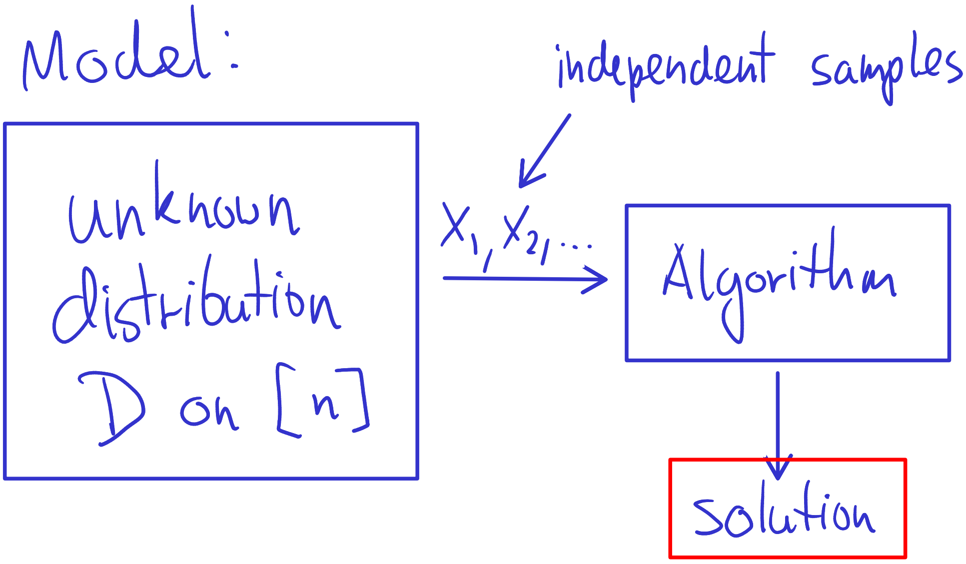
DS-563/CS-543 2024-04-04

Today: Continue uniformity testing

Review

Model:



Task: 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. $99/100$

$$\boxed{\text{Want to use as few samples as possible}}$$

Algorithm proposed last time

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

Analysis last time:

$$\mathbb{E}\left[Y/\binom{s}{2}\right] = \|D\|_2^2$$

distribution treated as n-dimensional vector of probabilities

(1): $D = \mathcal{U}_{[n]}$
$\Rightarrow \|D\|_2^2 = \frac{1}{n}$

(2): $d_{TV}(D, \mathcal{U}_{[n]}) \geq \varepsilon$
$\Rightarrow \|D\|_2^2 \geq \frac{1+2\varepsilon^2}{n}$

Our approach: show that $Y/\binom{s}{2}$ is a good estimator for $\|D\|_2^2$

Lemma: $$\mathrm{Var}[Y] \leq 7\left(\binom{s}{2}\|D\|_2^2\right)^{3/2}$$

Definition: $$\bar{Y}_{ij} = Y_{ij} - \mathbb{E}[Y_{ij}]$$

(of course, $\mathbb{E}[\bar{Y}_{ij}] = 0$)

Useful facts:

① $\mathbb{E}[\bar{Y}_{ij}\bar{Y}_{kl}] \leq \mathbb{E}[Y_{ij}Y_{kl}]$

Why? $$\mathbb{E}[\bar{Y}_{ij}\bar{Y}_{kl}] = \mathbb{E}\Big[\big(Y_{ij} - \overbrace{\mathbb{E}[Y_{ij}]}^{\|D\|_2^2}\big)\big(Y_{kl} - \overbrace{\mathbb{E}[Y_{kl}]}^{\|D\|_2^2}\big)\Big]$$

$$= \mathbb{E}\left[Y_{ij}Y_{kl} - Y_{ij}\|D\|_2^2 - Y_{kl}\|D\|_2^2 + \left(\|D\|_2^2\right)^2\right]$$

$$= \mathbb{E}[Y_{ij}Y_{kl}] - 2\left(\|D\|_2^2\right)^2 + \left(\|D\|_2^2\right)^2 \leq \mathbb{E}[Y_{ij}Y_{kl}]$$

② $\|D\|_3 \leq \|D\|_2$

This in an inequality on norms that follows from Hölder's inequality

(3) $\underline{s^2 \leq 3\binom{s}{2}}$

$$\Updownarrow$$

$$s^2 \leq \frac{3}{2}s(s-1)$$

$$\Updownarrow$$

$$s \leq \frac{3}{2}s - \frac{3}{2}$$

$$\Updownarrow$$

$3 \leq s \leftarrow$ $\boxed{\text{true if the constant } C \text{ defining } s \text{ is large enough}}$

(4) $\underline{\binom{s}{3} \leq \frac{s^3}{6}}$

$$\binom{s}{3} = \frac{s(s-1)(s-2)}{6} \leq \frac{s^3}{6}$$

Proof of the variance lemma:

$$Var[Y] = Var\left[\sum_{i<j} Y_{ij}\right] = Var\left[\sum_{i<j} \bar{Y}_{ij}\right]$$

$$= \mathbb{E}\left[\left(\sum_{i<j} \bar{Y}_{ij}\right)^2\right] - \underbrace{\left(\mathbb{E}\left[\sum_{i<j} \bar{Y}_{ij}\right]\right)^2}_{=0}$$

$$= \mathbb{E}\Bigg[\underbrace{\sum_{i<j} \bar{Y}_{ij}}_{\triangle 1} + \underbrace{\sum_{\substack{i<j \\ k<l \\ i,j,k,l \text{ distinct}}} \bar{Y}_{ij} \bar{Y}_{kl}}_{\triangle 2}$$

$$+ \underbrace{\sum_{\substack{i<j \\ i<l \\ i,j,l \text{ distinct}}} \bar{Y}_{ij} \bar{Y}_{il}}_{\triangle 3} + \underbrace{\sum_{\substack{i<j \\ k<j \\ i,j,k \text{ distinct}}} \bar{Y}_{ij} \bar{Y}_{kj}}_{\triangle 4}$$

$$+ \underbrace{\sum_{\substack{i<j \\ j<l}} \bar{Y}_{ij} \bar{Y}_{jl}}_{\triangle 5} + \underbrace{\sum_{\substack{i<j \\ k<i}} \bar{Y}_{ij} \bar{Y}_{ki}}_{\triangle 6} \Bigg]$$

We analyze the expectation term by term:

$\triangle 1$: $\mathbb{E}\left[\sum_{i<j} \bar{Y}_{ij}^2\right] \overset{(1)}{\leq} \mathbb{E}\left[\sum_{i<j} Y_{ij}^2\right]$

$$= \mathbb{E}\left[\sum_{i<j} Y_{ij}\right] = \binom{s}{2} \|D\|_2^2$$

$\triangle 2$: $\mathbb{E}\left[\sum_{\substack{i<j \\ k<l}} \bar{Y}_{ij} \bar{Y}_{kl}\right]$

i,j,k,l independent

$$\overset{\text{independence}}{=} \sum \underbrace{\mathbb{E}[\bar{Y}_{ij}]}_{=0} \underbrace{\mathbb{E}[\bar{Y}_{kl}]}_{=0} = 0$$

$\triangle 3$: $\mathbb{E}\left[\sum_{\substack{i<j \\ i<l}} \bar{Y}_{ij} \bar{Y}_{il}\right]$

i,j,l distinct

$$\overset{(1)}{\leq} \mathbb{E}\left[\sum Y_{ij} Y_{il}\right]$$

$$= \sum_{\substack{i<j \\ i<l \\ i,j,l \text{ distinct}}} \Pr[X_i = X_j = X_l]$$

$p_i =$ probability of drawing i from D

$$\leq \underbrace{2\binom{s}{3}}_{\substack{\text{because for any selection of three distinct elements we} \\ \text{have either } i<j<l \\ \text{or } i<l<j}} \underbrace{\sum_{i=1}^{n} p_i^3}_{=\|D\|_3^3}$$

$$\leq \underbrace{2 \cdot \frac{s^3}{6}}_{\text{due to } 4} \cdot \underbrace{\|D\|_2^3}_{\text{due to } 2}$$

$$\leq \sqrt{3}\left(\binom{s}{2}\|D\|_2^2\right)^{3/2}$$

$\boxed{\text{due to } 3,\ s^3 = (s^2)^{3/2} \leq \left(3\binom{s}{2}\right)^{3/2}}$

4 : same bound as 3

5 & 6 : $\leq \frac{\sqrt{3}}{2}\left(\binom{s}{2}\|D\|_2^2\right)^{3/2}$ each

Almost the same as 3 & 4, except for each triple selected from s options, there is a unique assignment to indices in the summation. So no factor of 2 needed.

Overall:

$$\mathrm{Var}[Y] \leq \binom{s}{2}\|D\|_2^2 + 3\sqrt{3}\left(\binom{s}{2}\|D\|_2^2\right)^{3/2}$$

If C in the definition of s is large enough, $\binom{s}{2} \geq n$. Then $\binom{s}{2}\|D\|_2^2 \geq n \cdot \frac{1}{n} \geq 1$, and therefore, $\binom{s}{2}\|D\|_2^2 \leq \left(\binom{s}{2}\|D\|_2^2\right)^{3/2}$.

Hence

$$\mathrm{Var}[Y] \leq \underbrace{\left(1+3\sqrt{3}\right)}_{\leq 7} \left(\binom{s}{2} \|D\|_2^2\right)^{3/2}$$

$\blacksquare$

(to be continued in the next lecture)