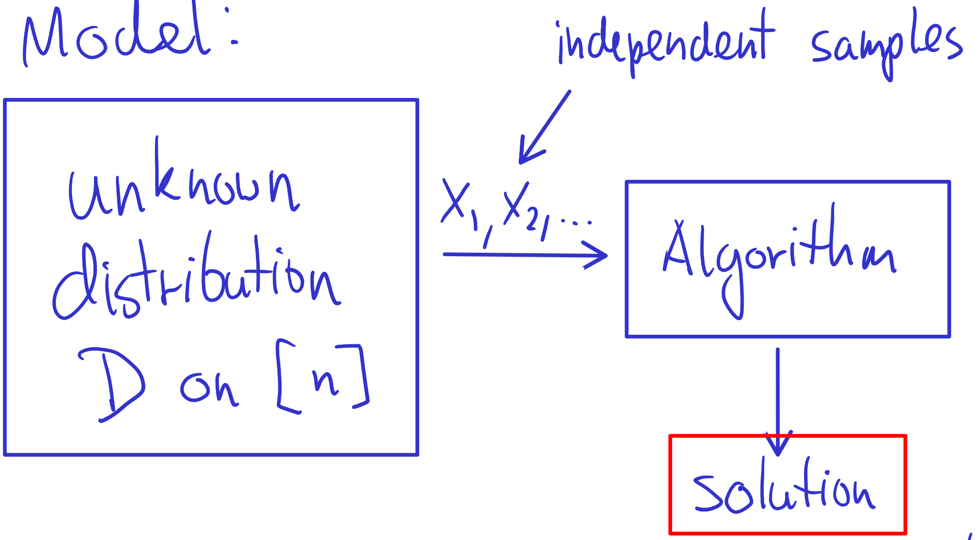
DS-563/CS-543 2024-04-02

Today:

- Logistics:
  - HW5 due tomorrow
  - Office Hours
  - Talk of interest (Friday)
  - Feedback on final project proposals
- Testing if a distribution is uniform

Review

Model:



Last time: estimating a distribution $\boxed{D'}$

s.t. $d_{TV}(D, D') \le \varepsilon$

w.p. $99/100$

Required number of samples: $\Theta(n/\varepsilon^2)$

Today: Uniformity testing

① If $D = U_{[n]}$, output $\boxed{\text{YES}}$ w.p. $99/100$

uniform distribution on $[n]$

② If $d_{TV}(D, U_{[n]}) \geq \varepsilon$, output $\boxed{\text{NO}}$ w.p. $99/100$

Intuition: Why $\Omega(\sqrt{n})$ samples are required?

It is difficult to distinguish

$$D = U_{[n]} \quad \text{from} \quad D = U_S$$

$$d_{TV}(\quad,\quad) = \frac{1}{2}$$

S is a random subset of $[n]$ of size $\sim n/2$

In both cases, you are unlikely to see the same number twice with $o(\sqrt{n})$ samples

[see "Birthday paradox" in the note on probabilistic inequalities]

Analysis of $\|D\|_2^2$

We treat our distributions as vectors in $[0,1]^n$, in which the i-th coordinate is the probability of $i \in [n]$

Notation for input distribution D: $D = (p_1, p_2, \ldots, p_n)$

Claim: $\|D - U_{[n]}\|_2^2 = \|D\|_2^2 - \frac{1}{n}$

Proof:
$$\|D - U_{[n]}\|_2^2 = \sum_{i=1}^{n} \left(p_i - \frac{1}{n}\right)^2$$
$$= \sum_{i=1}^{n} p_i^2 + 2\sum_{i=1}^{n} p_i \cdot \left(-\frac{1}{n}\right) + \sum_{i=1}^{n} \frac{1}{n^2}$$
$$= \|D\|_2^2 - \frac{2}{n} \cdot 1 + n \cdot \frac{1}{n^2} = \|D\|_2^2 - \frac{2}{n} + \frac{1}{n}$$
$$= \|D\|_2^2 - \frac{1}{n} \quad \blacksquare$$

How to interpret $\|D\|_2^2 = \sum p_i^2$?

It is the probability that two independent samples from D are identical.

= "collision"

What is $\|D\|_2^2$ in our two cases?

① $D = U_{[n]}$:

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

② $d_{TV}(D, U_{[n]}) \geq \varepsilon$

$$\Updownarrow$$

$$\|D - U_{[n]}\|_1 \geq 2\varepsilon$$

why?

$$\Downarrow$$

$$\|D - U_{[n]}\|_2^2 \geq n \cdot \left(\frac{2\varepsilon}{n}\right)^2 = \frac{4\varepsilon^2}{n}$$

From the Claim:

$$\|D\|_2^2 = \frac{1}{n} + \|D - U_{[n]}\|_2^2 \geq \frac{1}{n} + \frac{4\varepsilon^2}{n}$$

quadratic mean $\geq$ arithmetic mean

$\Delta_i = |p_i - \frac{1}{n}|$

$$\sqrt{\frac{1}{n}(\Delta_1^2 + \Delta_2^2 + \cdots + \Delta_n^2)} \geq \frac{\Delta_1 + \Delta_2 + \cdots + \Delta_n}{n} = \frac{\|D - U_{[n]}\|_1}{n}$$

$$\|D - U_{[n]}\|_2^2 = \sum_{i=1}^{n} \Delta_i^2 \geq n\left(\frac{\|D - U_{[n]}\|_1}{n}\right)^2$$

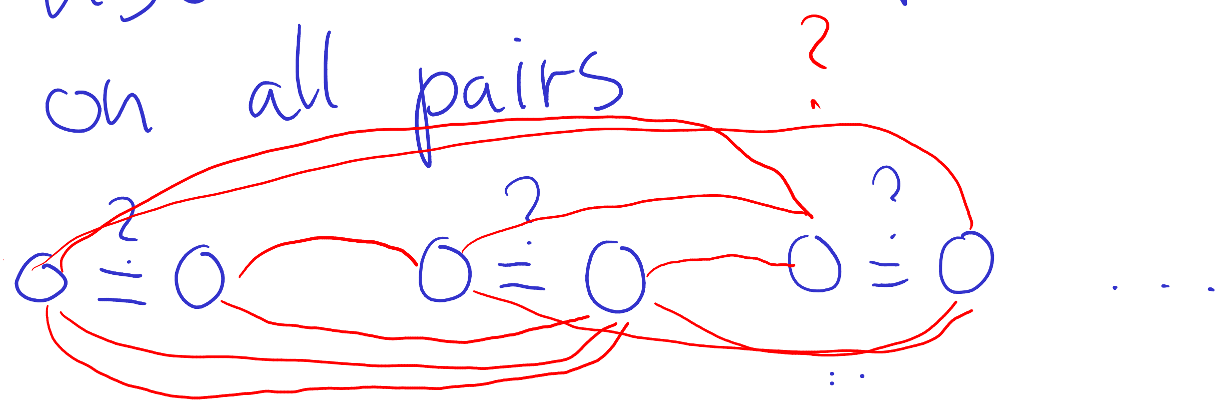
$\boxed{19-4}$

Direct algorithm:

- estimate $\|D\|_2^2$
- simplest approach:
  keep drawing pairs
  and see the number of
  collisions you get in them
- unfortunately: $\Omega(n)$ samples
  to see any
  collisions
  in $D = U_{[n]}$
  vs. $D = U_S$

Idea for better algorithm:

Draw s samples and
use the number of collisions
on all pairs



Example:

samples 3, 4, 2, 3, 3, 4

collisions

collisions

estimate of $\|D\|_2^2 = \dfrac{4}{\binom{6}{2}}$

Analysis more complicated because collisions are <u>not</u> independent

<u>Algorithm:</u>

sufficiently large constant

- collect $s = C \cdot \sqrt{n}/\varepsilon^4$ independent samples $X_1, X_2, \ldots, X_s$ from D
- count collisions:

$$Y = \sum_{i<j} Y_{ij}$$

$$Y_{ij} = \begin{cases} 1 & \text{if } X_i = X_j \\ 0 & \text{otherwise} \end{cases}$$

- if $Y/\binom{s}{2} \geq \frac{1}{n} + \frac{2\varepsilon^2}{n}$
  output NO

  else output YES