6.896 Sublinear Time Algorithms December 2, 2004

Lecture 22

Lecturer: Eli Ben-Sasson *Scribe: Rafael Pass*

1 Testing Proximity of Distributions

A distribution p on $[n] = \{1, 2, \ldots, n\}$ is given by the probabilities $(p_1, .., p_n)$, such that $\sum_{i=1}^{n} p_i = 1$, and $0 \leq p_i \leq 1$. We will consider algorithms that are given oracle access to a distribution p, i.e., each time we "press a button" we get a sample $i \in [n]$ with probability p_i. Our objective is to test if two distributions p, q over $[n]$ are "close".

1.1 Definitions of closeness of distributions

Ideally we would like to consider the L_1 distance between two distributions, defined as follows:

$$|p - q| = \sum_{i=1}^{n} |p_i - q_i|$$

Today we will instead focus on the (easier) L_2 distance (i.e., the Euclidean norm), defined as follows:

$$||p - q|| = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Later we will use this to estimate the L_1 distance. (A well known fact is that $|p-q| \leq \sqrt{n}||p-q||$. This fact will, however, not be enough to get a good estimate of the L_1 distance.)

1.2 The Theorem

We will prove the following theorem:

Theorem 1 (Batu, Fortnow, Rubinfeld, Smith, White [1]) *For every constant ϵ, and every distributions p, q over $[n]$, there exists a test that runs in time $O(\delta^{-4} \log(1/\epsilon))$ such that:*

- *If $||p - q|| < \delta/2$, then $\Pr[\text{test accepts}] \geq 1 - \epsilon$*
- *If $||p - q|| > \delta$, then $\Pr[\text{test accepts}] \leq \epsilon$*
- *The query complexity of the tester (i.e., the number of sample) is less than the running time (which is constant for constant δ, ϵ).*

Next lecture we show a tester for L_1 distance which uses a query complexity of $\approx n^{3/2}$ (which thus is super constant).

1.3 Why is it harder to test L_1 distance?

The following example shows why L_1 distance requires super-constant query complexity (even for constant δ, ϵ). Consider the following two cases:

- p, q are two uniform distributions on two equally large (unknown) *disjoint* subsets of $[n]$. Note that $|p - q| = 2$.
- $p = q = U_n$ (i.e., the uniform distribution on $[n]$). Note that (trivially) $|p - q| = 0$

In order to distinguish the above two cases we need to find a collision, which means that $\Omega(\sqrt{n})$ queries are needed.

Remark: Note that the above argument *does not* extend to the L_2 distance. The reason for this is that the L_2 distance between p and q in case 1, is small. In fact $||p-q|| = \sqrt{n(\frac{2}{n})^2)} = O(1/\sqrt{n})$. This example also shows that L_2 distance is *not* the right measure when is comes to distinguishing distributions. (However, it is a good starting point for obtaining bounds on L_1 distance, as we do in next lecture).

2 Proof of the Main Theorem 1

2.1 Definitions

Towards the goal of proving the theorem we define the following notions:

Definition 2 (Collision Probability) *Let the collision probability of* p, q *be*

$$CP(p,q) = \Pr[\textit{sample from } p = \textit{sample from } q] = \sum_{i=1}^{n} p_i q_i$$

Definition 3 (Self Collision Probability) *Let the self collision probability of* p *be*

$$SCP(p) = CP(p,p) = \Pr[\textit{two independent samples from } p \textit{ are equal}] = ||p||^2$$

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

$$||p-q||^2 = \sum_{i=1}^{n}(p_i - q_i)^2 = \sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} q_i^2 - 2\sum_{i=1}^{n} p_i q_i = SCP(p) + SCP(q) - 2CP(p,q) \tag{1}$$

The previous characterization will be used in order to approximate the L_2 distance.

2.2 The Test

Consider the following tester, $T_2(p, q, m, \delta, \epsilon)$:

- Initialize counter *bad* to 0
- Repeat $t = O(\log(1/\epsilon))$ times:
  - F_p = sample of size m from p; Let r_p = number of self collisions of F_p, i.e. number of pairs $1 \le i < j \le m$ such that i^{th} sample and j^{th} sample are equal.
  - F_q = sample of size m from q; Let r_q = number of self collisions of F_q.
  - Q_p, Q_q = samples of size m from p, q respectively; Let S_{pq} = number of collision in Q, i.e., number of i, j such that there is a collision between the i'th sample in Q_p and the j'th sample in Q_q.
  - Set $r = \frac{2m}{m-1}(r_p + r_q)$ and $s = 2S_{pq}$;
  - If $r - s > \frac{m^2\delta^2}{2}$ increment *bad* (i.e. $bad = bad + 1$).
- Reject iff $bad \ge t/2$ (otherwise accept)

Notice the running time (and query complexity) is $O(\log(1/\epsilon) \cdot m)$. We will set $m = O(\delta^{-4})$, thus achieving the running time stated in Theorem 1

What is the test doing? The test is estimating the L_2 distance by estimating the CP and the SCP.

2.3 Analysis of Tester

Overview We show (in Claim 4) the expectation of the random variable $r - s$ is equal to the quantity we wish to measure (namely, $||p-q||$). Then (in Lemma 6) we bound the variance of this random variable. Finally, we use Chebychev's inequality to show the value of the variable will typically be close to its expectation, thus completing the proof.

Claim 4 (Expectation) *The following equalities hold:*

- $E[r_p] = \binom{m}{2}||p||^2$
- $E[r_q] = \binom{m}{2}||q||^2$
- $E[S_{pq}] = m^2 CP(p,q)$, *i.e.,* $E[s] = 2m^2 CP(p,q)$
- $E[r-s] = m^2||p-q||^2$

Proof The first three parts follow from the definition of r_p, r_q, S_{pq} and the linearity of expectation. The last one follows from the first three and Equation (1). ■

We will now bound the variance of $r-s$. The following Lemma is presented without proof (it's proof can be found in [2]).

Lemma 5 *[2] The following equalities hold:*

- $\mathrm{Var}[r_p] \le 2(E[r_p])^{3/2}$
- $\mathrm{Var}[r_q] \le 2(E[r_q])^{3/2}$

Using Lemma 5 we prove our main lemma:

Lemma 6 *There exists a constant c such that for $b = \max(p_1, p_2, .., p_n, q_1, .., q_n)$ we have*

$$\mathrm{Var}[r-s] \le cm^3 b$$

Proof For $i, j \in [m]$ let the indicator variable $c_{i,j} = 1$ iff the i'th sample of Q_p equals the j'th sample of Q_q. Also let $\bar{c}_{i,j} = c_{i,j} - E[c_{i,j}]$ (i.e., $c_{i,j} - CP(p,q)$). Now,

$$\begin{aligned}
\mathrm{Var}\left[s\right] &= \mathrm{Var}\left[\sum_{i,j\in[m]} c_{i,j}\right] = E\left[\left(\sum_{i,j\in[m]} \bar{c}_{i,j}\right)^2\right] \\
&= E\left[\sum_{i,j\in[m]} \bar{c}_{i,j}^2 + \sum_{i,j,k,l\in[m],(i,j)\neq(k,l)} \bar{c}_{i,j}\bar{c}_{k,l}\right] \\
&= E\left[\sum_{i,j\in[m]} \bar{c}_{i,j}^2\right] + E\left[\sum_{i,j,k,l\in[m],(i,j)\neq(k,l)} \bar{c}_{i,j}\bar{c}_{k,l}\right] \qquad (2)
\end{aligned}$$

The second equality above follows from the definition of variance, and the last one is linearity of expectation. Let (a) denote the first term in Equation (2) and (b) the second. We start by bounding a single summand in the sum (a).

$$\begin{aligned}
E\left[\bar{c}_{i,j}^2\right] &= E\left[(c_{i,j} - CP(p,q))^2\right] = E\left[c_{i,j}^2 + CP(p,q)^2 - 2c_{i,j}CP(p,q)\right] \\
&= E\left[c_{i,j}^2\right] + CP(p,q)^2 - 2E\left[c_{i,j}\right]CP(p,q) \\
&= E\left[c_{i,j}\right] + CP(p,q)^2 - 2CP(p,q)^2 \\
&= CP(p,q) - CP(p,q)^2 \le CP(p,q)
\end{aligned}$$

(note that $c_{i,j}^2 = c_{i,j}$, by the fact that $c_{i,j}$ is zero/one valued). We can thus bound (a) by

$$(a) = E\left[\sum_{i,j\in[m]} \bar{c}_{i,j}^2\right] \le m^2 CP(p,q) = m^2 \sum_i p_i q_i \le m^2 \max\{p_i\} \sum_i q_i \le m^2 b$$

The last inequality follows because $\sum q_i = 1$ and $\max\{p_i\} \le b$. We turn to bounding (b). Towards this goal we start by showing the following facts.

Fact 7 *If $i \neq k$ and $j \neq l$, then $\bar{c}_{i,j}$ and $\bar{c}_{k,l}$ are independent, i.e., $E[\bar{c}_{i,j}\bar{c}_{k,l}] = 0$.*

Proof

$$\begin{aligned} E[\bar{c}_{i,j}\bar{c}_{k,l}] &= E[(c_{i,j} - CP(p,q))(c_{k,l} - CP(p,q))] \\ &= E[c_{i,j}c_{k,l}] - 2CP(p,q)E[c_{i,j}] + CP(p,q)^2 \end{aligned}$$

By the independence of the variables it follows that the first term equals $E[c_{i,j}]^2 = CP(p,q)^2$, which means that the whole expression equals 0. ■

Fact 8 $E[\bar{c}_{i,j}\bar{c}_{k,l}] \le E[c_{i,j}c_{k,l}]$.

Proof Follows by opening the expression. Details omitted. ■

Armed with the above two facts, we return to bounding (b). By Facts 7, 8 it follows that

$$\begin{aligned} (b) &= \sum_{i,j\neq l} E[\bar{c}_{i,j}\bar{c}_{i,l}] + \sum_{j,i\neq k} E[\bar{c}_{i,j}\bar{c}_{k,j}] \\ &\le \sum_{i,j\neq l} E[c_{i,j}c_{i,l}] + \sum_{j,i\neq k} E[c_{i,j}c_{k,j}] \\ &\le m^3 \sum_{\alpha\in[n]} p_\alpha q_\alpha^2 + p_\alpha^2 q_\alpha \\ &\le m^3 b^2 \sum_{\alpha\in[n]} q_\alpha = m^3 b^2 \end{aligned}$$

The last inequality above follows because $b = \max\{p_\alpha, q_\alpha\}$. We conclude that

$$\mathrm{Var}(s) \le (a) + (b) \le m^2 b + m^3 b^2$$

$$\mathrm{Var}(r) = (\frac{2m}{m-1})^2(\mathrm{Var}(r_p) + \mathrm{Var}(r_q))$$

The last equality follows because the variance of a sum of independent variables equals the sum of variances (this is known as "linearity of variances of independent variables"). Now, by Lemma 5,

$$\begin{aligned} \mathrm{Var}[r_p] &< (2E[r_p])^{3/2} \le 2(m^2||p||^2)^{3/2} = 2m^3(\sum_{i\in[n]} p_i^2)^{3/2} \\ &\le 2m^3(\max\{p_\alpha\} \cdot \sum_{i\in[n]} p_i)^{3/2} \le 2m^3 b^{3/2} \le 2m^3 b \end{aligned}$$

Thus, for large enough constant c we conclude the proof of our Lemma 6,

$$\mathrm{Var}[r-s] = \mathrm{Var}[r] + \mathrm{Var}[s] \le (a) + (b) + 2m^3 b \le c(m^2 b + m^3 b^2 + m^3 b) \le cm^3 b.$$

■

We have thus bounded the (expectation and) variance of $r - s$ and can now rely on Chebyschev's inequality to show the validity of the tester. This will be shown in the next lecture.

References

[1] Batu, T., Fortnow, L., Rubinfeld, R., Smith, W.D., White, P. "Testing the closeness of distributions". In *Proc. 41st IEEE Conference on Foundations of Computer Science*, 2000.

[2] O. Goldreich, D. Ron. On testing Expansion in bounded degree graphs. Technical Report TR00-020, ECCC, 2000.