6.896 Sublinear Time Algorithms February 6, 2007

Lecture 1

Lecturer: Ronitt Rubinfeld *Scribe: Chih-yu Chao*

1 Testing Global Properties of Distributions

We will begin by considering how to test properties of distributions over large domains. We assume that the only access we have to the distributions is via samples of the distribution, and we will focus on the number of samples required in order to answer our questions about the distribution. We will begin by studying the sample complexity of determining whether two distributions are identical, or far apart according to some measure of distance.

1.1 Closeness

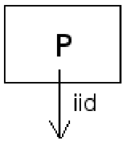

Figure 1: The black box that outputs the distribution of p

Here, we will first define the closeness of distributions. Suppose we are given a distribution p in domain D, where $|D| = n$, and probabilities $(p_1, \ldots p_n)$, such that
$p_x = \Pr[p \text{ outputs } x] \leftarrow$ unknown to the algorithm, $\sum_{x \in D} p_x = 1$.
We will consider algorithms that are given oracle access to a independent identical distribution p (see Figure 1) – each time we press the button on the black box, we get a sample i with probability p_i.

We would like to consider the L_1 distance between two distributions:

Definition 1 (L_1 distance)

$$|p - q|_1 = \sum_x |p_x - q_x|$$

Another common measure is the L_2 distance, defined as follows.

Definition 2 (L_2 distance)

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

Note that the L_1 norm is defined as

Definition 3 ($\mathbf{L_1}$ norm)

$$|p| = \sum_x |p_x|$$

and the L_2 norm is defined as

Definition 4 (L_2 norm)

$$\|p\| \equiv \sqrt{\sum_x p_x^2}$$

- Often we just look at $\|p-q\|_2^2$, and later we will use this to estimate the distance bewteen two probability distributions.
- A well-known fact is that $\|p-q\|_2 \leq |p-q|_1 \leq \sqrt{n}\|p-q\|_2$, but this is not enough to get a good estimate.

Other popular distance measures include

KL divergence: $p\|q = \sum\limits_i p_i \log \frac{p_i}{q_i}$ (not symmetric, not transitive)

and

Earth Movers Distance (EMD)

1.2 Testing Closeness of Two Distributions

We would like to test whether two probability distributions p and q are "close". The black box **p** in Figure 2 outputs the distribution p, while the black box **q** outputs the distribution q. However, the values of the probabilities of each element are not accessible to us; the only information we know is the samples from the black boxes.

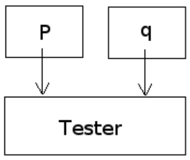

Figure 2: The tester output tells us whether the distributions p and q are close

We define $\mathcal{T}$ to be a "closeness tester" in $\begin{array}{c} \mathrm{L}_1 \\ (\mathrm{L}_2) \end{array}$ distance if it satisfies:

1. if $p = q$, $\Pr[\mathcal{T}$ outputs "pass"$] \geq \frac{2}{3}$
2. if $\begin{array}{c} |p-q|_1 > \varepsilon \\ (\|p-q\|_2 > \varepsilon) \end{array}$, $\Pr[\mathcal{T}$ outputs "fail"$] \geq \frac{2}{3}$

Comments:

1. $\frac{2}{3}$ is an arbitrary constant $> \frac{1}{2}$
   can repeat $O(\log \frac{1}{\beta})$ times and take majority to get error $< \beta$
2. Probabilities are over coin tosses of algorithm and samples
   not any assumptions on p, q
3. interest in sample/time complexity of $\mathcal{T}$

1.2.1 Naive Algorithm

We can use a naive algorithm to test whether two distributions p and q are close:

take m samples of p

take m samples of q

estimate $\hat{p_x} \leftarrow \frac{\# \text{ of samples of } p \text{ that are } x}{m}$

estimate $\hat{q_x} \leftarrow \frac{\# \text{ of samples of } q \text{ that are } x}{m}$

if $\begin{array}{c} \sum\limits_x |\hat{p_x} - \hat{q_x}| > \frac{\varepsilon}{2} \\ (\sum\limits_x (\hat{p_x} - \hat{q_x})^2 > \frac{\varepsilon}{2}) \end{array}$, then "fail"

We will need $\Omega(n)$ samples for this algorithm.

1.2.2 Collision Probability

Then, we define the following notions:

Definition 5 (collision probability of p, q)

$$CP(p,q) = \Pr[\textit{sample from } p \ = \ \textit{sample from } q] = \sum_x p_x q_x$$

Definition 6 (self-collision probability)

$$SCP(p) = \Pr[2 \textit{ samples from } p \textit{ are equal}] = \sum_x p_x^2 = \|p\|_2^2$$

The above definitions show that the L_2 distance can be expressed in terms of collision probabilities:

$$\|p-q\|_2^2 = \sum_x (p_x - q_x)^2 = \sum_x (p_x^2 - 2p_x q_x + q_x^2) = SCP(p) + SCP(q) - 2CP(p,q)$$

To estimate collision probability $CP(p,q)$, we can do the following:

Pick m samples $\begin{array}{ll} u_1 \dots u_m & \in P \\ v_1 \dots v_m & \in Q \end{array}$

$\sigma_i \leftarrow \begin{cases} 1 & \text{if } u_i = v_i \\ 0 & \text{otherwise} \end{cases}$

output $\frac{\sum \sigma_i}{m}$

Note that $CP(U_D, U_D) = \sum \left(\frac{1}{n}\right)^2 = \frac{1}{n}$.

1.2.3 The Tester

The tester is to estimate the L_2 distance by estimating the CP and the SCP.

Given $p, q, \sum$

$m \leftarrow O(\varepsilon^{-4})$

$$\begin{array}{c} SCP(p), \\ SCP(q) \end{array} \left\{ \begin{array}{l} S_p \leftarrow m \text{ samples of } p \\ S_q \leftarrow m \text{ samples of } q \\ r_p \leftarrow \#\text{ of self-collisions in } S_p \\ r_q \leftarrow \#\text{ of self-collisions in } S_q \end{array} \right.$$

$$CP(p,q) \left\{ \begin{array}{l} Q_p \leftarrow m \text{ samples from } p \\ Q_q \leftarrow m \text{ samples from } q \\ r_{pq} \leftarrow \#\ i\ j \text{ such that } i^{th} \text{ sample in } Q_p = j^{th} \text{ sample in } Q_q \end{array} \right.$$

$r \leftarrow \frac{2m}{m-1}(r_p + r_q)$

$S \leftarrow 2r_{pq}$

if $r - S > \frac{m^2\varepsilon^2}{2}$, reject

Theorem 1 $\forall\ \varepsilon, p, q\ \exists$ *an algorithm* $\mathcal{T}$ *that runs in* $O(\varepsilon^{-4})$ *sample & time complexity which satisfies:*

1. *if* $\|p-q\|_2 < \frac{\varepsilon}{2}$, $\Pr[\mathcal{T}$ *outputs "pass"*$] \geq \frac{2}{3}$
2. *if* $\|p-q\|_2 > \varepsilon$, $\Pr[\mathcal{T}$ *outputs "fail"*$] \geq \frac{2}{3}$

Claim 2 *If* $p = U_D$, *and* S *&* q *are chosen as follows:*

$S \leftarrow$ *random subset of* D *of size* $\frac{n}{2}$

$q \leftarrow U_s$

We need $\Omega(\sqrt{n})$ *queries to distinguish input* (p,p) *from input* (p,q).

Proof Idea WLOG, assume the algorithm gives the same result for any permutation of underlying domain. (can only use collision statistics, *i.e.* how many elements appear once, twice, ...)
With high probability, if m is $O(\sqrt{n})$, no collisions. ■

Note that for the p, q in the above proof,

$$|p-q|_1 = \sum \frac{1}{n} = 1 \qquad \text{(really far)}$$

$$\|p-q\|_2 = \sqrt{\sum \frac{1}{n^2}} = \frac{1}{\sqrt{n}} \qquad \text{(tiny)}$$

Thus, the proof says very little about a lower bound for estimating the L_2 distance.

1.2.4 Analysis

We would like to show that the expectation of the random variable $r-S$ is equal to the quantity we want to measure. In the next lecture, we will bound the variance of this random variable and use Chebychev's inequality to show that the value of the variable will be close to its expection.

Claim 3 (Expectation) *The following equalities hold:*

$E[r_p] = \binom{m}{2}\|p\|^2$

$E[r_q] = \binom{m}{2}\|q\|^2$

$E[r_{pq}] = m^2 CP(p,q)$

$E[S] = 2m^2 CP(p,q)$

$E[r-S] = \frac{2m}{m-1}\binom{m}{2}\left(\|p\|^2 + \|q\|^2\right) - 2m^2 CP(p,q) = m^2\|p-q\|_2^2$

Proof $E[r_p] = \binom{m}{2}\|p\|^2$

$$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } i^{th} \text{ sample in } S_p = j^{th} \text{ sample in } S_p \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr[\sigma_{ij} = 1] = E[\sigma_{ij}] = \|p\|^2$$

$$r_p = \sum \sigma_{ij}, \text{ so } E[r_p] = \binom{m}{2}\|p\|^2$$

■