6.896 Sublinear Time Algorithms	February 13, 2007

Lecture 3

Lecturer: Ronitt Rubinfeld — *Scribe: Jeremy Hurwitz*

In this lecture, we will prove the correctness of the equality tester for L_1.

L_1-Dist-Test(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

T is a tester described in the last lecture which passes p, q for which the L_2 distance is at most $\epsilon/2$ and fails p, q for which the L_2 distance is at least ϵ. We assume that the error probability is at most $1/6$.

1 Useful Claims, Observations, and a Lemma

We begin by proving a series of claims. We will then analyze the behavior of the algorithm.

Claim 1 $|S| \leq 2n^{2/3}(1 - \frac{\epsilon}{63})^{-1}$

Proof If $|S| > 2n^{2/3}(1 - \frac{\epsilon}{63})^{-1}$, then

$$\begin{aligned}\#samples &\geq |S| \cdot (1 - \epsilon)mn^{-2/3} \\ &> 2n^{2/3}(1 - \frac{\epsilon}{63})^{-1} \cdot (1 - \epsilon)mn^{-2/3} \\ &= 2m\end{aligned}$$

But we only have $2m$ samples. So we have a contradiction. ■

Lemma 2 *With probability greater than $1 - \frac{1}{6}$, for all i*

$$(\star)\begin{cases}\text{(1) If } p_i \geq \epsilon^2 n^{-2/3}, \text{ then } |\hat{p_i} - p_i| < \frac{\epsilon}{63}\max(p_i, n^{-2/3}) \\ \text{(2) If } p_i \leq \epsilon^2 n^{-2/3}, \text{ then } \hat{p_i} < (1 - \frac{\epsilon}{63}n^{-2/3})\end{cases}$$

This means that, most of the time, the "heavy" p_is are well approximated and "very light" p_is will be discarded from S. Anytime I refer to Lemma 2 "holding", I mean that (1) and (2) are satisfied by the current sampling.

Proof The lemma follows from Chernoff bounds. ■

Claim 3 *If $(\star)$ holds, then*

$$\left|\sum_{i\in S}|\hat{p}_i-\hat{q}_i|-\sum_{i\in S}|p_i-q_i|\right|\le\frac{\epsilon}{9}$$

Proof Since $(\star)$ holds, p_i and q_i are well approximated.

Therefore, given $i\in S$,

$$\begin{aligned}||\hat{p}_i-\hat{q}_i|-|p_i-q_i|| &< \frac{\epsilon}{63}(\max(p_i,n^{-2/3}))+\max(q_i,n^{2/3})\\ &< \frac{\epsilon}{63}(p_i+q_i+2n^{-2/3})\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{i\in S}||\hat{p}_i-\hat{q}_i|-|p_i-q_i|| &< \sum_{i\in S}\frac{\epsilon}{63}(p_i+q_i+2n^{-2/3})\\ &\le \frac{\epsilon}{63}(1+1+2n^{-2/3}\cdot|S|)\\ &\le \frac{\epsilon}{9}\end{aligned}$$

for sufficiently small ϵ, by Claim 1. ■

Observation 4

(a) For all $i\notin S$, $p'_i=p_i+\Pr[S]\frac{1}{|D|}$

(b) For all $i\in S$, $p'_i=\Pr[S]\frac{1}{|D|}$

(c) $\sum_{i\notin S}|p_i-q_i|=\sum_{i\notin S}|p'_i-q'_i|$

Claim 5 *If $(\star)$ holds, then for all $i\notin S$, $p'_i<\frac{1}{|D|}+n^{-2/3}$.*

Proof Suppose that $p'_i\ge\frac{1}{|D|}+n^{-2/3}$. Then $p_i\ge n^{-2/3}$.

Assuming that $(\star)$ holds, p_i is approximated well. So

$$\hat{p}_i\ge p_i-\frac{\epsilon}{63}\max(p_i,n^{-2/3})=p_i(1-\frac{\epsilon}{63})\qquad\ge n^{-2/3}(1-\frac{\epsilon}{63})$$

which would mean that i was in S, contradicting the requirement on S. ■

Finally, we consider the number of samples that the L_2 algorithm reqiures. Given parameters

$$\epsilon'=\frac{\epsilon}{2\sqrt{n}}\qquad b=n^{-2/3}+\frac{1}{|D|}<2n^{-2/3}$$

we need

$$m\ge c(b^2+\sqrt{b}\epsilon^2)\epsilon^{-4}=O(n^{2/3}\epsilon^{-4})$$

samples.

2 Behavior of the L_1 tester

We analyze the algorithm separately for the case $p=q$ and the case $|p-q|_1>\epsilon$.

2.1 $p = q$

Since $p = q$, $\sum_i |p_i - q_i| = 0$. By Claim 3, and assuming ($\star$) holds,

$$\sum_i |\hat{p}_i - \hat{q}_i| \leq \frac{\epsilon}{9} < \frac{\epsilon}{8}$$

which means that the distribution does not fail the first test with probability $1 - \frac{1}{6}$.

Since $p = q$ originally, $p' = q'$, so the L_2 test returns the correct answer with probability $1 - \frac{5}{6}$.

So the total failure rate is bounded by $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$, as desired.

2.2 $|p - q|_1 > \epsilon$

Since $|p - q|_1 > \epsilon$, at least one of

$$\text{(I)} \sum_{i \in S} |p_i - q_i| > \frac{\epsilon}{2} \quad \text{or} \quad \text{(II)} \sum_{i \notin S} |p_i - q_i| > \frac{\epsilon}{2}$$

is true.

Case I

If (I) holds, then assuming ($\star$) holds,

$$\sum_{i \in S} |\hat{p}_i - \hat{q}_i| \geq \frac{\epsilon}{2} - \frac{\epsilon}{9} > \frac{\epsilon}{8}$$

which means that the algorithm correctly returns "fail". So the algorithm returns correctly at least $1 - \frac{1}{6}$ of the time in this case.

Case II

If (II) holds, then

$$\sum_{i \notin S} |p_i - q_i| \geq \frac{\epsilon}{2}$$

which means that

$$\sum_{i \notin S} |p'_i - q'_i| \geq \frac{\epsilon}{2}$$

by Observation 4. Therefore,

$$||p' - q'||_2 \geq \frac{\epsilon}{2\sqrt{n}}$$

so the L_2 test runs correctly with probability $1 - \frac{5}{6}$.

So the total failure rate is again bounded by $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$, as desired.