Lecture 14

More on testing distributions:

- Poissonization
- Dealing with large L_2-norm

(+ by the way,... Testing Closeness)

Recall our setting:

Probability distributions: get samples (only)

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

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

$$\boxed{p} \Downarrow$$

this is all we see { iid samples

Distances

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

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

$$\|p\|_2 = \sqrt{\sum p_i^2}$$

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

Last time

Testing Uniformity

uniform dist on domain D

with prob $\geq 3/4$

goal: if $p = U_D$ then output PASS

if $\underbrace{\text{dist}}_{\ell_1 + \ell_2 \text{ distance measures}}(p, U_D) > \varepsilon$ then output FAIL

Generalizations: Given another distribution q,

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

today: focus on L_1-distance

q = uniform $O(\sqrt{n})$

for all q $O(\sqrt{n})$

1. "Identity Testing" $\boxed{p}$ $\boxed{q}$

q is known to algorithm, no samples of q needed

$\Theta(n^{2/3})$

2. "Closeness Testing" $\boxed{p}$ $\boxed{q}$

q is given via samples

Tolerant version:

$\|p-q\|_1 < \varepsilon$

$\|p-q\|_1 \geq \varepsilon'$ for $\varepsilon' \gg \varepsilon$ $\Big\}$ $\Theta\left(\frac{n}{\log n}\right)$

What is the sample complexity of these problems in terms of n

Recall: "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?

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

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 not indep

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

$Poi(\lambda)$: $Pr[x=k] = \frac{e^{-\lambda}\lambda^k}{k!}$

$E[x] = Var[x] = \lambda$

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

$\Longleftrightarrow$

For each $i \in [n]$

2

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

add X_i coples of i to sample

Randomly permute the sample

⋮

why equivalent?

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

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

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

$$= \frac{e^{-\cancel{m}} m^c p_i^c \, e^{m(\cancel{1}-p_i)}}{c!} = \frac{e^{-mp_i} (mp_i)^c}{c!} = \underset{X_i \sim \mathrm{Poi}(mp_i)}{\Pr[X_i = c]} = \underset{\text{process II}}{\Pr[X_i = c]}$$

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

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

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

use $\lambda = m$

need to check joint distributions same

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

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

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

"reduction" to small L_2-norm case

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

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

will work when q known & when given via samples

Transformation of p :

m = # samples "expected" by original alg

Diakonikolas Kane FOCS 2016

See also Sublinear Sp 2019 course notes by Maryam Aliakbarpour

$S \leftarrow$ Draw $m' = \mathrm{Poi}(m)$ samples from p over domain $[n]$

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

equivalently: $b_i \sim \mathrm{Poi}(p(i) \cdot m)$

$\forall i$ 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)

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

size $= m' + n$

Example: domain of p is $[5]$

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

$b_2 = b_3 = 2$

$b_5 = 1$

all other b_i's $= 0$

domain of p':

$\{ (1,1)$

$(2,1)$ $(2,2)$ $(2,3)$

$(3,1)$ $(3,2)$ $(3,3)$

$(4,1)$

$(5,1)$ $(5,2) \}$

Prob

$p(1)$

$\frac{p(2)}{3}$ $\frac{p(2)}{3}$ $\frac{p(2)}{3}$

$\frac{p(3)}{3}$ $\frac{p(3)}{3}$ $\frac{p(3)}{3}$

$p(4)$

$\frac{p(5)}{2}$ $\frac{p(5)}{2}$

$b_i \leftarrow$ # 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)

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

Claim: $E[\|p'\|_2^2] \leq \frac{1}{m}$ $\Big\}$ Problem? we don't know if q''s L_2 norm gets small

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

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

$$\leq \sum_i \frac{p(i)^{\cancel{2}}}{\underbrace{m \cdot \cancel{p(i)}}_{\lambda}} \leq \frac{1}{m} \cdot \sum p(i) \leq \frac{1}{m} \quad \blacksquare$$

$X \sim Poi(\lambda)$

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

$$E[X] = Var[X] = \lambda$$

Claim for $z \sim poi(\lambda)$ $E\left[\frac{1}{z+1}\right] \leq \frac{1}{\lambda}$

$$E\left[\frac{1}{z+1}\right] = \sum_{z=0}^{\infty} \frac{e^{-\lambda} \lambda^z}{(z+1) z!} = \frac{1}{\lambda} \sum_{z=0}^{\infty} \frac{e^{-\lambda} \lambda^{z+1}}{(z+1)!} = \frac{1}{\lambda} \underbrace{\sum_{z'=1}^{\infty} \frac{e^{-\lambda} \lambda^{z'}}{(z')!}}_{\text{prob of disjoint events so sum to } \leq 1}$$

$$\leq 1/\lambda$$

After transform $p \to p'$, $q \to q'$ using <u>same</u> S:

$$\|p-q\|_1 = \sum_x |p(x) - q(x)|$$

$$= \sum_x \sum_{y=1}^{b_x+1} \frac{|p(x) - q(x)|}{b_x+1}$$

$$= \sum_x \sum_{y=1}^{b_x+1} |p'(x) - q'(x)|$$

$$= \|p'-q'\|_1$$

Transformation of p:

$S \leftarrow$ Draw $\mathrm{Poi}(m)$ 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)

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

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

L_2 distance estimation between two distributions p & q:

easier when both $\|p\|_2^2$ & $\|q\|_2^2$ are small

Thm* given samples of p, q, distributions on $[n]$, s.t. $b \geq \max\{\|p\|_2, \|q\|_2\}$

Can distinguish $p = q$ from $\|p-q\|_1 > \varepsilon$ in $O(bn/\varepsilon^2)$ samples

Corr if $b = \min\{\|p\|_2, \|q\|_2\}$

Can distinguish $p = q$ from $\|p-q\|_1 > \varepsilon$ in $O(bn/\varepsilon^2)$ samples

Pf idea:

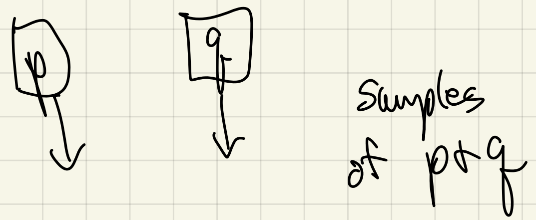
1. estimate $\|p\|_2$ & $\|q\|_2$ to mult factor of c (can do this in $O(\sqrt{n})$ samples)
2. if differ by $> c$ mult factor infer $p \neq q$ & reject
3. else use Thm * with $b' = c \cdot b$ $\blacksquare$

Testing Closeness:

Corr if $b \ge \min\{\|p\|_2^2, \|q\|_2^2\}$

can distinguish $p=q$ from $\|p-q\|_1 > \varepsilon$ in $O(bn/\varepsilon^2)$ samples

1. let $k = n^{2/3}\varepsilon^{-4/3}$
2. $S \leftarrow$ multiset of $\mathrm{Poi}(k)$ samples from q
3. run tester of Corr on p', q' w.r.t. S



Why does it work?

distinguishing p & q & p', q' are equivalent

how many samples needed?

- whp $|S|$ is $\Theta(k)$
- $E[\|q'\|_2^2] = O(\frac{1}{k})$ so whp $\|q'\|_2 = O(\frac{1}{\sqrt{k}})$
- $O(\underset{\substack{\uparrow \\ \text{picking } S}}{k} + \underset{\substack{\uparrow \\ \text{run tester on } p', q'}}{\frac{1}{\sqrt{k}} \cdot n \cdot \frac{1}{\varepsilon^2}}) = O(n^{2/3} \cdot \varepsilon^{-4/3} + \frac{1}{n^{1/3} \cdot \varepsilon^{-2/3}} \cdot n \cdot \frac{1}{\varepsilon^2})$

$= O(n^{2/3}\varepsilon^{-4/3})$ ☒