

4) Hoeffding

(same condition as (3))

$$\Pr \left[ \frac{s}{n} > p + \varepsilon \right] \leq e^{-2m\varepsilon^2}$$

$$\Pr \left[ \frac{s}{n} < p - \varepsilon \right] \leq e^{-2m\varepsilon^2}$$

## distribution testing

An  $(\epsilon, \delta)$ -tester for property  $P$

we have an unknown distribution  $d$

We aim to design an algorithm  $\mathcal{A}$

that distinguishes the following w.p.  $\geq 1 - \delta$ :

- if  $d \in P$ ,  $\mathcal{A}$  outputs **accept**

- if  $d$  is  $\epsilon$ -far from  $P$ ,  $\mathcal{A}$  outputs **reject**

what is a property?

$P$  = a set of distributions

$P = \{U_n\}$   $\rightarrow$  a uniform dist. on  $[n]$

$P = \{$  a set of unimodal distributions  $\}$

$d$  is  $\epsilon$ -far iff  $\text{dist}(d, P) > \epsilon$

$$\text{dist}(d, P) = \min_{d' \in P} \text{dist}(d, d')$$

Example distances:

$\ell_1$ -distance:  $\|d - d'\|_1 = \sum_{x \in \Sigma} |d(x) - d'(x)|$

$$\ell_2\text{-distance: } \|d - d'\|_2 = \sqrt{\sum_{x \in \Omega} (d(x) - d'(x))^2}$$

Total variation distance:  $\|d - d'\|_{TV} = \max_{E \subseteq \Omega} |d(E) - d'(E)|$   
 (statistical distance)

$E \subseteq \Omega$   
 ↳ every event

Turns out  $\|d - d'\|_{TV} = \frac{1}{2} \|d - d'\|_1$

Today's question: uniformity testing

Design algorithm A that receives  $n, \epsilon, \delta$ , and samples from  $d$  and outputs

- accept w.p.  $\geq 1 - \delta$  if  $d = u_n$

- reject w.p.  $\geq 1 - \delta$  if  $\|d - u_n\|_1 > \epsilon$

Q<sub>1</sub>: which one look like a real dice ?

2    3    1    4    6    1

4    6    4    3    4    5

Q<sub>2</sub> what did give it away?

A<sub>2</sub> repetitions!  $\rightsquigarrow$  samples from a uniform distribution looks "less" repeated.

Let's formalize this intuition...

collisions : two samples that are equal to each other

# collisions in the sample set , tells us if a distribution is uniform or not.

Algorithm:

Draw  $m$  samples from  $d: x_1, \dots, x_m$

$$\forall i < j \in [m]: \alpha_{ij} = \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{otherwise} \end{cases}$$

$$Y \leftarrow \sum_{i=1}^m \sum_{j>i}^m \alpha_{ij} / \binom{m}{2}$$

if  $Y < t$

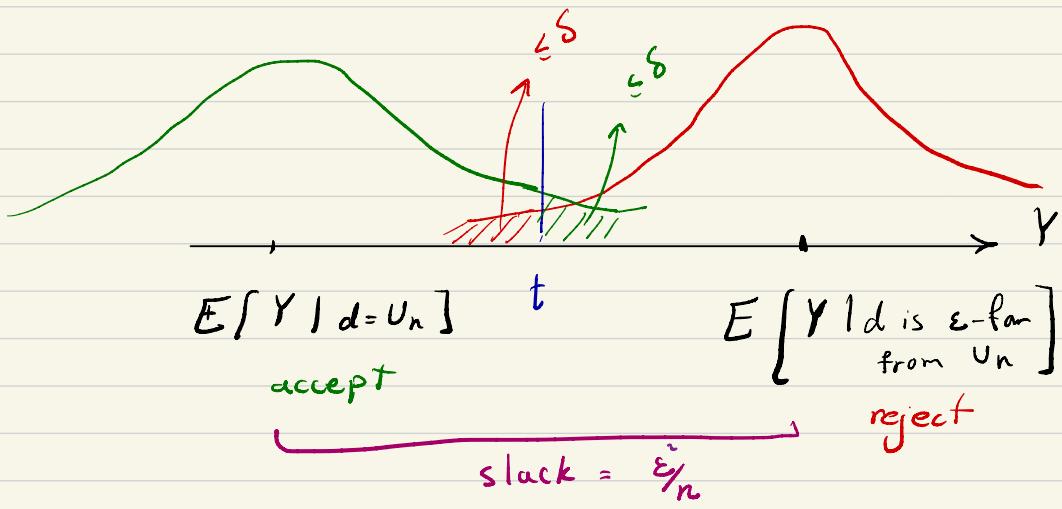
output accept

else

output reject

Our goal here: what should  $m$  &  $t$  be?

Visual description



First step : slack exists

$$E[\sigma_{ij}] = \sum_{a=1}^n \Pr[X_i=a] \cdot \Pr[X_j=a]$$

$$= \sum_{a=1}^n d_a^2 = \|d\|_2^2$$

$$E[Y] = \frac{1}{\binom{m}{2}} \sum_{i=1}^m \sum_{j=i+1}^m \sigma_{ij} = \|d\|_2^2$$

Case 1:  $d$  is uniform

$$\text{if } d = U_n : \|d\|_2^2 = \sum_{a=1}^n d_a^2 = n \times \frac{1}{n^2} = \frac{1}{n}$$

Case 2:  $d$  is  $\varepsilon$ -far from uniform

if  $\|d - U_n\|_1 > \varepsilon$ :

$$\|d\|_2^2 = \sum_{a=1}^n d_a^2 = \sum_{a=1}^n \left( \frac{1}{n} + (d_a - \frac{1}{n}) \right)^2$$

$$= \sum_{a=1}^n \frac{1}{n}^2 + \frac{2}{n} \left( d_a - \frac{1}{n} \right) + \left( d_a - \frac{1}{n} \right)^2$$

$$= \frac{1}{n} + \frac{2}{n} \underbrace{\left( \sum_{a=1}^n d_a - \frac{1}{n} \right)}_{=0} + \sum_{a=1}^n \left( d_a - \frac{1}{n} \right)^2$$

$$= \frac{1}{n} + \underbrace{\|d - U_n\|_2^2}_{\text{our slack}}$$

- Our conjecture is correct & "tends" to be larger when  $d$  is  $\epsilon$ -far from uniform.

How far?

$$\text{we know } \|d - v_n\|_1 \geq \epsilon \quad \left. \begin{array}{l} \\ \text{Cauchy-Schwarz: } (\sum x_i^2) \cdot (\sum y_i^2) \geq (\sum x_i y_i)^2 \end{array} \right\} \Rightarrow$$

$$\left( \sum_a \left( d_a - \frac{1}{n} \right)^2 \right) \cdot \left( \sum_{a=1}^n 1^2 \right) \geq \left( \sum |d_a - \frac{1}{n}| \right)^2$$

$\Rightarrow$

$$\|d - v_n\|_2^2 = \sum_{a=1}^n \left( d_a - \frac{1}{n} \right)^2 \geq \frac{\left( \sum |d_a - \frac{1}{n}| \right)^2}{n}$$

$$= \frac{\|d - v_n\|_1^2}{n} > \frac{\epsilon^2}{n}$$

$$E[Y | d = v_n]$$

$$E[Y | d \text{ is } \varepsilon\text{-far}]$$

$$\begin{array}{ccc} 1 & & 1 \\ \hline \frac{1}{n} & t = \frac{1 + \frac{\varepsilon}{2}}{n} & \geq \frac{1 + \varepsilon^2}{n} \end{array}$$

Next step : Concentration

Let set  $t$  to be in the middle :  $t \leftarrow \frac{1 + \frac{\varepsilon}{2}}{n}$

If we show the following, we get an

$(\varepsilon, \delta)$  - tester

①  $\Pr [ Y \geq \frac{1 + \frac{\varepsilon}{2}}{n} \mid d = v_n ] \leq \delta^\uparrow$   $\delta = 0.1$

②  $\Pr [ Y \leq \frac{1 + \frac{\varepsilon^2}{2}}{n} \mid d \text{ is } \varepsilon\text{-far from } v_n ] \leq \delta^\uparrow$   $\delta = 0.1$

$$Y = \sum_{\binom{m}{2}} \sigma_{ij} \quad \text{not a great candidate}$$

for Chernoff. bound

(why?)

Our plan : Using Chebyshov's

Lets compute the variance of Y

$$\text{Lemma 1} \quad \text{Var}(Y) = \frac{1}{\binom{m}{2}^2} \cdot \left( \binom{m}{2} \|d\|_2^2 + 6 \binom{m}{3} \|d\|_3^3 \right)$$

proof is deferred for now.

Case 1 :  $d = v_n$

$$\Pr \left[ |Y - E[Y]| \geq \frac{\epsilon^2}{2n} \right] \leq \frac{\text{Var}(Y)}{\left(\frac{\epsilon^2}{2n}\right)^2}$$

$$\leq \frac{1}{\binom{m}{2}^2} \cdot \left( \binom{m}{2} \|d\|_2^2 + 6 \binom{m}{3} \|d\|_3^3 \right) \cdot \frac{4n^2}{\epsilon^2}$$

$$= \Theta \left( \frac{n^2}{m^4 \epsilon^4} \cdot \left( m^2 \cdot \frac{1}{n} + \frac{m^3}{n^2} \right) \right)$$

$$= \Theta \left( \frac{n}{m^2 \epsilon^4} + \frac{1}{m \epsilon^4} \right) \leq 0.1$$

$$\text{if } m = c \cdot \left( \frac{1}{\epsilon^4} + \frac{\sqrt{n}}{\epsilon^2} \right)$$

for sufficiently large  $c$

Case 2:  $\|d - U_n\|_1 > \varepsilon$

The bound on the variance can be large.

$${m \choose 2} \|d\|_2^2 + 6 {m \choose 3} \|d\|_3^3$$

Could be problematic if we require  $|Y - E[Y]| \leq \frac{\varepsilon}{n}$

↳ adjust the length accordingly

$$\Pr [ Y - E[Y] \geq \frac{\varepsilon^2}{2} E[Y] ] \leq 4 \frac{\text{Var}[Y]}{\varepsilon^4 E[Y]^2}$$

$$\leq \frac{1}{{m \choose 2}^2} \cdot \frac{{m \choose 2} \|d\|_2^2 + 6 {m \choose 3} \|d\|_3^3}{\varepsilon^4 \|d\|_2^4} =$$

$$= \Theta \left( \frac{1}{m^2 \cdot \varepsilon^4 \|d\|_2^2} + \frac{\|d\|_3^3}{m \cdot \varepsilon^4 \|d\|_2^4} \right) \leq 0.1$$

$$m = C \cdot \frac{\sqrt{n}}{\varepsilon^4}$$

using  $\|d\|_3^3 \leq \|d\|_2^3$

$\uparrow$

$\ell_p$ -norm inequality  $\|d\|_3 \leq \|d\|_2$

$$\underline{\text{Lemma 1}} \quad \text{Var}(Y) = \frac{1}{\binom{m}{2}^2} \cdot \left( \binom{m}{2} \|d\|_2^2 + 6 \binom{m}{3} \|d\|_3^3 \right)$$

proof:

$$\text{Var}(Y) = \text{Var}\left(\frac{1}{\binom{m}{2}} \sum_{i < j} \alpha_{ij}\right)$$

$$= \frac{1}{\binom{m}{2}^2} \text{Var}\left(\sum_{i < j} \alpha_{ij}\right)$$

$$= \frac{1}{\binom{m}{2}^2} \left( E\left[ \left( \sum_{i < j} \alpha_{ij} \right)^2 \right] - \underbrace{\left( \sum_{i < j} E[\alpha_{ij}] \right)}_{\|d\|_2^2} \right)$$

$$= \frac{1}{\binom{m}{2}^2} E\left[ \sum_{i < j} \sum_{l < k} \alpha_{ij} \alpha_{lk} \right]$$

$$- \|d\|_2^4$$

$$\mathbb{E} \left[ \sigma_{ij}^2 \right] = \| d \|^2_2 \quad \textcircled{1} \quad |\{i, j, l, k\}|=2 \Rightarrow i=l, j=k$$

$$\mathbb{E} \left[ \sigma_{ij} \sigma_{lk} \right] = \| d \|^3_3 \quad \textcircled{2} \quad |\{i, j, l, k\}|=3$$

$\hookrightarrow \Pr \{ \text{three samples are equal} \}$

$$\mathbb{E} \left[ \sigma_{ij} \sigma_{lk} \right] = \mathbb{E}[\sigma_{ij}] \cdot \mathbb{E}[\sigma_{lk}] \quad \textcircled{3} \quad |\{i, j, l, k\}|=4$$

$$= \| d \|^4_2$$

$$\Rightarrow \text{Var}[Y] = \frac{1}{\binom{m}{2}^2} \left[ \binom{m}{2} \cdot \| d \|^2_2 + 6 \binom{m}{3} \| d \|^3_3 + \binom{m}{2} \binom{m-2}{2} \| d \|^4_2 - \binom{m}{2}^2 \| d \|^4_2 \right]$$

$$\leq \frac{1}{\binom{m}{2}^2} \left[ \binom{m}{2} \| d \|^2_2 + 6 \binom{m}{3} \| d \|^3_3 \right] \quad \square$$

Exercise: verify that

$$\binom{m}{2} + 6 \binom{m}{3} + \binom{m}{2} \binom{m-2}{2} = \binom{m}{2}^2$$