Lecture 4: Testing Properties of Distributions II

Lecturer: Ronitt Rubinfeld
Spring 2013
6.896: Sub-linear Algorithms
“Last” time: Testing Uniformity
Outbreak of diseases

- Similar patterns?
- Correlated with income level?
- More prevalent near large airports?
Worm detection

- find ``heavy hitters” – nodes that send to many distinct addresses
Today: Testing properties of distributions (part II)

- Testing closeness to a known distribution
- Survey of other properties
Recall 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?
    - $l_1$ distance: $||p-U||_1 = \sum |p_i - 1/n| > \varepsilon$
    - $l_2$ distance (squared): $||p-U||_2^2 = \sum (p_i - 1/n)^2 > \varepsilon^2$

Interested in sample complexity in terms of $\varepsilon, n$
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 identity via testing uniformity on subdomains:

- *(Relabel domain so that $q$ monotone)*
- Partition domain into $O(\log n)$ groups, so that each group almost “flat” --
  - differ by $<(1+\varepsilon)$ multiplicative factor
  - $q$ close to uniform over each group
- Test:
  - Test that $p$ close to uniform over each group
  - Test that $p$ assigns approximately correct total weights to each group
Bucketing

- How?
  - $R_0 = \{ j \mid q(j) < 1/(n \log n) \}$
  - $R_i = \{ j \mid (1 + \epsilon)^{i-1}/(n \log n) \leq q(j) \leq (1 + \epsilon)^i/(n \log n) \}$
  - Observe: $i < (2/ \log(1 + \epsilon)) \log n$ buckets enough

- A good approximation
  - Let $Z$ be the following distribution:
    - Pick bucket $i$ with probability $\sum_{j \in R_i} q(j)$
    - Pick an element of bucket $i$ uniformly
  - To show: $Z, q$ are $\epsilon$-close in $L_1$ distance and $\epsilon^2/n$-close in $L_2^2$ distance

- Total probability $O(1/\log n)$
- All probabilities within $(1+\epsilon)$ factor of each other
Notation:

- Let $q_i$ be $q$ conditioned on falling in bucket $i$
- Let $U_i$ be uniform distribution on bucket $i$
- Let $l$ be the number of elements in bucket $i$
Single Bucket: $q_i$ close to uniform

- **Bucket Lemma**: for any bucket $i > 0$, $U_i$ and $q_i$ are $\epsilon$-close according to $L_1$ distance and $\frac{\epsilon^2}{l}$-close according to $L_2^2$
  - For bucket $i$, let $x_1 \leq \cdots \leq x_l$ be the conditional probabilities corresponding to elements that fell in the bucket.
  - Averaging implies $x_1 \leq \frac{1}{l} \leq x_l$ and so $x_l \leq (1 + \epsilon)x_1 \leq \frac{1 + \epsilon}{l}$ and $x_1 \geq \frac{1}{l(1+\epsilon)} > \frac{1 - \epsilon}{l}$.
  - So $|x_j - \frac{1}{l}| \leq \frac{\epsilon}{l}$ and so total $L_1$ distance from is $U_i$ is $\Sigma |x_j - \frac{1}{l}| \leq \epsilon$ and $L_2^2$ at most $\Sigma (x_j - \frac{1}{l})^2 \leq \frac{\epsilon^2}{l}$ ... so $||q_i||_2^2 \leq \frac{1 + \epsilon^2}{l}$.
Testing identity via testing uniformity on subdomains:

- *(Relabel domain so that \( q \) monotone)*
- Partition domain into \( O(\log n) \) groups, so that each group almost “flat” --
  - differ by \(< (1+\varepsilon) \) multiplicative factor
  - \( q \) close to uniform over each group

Test:
- Test that \( p \) close to uniform over each group
- Test that \( p \) assigns approximately correct total weights to each group
Testing Identity - Algorithm

- Partition samples from $p$ into buckets according to $q$’s values
- Two levels of checks:
  - Each bucket has correct weight
  - Each bucket close to uniform

4, 2, 4, 12, 5, 1, 1, 15, 3, 7, 9, 8, 6, 6, 2, 10, 7...
Cheating assumption for today’s lecture:

- Assume (almost) no algorithmic estimation error!

- Note that we still have error from approximating $q$ via buckets of uniform distributions.
Possible problem:

- Tolerance of uniformity test:
  - Uniformity test only guaranteed to pass U, but p,q are only close to U

- We will directly use collision probability estimate
Single bucket: $q_i$ close to $p_i$?

- **Algorithm:** Estimate $||p_i||_2^2$ and fail if $> \frac{1+\epsilon^2}{|R_i|}$

- **Lemma:** if $||p_i||_2^2 \leq (1 + \epsilon^2) / |R_i|$ then $||p_i - q_i||_1 \leq 2\epsilon$

  - **Proof:** bound on $||p_i||_2^2$ implies $p_i$ $\epsilon$-close to uniform in L1 (lecture 2). Triangle inequality gives the lemma.
Algorithm (more details):

- Bucket \( q \)
- Calculate total weight \( q \) assigns to each bucket
- Estimate total weight \( p \) assigns to each bucket
- Fail if L1 distance between bucket weight vectors more than \( \epsilon \)
- For each bucket \( i \) with \( q \)-weight \( > \epsilon / 2k \)
  - Estimate collision probability of \( p_i \)
  - Fail if estimate bigger than \( (1+\epsilon^2)/|R_i| \)

\( O(\log n) \) samples

\( O(\sqrt{n} k \log n / \epsilon^2) \) samples of \( p \)
If test passes, $q$ and $p$ are close?

- If $q = p$,
  - By previous argument for $q$, the collision probability of $p$ in each bucket will be $\leq \frac{(1 + \epsilon^2)}{|R|}$
  - Whp the bucket weight vectors are close (standard Chernoff)
If $q$ and $p$ likely to pass test, must they be close?

- Total weight of skipped buckets at most $\epsilon$
- $p_i$ is close to $U_i$ in each bucket, so $\epsilon$-close to $q_i$ (triangle inequality) in each bucket
- Bucket weights of $p$ and $q$ are $\epsilon$-close,

Total distance: $3\epsilon$?

- Requires a lemma on “pasting” everything together
- (not really...remember our assumption on perfect estimation?)
Conclusion...

- Testing identity can be reduced to $O(\log n)$ uniformity tests
Testing closeness of two distributions:

Transactions of 20-30 yr olds

Transactions of 30-40 yr olds

trend change?
Testing closeness

Theorem: ([BFRSW] [P. Valiant])
Sample complexity of distinguishing $p=q$ from $|p-q|_1 > \varepsilon$ is $\tilde{\Theta}(n^{2/3})$
Naïve nearly linear time algorithm – “learning approach”

1. Estimate probabilities $p_i$ and $q_i$’s using frequencies

2. Estimate $L_1$ distance directly

Uses $O(n \log n)$ samples in general.
L₂ distance test

- Use $O(\varepsilon^{-4})$ algorithm for estimating L₂ distance to within $\varepsilon$

- Naïve use of relationship between L₁ and L₂ distance leads to $O(n^2)$ algorithm
Naïve nearly linear time algorithm – “learning approach” (revisited)

1. Estimate probabilities $p_i$ and $q_i$’s using frequencies

2. Estimate $L_1$ distance directly

  • Uses $O(n \log n)$ samples in general.

  • If minimum nonzero $p_i$ or $q_i$ is > $b$, then need only $O(1/b \ \log n)$ samples.
    • When $b=n^{-2/3}$, need $\tilde{O}(n^{2/3})$ samples
L₂ distance test (revisited)

- Use $O(\varepsilon^{-4})$ algorithm for estimating L₂ distance to within $\varepsilon$

- Naïve use of relationship between L₁ and L₂ distance leads to $O(n^2)$ algorithm ...

- Can analyze tighter variance bound for L₂ test in terms of maximum probability $b$ of $p_i$ and $q_i$’s.
  - When $b < n^{-2/3}$, need $\tilde{O}(n^{2/3})$ samples for L₁ distance
Filtering: two phase approach

- \(O(\varepsilon^{-4} n^{2/3} \log n)\) Algorithm:
  1. Sample to determine heavy elements (i.e. \(p_i, q_i > n^{-2/3}\))
  2. Estimate distance of heavy elements
    - Use naïve learning algorithm
  3. Filter out heavy elements to estimate distance of distribution restricted to light elements
    - Use \(L_2\) test
Theorem: [Batu Fortnow R. Smith White 00]

- There exists an algorithm which on input distributions $p$ and $q$ uses $O(\varepsilon^{-4} n^{2/3} \log n)$ samples and satisfies:
  - if $||p-q||_1 < \varepsilon/n^{1/3}$ outputs Pass
  - if $||p-q||_1 > \varepsilon$ outputs Fail
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]

Estimating $|p - q|_1$ requires $\Theta(n/\log n)$ samples [G. Valiant P. Valiant]
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 $p_{(a,b)} = (p_1)_a (p_2)_b$ for all $a,b$
Independence vs. product of marginals

Lemma: [Sahai Vadhan]

If $||P - P_1 \times P_2||_1 > \varepsilon$ then $\forall A, B, ||P - AxB||_1 > \varepsilon/3$
1st try: “Naïve” Algorithm

- Algorithm:
  - Approximate marginal distributions \( f_1 \approx P_1 \) and \( f_2 \approx P_2 \)
  - Use Identity testing algorithm to test that \( P \approx f_1 \times f_2 \)

- Number of queries: \( \tilde{O}(n+m) \)
  - But, if min nonzero probability prefix is bounded from below by \( b \), then can do \( \tilde{O}(1/b + m + (nm)^{1/2}) \)
  - *(also note: if \( n=m \), then this is very good!)*
2nd try: use closeness test

- Simulate $P_1$ and $P_2$, and check $\|P - P_1 \times P_2\|_1 > \varepsilon$.

- Behavior:
  - If $\|P - Q\|_1 < \varepsilon/n^{1/3}$ then PASS
  - If $\|P - Q\|_1 > \varepsilon$ then FAIL

- Sample complexity: $\tilde{O}((nm)^{2/3})$
  - Can do better if max probability element is bounded from above!
2-pronged approach: (more filtering)

- Divide prefixes into heavy/light:
  - For \( P_H = "P \text{ conditioned on heavy prefixes}" 
    - Further subdivide into nearly uniform blocks
    - Test independence using Naïve algorithm
  - For \( P_L = "P \text{ conditioned on light prefixes}" 
    - Further subdivide into nearly suffix uniform columns
    - Test independence using Closeness Test.
- Ensure all pieces can be pasted together
Bounding the max probability element:

- Split into buckets based on suffixes – need only $\tilde{O}(m)$ (which is $\tilde{O}(n^{2/3}m^{1/3})$) samples

- Consider conditional distribution in each bucket
  - if independent, max probability of an element is roughly max probability of prefix divided by size of bucket.
  - Prefix is light

- Test independence between buckets:
  - Test that distribution on prefixes is close in all buckets

- Total Complexity: $\tilde{O}(n^{2/3}m^{1/3})$ samples
Compressibility of data
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 \( \tilde{\Omega}(n^{1/\gamma^2}) \) samples (when entropy \( >2\gamma/\varepsilon \))
    - Based on filtering approach
  - requires \( \Omega(n^{1/\gamma^2}) \) [Valiant]
  - better bounds in terms of support size [Brautbar Samorodnitsky]
- Additive estimates require \( \theta(n \log n) \) samples [VV]
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] [Valiant Valiant]
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
More properties:

- **Limited Independence:** [Alon Andoni Kaufman Matulef R. Xie] [Haviv Langberg]
- **K-flat distributions** [Levi Indyk R.]
- **K-modal distributions** [Daskalakis Diakonikolas Servedio]
- **Poisson Binomial Distributions** [Daskalakis Diakonikolas Servedio]
- **Monotonicity over general posets** [Batu Kumar R.] [Bhattacharyya Fischer R. P. Valiant]
- **Properties of multiple distributions** [Levi Ron R.]
- **And more and more!**
Other properties?

- 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?