Testing that distributions are close

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, Patrick White

Seminar on Sublinear Time Algorithms

May 26, 2010

Anat Ganor

Introduction

Goal

Given two distributions over an n element set, we wish to check whether these distributions are statistically close

- The only allowed operation is independent sampling
- Sublinear time in the size of the domain
- There is a function $f(\epsilon)$ called the *gap* of the tester
- Closeness in L_1-norm
- No knowledge on the structure of the distributions

Outline

1. Related work
2. Algorithm that runs in time $O(n^{2/3}\epsilon^{-4}\log n)$
3. $\Omega(n^{2/3}\epsilon^{-2/3})$ lower bound
4. Applications to mixing properties of Markov processes
5. Further research

Related work

- Interactive setting

Theorem [A. Sahai, S. Vadhan]

Given distributions p and q, generated by poly-size circuits, the problem of distinguishing whether they are close or far in L_1-norm, is complete for statistical zero-knowledge

- Testing statistical hypotheses

The problem

Decide which of two known classes of distributions contains the distribution generating the examples

- Various assumptions on the distributions
- Various distance measures
- Testing closeness to a fixed, known distribution

Testing uniformity [O. Goldreich, D. Ron]

Testing the closeness of a given distribution to the uniform distribution

- **Running time**: $O(\sqrt{n})$ (tight)
- **Application**: testing closeness to being an expander
- **Idea**: Estimating the *collision probability*

Definition

The ***collision probability*** of p and q is the probability that a sample from each yields the same element

Key observations:

1. The self-collision probability of p is $||p||_2^2$
2. $||p - U||_2^2 = ||p||_2^2 - \frac{1}{n}$

Closeness in L_2-norm

Main idea: If p and q are close then the self-collision probability of each are close to the collision probability of the pair

$$||p - q||_2^2 = ||p||_2^2 + ||q||_2^2 - 2(p \cdot q)$$

Closeness in L_2-norm

Main idea: If p and q are close then the self-collision probability of each are close to the collision probability of the pair

$$||p - q||_2^2 = ||p||_2^2 + ||q||_2^2 - 2(p \cdot q)$$

The L_2-distance tester

1. Number of self-collisions taking m samples from $p \rightarrow r_p$
2. Number of self-collisions taking m samples from $q \rightarrow r_q$
3. Number of collisions taking m samples from each p and $q \rightarrow s_{pq}$
4. If $\frac{2m}{m-1}(r_p + r_q) - 2s_{pq} > \frac{m^2\epsilon^2}{2}$ then reject

Closeness in L_2-norm

Main idea: If p and q are close then the self-collision probability of each are close to the collision probability of the pair

$$||p - q||_2^2 = ||p||_2^2 + ||q||_2^2 - 2(p \cdot q)$$

The L_2-distance tester

1. Number of self-collisions taking m samples from $p \rightarrow r_p$
2. Number of self-collisions taking m samples from $q \rightarrow r_q$
3. Number of collisions taking m samples from each p and $q \rightarrow s_{pq}$
4. If $\frac{2m}{m-1}(r_p + r_q) - 2s_{pq} > \frac{m^2\epsilon^2}{2}$ then reject

Repeat $O(\log \frac{1}{\delta})$ times
Reject if the majority of iterations reject, accept otherwise

Closeness in L_2-norm

Main idea: If p and q are close then the self-collision probability of each are close to the collision probability of the pair

$$||p-q||_2^2 = ||p||_2^2 + ||q||_2^2 - 2(p \cdot q)$$

Theorem - closeness in L_2-norm

The tester runs in time $O(m \log \frac{1}{\delta})$
For $m = O(\epsilon^{-4})$ the following holds:

- If $||p-q||_2 \leq \frac{\epsilon}{2}$ then the test passes w.p. $\geq 1-\delta$
- If $||p-q||_2 > \epsilon$ then the test rejects w.p. $\geq 1-\delta$

Closeness in L_1-norm

Theorem - closeness in L_1-norm

Given parameters δ, ϵ and distributions p, q over $[n]$, there is a test which runs in time $O(n^{2/3}\epsilon^{-4} \log n \log \frac{1}{\delta})$ such that:

- If $||p - q||_1 \leq \max\{\frac{\epsilon^2}{32\sqrt[3]{n}}, \frac{\epsilon}{4\sqrt{n}}\}$ then the test passes w.p. $\geq 1 - \delta$
- If $||p - q||_1 > \epsilon$ then the test rejects w.p. $\geq 1 - \delta$

Naive approach

1^{st} attempt

1. For each distribution, sample enough elements to approximate the distribution
2. Compare the approximations

Naive approach

1^{st} attempt

1. For each distribution, sample enough elements to approximate the distribution
2. Compare the approximations

Problem We cannot learn a distribution using sublinear number of samples

Theorem

Suppose we have an algorithm $\mathcal{A}$ that draws $o(n)$ samples from an unknown distribution p and outputs a distribution $\mathcal{A}(p)$. There is some p for which p and $\mathcal{A}(p)$ have L_1-distance close to 1

Using the L_2-distance tester

Recall we can test L_2-distance in sublinear time such that:

- If $||p - q||_2 \leq \frac{\epsilon}{2}$ then the test passes w.p. $\geq 1 - \delta$
- If $||p - q||_2 > \epsilon$ then the test rejects w.p. $\geq 1 - \delta$

Using the L_2-distance tester

Recall we can test L_2-distance in sublinear time such that:

- If $||p - q||_2 \leq \frac{\epsilon}{2}$ then the test passes w.p. $\geq 1 - \delta$
- If $||p - q||_2 > \epsilon$ then the test rejects w.p. $\geq 1 - \delta$

2^{nd} attempt

Test for L_2-distance

Using the L_2-distance tester

Recall we can test L_2-distance in sublinear time such that:

- If $||p - q||_2 \leq \frac{\epsilon}{2}$ then the test passes w.p. $\geq 1 - \delta$
- If $||p - q||_2 > \epsilon$ then the test rejects w.p. $\geq 1 - \delta$

2^{nd} attempt

Test for L_2-distance

Problem L_2-distance does not in general give a good approximation to L_1-distance

Example

Two distributions can have disjoint support and still have small L_2-distance

Using the L_2-distance tester

Recall that $||v||_1 \leq \sqrt{n} \cdot ||v||_2$

3^{rd} attempt

Use the L_2-distance tester with $\epsilon' = \Theta(\frac{\epsilon}{\sqrt{n}})$

Using the L_2-distance tester

Recall that $||v||_1 \leq \sqrt{n} \cdot ||v||_2$

3^{rd} attempt

Use the L_2-distance tester with $\epsilon' = \Theta(\frac{\epsilon}{\sqrt{n}})$

Problem: Takes too much time

Theorem

Given parameters δ, ϵ and distributions p, q over $[n]$, there is a test which runs in time $O(\epsilon^{-4} \log \frac{1}{\delta})$ such that:

- If $||p - q||_2 \leq \frac{\epsilon}{2}$ then the test passes w.p. $\geq 1 - \delta$
- If $||p - q||_2 > \epsilon$ then the test rejects w.p. $\geq 1 - \delta$

Using the L_2-distance tester

When can we use the L_2-distance tester with ϵ' and run sublinear time?

Key observation:

Let $b = \max\{||p||_\infty, ||q||_\infty\}$ then $b^2 \leq ||p||_2^2, ||q||_2^2 \leq b$

Theorem - closeness in L_2-norm, revised

The tester runs in time $O(m \log \frac{1}{\delta})$

For $m = O((b^2 + \epsilon^2 \sqrt{b})\epsilon^{-4})$ the following holds:

- If $||p - q||_2 \leq \frac{\epsilon}{2}$ then the test passes w.p. $\geq 1 - \delta$
- If $||p - q||_2 > \epsilon$ then the test rejects w.p. $\geq 1 - \delta$

Corollary

If $b = O(n^{-\alpha})$ then applying the test with ϵ' takes time $O((n^{1-\alpha/2} + n^{2-2\alpha})\epsilon^{-4} \log \frac{1}{\delta})$

The L_1-distance tester

The L_1-distance tester

1. Identify the “big" elements of $p, q \rightarrow S_p, S_q$
2. Measure the distance corresponding to the “big" elements via straightforward sampling
3. Modify the distributions so that the distance attributed to the “small" elements can be estimated using the L_2-distance test

Theorem - closeness in L_1-norm

The tester runs in time $O(n^{2/3}\epsilon^{-4} \log n \log \frac{1}{\delta})$ and the following holds:

- If $||p - q||_1 \leq \max\{\frac{\epsilon^2}{32\sqrt[3]{n}}, \frac{\epsilon}{4\sqrt{n}}\}$ then the test passes w.p. $\geq 1 - \delta$
- If $||p - q||_1 > \epsilon$ then the test rejects w.p. $\geq 1 - \delta$

Proof outline

- The "big" elements are identified correctly w.h.p
- Let Δ_1 be the L_1-distance attributed to the "big" elements
  By Chernoff's inequality, Δ_1 is approximated upto a small additive error w.h.p using $O(n^{2/3}\epsilon^{-2}\log n)$ samples
- Let Δ_2 be the L_1-distance of p' and q'
  Since $\max\{||p'||_\infty, ||q'||_\infty\} < n^{-2/3} + n^{-1}$ we can apply the L_2-distance tester on $\frac{\epsilon}{2\sqrt{n}}$ with $O(n^{2/3}\epsilon^{-4}\log n \log\frac{1}{\delta})$ samples and get a good approximation of Δ_2
- It holds that $\Delta_1, \Delta_2 \leq ||p-q||_1 \leq 2\Delta_1 + \Delta_2$

$\Omega(n^{2/3}\epsilon^{-2/3})$ lower bound

Theorem

Given any test using $o(n^{2/3})$ samples, there exist distributions $\bar{a}, \bar{b}$ of L_1-distance 1 such that the test will be unable to distinguish the case where one distribution is $\bar{a}$ and the other is $\bar{b}$ from the case where both distributions are $\bar{a}$

$\Omega(n^{2/3}\epsilon^{-2/3})$ lower bound

Theorem

Given any test using $o(n^{2/3})$ samples, there exist distributions $\bar{a}, \bar{b}$ of L_1-distance 1 such that the test will be unable to distinguish the case where one distribution is $\bar{a}$ and the other is $\bar{b}$ from the case where both distributions are $\bar{a}$

Define $\bar{a}, \bar{b}$ as follows:

$1 \le i \le n^{2/3}$: $a_i = b_i = \frac{1}{2n^{2/3}}$ (heavy elements)

$n/2 < i \le 3n/4$: $a_i = \frac{2}{n}$, $b_i = 0$ (light elements of a)

$3n/4 < i \le n$: $b_i = \frac{2}{n}$, $a_i = 0$ (light elements of b)

For the remaining i: $a_i = b_i = 0$

Proof outline

- We can assume that the tester is symmetric
- None of the light elements occur more than twice w.h.p
- Let H be the number of collisions among the heavy elements
  Let L be the number of collisions among the light elements
  The L_1-distance between H and $H + L$ is $o(1)$
  $\Rightarrow$ No statistical test can distinguish H and $H + L$ with non-trivial probability

Rapidly mixing Markov processes

Definition

A Markov chain M is (ϵ, t)-***mixing*** if there exists a distribution $\bar{s}$ such that for all states u, $||\bar{e_u} M^t - \bar{s}||_1 \leq \epsilon$

Rapidly mixing Markov processes

Definition

A Markov chain M is (ϵ, t)***-mixing*** if there exists a distribution $\bar{s}$ such that for all states u, $||\bar{e_u} M^t - \bar{s}||_1 \leq \epsilon$

Theorem

There exists a test with time complexity $\tilde{O}(nt \cdot T(n, \epsilon, \delta/n))$ such that:

- *If M is $(f(\epsilon)/2, t)$-mixing then the test passes w.p. $\geq 1 - \delta$*
- *If M is (ϵ, t)-mixing then the test rejects w.p. $\geq 1 - \delta$*

Rapidly mixing Markov processes

The mixing tester

Use the L_1-distance tester to compare each distribution $\bar{e}_u M^t$ with the average distribution after t steps:

$$s_{\bar{M},t} = \frac{1}{n} \sum_u \bar{e}_u M^t$$

Rapidly mixing Markov processes

The mixing tester

Use the L_1-distance tester to compare each distribution $\bar{e_u}M^t$ with the average distribution after t steps:

$$s_{\bar{M},t} = \frac{1}{n}\sum_u \bar{e_u}M^t$$

Proof sketch:

- Assume that every state is $(f(\epsilon)/2, t)$-close to some distribution $\bar{s}$
  $\Rightarrow s_{\bar{M},t}$ is $f(\epsilon)/2$-close to $\bar{s}$
  $\Rightarrow$ every state is $(f(\epsilon), t)$-close to $s_{\bar{M},t}$
- If there is no distribution that is (ϵ, t)-close to all states then, in particular, $s_{\bar{M},t}$ is not (ϵ, t)-close to at least one state

Other variants of testing mixing properties

1. Test that *most* states reach the same distribution after t steps
   The idea: pick $O(\frac{1}{\rho} \cdot \log \frac{1}{\delta})$ starting states uniformly at random
2. Test if it is possible to change ϵ fraction of the matrix to turn it into a (ϵ, t)-mixing Markov chain
3. Extension to sparse graphs and uniform distributions

Further research

- Other distance measures
- Weighted distances
- Non independent samples
- Tighter bounds