Lecture 13

Testing distributions:

the case of uniformity (cont)

A new model:

Probability distributions: get samples

$$\boxed{p}$$
$$\Downarrow$$

this is all we see $\{$ iid samples

Discrete Domain D s.t. $|D| = n$ ← known

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

Examples: lottery data
shopping choices
experimental outcomes
⋮

What do we need to know? is it uniform?
high entropy?
large support? (many distinct elts with >0 probability)
monotone increasing, k-modal?
k-histogram?

Methods?

learn distribution

χ^2- test

plug-in estimate

Max likelihood estimate

Goal: sample complexity sublinear in n

domain size

Testing Uniformity

uniform dist on domain D

with prob $\geq 3/4$

goal: if $p \equiv U_D$ then output PASS

if $\underbrace{\text{dist}(p, U_D)}_{\text{which measure of distance?}} > \varepsilon$ then output FAIL

ℓ_1, ℓ_2, KL-divergence, Earthmover, Jensen-Shannon ....

today's focus

Distances

ℓ_1-distance: $\quad \|p-q\|_1 = \sum_{i \in D} |p_i - q_i|$

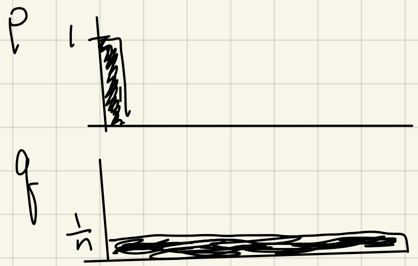
ℓ_2-distance: $\quad \|p-q\|_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}$

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

examples:

① $p = (1, 0, 0, 0, \dots, 0)$

$q = \left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right)$

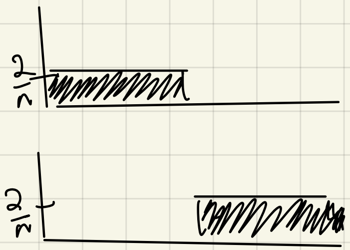


$\|p-q\|_1 = \left(1 - \frac{1}{n}\right) + (n-1)\left(\frac{1}{n}\right) \approx 2$

$\|p-q\|_2 = \left(1-\frac{1}{n}\right)^2 + (n-1)\left(\frac{1}{n^2}\right) \approx 1$

② $p = \left(\frac{2}{n}, \frac{2}{n}, \frac{2}{n}, \dots \frac{2}{n}, 0, 0, \dots 0\right)$

$q = \left(0, 0, \dots 0, \frac{2}{n}, \frac{2}{n}, \dots \frac{2}{n}\right)$



$\|p-q\|_1 = n \cdot \frac{2}{n} = 2$

$\|p-q\|_2^2 = n \cdot \left(\frac{2}{n}\right)^2 = \frac{4}{n}$ so

$\|p-q\|_2 = \frac{2}{\sqrt{n}}$ tiny even though L_1 is big

Via "Plug-in" Estimate:

- take m samples from p
- estimate $p(x)$ $\forall x$ via $\hat{p}(x) = \frac{\#\text{ times } x \text{ occurs in sample}}{m}$
- if $\sum_x |\hat{p}(x) - \frac{1}{n}| > \varepsilon$ reject

  else accept

How many samples?

can "learn" (approximately) any distribution w.r.t. L_1 distance in $\Theta\left(\frac{n}{\varepsilon^2}\right)$ samples

Let's consider L_2-distance (squared):

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

$\underbrace{U}_{\text{uniform on } 1 \dots n}$

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

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

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

$$= \underbrace{\sum p_i^2}_{\text{collision prob of } p:\ \|p\|_2^2 = \Pr_{s,t \in p}[s=t] = \sum p_i^2} - \frac{1}{n}$$

$\frac{1}{n}$: collision prob of uniform distribution $= \|U_{[n]}\|_2^2$

we know this since we know n

$$= \|p\|_2^2 - \|U_{[n]}\|_2^2$$

Algorithm to estimate:

- take s samples of p
- let $\hat{c} \leftarrow$ estimate of $\|p\|_2^2$ from sample
- if $\hat{c} < \frac{1}{n} + \delta$ pass

  else fail

1. how big is s?
2. how to estimate?
3. what should δ be

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

i.e. what should δ be ?

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

will take enough samples s.t. this holds with prob $\geq 3/4$

this is our parameter that determines whether our approximation is good.

recall:

$$\|p - U_{[n]}\|_2^2 = \|p\|_2^2 - \|U_{[n]}\|_2^2$$

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

• if $p = U_{[n]}$ then $\hat{C} < \|U_{[n]}\|_2^2 + \frac{\varepsilon^2}{2} \leq \frac{1}{n} + \frac{\varepsilon^2}{2}$ so if use $\delta = \frac{\varepsilon^2}{2}$ test should PASS

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

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

& ☆ $\Rightarrow$ $\hat{C} > \left(\varepsilon^2 + \frac{1}{n}\right) - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{n} + \frac{1}{n}$ so if we use $\delta = \frac{\varepsilon^2}{2}$ test should FAIL

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

Naive idea: • repeat several times:

• take two samples & set $X_i \leftarrow \begin{cases} 1 & \text{if two samples equal} \\ 0 & \text{o.w.} \end{cases}$

• output average of X_i's

Better idea: "recycle" use all pairs in sample

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

• Take s samples from p: $x_1, \dots x_s$

• For each $1 \le i < j \le s$

$$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{o.w.} \end{cases}$$

σ_{ij}'s are not independent

$\Rightarrow$ can't use Chernoff

• Output $\hat{c} \leftarrow \dfrac{\sum_{i<j} \sigma_{ij}}{\binom{s}{2}}$

Analysis: $E[\hat{c}] = \frac{1}{\binom{s}{2}} \cdot E\left[\sum_{i<j} \sigma_{ij}\right] = \frac{1}{\binom{s}{2}} \sum_{i<j} E[\sigma_{ij}] = \frac{\binom{s}{2}}{\binom{s}{2}} E[\sigma_{ij}] = \Pr[\sigma_{ij}=1]$

$= \|p\|_2^2$

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

Chebyshev's $\neq$

recall $Var[x] = E[(x - E[x])^2]$

$$Var[\hat{c}] = \frac{1}{\binom{s}{2}^2} \underbrace{Var\left[\sum_{i<j} \sigma_{ij}\right]}_{\text{need to bound}}$$

by fact: $Var[aX] = a^2 Var[x]$

need to bound

difficulty: σ_{ij}'s not independent

<u>Lemma</u> $Var\left[\sum_{i<j} \sigma_{ij}\right] \le \binom{s}{2} \|p\|_2^2 + 4\left(\binom{s}{2} \|p\|_2^2\right)^{3/2}$

so $Var[\hat{c}]$ is $O\left(\frac{\|p\|_2^2}{s^2} + \frac{\|p\|_2^3}{s}\right)$

Lemma $\mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] \le \binom{s}{2}\|p\|_2^2 + 4\cdot\left[\binom{s}{2}\|p\|_2^2\right]^{3/2}$

Proof

def $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$ $\Longleftarrow$ trick: rewrite variance as $E\left[\sum \bar{\sigma}_{ij}^2\right]$

why? $\mathrm{Var}\left[\sum \bar{\sigma}_{ij}\right] = E\left[\left(\sum \bar{\sigma}_{ij} - \underbrace{E\left[\sum \bar{\sigma}_{ij}\right]}_{=0}\right)^2\right]$

$= E\left[\left(\sum \sigma_{ij} - E[\sigma_{ij}]\right)^2\right]$

$= \mathrm{Var}\left[\sum \sigma_{ij}\right]$

So can equivalently bound $\mathrm{Var}\left[\sum \bar{\sigma}_{ij}\right]$

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

Facts:

- $E[\bar{\sigma}_{ij}\bar{\sigma}_{k\ell}] \le E[\sigma_{ij}\sigma_{k\ell}]$
- $\left(\sum_x p(x)^3\right)^{1/3} \le \left(\sum_x p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

(Verify @ home)

Lemma $\quad \mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] \le \binom{s}{2}\|p\|_2^2 + 4\cdot\left[\binom{s}{2}\|p\|_2^2\right]^{3/2}$

$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{o.w.} \end{cases}$

def $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$

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

Facts:

- $E[\bar{\sigma}_{ij}\bar{\sigma}_{kl}] \le E[\sigma_{ij}\sigma_{kl}]$
- $\left(\sum_x p(x)^3\right)^{1/3} \le \left(\sum_x p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

Proof

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

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

$$+ \underbrace{\sum_{i<j<l}\bar{\sigma}_{ij}\bar{\sigma}_{jl}}_{5}\Bigg]$$

Lets bound each term:

$\sigma_{ij}^2 = \sigma_{ij}$ since indicator var

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

(2) $$E\Big[\sum_{\substack{i<j \\ k<l \\ \text{all distinct}}} \bar{\sigma}_{ij} \cdot \bar{\sigma}_{kl}\Big] \le \underbrace{\sum E[\bar{\sigma}_{ij}] \cdot E[\bar{\sigma}_{kl}]}_{} = 0$$

this is where the trick helps — gets rid of lots of terms

$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{o.w.} \end{cases}$

def $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$

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

Facts:

- $E[\bar{\sigma}_{ij}\bar{\sigma}_{kl}] \le E[\sigma_{ij}\sigma_{kl}]$
- $\left(\sum_x p(x)^3\right)^{1/3} \le \left(\sum_x p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

(3) (+ similarly (4) + (5))

$$E\Big[\sum_{\substack{i<j \\ i,j,l \text{ distinct}}} \bar{\sigma}_{ij}\, \bar{\sigma}_{il}\Big] \le E\Big[\sum \sigma_{ij}\, \sigma_{il}\Big] = \sum_{i,j,l \text{ distinct}} \Pr[x_i = x_j = x_l]$$

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

expected # 3-way collisions

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

by facts.

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

So,

$$\mathrm{Var}\Big[\sum_{i<j} \sigma_{ij}\Big] = \mathrm{Var}\Big[\sum_{i<j} \bar{\sigma}_{ij}\Big]$$

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

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

$\blacksquare$

$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{o.w.} \end{cases}$

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

where $s = \#$ samples

We have:

$$\mathrm{Var}(\hat{C}) = O\left(\frac{\|p\|_2^2}{s^2} + \frac{\|p\|_2^3}{s} \right)$$

Put into Chebyshev with $\rho = \frac{\varepsilon^2}{2}$:

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

$$\le \frac{\mathrm{const} \cdot \overbrace{\|p\|_2^2}^{\le 1}}{\underbrace{\varepsilon^4 \cdot s^2}_{\text{want this } \le 1}} + \underbrace{\mathrm{Const} \cdot \frac{1}{\varepsilon^4} \cdot \frac{1}{s}}_{\text{also want this } \ll 1} \cdot \underbrace{\|p\|_2^3}_{\le 1}$$

need $s = \Omega(1/\varepsilon^2)$

so pick $s = \Omega\left(\frac{1}{\varepsilon^4}\right)$

BIGGER CONSTRAINT

Note can get better bounds

s is independent of n!

recall: $\|p-U_{[n]}\|_2^2 = \|p\|_2^2 - \|U_{[n]}\|_2^2$

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 $\|p-U\|_1 > \varepsilon \Rightarrow \|p-U\|_2 > \frac{\varepsilon}{\sqrt{n}}$

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

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

So either additive estimate of $\|p\|_2^2$ to within $\frac{\varepsilon^2}{2n}$

or mult estimate of $\|p\|_2^2$ to within $\left(1 \pm \frac{\varepsilon^2}{3}\right)$

suffices

turns out that picking # samples $S \gg \frac{\sqrt{n}}{\varepsilon^4}$ suffices

$S = O(\sqrt{n})$

Generalizations: Given another distribution q,

is $p = q$ or is p "far" from q?

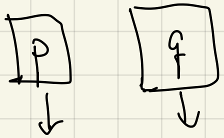
focus on L_1 distance

1. "Identity Testing"

q is known to algorithm, no samples of q needed

} focus on sample complexity but runtime can be made similar

2. "Closeness Testing"



q is given via samples

Will see more on these soon (e.g. pset, lecture ...)

What is complexity in terms of n??

A difficulty in analyzing distribution testers:

typical algorithm:

take m samples $\{s_1 \dots s_m\} = S$

let $X_i :=$ # times i occured in sample

⋮

problem:
X_i's are <u>not</u> independent

e.g. if $X_1 = \frac{m}{2}+1$
then $X_2 < \frac{m}{2}$

e.g. $S = \{2, 5, 3, 2, 3\}$

$X_2 = X_3 = 2$

$X_5 = 1$

all other $X_i = 0$

$$\boxed{\begin{array}{l} Poi(\lambda): \Pr[x=k] = \frac{e^{-\lambda}\lambda^k}{k!} \\ E[x] = Var[x] = \lambda \end{array}}$$

Can we make the X_i's independent? Poissonization

new algorithm:

1

$\hat{m} \leftarrow Poi(m)$

Take $\hat{m}$ samples to get $\hat{S}$

let $X_i :=$ # times i occured in $\hat{S}$

⋮

equivalent
$\Longleftrightarrow$

2

For each $i \in [n]$

$X_i \leftarrow Poi(m \cdot p_i)$

add X_i coples of i to sample

Randomly permute the sample

⋮

why equivalent?

$$\Pr[X_i = c \text{ according to } (1)] = \sum_{k=c}^{\infty} \Pr[\hat{m} = k] \cdot \binom{k}{c} \cdot p_i^c \cdot (1-p_i)^{k-c}$$

$X \sim \mathrm{Poi}(\lambda)$

$\Pr[X=k] = \frac{e^{-\lambda}\lambda^k}{k!}$

$E[X] = \mathrm{Var}[X] = \lambda$

use $\lambda = m$

$$= \sum_{k=c}^{\infty} \frac{e^{-m} m^k}{\cancel{k!}} \cdot \frac{\cancel{k!}}{(k-c)! \cdot c!} \cdot p_i^c \cdot (1-p_i)^{k-c}$$

$$= \frac{e^{-m} m^c p_i^c}{c!} \underbrace{\sum_{k=c}^{\infty} \frac{m^{k-c} (1-p_i)^{k-c}}{(k-c)!}}_{= \sum_{k'=0}^{\infty} \frac{(m(1-p_i))^{k'}}{k'!} = e^{m(1-p_i)}}$$

$$= \frac{e^{-m+m(1-p_i)} (mp_i)^c}{c!} = \frac{e^{mp_i} (mp_i)^c}{c!} = \Pr[X_i = c \text{ when } X_i \sim \mathrm{Poi}(mp_i)]$$

$$= \Pr[X_i = c \text{ according to } (2)]$$

Another difficulty: $\|p\|_2$ can be large

e.g. uniformity test statistic $\quad Var[\hat{C}] = O\left(\frac{\|p\|_2^2}{s^2} + \frac{\|p\|_2^3}{s}\right)$

on new domains

Goal: transform distributions p, q into p', q' s.t. $\|p'\|_2$ & $\|q'\|_2$ small

& $\quad p = q \Rightarrow p' = q'$

$\|p-q\|_1 > \varepsilon \Rightarrow \|p'-q'\| > \varepsilon$

give reduction to small L_2-norm

Comment: q may be "known" or given via samples

Transformation of p:

$S \leftarrow$ Draw $Poi(k)$ samples from p over domain $[n]$

$b_i \leftarrow$ # times i appears in S $\quad \forall i \in [n]$

$\forall i$ add $b_i + 1$ elements to new domain

(i, j) where $j \in [b_i + 1]$

New distribution p':

pick $i \in_R p$

pick $j \in_R [b_i + 1]$

output (i, j)

$$p'(i,j) = \frac{p(i)}{b_i + 1}$$

\# samples

$\downarrow$

size $m + n$

Example:

domain of p is $[5]$ $\quad n = 5$

e.g. $S = \{2, 5, 3, 2, 3\}$

$X_2 = X_3 = 2$

$X_5 = 1$

all other $X_i = 0$

domain of $p' = \{ (1,1),$
$(2,1)\,(2,2),\ (2,3),$
$(3,1),(3,2),\ (3,3),$
$(4,1),$
$(5,1),(5,2)\}$

p' is "mixture" of uniform + observed distribution: $p' \approx \alpha \cdot U + (1-\alpha) \cdot \hat{p}(i)$

$b_i \leftarrow$ # times i appears in S' $\quad \forall i \in [n]$

p' :

pick $i \in_R p$

pick $j \in_R [b_i + 1]$

output (i, j)

$$\left\{ \; p'(i,j) = \frac{p(i)}{b_i + 1} \right.$$

<u>Claim</u> : $\quad E[\|p\|_2^2] \leq \frac{1}{m}$

Why? $\quad E[\|p\|_2^2] = E\left[\sum_{i=1}^{n} \sum_{j=1}^{b_i+1} p'(i,j)^2 \right] = E\left[\sum_{i=1}^{n} \sum_{j=1}^{b_i+1} \frac{p(i)^2}{(b_i+1)^2} \right]$

$$= E\left[\sum_i \frac{p(i)^2}{(b_i+1)} \right]$$

$$\underset{\star}{\leq} \sum_i \frac{p(i)^2}{k \cdot p(i)} = \frac{1}{k}$$