Sublinear Algorithms

LECTURE 25

Last time

- Local Computation Algorithms (LCAs)
- Distributed LOCAL model
- Maximal Independent Set (MIS)

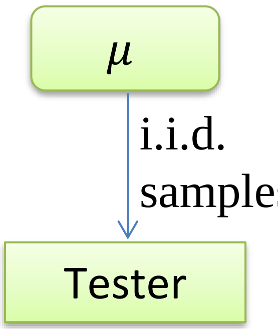
Today

- Testing properties of distributions
- Uniformity testing
- Presentation tips

Sofya Raskhodnikova;Boston University (based on lecture notes by Ronitt Rubinfeld and Sepehr Assadi

Testing Properties of Distributions

Motivation: Can we decide if a distribution (over a finite domain) satisfies a given property by examining just a few samples?



- Fix a domain $[n]$
- The tester gets access to i.i.d. samples from an unknown distribution μ over $[n]$
- It has to **accept** if μ has property $\mathcal{P}$ and **reject** if $dist(\mu, \mathcal{P}) \geq \varepsilon$ $\Big\}$ with probability $\geq \frac{2}{3}$
- Distance measure: *total variation distance*

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{i \in [n]} |\mu(i) - \nu(i)|$$

$$= \frac{1}{2} |\mu - \nu|_1$$

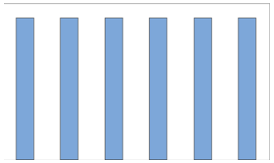
$$= \max_{\Omega \subseteq [n]} \{\mu(\Omega) - \nu(\Omega)\}$$

Here distributions are viewed as n-element vectors of probabilities

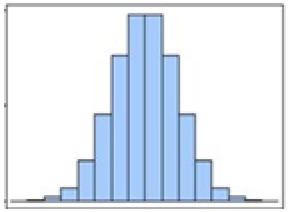
Measure of distinguishability of two distributions

Examples of Properties of Distributions

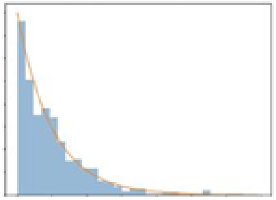
- **Uniformity:** Is μ uniform over $[n]$?



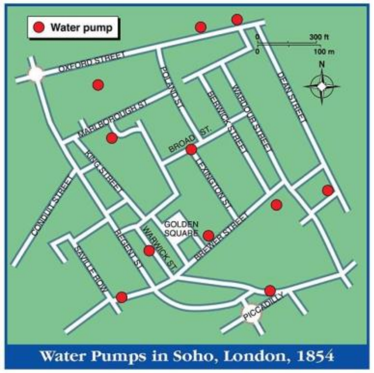
- **Identity:** Is μ equal to a specific distribution (e.g. Binom$(n, 1/2)$)?



- **Closeness:** Are two unknown distributions equal? (Samples from both distributions are given)
- **Monotonicity:** Is μ monotone?
- **k-modality:** Does μ have at most k modes?



Example settings: lottery data, shopping choices, experimental outcomes, cases of cholera as a function of the distance to water sources



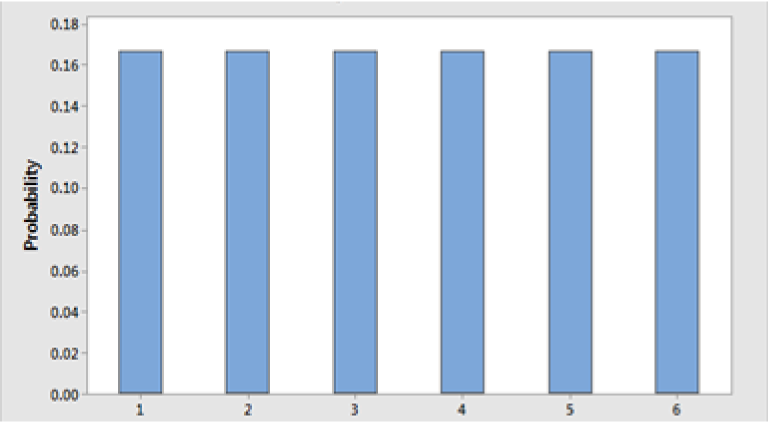
Testing Uniformity

- Let U_n be the uniform distribution over [n]
- Given access to i.i.d. samples from distribution μ over $[n]$, distinguish $\mu = U_n$ from $d_{TV}(\mu, U_n) \geq \varepsilon$

Known: $\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples are necessary and sufficient [Paninski]

Today: $O\left(\frac{\sqrt{n}}{\varepsilon^4}\right)$ samples are sufficient [Goldreich Ron]

$\Omega(\sqrt{n})$ samples are necessary



Norms and L_p-distances

Facts about norms

For all vectors $x \in \mathbb{R}^n$

1. $||x||_1 \leq \sqrt{n}\, ||x||_2$
2. $||x||_p \leq ||x||_q$ for all integers $p \geq q$

Main Idea in the Tester

Idea: Count the number of collisions, i.e., the pairs of equal samples.

- What is the probability of two samples colliding under U_n?
- In general?

$$\Pr_{x,y\sim\mu}[x = y] =$$

Collisions Theorem [for far distributions]

If distribution μ satisfies $d_{TV}(\mu, U_n) \geq \varepsilon$ then $||\mu||_2^2 \geq (1 + 4\varepsilon^2)\frac{1}{n}$

Proof of Collision Theorem

Collisions Theorem [for far distributions]

If distribution μ satisfies $d_{TV}(\mu, U_n) \geq \varepsilon$ then $||\mu||_2^2 \geq (1 + 4\varepsilon^2)\frac{1}{n}$

Proof: We first consider $||\mu - U_n||_2^2$ and then use the relationships between the norms.

$$||\mu - U_n||_2^2 = \sum_{i \in [n]} \left(\mu(i) - \frac{1}{n}\right)^2 = \sum_{i \in [n]} \mu(i)^2 - 2\sum_{i \in [n]} \frac{\mu(i)}{n} + \sum_{i \in [n]} \frac{1}{n^2}$$

$$= ||\mu||_2^2 - \frac{2}{n} + \frac{1}{n} = ||\mu||_2^2 - \frac{1}{n}$$

$||x||_1 \leq \sqrt{n}\,||x||_2$ for all vectors $x \in \mathbb{R}^n$

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

$$\geq \frac{1}{n} \cdot ||\mu - U_n||_1^2 + \frac{1}{n} = \frac{1}{n} \cdot \left(2d_{TV}(\mu, U_n)\right)^2 + \frac{1}{n}$$

$$\geq \frac{4\varepsilon^2}{n} + \frac{1}{n}$$

Algorithm for Testing Uniformity

Uniformity Tester

1. Sample $x_1, \ldots, x_s$, where $s = const \cdot \frac{\sqrt{n}}{\varepsilon^4}$
2. For all indices $i, j \in [s]$, where $i < j$, let Y_{ij} be the indicator for $x_i = x_j$
3. Set $Y \leftarrow \frac{\sum_{i,j\in[s]:\, i<j} Y_{ij}}{\binom{s}{2}}$
4. If $Y \leq \left(1 + \frac{\varepsilon^2}{2}\right) \cdot \frac{1}{n}$, **accept**; otherwise, **reject**.

Analysis: Suppose Y estimates $||\mu||_2^2$ within a factor of $1 \pm \frac{\varepsilon^2}{2}$

If $\boldsymbol{\mu = U_n}$, then $||\mu||_2^2 = \frac{1}{n}$ and $Y \leq \left(1 + \frac{\varepsilon^2}{2}\right) \cdot \frac{1}{n}$

The tester correctly accepts.

If $\boldsymbol{d_{TV}(\mu, U_n) \geq \varepsilon}$, then $||\mu||_2^2 \geq (1 + 4\varepsilon^2)\frac{1}{n}$ and $Y \geq \left(1 - \frac{\varepsilon^2}{2}\right) \cdot (1 + 4\varepsilon^2)\frac{1}{n}$

$$\left(1 - \frac{\varepsilon^2}{2}\right) \cdot (1 + 4\varepsilon^2) = 1 + 4\varepsilon^2 - \frac{\varepsilon^2}{2} - 2\varepsilon^4 > 1 + \frac{\varepsilon^2}{2}$$

The tester correctly rejects.

It remains to show that Y estimates $||\mu||_2^2$ within a factor of $1 \pm \frac{\varepsilon^2}{2}$ w.p.$\geq \frac{2}{3}$

Analyzing the Collision Estimator

Lemma (Accuracy of Collision Estimator)

$$\Pr\left[\left|Y - ||\mu||_2^2\right| > \frac{\varepsilon^2}{2}||\mu||_2^2\right] \le \frac{1}{3}$$

$Y = \frac{\sum_{i,j\in[s]:\, i<j} Y_{ij}}{\binom{s}{2}}$; $\quad s = const \cdot \frac{\sqrt{n}}{\varepsilon^4}$

Y_{ij} is the indicator for $x_i = x_j$

Proof: Calculate $\mathbb{E}[Y]$

Upper bound $\mathrm{Var}[Y]$

- Let $X = \sum_{i,j\in[s]:\, i<j} Y_{ij}$

$$\mathbb{E}[X] = \binom{s}{2}\mathbb{E}[Y]$$

$$\mathrm{Var}[X] = \binom{s}{2}^2 \mathrm{Var}[Y]$$

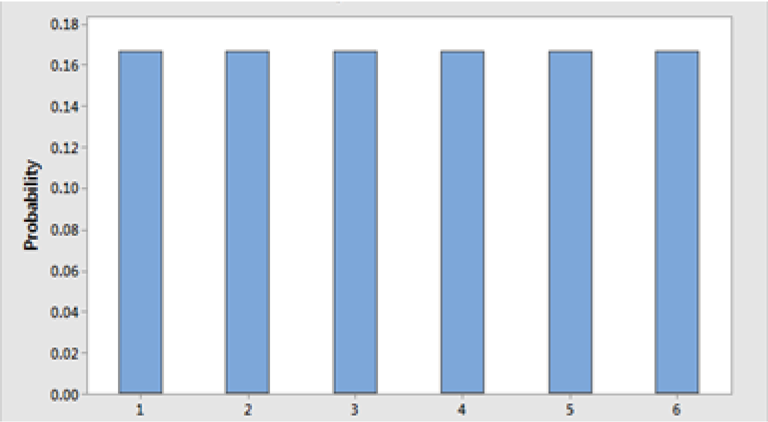
Testing Uniformity

- Let U_n be the uniform distribution over [n]
- Given access to i.i.d. samples from distribution μ over $[n]$, distinguish $\mu = U_n$ from $d_{TV}(\mu, U_n) \geq \varepsilon$

Known: $\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples are necessary and sufficient [Paninski]

Today: $O\left(\frac{\sqrt{n}}{\varepsilon^4}\right)$ samples are sufficient [Goldreich Ron]

$\Omega(\sqrt{n})$ samples are necessary



Presentation Tips: Motivation

Motivate like you're pitching to a venture capitalist… who only funds algorithms that don't look at the input

- ✓ Explain what the goals of the project are.
- ✓ Provide motivation.

Presentation Tips: Structure Your Talk

Structure your talk like a sublinear algorithm: skip the boring parts

Break it into 3-5 sections:

1. Problem & motivation
2. Model & definitions
3. Previous work
4. Theorems/results
5. Open questions / regrets / existential uncertainty

✓ Include a roadmap slide.

Presentation Tips: Designing Each Slide

Keep your slides sublinear

- ✓ Don’t crowd slides
- ✓ Keep the color scheme consistent
  (use colors to help you, but not too many colors)
- ✓ One picture is worth a thousand formulas

Presentation Tips: Tailor to Your Audience

Make jokes about being ε-far from confused

- ✓ Explain so that your fellow students can understand
- ✓ Don’t explain things they already know
- ✓ Stress ideas they would find interesting

Presentation Tips: Speaking

Your audience deserves to understand every ε fraction of your talk

- ✓ **Don’t rush**: If you're faster than your slides, you're in an unsimulated complexity class.
- ✓ **Enunciate**: Don’t let “ε” sound like “δ”—this isn’t an adversarial channel.
- ✓ **Project your voice**
- ✓ **Look up**: Talk to the room, not your laptop (unless your laptop is enrolled in the class).
- ✓ **Practice out loud**

Presentation Tips: How to Fight the Fear

Today from ``BU Today'': Roughly 1/3 of American adults say they fear public speaking more than insects, needles – even murder.

- ✓ Practice
- ✓ Learn the first minute by heart
- ✓ Tell yourself that your audience is here to learn,
  not to judge you
- ✓ Make eye contact with at least one nonintimidating human
- ✓ If you know something helps you, make it likely to happen
  (E.g., questions from the audience help me, so I tell everybody about it, hoping that people will ask questions. 😉
  If you like explaining to your teddy bear, bring it along.)
- ✓ Help out your presentation partner: laugh at their jokes, give them credit, etc. You are on the same team!