Today: Testing properties of distributions

- Finish uniformity
- Survey of other properties
Testing Uniformity
The model:

- $[n] = \{1, \ldots, n\}$
- $p$ is a black-box distribution over $[n]$, generates iid samples.
- $p_i = \text{Prob}[p \text{ outputs } i]$
Testing uniformity

- The goal:
  - pass uniform distribution
  - fail distributions that are $\varepsilon$-far from uniform
    - what measure of distance?
      - $\ell_1$ distance: $||p-U||_1 = \sum|p_i - 1/n| > \varepsilon$
      - $\ell_2$ distance (squared): $||p-U||_2^2 = \sum(p_i - 1/n)^2 > \varepsilon^2$
  - Interested in sample complexity in terms of $\varepsilon, n$
Focus on $\ell_2$ for a minute....

- Is it easier?

- Known relationship between $\ell_1$ and $\ell_2$:
  $$|| p-q ||_2 < ||p-q||_1 < n^{1/2} || p - q ||_2$$

- So if $p=q$, $|| p-q ||_2 = ||p-q||_1 = 0$

  and if $||p-q||_1 > \varepsilon$ then $|| p-q ||_2 > \varepsilon / n^{1/2}$

How many samples do you need to determine this???
Estimating $\ell_2$ distance via collisions

- Collision probability of $p$:
  \[ \Pr_{s,t \in p} [s=t] = \sum_{a \in [n]} p(a)^2 = (\|p\|_2)^2 \]

- Observation: (last time)
  \[ (\|p-u\|_2)^2 = \|p\|_2^2 - 1/n \]

So the collision probability of $p$ yields $p$’s distance from uniform!
How well do we need to estimate the collision probability?

- Achieving additive error $\varepsilon$ for approximating $l_2$: getting additive $l_2^2$ error $\varepsilon^2/2$ for collision probability sufficient
How well do we need to estimate the collision probability?

- Distinguishing $p=U$ from $\|p-U\|_1 > \varepsilon$:
  
  Getting multiplicative error $\varepsilon^2/3$ on collision probability estimate is sufficient

What is the sample complexity required for this?
Collision Probability Estimate

- Take sample of size $s$ from $p$: $x_1, \ldots, x_s$
- For each $1 \leq i < j \leq s$, set $\sigma_{i,j}$ to 1 if $x_i = x_j$ and 0 otherwise
- Output $A \equiv (\sum_{i<j} \sigma_{i,j}) / \binom{s}{2}$
Chebyshev's Inequality

- \( \Pr[|A - E[A]| > \rho] \leq \frac{\text{Var}[A]}{\rho^2} \)
- Here \( A \equiv (\sum_{i<j} \sigma_{i,j})/(s \text{ choose } 2) \)
  So \( E[A] = \|p\|_2^2 \)

- What will we use for \( \rho \)?
  - \( \varepsilon ? \) Gives additive error \( \varepsilon \)
  - how about \( \varepsilon \|p\|_2^2 ? \) Gives multiplicative error
    i.e., \( \|p\|_2^2 (1 - \varepsilon) \leq \text{output} \leq \|p\|_2^2 (1 + \varepsilon) \)
Giving us...

\[
\Pr[|A - \|p\|_2^2| > \varepsilon] \leq \frac{\text{Var}[A]}{\varepsilon^2}
\]
\[
\Pr[|A - \|p\|_2^2| > \varepsilon\|p\|_2^2] \leq \frac{\text{Var}[A]}{\varepsilon^2\|p\|_2^4}
\]

(Last time: using \(\varepsilon = \varepsilon_0^{2/3}\) gives good \(\ell_1\) test for uniformity)

But are these probabilities small?

we need to figure out how many samples needed to bound \(\text{Var}[A]\)
See blackboard calculations:

$$\text{Var } [A] / \varepsilon^2 \|p\|_2^4 \ll 1 \text{ if } \Omega(n^{1/2}/\varepsilon^2) \text{ samples}$$

Note: Can get better bound on variance if have bound on max probability element

(e.g. if max probability is $O(1/ n^{1/2})$)
\( \ell_2 \) distance estimation:

- **Theorem:** There is an algorithm which on input distribution \( p \) satisfies:
  - if \( \|p-U\|_2^2 < \epsilon /2 \) output Pass (whp)
  - if \( \|p-U\|_2^2 > \epsilon \) output Fail (whp)
  - sample, time complexity \( O(\epsilon^{-4}) \)

- No dependence on \( n \) !?!
Multiplicative estimate of $\ell_2$ distance

Theorem: Can estimate $\|p\|_2^2$ to within multiplicative factor of $1 + \varepsilon$ using $O(n^{1/2} / \varepsilon^2)$ samples (error probability $< \frac{1}{4}$)
\textbf{$l_1$ distance test:}

- **Theorem:** ([Goldreich Ron][Batu Fortnow R. Smith White] [Paninski]) Sample complexity of distinguishing $p=U$ from $|p-U|_1>\epsilon$ is $\theta(n^{1/2})$
Survey of what is known on testing properties of distributions
We already saw...

- Estimating support size
- Testing uniformity

- What else?
Testing Independence:

Shopping patterns:

Independent of zip code?
Testing closeness of two distributions:

Transactions of 20-30 yr olds

Transactions of 30-40 yr olds

Trend change?
Outbreak of diseases

- Similar patterns?
- Correlated with income level?
- More prevalent near large airports?
Information in neural spike trails

[Strong, Koberle, de Ruyter van Steveninck, Bialek ’98]

- Each application of stimuli gives sample of signal (spike trail)
- Entropy of (discretized) signal indicates which neurons respond to stimuli
Compressibility of data
Worm detection

- find "heavy hitters" – nodes that send to many distinct addresses
Some properties

- Similarities of distributions:
  - Testing uniformity
  - Testing identity
  - Testing closeness

- Entropy estimation

- Support size

- Independence properties

- Monotonicity
Similarities of distributions

- Are $p$ and $q$ close or far?
  - $q$ is known to the tester
    - $q$ is uniform
  - $q$ is given via samples
Is $p$ uniform?

**Theorem:** ([Goldreich Ron][Batu Fortnow R. Smith White][Paninski]) Sample complexity of distinguishing $p=U$ from $|p-U|_1 > \varepsilon$ is $\Theta(n^{1/2})$.

Nearly same complexity to test if $p$ is any *known* distribution [Batu Fischer Fortnow Kumar R. White]: “Testing identity”
Testing closeness

Theorem: ([BFRSW] [P. Valiant])
Sample complexity of distinguishing $p=q$ from $|p-q|_1 > \varepsilon$ is $\tilde{\Theta}(n^{2/3})$
Approximating the distance between two distributions?

Distinguishing whether $|p-q|_1 < \varepsilon$ or $|p - q|_1$ is $\Theta(1)$ requires nearly linear samples [P. Valiant 08]
Can we approximate the entropy? [Batu Dasgupta R. Kumar]

- In general, not to within a multiplicative factor...
  - \( \approx 0 \) entropy distributions are hard to distinguish (even in superlinear time)

- What if entropy is big (i.e. \( \Omega(\log n) \))?
  - Can \( \gamma \)-multiplicatively approximate the entropy with \( \tilde{O}(n^{1/\gamma^2}) \) samples (when entropy \( > 2\gamma/\epsilon \))
  - requires \( \Omega(n^{1/\gamma^2}) \) [Valiant]
  - better bounds in terms of support size [Brautbar Samorodnitsky]
Estimating Compressibility of Data

[Raskhodnikova Ron Rubinfeld Smith]

- General question undecidable
- Run-length encoding
- Huffman coding
  - Entropy
- Lempel-Ziv
  - "Color number" = Number of elements with probability at least $1/n$
  - Can weakly approximate in sublinear time
  - Requires nearly linear samples to approximate well [Raskhodnikova Ron Shpilka Smith]
Testing Independence:

Shopping patterns:

Independent of zip code?
Independence of pairs

- $p$ is joint distribution on pairs $<a,b>$ from $[n] \times [m]$ (wlog $n \geq m$)

- Marginal distributions $p_1, p_2$

- $p$ independent if $p = p_1 \times p_2$, that is $p_{(a,b)} = (p_1)_a \cdot (p_2)_b$ for all $a, b$
Theorem: [Batu Fischer Fortnow Kumar R. White]

There exists an algorithm for testing independence with sample complexity $O(n^{2/3}m^{1/3}\text{poly}(\log n, \varepsilon^{-1}))$ s.t.

- If $p = p_1 \times p_2$, it outputs PASS
- If $||p - q||_1 > \varepsilon$ for any independent $q$, it outputs FAIL
An open question:

- What is the complexity of testing independence of distributions over $k$-tuples from $[n_1] \times \ldots \times [n_k]$?

- Easy $\Omega(\prod n_i^{1/2})$ lower bound
Testing the monotonicity of distributions:

Does the occurrence of cancer decrease with distance from the nuclear reactor?
Monotone distributions

- $p$ is monotone if $i < j$ implies $p_i \leq p_j$
- Many distributions are monotone or are “made of” small number of monotone distributions
Lemma: Testing monotonicity requires $\Omega(\sqrt{n})$ samples

Lower bound [Batu Kumar R.]
Other properties?

- $K$-flat distributions
- Mixtures of $k$ Gaussians
- “Junta”-distributions
- Generated by a small Markovian process
- ...


Getting past the lower bounds

- Special distributions
  - e.g., uniform on a subset, monotone
- Other query models
  - Queries to probabilities of elements
- Other distance measures
Flat distributions

Entropy can be estimated somewhat faster when distribution is uniform on a subset of the elements [Batu Dasgupta Kumar R.][Brautbar Samorodnitsky]
Monotone distributions over totally ordered domains

- Test uniformity with $O(1)$ samples [Batu Kumar R.]
- Other tasks doable with polylogarithmic samples: [Batu Dasgupta Kumar R.][BKR]

  **Examples:**
  - Testing closeness
  - Testing independence
  - Estimating entropy
Other query models:

- Distribution given explicitly [BDKR]
- Distribution given both by samples and oracle for $p_i$’s [BDKR][RS]
  - Can estimate entropy in polylog(n) time
Other distance measures:

- **Earth Mover Distance** [Doba Nguyen² R.]
  - Measures min weight matching to some distribution with the property
  - Can estimate distance between distributions, independence over $[0,1]^N$, in time independent of domain size
  - Still exponential in $N$
    - Can improve over highly clusterable distributions
Conclusions and Future Directions

- Distribution property testing problems are everywhere
- Several useful techniques known
- Other properties for which sublinear tests exist?
- Special classes of distributions?
- Time vs. query complexity
- Other query models?
- Non-iid samples?