

We test uniformity via:

$$c < \frac{1+\alpha}{n} \Rightarrow \text{"uniform"}, \qquad c \geq \frac{1+\alpha}{n} \Rightarrow \text{"far"}.$$

Let $d = \|D\|_2^2$.

Claim 1. *The expectation of the collision count equals d:*

$$\mathbb{E}[c] = d.$$

Claim 2.

$$\|D - U_n\|_2^2 \geq \frac{\|D - U_n\|_1^2}{n}.$$

Fact 3.

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

and this is the minimizer over all distributions.

2.2 Variance Calculation

We use:

$$\mathrm{Var}[C] \le \frac{1}{\binom{m}{2}^2} \left[\binom{m}{2} d + 2m^3 \|D\|_3^3 \right]$$

and after bounding terms:

$$\mathrm{Var}[C] \le \Theta(\varepsilon^4 d^2).$$

For

$$m = \Theta\left(\frac{\sqrt{n}}{\varepsilon^4}\right),$$

Chebyshev's inequality gives:

$$\Pr\left[|c - d| \le \frac{\varepsilon^2}{3n} \right] \le 0.1,$$

hence with probability at least 0.9 we correctly distinguish the cases.

2.3 Completeness and Soundness

If $D = U_n$:

$$c \le d + \frac{\varepsilon^2 d}{3} = \frac{1}{n}\left(1 + \frac{\varepsilon^2}{3}\right).$$

If $\|D - U_n\|_1 > \varepsilon$:

$$c \ge \frac{1}{n} + \frac{\varepsilon^2}{2n},$$

for $n > \frac{2}{\varepsilon^2}$.

Setting the threshold $d = \varepsilon^2/2$ suffices.

3 Extensions of Distribution Testing

3.1 Identity Testing

We are given a known distribution Q over $[n]$. Given samples from D, we must distinguish:

$$D = Q \quad \text{vs.} \quad \|D - Q\|_1 \ge \varepsilon.$$

Theorem 4. *Identity testing can be solved with*

$$\Theta\left(\frac{\sqrt{n}}{\varepsilon^4}\right)$$

samples.

3.2 Reduction to Uniformity Testing

We reduce the problem to uniformity by constructing domain

$$S = \bigcup_{i:Q_i>0} \{(i,j) : j = 1, \dots, N \cdot Q_i\},$$

so

$$|S| = n.$$

Define:

$$Q' = U_S, \qquad D'(i,j) = \frac{D_i}{nQ_i}.$$

If $D = Q$ then $D' = Q'$. If $\|D - Q\|_1 \geq \varepsilon$:

$$\|D' - Q'\|_1 \geq \varepsilon.$$

We can simulate samples from D' using samples from D.

Thus identity testing reduces to uniformity testing on a domain of size n.

3.3 Instance Optimality

For a known Q, testing complexity becomes:

$$\Theta\big(\text{poly}(1/\varepsilon) \cdot C_Q\big),$$

and this is optimal.

3.4 Related Problems

- **Closeness Testing**: D and Q unknown. Lower bound $\Theta(n^{2/3})$.
- **Independence Testing**: Is $D(i,j) = p_i q_j$?
- **Tolerance Testing**: Distinguish $D = U_n$ vs. $\|D - U_n\|_1 \geq \varepsilon$.
- **Robust Statistics**: Ignore ε-fraction of corrupted points.

3.5 Statistical vs. Algorithmic View

Statistics: closed-form tests (e.g., chi-square). Algorithms: design efficient sample-optimal testers.

Pearson's χ^2:

$$\sum_{i=1}^{n} \frac{(mD_i - mQ_i)^2 - mD_i}{Q_i}.$$

Valiant–Valiant (2014) introduced an improved χ^2-style statistic:

$$\sum_{i} \frac{(mD_i - mQ_i)^2 - mD_i}{Q_i^{2/3}}.$$

3.6 Learning-Augmented Algorithms (LAA)

A learning model provides "hints" to the algorithm.

Requirements:

- If the hint is good, the algorithm should improve.
- If the hint is bad or hallucinated, performance should not degrade compared to worst-case.