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. $|D| = n$

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

That is all we see

iid samples

Distances

$\ell_1$-distance: $\|p - q\|_1 = \sum_{i=0}^{\infty} |p_i - q_i|$

$\ell_2$-distance: $\|p - q\|_2 = \sqrt{\sum_{i=0}^{\infty} (p_i - q_i)^2}$

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

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

Testing Uniformity

Goal: if \( p = U_D \) then output PASS

if \( \text{dist} (p, U_D) > \varepsilon \) then output FAIL

\( \ell_1 + \ell_2 \) distance measures
Generalizations: Given another distribution $q$, is $p = q$ or is $p$ "far" from $q$?

$q = \text{uniform } 0(\sqrt{n})$ for all $q$.

1. "Identity Testing"
   - $q$ is known to algorithm, no samples of $q$ needed.

2. "Closeness Testing"
   - $q$ is given via samples.

Tolerant version:

$||p-q||_1 \leq \epsilon$ for $\epsilon \leq 3$.

$||p-q||_\infty \geq \epsilon'$.

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 \( S_1, \ldots, S_m = S \)
- Let \( X_i = \# \) times \( i \) occurred in sample

**problem:**
- \( X_i \)’s not independent
- \( i = 1, \ldots, m \)
- \( X_1 = 1 \)
- \( X_i < \frac{m}{2} \)

Can we make the \( X_i \)’s independent? **Poissonization**

**new algorithm:**
- \( \hat{m} \leftarrow \text{Poi}(m) \)
- Take \( \hat{m} \) samples to get \( \hat{S} \)
- Let \( X_i = \# \) times \( i \) occurred in \( \hat{S} \)

**Poi(\( \lambda \))**
- \( \text{Pr}(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \)
- \( E[X] = Var(X) = \lambda \)

Equivalent:
- For each \( i \in [n] \)
- \( X_i \leftarrow \text{Poi}(m \cdot p_i) \)
- Add \( X_i \) copies of \( i \) to the sample
- Randomly permute the sample
Why equivalent?

\[ \text{Pr}(X_i = C \text{ according to } (1)) = \sum_{k=C}^{\infty} \text{Pr}[\hat{m} = k] \cdot (k)_{p_x}^c (1-p_x)^{k-c} \]

\[ = \sum_{k=C}^{\infty} \frac{e^{-m} m^k}{k!} \cdot \frac{k^c}{c!(k-c)!} \cdot (1-p_x)^{k-c} \]

\[ = \frac{e^{-m} m^c}{c!} \cdot \sum_{k=C}^{\infty} \frac{m^{k-c} (1-p_x)^{k-c}}{(k-c)!} \]

\[ = \frac{m^c (1-p_x)^C}{c!} \cdot \sum_{k'=0}^{\infty} \frac{m^{k'} (1-p_x)^{k'}}{(k')!} \]

\[ = e^{-mp_x} (mp_x)^c \cdot \frac{e^{m(1-p_x)}}{c!} = \text{Pr}[X_i = C] = \text{Pr}[\hat{m} = C|X_i = C] \]

Need to check joint distributions.
Another difficulty: \( \|p\|_2 \) can be large

e.g. uniformity test statistic

\[
\text{Var} \left( \hat{C} \right) = 0 \left( \frac{\|p\|_2^2}{s^2} + \frac{\|q\|_2^3}{s} \right)
\]

Goal: transform distributions \( p, q \) into \( p', q' \) such that \( \|p\|_2 + \|q\|_2 \) small

```
\|p - q\|_2 > \varepsilon \Rightarrow \|p' - q'\|_1 > \varepsilon
```

"Reduction" to small \( L_2 \)-norm case will work when \( q \) known & when given via samples
Transformation of $p$:

$m = \# \text{ samples } $ \text{"expected" by original alg}$

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

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

$\forall \ i \ \text{ add } b_i+1 \text{ elements to new domain}$

$(i, j) \ \text{ where } j \in [b_i+1]$

New distribution $p'$:

pick $i \in \text{R}$

pick $j \in \text{R} \ [b_i+1]$

output $(i, j)$

$\mathcal{P}'(ij) = \frac{p(i)}{b_i+1}$

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 $p'$:

$\sum (i, j)$

$\begin{array}{cccc}
(1,1) & (2,1) & (2,2) & (2,3) \\
(3,1) & (3,2) & (3,3) \\
(4,1) & \\
(5,1) & (5,2)
\end{array}$

$\text{Prob}$

$\begin{array}{cccc}
p(1) & \frac{p(2)}{3} & \frac{p(3)}{3} & \frac{p(3)}{3} \\
\frac{p(2)}{3} & \frac{p(3)}{3} & \frac{p(3)}{3} \\
\frac{p(3)}{3} & \frac{p(3)}{3} & \frac{p(5)}{2}
\end{array}$
Problem 2: we don't know if $q$'s $l_2$ norm gets small.

Claim: \[ E[\|p\|_2^2] \leq \frac{1}{m} \]

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

Claim for $Z \sim \text{Poi}(\lambda)$, \[ E\left[\frac{1}{Z+1}\right] \leq \frac{1}{\lambda} \]
\[
E\left[\frac{1}{Z+1}\right] = \sum_{Z=0}^{\infty} \frac{e^{-\lambda} \lambda^Z}{(Z+1)Z!} = \frac{1}{\lambda} \sum_{Z=0}^{\infty} \frac{e^{-\lambda} \lambda^{Z+1}}{(Z+1)} = \frac{1}{\lambda} \sum_{Z=1}^{\infty} \frac{e^{-\lambda} \lambda^Z}{Z!}
\]
\[
\leq \frac{1}{\lambda}
\]

$X \sim \text{Poi}(\lambda)$.
\[
\Pr(X=K) = \frac{e^{-\lambda} \lambda^K}{K!}
\]
$E[X] = \text{Var}[X] = \lambda$
After transform $p \rightarrow p'$, using same $S$:

\[ \tilde{y} = \sum_{i=1}^{b+1} \frac{p(x_i)}{b+1} \]

Claim:

\[ \mathbb{E}\left|\tilde{\Delta}_M\right| = \frac{1}{m} \]
$L_2$ distance estimation between two distributions $p$, $q$:

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

**Theorem**

Given samples of $p$, $q$, distributions on $[n]$, s.t. $b = \max \{\|p\|_2^2, \|q\|_2^2\}$, can distinguish $p=q$ from $\|p-q\|_1 > \varepsilon$ in $O(bn/\varepsilon^2)$ samples.

**Corollary**

If $b = \min \{\|p\|_2^2, \|q\|_2^2\}$, can distinguish $p=q$ from $\|p-q\|_1 > \varepsilon$ in $O(bn/\varepsilon^2)$ samples.

**Proof Idea:**

1. Estimate $\|p\|_2^2 + \|q\|_2^2$ to mutl factor of $c$ (can do this in $O(\sqrt{n})$ samples)

2. If differ by $> c$ mutl factor, infer $p \neq q$ and reject

3. Else use Thm * with $b' = c \cdot b$
Testing Closeness

1. Let $k = \frac{n^3}{\varepsilon}$
2. $S \leftarrow$ multiset of $\text{Po}(k)$ samples from $q$
3. Run tester of $\text{Corr}$ on $p, q'$ w.r.t. $S$

Why does it work?

- Distinguishing $p+q$ and $p', q'$ are equivalent
- How many samples needed?

- Why $|S| = \Theta(k)$
- $\mathbb{E}[\|q'\|^2] = O(\frac{1}{k})$ so w.h.p. $\|q'\|_2 = O(\frac{1}{\sqrt{k}})$
- $O\left(k + \frac{1}{\sqrt{k}} - n \cdot \frac{1}{\varepsilon^2}\right) = O\left(n^{2/3} \varepsilon^{1/3} + \frac{1}{n^{1/3}} \cdot \varepsilon^{-1/3} \cdot n \cdot \frac{1}{\varepsilon^2}\right)$

Run tester on $p, q'$

$= O\left(n^{2/3} \varepsilon^{1/3}\right)$