Lecture 12

Testing distributions:

the case of uniformity

A new model:

Probability distributions: get samples

$\boxed{p}$

$\Downarrow$

this is all we see $\Big\{$ iid samples

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

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

Examples: lottery data
shopping choices
experimental outcomes
$\vdots$

What do we need to know? is it uniform? e.g. lottery
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

Testing Uniformity

uniform dist on D

goal: if $p = U_D$ then output PASS

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

with prob $\geq 3/4$

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

today's focus

Distances

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

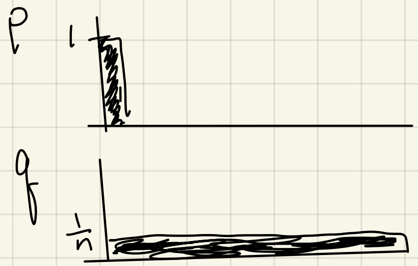
ℓ_2-distance: $\qquad \|p-q\|_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}$

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

examples:

① $p = (1, 0, 0, 0, \ldots, 0)$

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

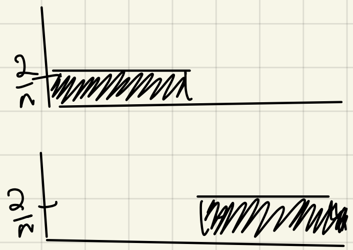


$\|p-q\|_1 = \frac{n-1}{n} + (n-1) \cdot \frac{1}{n} \approx 2$

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

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

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



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

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

small even though p, q very different

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

Naive Analysis: (better analyses exist)

pick m s.t. $\forall x$ $|\hat{p}(x) - p(x)| < \frac{\varepsilon}{n}$ $\Rightarrow$ $\|\hat{p} - p\|_1 < \varepsilon$ ← so if $p = U$, likely to pass

by $\Delta \neq$, if $\|p - \hat{p}\|_1 < \varepsilon$ & $\|\hat{p} - U\|_1 < \varepsilon$ then $\|p - U\|_1 < 2\varepsilon$

so if $\|p - U_n\|_1 > 2\varepsilon$ likely to fail

how big should m be?

maybe need to see each x enough to get $\frac{\varepsilon}{n}$-additive error on $\hat{p}(x)$ $\Rightarrow$ $\Omega(n)$? $\Omega(n/\varepsilon)$? $\Omega(n/\varepsilon^2)$? $\Omega(\frac{n}{\varepsilon^3})$?

Can do better - don't need good approx for all x

Claim $E[\|\hat{p}-p\|_1] \le \sqrt{\frac{n}{m}}$

so pick $m = \Theta(\frac{n}{\varepsilon^2})$ gives $E[\|\hat{p}-p\|_1] \le \frac{\varepsilon}{c}$

$\Downarrow$ by Markov's $\neq$, with prob $1-\frac{1}{c}$

$\|\hat{p}-p\|_1 \le \varepsilon$

Pf of claim

$$E[\|\hat{p}-p\|_1] = \sum_x E[|\hat{p}(x)-p(x)|]$$

$$E[\hat{p}(x)] = \frac{1}{m} E\left[\sum_{i=1}^m 1_{i^{th}\text{ sample is } x}\right] = \frac{1}{m}\sum_{i=1}^m E[1_{i^{th}\text{ sample is } x}] = \frac{m}{m}\cdot p(x) = p(x)$$

$$\le \sum_x \sqrt{E[(\hat{p}(x)-p(x))^2]}$$

Jensens $\neq$

$$= \sum_x \sqrt{Var(\hat{p}(x))}$$

$$\le \sum_x \sqrt{\frac{p(x)}{m}}$$

$$Var(\hat{p}(x)) = \frac{1}{m^2}\cdot m \cdot p(x)\cdot(1-p(x)) \le \frac{p(x)}{m}$$

$$\le \frac{1}{\sqrt{m}}\cdot\sqrt{n}$$

since $\max_p \sum \sqrt{p(x)}$ is $\sqrt{n}$

p prob distributions over domain of size n

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

Let's consider L_2-distance (squared):

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

$$= \sum p_i^2 - \frac{2}{n} \underbrace{\sum p_i}_{=1} + \underbrace{\sum \left(\frac{1}{n}\right)^2}_{=\frac{1}{n}} = \underbrace{\sum p_i^2}_{\text{"collision" probability of } p} - \frac{1}{n}$$

"collision" probability of p: $\|p\|_2^2 = \Pr_{s,t \in p}[s = t] = \sum p_i^2$

$$= \underbrace{\|p\|_2^2}_{\text{we can estimate this}} - \underbrace{\|U_{[n]}\|_2^2}_{\text{we know this since we know } n \text{ (it is } \frac{1}{n})}$$

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

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

Algorithm to estimate:

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

   else fail

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

i.e. what should δ be ?

Assumption $\star$: $|\hat{c} - \|p\|_2^2| < \Delta$

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 $\star$ holds with $\Delta = \frac{\varepsilon^2}{2}$?

- if $p = U_{[n]}$ then $\hat{c} \leq \|U_{[n]}\|_2^2 + \Delta = \frac{1}{n} + \frac{\varepsilon^2}{2}$ so test will PASS

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

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

  $\& \star \Rightarrow \hat{c} > \|p\|_2^2 - \Delta \geq \varepsilon^2 + \frac{1}{n} - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{2} + \frac{1}{n}$ so test will 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

$\Theta(k)$ samples of collisions from k samples of p

Better idea: "recycle" use <u>all</u> 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 $1 \le i < j \le s$

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

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

σ_{ij}'s not independent so can't use Chernoff to argue that $\hat{c}$ is close to $E[\sigma_{ij}]$

Analysis: $E[\hat{C}] = \frac{1}{\binom{s}{2}} \cdot E[\sum_{i<j} \sigma_{ij}] = \frac{\binom{s}{2}}{\binom{s}{2}} E[\sigma_{ij}] = \|p\|_2^2$

$\Pr[|\hat{C} - \|p\|_2^2| > \rho] \leq \frac{Var[\hat{C}]}{\rho^2}$ Chebyshev's ‡

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

Fact: $Var[aX] = a^2 \cdot Var[X]$

So $Var[\hat{C}] = Var[\frac{1}{\binom{s}{2}} \sum_{i<j} \sigma_{ij}]$

$= \frac{1}{\binom{s}{2}^2} Var[\sum_{i<j} \sigma_{ij}]$ } need to bound this but σ_{ij}'s not independent

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

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

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 $E[\Sigma \bar\sigma_{ij}^2]$

why? $\mathrm{Var}[\Sigma \bar\sigma_{ij}] = E[(\Sigma\bar\sigma_{ij} - \underbrace{E[\Sigma \bar\sigma_{ij}]}_{=0})^2]$

$= E[(\Sigma \sigma_{ij} - E[\sigma_{ij}])^2]$

$= \mathrm{Var}[\Sigma \sigma_{ij}]$

So can equivalently bound $\mathrm{Var}[\Sigma \bar\sigma_{ij}]$

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

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

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

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

Facts:

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

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

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

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, $\quad \mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] = \mathrm{Var}\left[\sum_{i<j} \bar{\sigma}_{ij}\right]$

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

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

$\blacksquare$

We have:

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

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

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

where $s = \#$ samples

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

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

$$\leq \frac{\text{const} \cdot \overbrace{\|p\|_2^2}^{\leq 1}}{\underbrace{\varepsilon^4 \cdot s^2}_{\substack{\text{want this} \\ \leq 1 \\ \text{need } s = \Omega(1/\varepsilon^2)}}} + \underbrace{\text{Const} \cdot \frac{1}{\varepsilon^4} \cdot \frac{1}{s}}_{\substack{\text{also want this} \\ \ll 1 \\ \text{so pick } s = \Omega\left(\frac{1}{\varepsilon^4}\right)}} \cdot \underbrace{\|p\|_2^3}_{\leq 1}$$

BIGGER CONSTRAINT

<u>Note</u> can get better bounds

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

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

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

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