Testing Closeness of distributions

- outline 8

- Poissonization method
- Flattening Technique
- Estimating L2 distance
- L1 closeness tester

Problem: sample access to p and q

Test whether $\begin{cases} p = q \\ \text{vs} \\ \|p-q\|_1 \geq \varepsilon \end{cases}$

$\boxed{\text{Poissonization method}}$ ← A general method that facilitates the analysis of distribution testing algorithm by making the numbers of instances of different elements independent.

Sample set: $S = \{ s_1, \ldots, s_m \}$ $X_i :=$ # instances of i in S

- Example $S = \{2, 5, 3, 2, 3\}$ $X_2 = X_3 = 2$, $X_5 = 1$

Main Difficulty:

For a fixed m, X_i's are $\boxed{\text{not}}$ independent

e.g. if $X_4 = \frac{m}{2}$, then $X_3 \leq \frac{m}{2}$.

- Can we make X_i's independent? Yes, via Poissonization...

(2) For $i = 1, \ldots, n$

$X_i \sim Poi(s \cdot p_i)$

independent X_i's

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

For $i = 1, \ldots, \hat{m}$

$s_i \leftarrow$ Draw a sample from p.

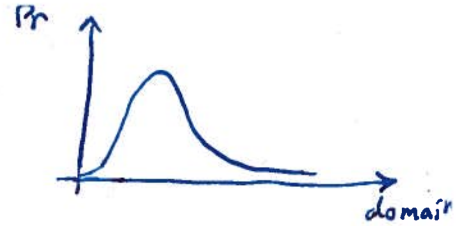
$S = \{ s_1, s_2, \ldots, s_{\hat{m}} \}$

Compute X_i's from S.

=

* These two processes result in the same distribution on X_i's

Recall:



$X \sim Poi(\lambda)$

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

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

Proof of * For any $c = 0, 1, 2, \ldots$

$Pr[X_i = c$ according to (1)]

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

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

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

$$= \frac{e^{-m+m(1-p_i)}(mp_i)^c}{c!} = \frac{e^{mp_i}(mp_i)^c}{c!}$$

Probability of observin $X_i = c$ when $X_i \sim Poi(mp_i)$

$= Pr[X_i = c$ according to (2)]

Done!

$\boxed{L_p \text{ norm}}$

$$\|q\|_p = \left(\sum_i (q(i))^p\right)^{1/p}$$

$\boxed{L_p \text{ distance}}$

$$\|q_1 - q_2\|_p = \left(\sum_i (q_1(i) - q_2(i))^p\right)^{1/p}$$

Reducing the L2 norm of distributions

Goal: transform a distribution, p, to another distributio, p', such that $\|p'\|_2^2$ is low.

we use the following randomized process:

$S \leftarrow$ Draw $Poi(k)$ samples from p

$b_x \leftarrow$ the number of instances of $x \in S$ $\quad \forall x = 1, \dots, n$

For each element x in the domain of, we assign $b_x + 1$ elements in the domain of p' to x

To generate a sample from p':

1) Draw $x \sim p$
2) Pick y uniformly at random from $[b_i + 1]$
3) Output (x, y)

Example: $S = \{3, 3, 1\}$

$P =$

0.2	0.05	0.75
1	2	3

$p' =$

0.1	0.1	0.05	0.25	0.25	0.25
(1,1)	(1,2)	2	(3,1)	(3,2)	(3,3)

Facts about p'

- Domain $= \{(x, y) \mid x \in [n] \wedge y \in [b_x + 1]\}$
- Domain size $= n + k$.
- $p'(x, y) = \dfrac{p(x)}{b_i + 1}$

- $E[\|P'\|_2^2] \le \frac{1}{k}$ (over the randomness of S)

$$E[\|P'\|_2^2] = E\Big[\sum_{x=1}^{n}\sum_{y=1}^{b_x+1} p'(x,y)^2\Big] = E\Big[\sum_x \sum_y \frac{P(x)^2}{(b_x+1)^2}\Big]$$

$$= E\Big[\sum_x \frac{P(x)^2}{b_x+1}\Big] \overset{*}{\le} \sum_x \frac{P(x)^{\cancel{2}}}{k.\cancel{P(x)}} = \frac{1}{k}$$

Proof of * in [DK '16]

Let $Z \sim Poi(\lambda)$, then

$$E[a^Z] = \sum_{Z=0}^{\infty} \frac{e^{-\lambda}(\lambda a)^Z}{Z!} = e^{\lambda(a-1)} \sum_{Z=0}^{\infty} \frac{e^{-\lambda a}(\lambda a)^Z}{Z!} = e^{\lambda(a-1)}$$

$$E\Big[\frac{1}{Z+1}\Big] = E\Big[\int_0^1 a^Z da\Big] = \int_0^1 E[a^Z] da = \int_0^1 e^{\lambda(a-1)} da$$

$$= \frac{1}{\lambda} e^{\lambda(a-1)}\Big|_{a=0}^{1} = \frac{1}{\lambda}(1-e^{-\lambda}) \le \frac{1}{\lambda}$$

Alternative proof of *

$$E\Big[\frac{1}{Z+1}\Big] = \sum_{Z=0}^{\infty} \frac{e^{-\lambda}\lambda^Z}{(Z+1)!} = \frac{1}{\lambda} \times \sum_{Z=0} \frac{e^{-\lambda}\lambda^{Z+1}}{(Z+1)!} = \frac{1}{\lambda}\sum_{Z'=1}^{\infty} \frac{e^{-\lambda}\lambda^{Z'}}{Z'!}$$

$$= \frac{1}{\lambda}\cdot\Big(\sum_{Z'=0}^{\infty} \frac{e^{-\lambda}\lambda^{Z'}}{Z'!}\Big) - \frac{e^{-\lambda}}{\lambda} = \frac{1-e^{-\lambda}}{\lambda}$$

- Let q' be the transformed version of q with samples in S.

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

$$\|P'-q'\|_1 = \sum_{x=1}^{n}\sum_{y=1}^{b_x+1} |p'(x) - q'(x)| = \sum_x \sum_y \frac{|P(x)-q(x)|}{b_x+1}$$

$$= \sum_x |P(x) - P(y)| = \|P-q\|_1$$

- For a known distribution q:

$$b_x = \lfloor n\, q(x) \rfloor$$

Then, we have

$$\|q'\|_2^2 = \sum_{x=1}^{n} \sum_{y=1}^{b_x+1} q'(x)^2 = \sum_{x} \frac{q(x)^2}{\lfloor n q_i \rfloor + 1}$$

$$\leq \sum_{\substack{x \\ q(x)\neq 0}} \frac{q(x)^2}{n\, q(x)} \leq \frac{1}{n}$$

new domain size $\sum_{x=1}^{n} b_x + 1 = n + \sum_{x} \lfloor n q_i \rfloor$

$$\leq n + n \cdot \sum_{x} p(x) \leq 2n$$

$$\boxed{\text{General Framework of [DK'16]}}$$

$p', q' \leftarrow$ Flatten p and q using $\text{Poi}(k)$ from p

Estimate $\|p\|_2^2$ and $\|q\|_2^2$ within constant factor error

if the estimation of $\|p\|_2^2$ and $\|q\|_2^2$ are more than a constant apart from each other:

infer $p \neq q$ and reject.

Else

Test $p' = q'$ Given that $\underbrace{\|p\|_2^2 \;\&\; \|q\|_2^2}$ $\leq \theta(\frac{1}{k})$

Note that $\|p\|_2^2$ is low due to flattening

and $\|q\|_2^2$ is low because it is within a constant factor of $\|p\|_2^2$

L_2 distance estimator.

[Chan, Diakonikolas, Valiant, Valiant '14]

Let p, q be two distribution that we have sample access to

They provide a statistic Z where

$$\Pr\left[\left| Z/m^2 - \|p-q\|_2^2 \right| \geq \varepsilon^2 \right] \leq 2/3$$

- How to compute the estimation of $\|p-q\|_2^2$?

  - Draw Poi(m) sample from p and q
  - Let $\begin{cases} X_i & \text{be \# instances of } i \text{ in the samples from } p. \\ Y_i & \text{" " " " " " " " } q. \end{cases}$
  - $Z = \sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i$
  - output Z/m^2.

- steps of the analysis

b is $\max(\|p\|_2^2, \|q\|_2^2)$

1) $E[Z] = m^2 \|p-q\|_2^2$, $Var[Z] \leq 8m^3 \|p-q\|_4^2 \sqrt{b} + 8m^2 b$

next page.

2) Chebyshev's inequality :

$$\Pr\left[\left|\frac{Z}{m^2} - \|p-q\|_2^2\right| \geq \varepsilon^2\right] = \Pr\left[\left|Z - E[Z]\right| \geq m^2\varepsilon^2\right]$$

$$\leq \frac{Var[Z]}{m^4 \varepsilon^4} \leq \frac{1}{10}$$

$$\Rightarrow \text{for } m \geq \Theta\left(\frac{\sqrt{b}\,\|p-q\|_4^2}{\varepsilon^4} + \frac{\sqrt{b}}{\varepsilon^2}\right)$$

Let's analyze Z: $\begin{cases} X_i \sim Poi(m p_i) \\ Y_i \sim Poi(m q_i) \end{cases}$

$$E[2X_iY_i] \underset{\text{indep.}}{\overset{\text{by}}{=}} 2E[X_i].E[Y_i] = 2(m p_i)\cdot(m q_i)$$

$$E[Z] = \sum_{i=1}^{n} E[(X_i - Y_i)^2 - X_i - Y_i]$$

$$= \sum_{i=1}^{n} \underbrace{E[X_i^2 - X_i]} - \overbrace{2E[X_i].E[Y_i]} + E[\underbrace{Y_i^2 - Y_i}]$$

$$= \sum_{i=1}^{n} m^2 p_i^2 - 2m^2 p_i q_i + m^2 q_i^2$$

$$= \sum_{i=1}^{n} m^2 (p_i - q_i)^2 = m^2 \|p - q\|_2^2$$

$$\boxed{\begin{aligned} &\text{if } X \sim Poi(\lambda) \\ &E[X^2 - X] = E[X^2] - \lambda \\ &= Var[X] + E[X]^2 - \lambda \\ &= \lambda + \lambda^2 - \lambda = \lambda^2 \end{aligned}}$$

So, $E[Z/m^2]$ is exactly the distance that we want!

$$Var[Z] \overset{\text{by independence}}{=} \sum_{i=1}^{n} Var[(X_i - Y_i)^2 - X_i - Y_i]$$

$$= \sum_{i=1}^{n} E[((X_i - Y_i)^2 - X_i - Y_i)^2] - E[(X_i - Y_i)^2 - X_i - Y_i]^2$$

$\vdots$

$= \sum_{i=1}^{n}$ bunch of term lik $E[X_i^\ell]$ where $\ell = 1, 2, 3, 4$

$\vdots$ $\rightarrow$ moments of $Poi(\lambda)$ $\rightarrow$ we can look them up!

$$= \sum_{i=1}^{n} 4m^3 (p_i - q_i)^2 (p_i + q_i) + 2m^2 (p_i + q_i)^2$$

$$\leq \sum_{i=1}^{n} 4m^3 \sqrt{\sum_{i=1}^{n} (p_i - q_i)^4 \sum_i (p_i + q_i)^2} + 2m^2 \underbrace{(p_i + q_i)^2}$$

Cauchy-Schwarz

$$\leq \sum_i 2(p_i^2 + q_i^2)$$

$$\leq \underbrace{4\max(\|P\|_2^2, \|q\|_2^2)}_{b}$$

Let's call this b

$$\leq 8m^3 \|P - q\|_4^2 \sqrt{b} + 8m^2 b$$

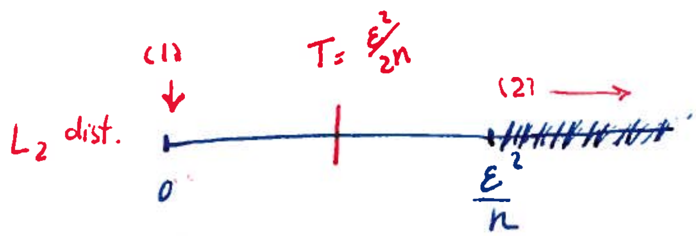
L_1 closeness tester base on $\|p-q\|_2^2$ estimation

How L_1 distance is related to L_2 distance?

$$\|p-q\|_2^2 \underset{\text{Lp norm inequality}}{\leq} \|p-q\|_1^2 \underset{\text{Cauchy-schwarz}}{\leq} n \cdot \|p-q\|_2^2$$

L_1 distance closeness tester distinguishes:

$$\begin{cases} (1)\ \|p-q\|_1 = 0 & \Rightarrow \|p-q\|_2^2 = 0 \\ (2)\ \|p-q\|_1 \geq \varepsilon & \Rightarrow \|p-q\|_2^2 \geq \dfrac{\varepsilon^2}{n} \end{cases}$$



with probability 0.9 for large m

1) $\Rightarrow$ $Z/m^2 \leq T$

2) $\Rightarrow$ $Z/m^2 > T$

(1) $$\Pr\left[\left|\frac{Z}{m^2} - 0\right| \geq \frac{\varepsilon^2}{2n}\right] \leq \frac{n^2 \operatorname{Var}[Z]}{m^4 \varepsilon^4}$$

$$\leq \frac{n^2}{m^4\varepsilon^4} \cdot \left(8m^3 \underbrace{\|p-q\|_4^2}_{=0} \sqrt{b} + 8m^2 b\right)$$

$$\leq \frac{n^2 b}{\varepsilon^4 m^2} \leq \frac{1}{10} \Leftarrow m = \Theta\left(\frac{n\sqrt{b}}{\varepsilon^2}\right)$$

(2) $$\Pr\left[\frac{Z}{m^2} \geq \frac{\varepsilon^2}{2n}\right] \leq \Pr\left[\left|\frac{Z}{m^2} - \|p-q\|_2^2\right| > \frac{\|p-q\|_2^2}{2}\right]$$

$$\leq \frac{4 \operatorname{Var}[Z]}{m^4 \|p-q\|_2^4} \leq \frac{32 \cancel{\|p-q\|_4^2} \sqrt{b}}{m \|p-q\|_2^{\cancel{4}\,2}} + \frac{32\, b}{m^2 \|p-q\|_2^4}$$

$$\leq \frac{32 n \sqrt{b}}{m\, \varepsilon^2} + \frac{32\, b\, n^2}{m^2\, \varepsilon^4} \leq \frac{1}{10} \Leftarrow m = \Theta\left(\frac{n\sqrt{b}}{\varepsilon^2}\right)$$

* Putting everything together ....

by Markov $\Pr\left[\|P\|_2^2 \geq \frac{10}{k}\right] \leq \frac{1}{10}$.

Let say with probability 0.9 we estimate $\|P\|_2^2$ and $\|q\|_2^2$ correctly

with probability 0.9 L1 tester works!

$\Rightarrow$ with probability $0.7 \geq 2/3$ the tester works

(union bound)

Sample complexity

$$\Theta\left(k + \frac{n\sqrt{b}}{\varepsilon^2}\right) = \Theta\left(k + \frac{n}{\sqrt{k}\,\varepsilon^2}\right) \xrightarrow[\text{for } k]{\text{optimize}} \quad k \leq n$$

$$\Rightarrow \boxed{\Theta\left(\frac{n^{2/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2}\right)}$$

optimal sample complexity for testing closeness