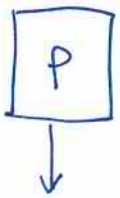
# Lecture 3

## Uniformity testing

## Model




Outputs iid samples

domain $D$, $|D| = n$ ← known

$D = \{1 \ldots n\} = [n]$ for today

$p_i = \Pr[p \text{ outputs } i]$ — unknown

## Testing Uniformity

- if $P \equiv U_{[n]}$ tester outputs PASS ← with prob $\geq 3/4$
- if $\text{dist}(P, U_{[n]}) > \varepsilon$ then tester outputs FAIL

distance measures:

$$\ell_1 : \|p-q\|_1 = \sum_{i \in D} |p_i - q_i|$$

$$\ell_2 : \|p-q\|_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}$$

$$\|p-q\|_2 \leq \|p-q\|_1 \leq \sqrt{n}\, \|p-q\|_2$$

Last time: "plug in" estimate seems to need $\Omega(n)$ samples

Examples

- NJ lottery
- Multilotek

## $L_2$ - Distance (squared):

$$\|p - U_{[n]}\|_2^2 = \sum_{i \in [n]} \left(p_i - \frac{1}{n}\right)^2$$

$$= \sum p_i^2 - \frac{2}{n}\underbrace{\sum p_i}_{=1} + \underbrace{\sum \left(\frac{1}{n}\right)^2}_{=\frac{1}{n}}$$

$$= \underbrace{\sum p_i^2}_{} - \frac{1}{n}$$

Collision probability of $p$:

$$\|p\|_2^2 \equiv \Pr_{s,t \in p}[s = t] = \sum p_i^2$$

for $p = U$, $\|p\|_2^2 = \frac{1}{n}$

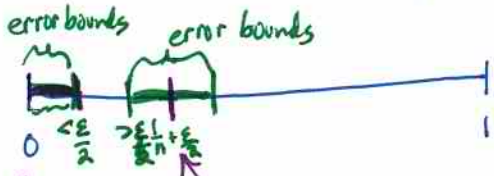
for $p \neq U$, $\|p\|_2^2 > \frac{1}{n}$

$$= \underbrace{\|p\|_2^2}_{\text{we can estimate this}} - \underbrace{\|U_{(n)}\|_2^2}_{\text{we know this since we know } n}$$

## Algorithm

1. take $s$ samples from $p$ ← how many samples?
2. let $\hat{c} \leftarrow$ estimate of $\|p\|_2^2$ from sample ← how?
3. if $\hat{c} < \frac{1}{n} + \delta$ pass
   else fail ← what should $\delta$ be?

## How well do we need to estimate $\|p\|_2^2$?




U has collision prob 0, should output PASS even if have sampling error

if $p'$ is $\varepsilon^2$-far from U in $(L_2\text{-distance})^2$, $\|p'\|_2^2$ is $> \frac{1}{n} + \varepsilon^2$ should output FAIL even if have sampling error

- use $\delta \leq \frac{\varepsilon^2}{2}$ (we used $\delta = \frac{\varepsilon^2}{4}$ on board in class)
- assume with prob $\geq 3/4$ estimate of collision prob $\|p\|_2^2$ is within $\pm\delta$ for $\delta \leq \frac{\varepsilon^2}{2}$

Proper Behavior:

- Then if $p = U$, $\hat{c} < \frac{1}{n} + \underbrace{\frac{\varepsilon^2}{2}}_{\text{error bound}} \Rightarrow$ PASS
- if $\|p - U\|_2^2 > \varepsilon^2$ then $\hat{c} > \frac{1}{n} + \varepsilon^2 - \underbrace{\frac{\varepsilon^2}{2}}_{\text{error bound}} = \frac{1}{n} + \frac{\varepsilon^2}{2}$

$\Rightarrow$ FAIL

- But how many samples to get prob $\geq 3/4$ of good estimate?

[• What if want better estimate of $\|p\|_2^2$, to within $\frac{\varepsilon^2}{3n}$]

How to estimate $\|p-u\|_1$?

1) $\|p-u\|_1 = 0 \iff \|p-u\|_2^2 = 0 \Rightarrow \|p\|_2^2 = \frac{1}{n}$

2) if $\|p-u\|_1 > \varepsilon \Rightarrow \|p-u\|_2 > \frac{\varepsilon}{\sqrt{n}}$

$\Rightarrow \|p-u\|_2^2 > \frac{\varepsilon^2}{n}$

$\Rightarrow \|p\|_2^2 \geq \frac{1}{n} + \frac{\varepsilon^2}{n}$

either additive estimate with error $\leq \frac{\varepsilon^2}{2n}$

or mult error $\leq \left(1 \pm \frac{\varepsilon^2}{3n}\right)$

suffices

First:
How to estimate $\|p\|_2^2$ ?

Naive idea:

take two new samples:

$$X_i \leftarrow \begin{cases} 1 & \text{if samples are equal} \\ 0 & \text{o.w} \end{cases}$$

"gives $\Theta(k)$ samples of collision probability from $k$ samples of $p$"

Better idea: recycle - use all pairs in sample

"gives $\Theta(k^2)$ samples of collision probability from $k$ samples of $p$"

Estimate by recycling:

- Take $s$ samples from $p$: $X_1 \cdots X_s$
- for each $1 \le i < j \le s$

$$b_{ij} \leftarrow \begin{cases} 1 & \text{if } X_i = X_j \\ 0 & \text{if } X_i \neq X_j \end{cases}$$

- Output $\hat{C} \leftarrow \dfrac{\sum_{i<j} b_{ij}}{\binom{s}{2}}$

$b_{ij}$'s not independent so can't use Chernoff

Analysis: $E[\hat{C}] = \frac{1}{\binom{s}{2}} \cdot \binom{s}{2} \cdot E[b_{ij}]$

$= \|p\|_2^2$

## Analysis

$$E[\delta_{ij}] = \Pr[\delta_{ij}] = 1$$
$$= \|p\|_2^2$$

$$E[\hat{c}] = \frac{1}{\binom{s}{2}} \binom{s}{2} E[\delta_{ij}] = \|p\|_2^2$$

$$\Pr[\,|\hat{c} - \|p\|_2^2| > \rho\,] \le \frac{\mathrm{Var}[\hat{c}]}{\rho^2}$$ Chebyshev ‡

**Fact** $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$

$$\text{So} \quad \mathrm{Var}[\hat{c}] = \mathrm{Var}\left[\frac{1}{\binom{s}{2}} \cdot \sum_{i<j} \delta_{ij}\right]$$
$$= \frac{1}{\binom{s}{2}^2} \mathrm{Var}\left[\sum_{i<j} \delta_{ij}\right]$$

**Lemma** $\mathrm{Var}[\sum \delta_{ij}] \le 2\left(\binom{s}{2} \|p\|_2^2\right)^{3/2}$

**Why? (proof...)**

def. $\bar{\delta}_{ij} = \delta_{ij} - E[\delta_{ij}]$

so $E[\bar{\delta}_{ij}] = 0$

Also: $E[\bar{\delta}_{ij}\,\bar{\delta}_{kl}] \le E[\delta_{ij}\,\delta_{kl}]$

verify at home? (or trust...)

- $\left(\sum p(x)^3\right)^{1/3} \le \left(\sum p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

e.g. $(a^3+b^3)^2 \le (a^2+b^2)^3$
$a^6 + 2a^3b^3 + b^6 \le a^6 + b^6 + 3a^4b^2 + 3a^2b^4$

So

$$\mathrm{Var}\left[\sum_{i<j}\bar{\sigma}_{ij}\right] = E\left[\left(\sum_{i<j}\bar{\sigma}_{ij} - E\left[\sum_{i<j}\bar{\sigma}_{ij}\right]\right)^2\right]$$

$$= E\left[\left(\sum_{i<j}\bar{\sigma}_{ij}\right)^2\right]$$

$$= E\Big[\underbrace{\sum_{i<j}\bar{\sigma}_{ij}^2}_{①} + \underbrace{\sum_{\substack{i<j\\ k<\ell\\ i,j,k,\ell \text{ distinct}}}\bar{\sigma}_{ij}\,\bar{\sigma}_{k\ell}}_{②} + \underbrace{\sum_{\substack{i<j\\ k=\ell\\ i,j,\ell \text{ distinct}}}\bar{\sigma}_{ij}\,\bar{\sigma}_{k\ell}}_{③} + \underbrace{\sum_{\substack{i<j\\ k<\ell\\ i,k,j \text{ distinct}}}\bar{\sigma}_{ij}\,\bar{\sigma}_{kj}}_{④}\Big]$$

① $E\left[\sum_{i<j}\bar{\sigma}_{ij}^2\right] \le E\left[\sum\sigma_{ij}^2\right] = \binom{s}{2}\|p\|_2^2$

② ($\bar{\sigma}_{ij}$ independent of $\bar{\sigma}_{k\ell}$)

$$E\left[\sum_{\substack{i<j\\ k<\ell\\ \text{all 4 distinct}}}\bar{\sigma}_{ij}\,\bar{\sigma}_{k\ell}\right] \le \sum E[\bar{\sigma}_{ij}]\,E[\bar{\sigma}_{k\ell}] = 0$$

③ $$E\left[\sum\bar{\sigma}_{ij}\,\bar{\sigma}_{i\ell}\right] \le E\left[\sum_{\substack{i,j,\ell\\ \text{distinct}}}\sigma_{ij}\cdot\sigma_{i\ell}\right]$$

$$\le \binom{s}{3}\sum_x p(x)^3$$

expected # 3-way collisions

$$\le \frac{s^3}{6}\left(\sum_x p(x)^2\right)^{3/2}$$

$$\le \frac{\sqrt{3}}{2}\binom{s}{2}^{3/2}\left(\|p\|_2^2\right)^{3/2}$$ by the facts

$\frac{1}{6}\left(s^2\right)^{3/2} < \frac{\left(3\binom{s}{2}\right)^{3/2}}{6} = \frac{\sqrt{3}}{2}\binom{s}{2}^{3/2}$

(4) same as (3)

In total: $\mathrm{Var}\left[\sum_{i<j} \delta_{ij}\right] \le \mathrm{Var}\left[\sum_{i<j} \bar{\delta}_{ij}\right]$

$$\le \binom{s}{2}\|p\|_2^2 + 0 + 2\cdot\frac{\sqrt{3}}{2}\left(\binom{s}{2}\|p\|_2^2\right)^{3/2}$$

$$\le 2\left[\binom{s}{2}\|p\|_2^2\right]^{3/2}$$

Putting lemma into Chebyshev:

2) use $\rho = \frac{\varepsilon}{2}E[\hat{C}]$ ; $\Pr\left[\left|\hat{C} - \|p\|_2^2\right| > \frac{\varepsilon}{2}\|p\|_2^2\right]$

$$\le \frac{\mathrm{Var}\left[\sum \delta_{ij} / \binom{s}{2}\right]}{\left(\frac{\varepsilon}{2}\right)^2 E\left[\sum \delta_{ij} / \binom{s}{2}\right]^2}$$

$$\le \frac{2\left[\binom{s}{2}\|p\|_2^2\right]^{3/2}}{\left(\frac{\varepsilon}{2}\right)^2\left(\|p\|_2^2\right)^2} \cdot \frac{1}{\binom{s}{2}^2}$$

$$\le \frac{8}{\binom{s}{2}^{1/2}}\varepsilon^{-2}\left(\|p\|_2^2\right)^{-1} \le \frac{8}{\varepsilon^2\sqrt{\binom{s}{2}\|p\|_2^2}}$$

Pick $s \ge \frac{\text{const.}\cdot\sqrt{n}}{\varepsilon^2} \Rightarrow \le \frac{8}{\varepsilon^2\left(\frac{c'\cdot n}{\varepsilon^4}\cdot\|p\|_2^2\right)^{1/2}} = \frac{8}{\sqrt{n}\cdot c''\cdot\left(\|p\|_2^2\right)^{1/2}} \le \frac{1}{4}$ for $c'' \gg 32$

1) use $p = \frac{\varepsilon^2}{2}$

$$\Pr\left[\left|\hat{c} - \|p\|_2^2\right| > \frac{\varepsilon^2}{2}\right] \leq \frac{\mathrm{Var}[\hat{c}]}{\varepsilon^4} \cdot 4$$

$$\leq \frac{2\left[\binom{s}{2}\|p\|_2^2\right]^{3/2}}{\varepsilon^4} \cdot 4 = \frac{8}{\varepsilon^4} \cdot s^3 \cdot \|p\|_2^3$$

Pick $s \geq \frac{1}{\varepsilon^{4/3}}$

Note: Can get better bnd if have bound on $\|p\|_\infty$ ↑ max prob element

In homework:

1) Testing closeness to any known distribution — reduce to uniform case!

2) lower bound

Some other extensions:

What if $p, q$ both unknown?

$L_2$ distance is similar, but what does it say?

$L_2$ distance: $\|p-q\|_2^2 = \sum_i (p_i - q_i)^2$

$= \sum p_i^2 - 2\sum p_i q_i + \sum q_i^2$

- $\sum p_i^2 = \|p\|_2^2$ — self-collision prob of $p$
- $\sum p_i q_i$ — cross-collision probability of $p$ & $q$
- $\sum q_i^2 = \|q\|_2^2$ — self-collision prob of $q$

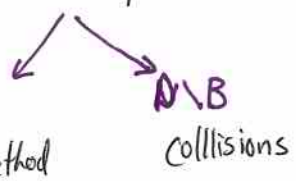
- can bound variance of $\|p\|_2^2$, $\sum p_i q_i$ & $\|q\|_2^2$ estimators if <u>max prob element is bounded by $b$</u>
- what about other case?

use naive method on elements whose prob $\geq b$ ($\leq \frac{1}{b}$ of these)

Filtering algorithm:

learn $B$ = domain elements with prob $\geq b$ ← $O(\frac{1}{b}\log\frac{1}{b})$ samples

filter rest of samples




$O(\frac{1}{\varepsilon^2}\cdot\frac{1}{b}\log\frac{1}{b})$ samples

Note strange dependence on $n$!

$n^{2/3}$ is tight!! → Turns out $O(\frac{1}{\varepsilon^4} n^{2/3}\log n)$ samples suffice [recent improvements on $\varepsilon$, $\log n$ known]