Model

Outputs iid samples

domain D, $|D| = n$ ← known

$D = \{1 \dots n\} = [n]$ 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

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

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

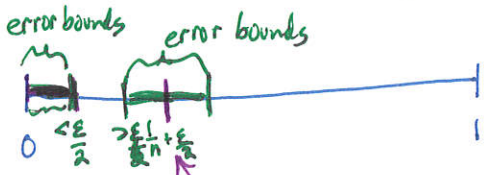
Algorithm

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

   else fail ← (3) what should δ be?

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

actual collision probability



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

if p' is ε^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 $\frac{\varepsilon^2}{3n}$]

[see later for estimate of L_1]

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

Assumption ✡: $|\hat{C} - \|p\|_2^2| < \Delta$

will take enough samples so that this holds with prob $\geq 3/4$

↑ this is our parameter that determines whether our approximation is good. Spoiler: will set $\Delta = \frac{\varepsilon^2}{2}$

What happens if ✡ holds with $\Delta = \frac{\varepsilon^2}{2}$?

- if $p = U_{[n]}$ then $\hat{C} \leq \|U\|_2^2 + \Delta = \frac{1}{n} + \frac{\varepsilon^2}{2}$

  so test will PASS

- if $\|p - U_{[n]}\|_2 > \varepsilon$ then $\|p - U_{[n]}\|_2^2 > \varepsilon^2$

  but then $\|p\|_2^2 = \|p - U_{[n]}\|_2^2 + \frac{1}{n}$

  $> \varepsilon^2 + \frac{1}{n}$

  + $\hat{C} > \|p\|_2^2 - \Delta$

  $\geq \varepsilon^2 + \frac{1}{n} - \Delta = \varepsilon^2 + \frac{1}{n} - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{2} + \frac{1}{n}$

  so test will FAIL

How many samples do we need to estimate $\hat{C}$ to within Δ?

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

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

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

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

Fact $Var[aX] = a^2 Var[X]$

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

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

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...)

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

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

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

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

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

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

$E[\delta_{ij}] = E[\delta_{ij}^2]$ since δ_{ij} is indicator var

② independent

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

③ $E\left[\sum\bar{\delta}_{ij}\,\bar{\delta}_{i\ell}\right] \le E\left[\sum_{\substack{i,j,\ell\\ \text{distinct}}}\delta_{ij}\,\delta_{i\ell}\right] = \sum_{\substack{i,j,\ell\\ \text{distinct}}}\Pr[X_i=X_j=X_\ell]$

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

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

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

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

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

($\frac{8}{\varepsilon^4}$: also want this to be ≤ 1; $\|p\|_2^3 \leq 1$)

So pick $s \geq \frac{1}{\varepsilon^4}$

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

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

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

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

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

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

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

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

suffices

would have this if have additive error $\leq \frac{\varepsilon^2}{3n} \cdot \|p\|_2^2$

to get additive error $\leq \frac{\varepsilon^2}{3n} \|p\|_2^2$

suffices to have

$$s \geq \frac{\text{const} \cdot \sqrt{n}}{\varepsilon^2} \quad \text{samples}$$

Some other extensions:

What if p, q both unknown?

L_2 distance is similar, but what does it say?

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

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

- p_i^2: $\|p\|_2^2$, self-collision prob of p
- $\sum p_i q_i$: cross-collision probability of p & q
- 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 max prob element is bounded by b

- 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

- B: naive method, $O(\frac{1}{\varepsilon^2}\cdot\frac{1}{b}\log\frac{1}{b})$ samples
- $D \setminus B$: collisions

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 ε, $\log n$ known]