- Closeness Testing (p, q unknown)
- Learning & Testing monotone distributions

Some other extensions:

What if p, q both unknown? "Closeness testing"

L_2 distance is similar, but what does it say?

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

$$= \underbrace{p_i^2}_{\substack{\|p\|_2^2 \\ \text{self-collision prob of } p}} - 2 \underbrace{\sum p_i q_i}_{\substack{\text{cross-collision probability} \\ \text{of } p+q}} + \underbrace{q_i^2}_{\substack{\|q\|_2^2 \\ \text{self-collision prob of } q}}$$

- 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 $\underbrace{\text{prob} \geq b}_{\leq \frac{1}{b} \text{ of these}}$

One way: Filtering algorithm:

learn B = domain elements with prob $\geq b$ ← $O(\frac{1}{b})$ samples

filter rest of samples

B: naive method $O(\frac{1}{\varepsilon^2} \cdot \frac{1}{b})$ samples

$D \setminus B$: collisions

Note strange dependence on n!

$n^{2/3}$ is tight!! ⟹ Turns out $O\left(\frac{1}{\varepsilon^4} n^{2/3}\right)$ samples suffice [recent improvements on ε known]

Sketch of l.b. for p, q given by samples $\Leftarrow$ "closeness testing"

Thm Closeness testing requires $\Omega(n^{2/3})$ samples

Proof idea:

$p_0 =$

$n^{2/3}$ heavy elements weight $\frac{1}{2n^{2/3}}$

$\frac{n}{4}$ light elements weight $\frac{2}{n}$

light elements are disjoint

q_0

$n^{2/3}$ heavy elements

$\frac{n}{4}$ light elements

Positive pairs

ℓ_1 dist $= 0 \Rightarrow (\pi(p_0), \pi(p_0)) \; \forall \pi$

Negative pairs

$(\pi(p_0), \pi(q_0)) \; \forall \pi \Leftarrow \ell_1$ dist $= 1$

where $\pi(p)$ relabels domain elts randomly

$\pi(p_0), \pi(p_0)$ applies same relabeling to both

Main idea: $\boxed{\text{Only Collision Statistics matter!}}$

for positive pairs have collisions in both heavy + light elts

for negative pairs have collisions only in heavy elts

when see a collision, usually can't tell if it was a heavy or light element!

After $o(n^{2/3})$ samples:

probability see any small element twice really small ⇐ happens, but not too often

probability see any heavy element 3X is small ⇐

probability see any small elt 3X is tiny ⇐ unlikely to happen

heavy " 4X is tiny ⇐

So, what collision statistics could we have?

how many elts in domain appear n_p times, n_q times in p, q?

p	0	0	1	0	2	1	0	3	1	2	4	0	3	1	2
q	0	1	0	2	0	1	3	0	2	1	0	4	1	3	2

#domain elts

will happen less in pos pairs than in neg pairs?

will happen more in pos pairs than in neg pairs

only heavy elements - same distribution for pos + neg pairs

·unlikely - can ignore

when you see collision, you don't know if it came from heavy or light element

m = # samples

H = # heavy collisions ← same distribution for pos + neg pairs

L = # light collisions (1 from each dist) ⇐ = 0 when neg pair

$E[\#\text{collisions in pos pair}] = E[H] + E[L] = \frac{m^2}{2n^{2/3}} + \frac{m^2}{n} \approx \frac{m^2}{2n^{2/3}}$

$E[\#\text{collisions in neg pair}] = E[H] = \frac{m^2}{2n^{2/3}}$

Need to show something a bit stronger - can't distinguish the random variables!

$E[H] = \frac{m^2}{4n^{2/3}}$ $\qquad$ $\binom{m}{2}$ pairs, each collides with prob $\frac{1}{2n^{2/3}}$

$\mathrm{Var}[H] \approx \frac{m^2}{n^{2/3}}$

$E[L], \mathrm{Var}[L] \approx \frac{m^2}{n}$ $\qquad$ $\binom{m}{2}$ pairs, each collides with prob $\frac{2}{n}$

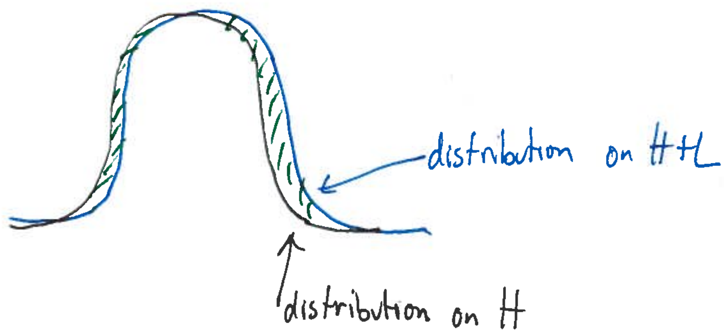
L_1 distance small

$\Downarrow$

almost same distribution

$\Downarrow$

hard to distinguish!



how do we show L_1 dist is small?

if they were Gaussian,

could show that $\sqrt{\mathrm{Var}(H)} \leq E[L]$

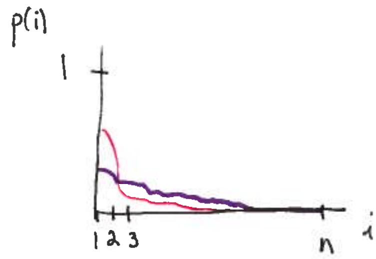
they aren't quite, so it's more difficult.

$$\Leftrightarrow \frac{m}{n^{1/3}} \leq \frac{m^2}{n}$$

$$\Leftrightarrow m \geq n^{2/3}$$

Testing & Learning Monotone Distributions (over totally ordered domain)

<u>Def.</u> p over $[n]$ is "monotone decreasing"
if $\forall\, i \in [n-1]$ $\quad p(i) \geq p(i+1)$



Monotonicity Tester:

- if p monotone increasing, Pass with prob $\geq 3/4$
- if p ε-far in L_1 dist from mon increasing, Fail with prob $\geq 3/4$

Useful tool: "Birge Decomposition"

(note: this is a different decomposition than in homework
in particular, it is <u>oblivious</u>!)

decompose domain $1..n$ into $\ell = \Theta\left(\frac{\log \varepsilon n}{\varepsilon}\right) \approx \Theta\left(\frac{\log n}{\varepsilon}\right)$ intervals

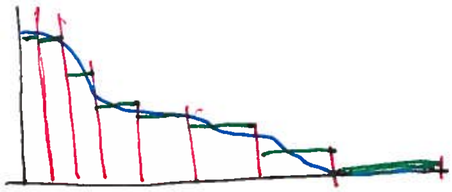
$$I_1^\varepsilon, I_2^\varepsilon, \ldots, I_\ell^\varepsilon \quad \text{s.t.}$$

$$|I_{k+1}^\varepsilon| = \lceil (1+\varepsilon/2) \cdot |I_k^\varepsilon| \rceil$$

will drop ε in notation once it is fixed

so $|I_1^\varepsilon| = 1$ $\quad$ but then at some point the sizes grow exponentially
$|I_2^\varepsilon| = 2$
$|I_3^\varepsilon| = 3$

define "flattened distribution"



$\forall\ 1 \le j \le \ell$

$\forall\ i \in I_j$ $\quad \tilde{q}_\varepsilon(i) = \dfrac{q(I_j)}{|I_j|}$

← assign all elements in same interval the same probability

note: $q(I_j) = \tilde{q}_\varepsilon(I_j)$

<u>Thm</u> if q mon decreasing then $\|\tilde{q}_\varepsilon - q\|_1 < \varepsilon$

<u>Corrl</u> if q ε-close to mon decreasing then $\|\tilde{q}_\varepsilon - q\|_1 < O(\varepsilon)$

Testing Algorithm:

Take samples of q

do uniformity test for each partition (using samples that fell in it) fail if any partition fai[ls]
(if not enough samples then pass)

$w_j \leftarrow$ # samples that fell in partition j

use LP to verify w close to monotone

↑ note this is LP on $O(\log n)$ vars

How many samples?

for each partition with enough weight, say $\frac{\varepsilon}{\log n}$, need $\frac{\sqrt{n}}{\varepsilon^2}$ samples

↑ need $\frac{\sqrt{n} \cdot \log n}{\varepsilon^3}$ for each o[f]

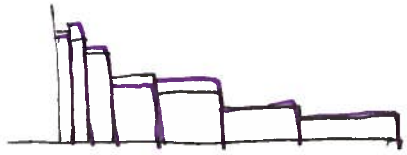
need another $\log\log n$ for union bound

$\approx O(\sqrt{n}\ \text{polylog}\ n \cdot \text{poly}\ \frac{1}{\varepsilon})$

(note: this can be improved!!)

Last step:

difficulty



purple is not monotone but is close

good thing: only $\frac{\log n}{\sqrt{\varepsilon}}$ variables!

Can be solved via brute force
LP (actually quite efficient)
⋮

Slightly changing perspective...

What if we know dist q is monotone, can we learn it?

Yes! use sampling to estimate $\tilde{q}_\varepsilon(I_j)$'s

Birge's Thm Can learn monotone distributions to w/in $\le\varepsilon$ L_1 error in $\Theta(\frac{1}{\varepsilon^3}\log n)$ samples.