6.896 Sublinear Time Algorithms	February 8, 2007
Lecture 2	
Lecturer: Ronitt Rubinfeld	*Scribe: Adriana Lopez*

In this lecture, we will prove a theorem from last lecture that states that there exists a *closeness tester* in L_2. We will also give an algorithm for L_1-distance testing.

Recall the following definitions and facts from last lecture:

Definition 1 $p_x = \Pr[p \text{ outputs } x]$

Definition 2 *(collision probability of p and q)*
$CP(p,q) = \Pr[\text{sample from } p \text{ equals sample from } q] = \sum_x p_x q_x$

Definition 3 *(self-collision probability of p)*
$SCP(p) = \Pr[\text{2 samples from } p \text{ are equal}] = \sum_x p_x^2 = \|p\|_2^2$

1 L_2-Distance Testing

Theorem 4 *For all ϵ, p, q, there exists an $O(\epsilon^{-4})$ sample tester T such that*

- *If* $\|p-q\|_2 < \frac{\epsilon}{2}$ *then* $\Pr[T \text{ accepts}] \geq \frac{2}{3}$.
- *If* $\|p-q\|_2 > \epsilon$ *then* $\Pr[T \text{ rejects}] \geq \frac{2}{3}$.

We will prove that tester T from last lecture (reproduced below) satisfies the conditions stated in Theorem 4.

$T(p,q,\epsilon)$:
  $m \leftarrow O(\epsilon^{-4})$
  $S_p \leftarrow m$ samples from p
  $S_q \leftarrow m$ samples from q
  $r_p \leftarrow$ number of self-collisions in S_p
  $r_q \leftarrow$ number of self-collisions in S_q
  $Q_p \leftarrow m$ (new) samples from p
  $Q_q \leftarrow m$ (new) samples from q
  $r_{pq} \leftarrow$ number of pairs i,j s.t. ith sample in Q_p is equal to jth sample from Q_q
  $r \leftarrow \frac{2m}{m-1}(r_p + r_q)$
  $s \leftarrow 2r_{pq}$
  if $r - s > \frac{m^2\epsilon^2}{2}$
    then reject
    else accept

Also recall the following fact from last lecture: $\mathbb{E}[r-s] = m^2\|p-q\|_2^2$

Lemma 5 *Let* $b = \max_x p_x, q_x$. *Then* $\mathrm{Var}[r-s] \leq O(m^3b^2 + m^2b)$.

Proof First notice that since r and s are independent, we have that $\mathrm{Var}[r-s] = \mathrm{Var}[r] + \mathrm{Var}[s]$. We start by computing $\mathrm{Var}[s]$. Let c_{ij} be the indicator random variable defined as follows:

$$c_{ij} = \begin{cases} 1 & \text{if the } i\text{th sample from } Q_p \text{ is equal to the } j\text{th sample from } Q_q \\ 0 & \text{otherwise} \end{cases}$$

Notice that $\mathbb{E}[c_{ij}] = CP(p,q)$. Define $\bar{c_{ij}} = c_{ij} - \mathbb{E}[c_{ij}]$, and notice that $\mathbb{E}[\bar{c_{ij}}] = 0$. It is easy to verify that for all i,j,k,l we have $\mathbb{E}[\bar{c_{ij}}\bar{c_{kl}}] = \mathbb{E}[c_{ij}c_{kl}] - \mathbb{E}[c_{ij}]\mathbb{E}[c_{kl}]$, and thus $\mathbb{E}[\bar{c_{ij}}\bar{c_{kl}}] \leq \mathbb{E}[c_{ij}c_{kl}]$. We will use this fact later in the proof.

From the definition of s, we get

$$\begin{aligned}
\mathrm{Var}[s] &= 4\mathrm{Var}[\sum_{i,j} c_{ij}] = 4\mathbb{E}[(\sum_{i,j} c_{ij} - \mathbb{E}[\sum_{i,j} c_{ij}])^2] \\
&= 4\mathbb{E}[(\sum_{i,j}(c_{ij} - \mathbb{E}[c_{ij}]))^2] = 4\mathbb{E}[(\sum_{i,j}\bar{c_{ij}})^2] \\
&= 4\mathbb{E}[\sum_{i,j}\bar{c_{ij}}^2] + 4\mathbb{E}[\sum_{\substack{i,j,k,l \\ (i,j)\neq(k,l)}} \bar{c_{ij}}\bar{c_{kl}}] \qquad (1)
\end{aligned}$$

We will bound each of the summands in 1 separately. Using the fact that we proved above, we get that

$$\mathbb{E}[\bar{c_{ij}}^2] \leq \mathbb{E}[c_{ij}^2] = \mathbb{E}[c_{ij}] = CP(p,q)$$

This implies that

$$\begin{aligned}
\mathbb{E}[\sum_{i,j}\bar{c_{ij}}^2] &\leq m^2 CP(p,q) = m^2\sum_x p_x q_x \\
&\leq m^2\sum_x b q_x = m^2 b\sum_x q_x \\
&= m^2 b \qquad (2)
\end{aligned}$$

To bound the second summand, first notice that in the case that $i \neq k, j \neq l$ we know that $\bar{c_{ij}}$ and $\bar{c_{kl}}$ are independent, so that $\mathbb{E}[\bar{c_{ij}}\bar{c_{kl}}] = \mathbb{E}[\bar{c_{ij}}]\mathbb{E}[\bar{c_{kl}}] = 0$. This means that

$$\begin{aligned}
\mathbb{E}[\sum_{\substack{i,j,k,l \\ (i,j)\neq(k,l)}} \bar{c_{ij}}\bar{c_{kl}}] &= \sum_{i,j\neq l}\mathbb{E}[\bar{c_{ij}}\bar{c_{il}}] + \sum_{j,i\neq k}\mathbb{E}[\bar{c_{ij}}\bar{c_{kj}}] \\
&\leq \sum_{i,j\neq l}\mathbb{E}[c_{ij}c_{il}] + \sum_{j,i\neq k}\mathbb{E}[c_{ij}c_{kj}] \\
&\leq \sum_{i,j\neq l}\sum_x p_x q_x^2 + \sum_{j,i\neq k}\sum_x p_x^2 q_x \\
&= m\binom{m}{2}(\sum_x p_x q_x^2 + \sum_x p_x^2 q_x) \\
&\leq 2m^3 b^2(\sum_x q_x + \sum_x p_x) \\
&= 4m^3 b^2 \qquad (3)
\end{aligned}$$

Plugging in 2 and 3 in 1, we get that $\mathrm{Var}[s] \leq 4m^2 b + 16m^3 b^2 = O(m^2 b + m^3 b^2)$. Similarly, we can prove that $\mathrm{Var}[r] = O(m^2 b + m^3 b^2)$ and thus $\mathrm{Var}[r-s] = \mathrm{Var}[r] + \mathrm{Var}[s] = O(m^2 b + m^3 b^2)$. ∎

Proof of Theorem 1: We use Chebyshev's inequality to bound the probability that $r - s$ is far from its expectation. Let Event A be the event that $|(r-s) - \mathbb{E}[r-s]| > \frac{m^2\epsilon^2}{4}$. Using Chebyshev and lemma 5, we have that

$$\Pr[A] \leq O\left(\frac{m^3 b^2 + m^2 b}{m^4\epsilon^4}\right)$$

If we pick m such that $m > \frac{c(b^2+\epsilon^2\sqrt{b})}{\epsilon^4}$ with large enough c, we can get $O\left(\frac{m^3b^2+m^2b}{m^4\epsilon^4}\right) \leq \frac{1}{3}$, or in other words, there exists constant c such that picking $m > \frac{c(b^2+\epsilon^2\sqrt{b})}{\epsilon^4}$ yields $O\left(\frac{m^3b^2+m^2b}{m^4\epsilon^4}\right) \leq \frac{1}{3}$, and therefore $\Pr[A] \leq \frac{1}{3}$.

Suppose that $\|p-q\|_2 \leq \frac{\epsilon}{2}$. Then $\mathbb{E}[r-s] = m^2\|p-q\|_2^2 \leq \frac{m^2\epsilon^2}{4}$. This means that

$$\Pr\left[|(r-s) - \frac{m^2\epsilon^2}{4}| \leq \frac{m^2\epsilon^2}{4}\right] = 1 - \Pr[A] \geq \frac{2}{3}$$

$$\Pr[r-s < \frac{m^2\epsilon^2}{2}] \geq \frac{2}{3}$$

Therefore, if $\|p-q\|_2 \leq \frac{\epsilon}{2}$, T will accept with probability at least $\frac{2}{3}$.

Now suppose $\|p-q\|_2 > \epsilon$. Then $\mathbb{E}[r-s] = m^2\|p-q\|_2^2 > m^2\epsilon^2$. This means that

$$\Pr\left[|(r-s) - m^2\epsilon^2| \leq \frac{m^2\epsilon^2}{4}\right] = 1 - \Pr[A] \geq \frac{2}{3}$$

$$\Pr\left[r-s > \frac{3m^2\epsilon^2}{4}\right] \leq \frac{2}{3}$$

Therefore, if $\|p-q\|_2 > \epsilon$ then T will fail with probability at least $\frac{2}{3}$. ■

2 L_1-Distance Testing

Recall from last lecture that $\|p-q\|_2 \leq |p-q|_1 \leq \sqrt{n}\|p-q\|_2$. An idea would be to use the L_2-distance tester T from above with $\epsilon = \frac{\epsilon_0}{\sqrt{n}}$. If $|p-q|_1 = 0$, then $\|p-q\|_2 = 0$ and T will accept with at least probability $\frac{2}{3}$. If $|p-q|_1 > \epsilon_0$ then $\|p-q\|_2 \geq \frac{|p-q|_1}{\sqrt{n}} \geq \frac{\epsilon_0}{\sqrt{n}}$ and T will fail with probability at least $\frac{2}{3}$. So running T with $\epsilon = \frac{\epsilon_0}{\sqrt{n}}$ gives a valid L_1-distance tester. However, the running time of such tester is $O(n^2)$, which is larger than the running time of the naive tester!

What we will do is combine the naive tester with the L_2-distance tester. We will first use the naive tester to discard high probability elements, and then we'll run T on the remaining elements.

L_1-Distance Tester (p, q, ϵ)
- $m \leftarrow O(\max(\epsilon{-}2, 4)n^{\frac{2}{3}}\log n)$
- $S_p \leftarrow m$ samples from p
- $S_q \leftarrow m$ samples from q
- $S \leftarrow S_p \cup S_q$
- discard any element that appears less than $(1 - \frac{\epsilon}{63})mn^{-\frac{2}{3}}$ times in S
- **if** $S \neq \emptyset$
  - $\hat{p_i} \leftarrow$ number of times i appears in S_p
  - $\hat{q_i} \leftarrow$ number of times i appears in S_q
  - **fail if** $\displaystyle\sum_{i \in S} |\hat{p_i} - \hat{q_i}| > \frac{\epsilon m}{8}$
  - define p' as output of following process
    - sample $x \in p$
    - **if** $x \notin S$ output x
    - **else** output $x \in U_D$
  - define q' similarly
- **else** $p' \leftarrow p$, $q' \leftarrow q$

run $T(p', q', \frac{\epsilon}{2\sqrt{n}})$ with $O(\frac{n^{\frac{2}{3}}}{\epsilon^4})$ samples

Notice that S contains high-probability elements. We will call them "heavy" elements. It is possible to prove that with probability at least $1 - \frac{1}{n}$, S contains all "heavy" elements. We formalize this as follows.

Definition 6 *An element $i \in D$ is* heavy *if $p_i > n^{-\frac{2}{3}}$ or if $q_i > n^{-\frac{2}{3}}$.*

Lemma 7 *With probability at least $1 - \frac{1}{n}$, for every i:*

1. *If $p_i \geq \epsilon^2 n^{-\frac{2}{3}}$ then $|\hat{p}_i - p_i| < \frac{\epsilon}{63} \max(p_i, n^{-\frac{2}{3}})$, or in other words, if i is heavy then $i \in S$.*

2. *If $p_i < \epsilon^2 n^{-\frac{2}{3}}$ then $\hat{p}_i < (1 - \frac{\epsilon}{63}) n^{-\frac{2}{3}}$, or in other words, if i is not heavy then $i \notin S$.*

Proof Idea Use Chernoff bounds. ■