• Closeness Testing \((p, q \text{ unknown})\)

• 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:

$$\|p - q\|_2^2 = \sum_x (p_x - q_x)^2$$

$$= p_x^2 - 2p_x q_x + q_x^2$$

Cross-collision probability

Self-collision probability of $p$

Self-collision probability of $q$

Can bound variance of $\|p\|_2^2$, $\|q\|_2^2$, and $\|p - q\|_2^2$ if $\max_p \Pr \text{ element is bounded by } b$

What about other case?

Use naive method on elements whose $\Pr \geq \frac{1}{b}$ of these

One way: Filtering algorithm:

Learn $B$ = domain elements with $\Pr \geq b \leq O\left(\frac{1}{b}\right)$

Filter rest of samples

$$B \leftarrow N \setminus B$$

Naive method

Collisions $O\left(\frac{1}{b} \cdot n^2\right)$ samples

Note: Strange dependence on $n^{2/3}$ is tight. Turns out $O\left(\frac{1}{\varepsilon^4} n^{2/3}\right)$ samples suffice on $\varepsilon$. Recent improvements known.
Sketch of lb. for \( p,q \) given by samples \( \leq \) "closeness testing"

Thin closeness testing requires \( \Omega(n^{3/8}) \) samples

Proof idea:

\[
\begin{align*}
p_0 & = \\
n^{3/8} \text{ heavy elements} & \quad \frac{n}{n} \text{ light elements} \\
\text{weight} \frac{1}{2n^{1/3}} & \quad \text{weight} \frac{2}{n} \\

q_0 & = \\
n^{3/8} \text{ heavy elements} & \quad \frac{n}{n} \text{ light elements}
\end{align*}
\]

Positive pairs \( \iff \) \( (\Pi(p_0), \Pi(q_0)) \neq \Pi \)

Negative pairs \( \iff \) \( (\Pi(p_0), \Pi(q_0)) \neq \Pi \)

\( \iff \dist = 1 \)

where \( \Pi(p) \) relabels domain els. randomly

\( \Pi(p_0), \Pi(p) \) applies same relabeling to both

Main idea: for positive pairs have collisions in both heavy + light els

for negative pairs have collisions only in heavy els

when see a collision, usually can't tell if it was a heavy or light element!
After $o(n^{2/3})$ samples:

- probably see any small element twice really small
- probability see any heavy element $3X$ is small happens, but not too often
- probability see any small elt $3X$ is tiny heavy $4X$ is tiny unlikely to happen

So, what collision statistics could we have?

How many els in domain appear $p$ times, $q$ times in $p,q$?

\[
\begin{array}{cccccccccccc}
\text{#domain} & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 2 & 1 & 1 & 3 & 0 & 1 & 2 & 1 & 4 & 0 & 3 & 1 & 3 & 2
\end{array}
\]

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

\[
m = \# \text{ samples}
\]
\[
H = \# \text{ heavy collisions}
\]
\[
L = \# \text{ light collisions (1 from each dist)}
\]

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

\[
E[\# \text{ collisions in neg pair}] = E[H] = \frac{m^2}{2n^{2/3}}
\]
Need to show something a bit stronger - can’t distinguish the random variables!

\[ E[H] = \frac{m^2}{4n^{3/4}} \]

\( \text{Var}[H] \propto \frac{m^2}{n^{3/4}} \)

\[ E[L], \text{Var}[L] \sim \frac{m^2}{n} \]

\( \binom{m}{2} \text{ pairs, each collides with prob } \frac{1}{2n^{3/4}} \)

\( \binom{m}{3} \text{ pairs, each collides with prob } \frac{1}{n} \)

\( L_1 \) distance small

almost same distribution

hard to distinguish!

how do we show \( L_1 \) dist is small?

if they were Gaussian, could show that \( \sqrt{\text{Var}(H)} \leq E[L] \)

\[ \iff \frac{m}{n^{3/4}} \leq \frac{m^2}{n} \]

\[ \iff m \geq n^{3/4} \]

\( \iff \) they aren’t quite, so it’s more difficult.
Testing & Learning Monotone Distributions (over totally ordered domain)

Def: \( p \) over \([n]\) is "monotone decreasing" if \( \forall i \in [n-1] \), \( p(i) \geq p(i+1) \).

Monotonicity Tester:
- if \( p \) monotone increasing, Pass with prob \( \geq \frac{3}{4} \).
- if \( p \) $\epsilon$-far in \( L_1 \) dist from non-increasing, Fail with prob \( \geq \frac{3}{4} \).

Useful tool: "Birge Decomposition"

(note: this is a different decomposition than in homework.

In particular, it is oblivious!)

Decompose domain \([1..n]\) into \( I = \Theta \left( \frac{\log n}{\epsilon} \right) \times \Theta \left( \frac{\log n}{\epsilon} \right) \) intervals

\[
I_1^\epsilon, I_2^\epsilon, \ldots, I_k^\epsilon \quad \text{s.t.} \quad |I_k^\epsilon| = \Gamma \left( \left( 1 + \frac{\epsilon}{2} \right)^{-1} \right) \cdot |I_k^\epsilon| \]

So \( |I_1^\epsilon| = 1 \) but then at some point the size grow exponentially.

\( |I_2^\epsilon| = 2 \)

\( |I_3^\epsilon| = 3 \)
Define "flattened distribution"

\[ q_e(i) = \frac{q(I_j)}{|I_j|} \]

Note: \( q(I_j) = \hat{q}_e(I_j) \)

If \( q \) monotonically decreasing then \( \|q_e - q\|_1 \leq \varepsilon \)

Corollary: If \( q \) \( \varepsilon \)-close to \( q \) monotonically decreasing then \( \|q_e - q\|_1 \leq O(\varepsilon) \)

Testing Algorithm:

Take samples of \( q \)
do uniformity test for each partition (using samples that fell in it)
(if not enough samples then pass)

\( w \) are samples that fell in partition \( j \)
use LP to verify \( w \) close to monotone

\( \text{Note: This is LP on} \ O(\log n) \text{ vars} \)

How many samples?

For each partition with enough weight, say \( \frac{\varepsilon}{\log n} \), need \( \frac{\sqrt{n}}{\varepsilon^2} \) samples

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

(Note: This can be improved!!)
Last step:

difficulty

purple is not monotone, but is close

good thing: only \( \frac{\log n}{\varepsilon} \) 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 \( \hat{q}_x(I_j) \)'s

Birge's Thm: Can learn monotone distributions to within \( \varepsilon \) error in \( \Theta(\frac{1}{\varepsilon^3 \log n}) \) samples.