0368.416701 Sublinear Time Algorithms November 3, 2014

Lecture 2

Lecturer: Ronitt Rubinfeld *Scribe: O. Salzman, D. Shaharabani and R. Hollander*

1 Property Testing for the Element Distinctness Problem

We discuss the *Element Distinctness Problem*: Given a set of elements $X = \{x_1, \ldots, x_n\}$, determine if the elements are distinct. I.e. does $x_i \neq x_j$ hold for every $i \neq j$. Clearly, to answer this problem exactly we need at least linear time as we need to look at every input.

Instead of answering the above problem, we will solve the following *property test*:

- If all elements of X are distinct, output **PASS**.
- If all more than εn duplicates occur, output **FAIL** with probability larger than $\frac{3}{4}$.
- Otherwise, any answer is acceptable.

1.1 Sub-linear time algorithm

To solve the property testing problem, we run the following algorithm,

Algorithm 1 Property Testing for Element Distinctness Problem

```
1: Sample c√n/ε independent samples from X
2: Test for duplicates among the sampled elements
3: if Found duplicate then
4:    return FALSE
5: else
6:    return TRUE
```

Before starting the analysis of Alg. 3, we note the following:

Remark Testing for duplicates can be done in time linear in the amount of samples using hashing. Thus, the total time-complexity of the algorithm is $\frac{c\sqrt{n}}{\varepsilon}$.

Remark The above property (and the algorithm) does not depend on the *order* of the elements. This is called a *symmetric* property. Testing if a list of elements is sorted (see Lecture 1) is an example of a property that is not symmetric. Usually, when dealing with symmetric problems, we can not do any better than random sampling of the input.

Remark Clearly, if the set of elements X are distinct, Alg. 3 will output TRUE. Thus, it can only fail to report a correct answer if the elements are not distinct (i.e. there are duplicates) and it returns TRUE. This is an example of a *one-sided error*.

1.2 Analysis of Alg. 3

As mentioned, if the set of elements X are distinct, the algorithms output is indeed correct. Thus, we wish to show that if there are more than εn elements, the algorithm will output FAIL with high probability (w.h.p). To do so, our analysis will:

1. Pair off duplicate elements of X.
2. Argue that as we sample $\frac{c\sqrt{n}}{\varepsilon}$ elements, we are likely to hit both members of a certain pair.

1.2.1 Pair off duplicate elements of X

We divide the list into *ordered pairs* of duplicate elements and let P denote the set of ordered pairs. Note that (i) elements without duplicates do not get paired and (ii) if there is an odd number of duplicates, the last one does not get paired.

For example, let $X = 1, 1, 1, 2, 3, 4, 4, 5, 5, 5, 5, 5$. The pairs are: $(1, 2), (6, 7), (8, 9), (10, 11)$. Why? There are three elements whose value equals to one: x_1, x_2 and x_3. We pair the first and the second (denoted by the pair $(1, 2)$ due to their indices) while the third is not paired. Elements x_4, x_5 are distinct and thus they are not paired. There are two elements whose value equals to four: x_6, x_7. They are paired (denoted by the pair $(6, 7)$ due to their indices). Finally, there are five elements whose value equals to five: x_8, x_9, x_{10}, x_{11} and x_{12}. We pair the first four elements (denoted by the pairs $(8, 9)$ and $(10, 11)$) while the fifth is not paired.

Claim 1 *If the number of distinct elements is less than* $(1 - \varepsilon)n$*, then* $|P| \geq \frac{\varepsilon n}{3}$.

I.e. if there are many duplicates (and hence the algorithm should output false), the number of ordered pairs of duplicates is large (more than $\frac{\varepsilon n}{3}$)

The intuition behind the claim is as follows: if there were only an even number of duplicates, all of them would be paired and then $|P| = \frac{\varepsilon n}{2}$. Now, the fact that we do not pair all of the elements means that $|P|$ is smaller. What the claim says is that it is not much smaller.

Formally, we say that a value c is duplicated if there exists at least two indices $i \neq j$ such that the elements x_i, x_j equal c ($x_i = x_j = c$). We denote by I_c be the set of indices i such that $x_i = c$. Clearly, if $|I_c| > 1$ (i.e. the value c is duplicated), the number of indices in I_c that are not paired is zero if $|I_c|$ is even and one if $|I_c|$ is odd.

Thus, the number of elements who's value is duplicated that are not paired is equal to the number of elements c such that (i) $|I_c| > 1$ and (ii) $|I_c|$ is odd. This means that for all the above values, $|I_c| \geq 3$. From this, one can deduce that at least $\frac{2}{3}$ of the above values are paired. Recall that P counts the number of pairs (and not elements - each pair contains two elements) and that there are εn duplicate elements, thus $|P| = \frac{\varepsilon n}{3}$.

1.2.2 Both Members of a Pair are Hit w.h.p

To argue that as we sample $\frac{c\sqrt{n}}{\varepsilon}$ elements, we are likely to hit both members of a certain pair, we will reformulate Alg. 3 (which allows for an easier analysis).

Algorithm 2 Property Testing for Element Distinctness Problem (reformulation)

```
1: Sample S_1 = √n independent samples from X
2: Test for duplicates among S_1
3: if Found duplicate then
4:     return FALSE
5: Sample S_2 = c√n/ε independent samples from X
6: Test for duplicates among S_1 ∪ S_2
7: if Found duplicate then
8:     return FALSE
9: else
10:    return TRUE
```

Remark If we prove that Alg. 2 solves the property testing problem then Alg. 3 solves it as well.

Our analysis of Alg. 2 will look at each phase independently. We will argue that the samples in S_1 (line 1) include many elements x_i such that i is in the first part of an ordered pair (we say that these

pairs were *hit*). Then, we will claim that the samples in S_2 (line 5) include with high probability an element x_j such that j is in the second part of an ordered pair among the ones that were hit.

Formally, let $F = \{i \mid i \text{ is the 1st number of an ordered pair and } i \text{ is hit by } S_1\}$. The following Lemma shows that F will be large enough with high probability.

Lemma 2 *$|F| \geq \frac{\varepsilon\sqrt{n}}{12}$ with probability greater than $\frac{5}{6}$.*

Proof

We begin with an attempt at a proof, which we will later refine in order to get a correct proof. Observe that by Claim 1, we get that

$$\Pr[\text{Some sample in } S_1 \text{ hits a first element of a pair in } P] \geq \frac{|P|}{n} \geq \frac{\varepsilon}{3} .$$

We define an indicator variable y_k,

$$y_k = \begin{cases} 1, & \text{sample } k \text{ hits a first element of a pair in } P \\ 0, & \text{otherwise} \end{cases}$$

We get that the expected value of y_k is

$$\mathrm{E}[y_k] = \Pr[y_k = 1] \geq \frac{\varepsilon}{3} .$$

Define $Y \equiv \sum_{k \in S_1} y_k$, then the expected value of Y is

$$\mathrm{E}[Y] = \sum \mathrm{E}[y_k] \geq |S_1| \frac{\varepsilon}{3} = \frac{\varepsilon\sqrt{n}}{3} ,$$

where the first equality follows by linearity of expectation.

Note that Y describes the total number of times that a first element of a pair in P was hit by a sample in S_1. The problem with that is that it includes repetitions, that is, if the same pair was hit several times, it would be counted several times in Y. Instead, we would like to observe the total number of pairs whose first element was hit by a sample in S_1. In addition, so far we have only shown that the expected value of Y is large, while in fact, we would like to show that Y is large enough with high probability.

Therefore, in order to complete the proof we will show that: (i) Y is large enough with high probability; (ii) $|F| \geq Y - 6$ with high probability.

We begin with showing (i). Notice that $y_1, \ldots, y_k$ are independent random variables, and we can therefore use the Chernoff bound, getting

$$\Pr[Y < \frac{\mathrm{E}[Y]}{2}] < \mathrm{e}^{-\frac{1}{4} \cdot \mathrm{E}[Y] \cdot \frac{1}{2}} = \mathrm{e}^{-\frac{1}{4} \cdot \frac{\varepsilon\sqrt{n}}{3} \cdot \frac{1}{2}} = \mathrm{e}^{-\frac{1}{8} \cdot \frac{\varepsilon\sqrt{n}}{3}} \ll \frac{1}{12} .$$

We conclude that

$$\Pr[Y \geq \frac{\varepsilon\sqrt{n}}{6}] \geq \frac{11}{12} ,$$

thus proving (i).

For proving (ii), let (k,l)-*sample collision* indicate whether samples k and l hit the same element of the input, and define the indicator variable

$$Z_{kl} = \begin{cases} 1, & \text{there is a } (k,l)\text{-sample collision} \\ 0, & \text{otherwise} \end{cases}$$

It is easy to see that $Z = \sum_{k<l} Z_{kl}$ is an upper bound on the total number of repetitions in Y, that is, the number of times where a pair was hit twice or more. In fact, this upper bound is very crude – if a pair was hit m times, it will be counted $\binom{m}{2}$ times. Therefore, $|F| \geq Y - Z$. In addition, $\mathrm{E}[Z_{kl}] = \Pr[Z_{kl}] = \frac{1}{n}$, and by linearity of expectation we get that $\mathrm{E}[Z] = \binom{\sqrt{n}}{2} \cdot \frac{1}{n} = \frac{\sqrt{n}(\sqrt{n}-1)}{2} \cdot \frac{1}{n} = \frac{1}{2} - \frac{1}{2\sqrt{n}} < \frac{1}{2}$. Using Markov's inequality, we get that $\Pr[Z \geq 6] \leq \frac{\mathrm{E}[Z]}{6} < \frac{1}{12}$, thus showing that $|F| \geq Y - 6$ with probability at least $\frac{11}{12}$.

From (i), we get that $\Pr[Y < \frac{\varepsilon\sqrt{n}}{6}] < \frac{1}{12}$, and from (ii) we get that $\Pr[Z \geq 6] < \frac{1}{12}$. Using union bound, the probability of something bad happening, that is, either Y is too small, or Z is too large, is at most $\frac{1}{6}$. Therefore, we get that $|F| \geq Y - 6 \geq \frac{\varepsilon\sqrt{n}}{6} - 6 \geq \frac{\varepsilon\sqrt{n}}{12}$ with probability at least $\frac{5}{6}$, thus proving the lemma. ■

We will now show that if F is large enough, then the second sampling step is likely to hit the second element of an element in F.

Lemma 3 *If $|F| \geq \frac{\varepsilon\sqrt{n}}{12}$ then S_2 samples an element which is associated with the second part of an ordered pair among the ones that were hit with probability greater than $\frac{11}{12}$.*

Proof We observe some sample $x_j \in S_2$. Then since $|F| \geq \frac{\varepsilon\sqrt{n}}{12}$ we get that

$$\Pr[x_j \text{ is paired with some element in } F] \geq \frac{\frac{\varepsilon\sqrt{n}}{12}}{n} = \frac{\varepsilon}{12\sqrt{n}} .$$

Therefore, the probability that there is some sample in S_2 that is paired with an element in F is

$$\Pr[\text{some sample in } S_2 \text{ is paired with some element in } F] \geq$$
$$1 - (1 - \frac{\varepsilon}{12\sqrt{n}})^{\frac{c\sqrt{n}}{\varepsilon}} = 1 - (1 - \frac{\varepsilon}{12\sqrt{n}})^{\frac{12\sqrt{n}}{\varepsilon}\frac{c}{12}} \geq 1 - \mathrm{e}^{-\frac{c}{12}} \geq \frac{11}{12} ,$$

where the last inequality follows by picking c such that $\mathrm{e}^{-\frac{c}{12}} < \frac{1}{12}$. ■

From the previous lemmas, the following theorem easily follows.

Theorem 4 *Alg. 2 is a property tester with sample complexity of $O(\frac{\sqrt{n}}{\varepsilon})$.*

Proof It suffices to show that if there are at least εn duplicates in the input, then the property tester will output the correct result with probability at least $\frac{3}{4}$. From Lemma 2, the property tester will fail at hitting enough pairs in the first sampling step with probability at most $\frac{1}{12}$. From Lemma 3, and assuming that the first sampling step hits enough pairs, the second sampling step will fail at finding a duplicate with probability at most $\frac{1}{6}$. Therefore, using union bound, the probability of failure is at most $\frac{1}{6} + \frac{1}{12} = \frac{3}{12}$, and the probability of success is at least $\frac{3}{4}$, as required. ■

2 Turning to A New Model

Let $D = \{1, 2, ..., n\} = [n]$ be a domain, $|D| = n$.
Denote by P be a distribution and $P_i = Pr[P \text{ outputs } i] \leftarrow$ unknown.
P outputs iid samples (this is all we can learn from).
Examples: Lottery data, shopping choices, experimental outcomes, etc.
What do we want to know?

1. Is it uniform?

2. Is it high enthropy?

3. Is p monotone increasing, k-model, monotone hazard rate...?

How can we do it?

1. χ^2 test.

2. Plug in estimate.

3. Learn distribution, maximum likelihood estimates.

Goal: Sample complexity SUBLINEAR in n,

2.1 Testing uniformly

Let U be a uniform distribution on the domain D and let P be a given distribution. Let $dist(P,U)$ be a distance measure between U and P (we will see two possible definitions for the distance later).

The goal is to output the following with probability $\geq \frac{3}{4}$:

1. If P =U[n] (P distributed uniformly), then tester outputs PASS.

2. If $dist$(P,U[n]) $> \Delta$, then tester outputs FAIL.

Denote by l the distance between two elements in D.

For distributions p and q over domain D we define the distances l_1 and l_2 between p and q. Both l_1 and l_2 keep the symmetry property and the triangle inequality (other known distances: KL-divergence, Earthmover, Jensen-Shammon).

Definition 5 *Let* $l_1 \equiv ||p-q||_1 = \Sigma_{i\in D}|p_i - q_i|$.

Definition 6 *Let* $l_2 \equiv ||p-q||_2 = \sqrt{\Sigma_{i\in D}(p_i - q_i)^2}$.

Property of l_1 and l_2:

$$||p-q||_2 \leq ||p-q||_1 \leq \sqrt{n}||p-q||_2 \ .$$

Example 1 *Let* $p = (1, 0, ..., 0)$, $q = (\frac{1}{n}, ..., \frac{1}{n})$.
Then,

$$l_1 = \frac{n-1}{n} + (n-1)\frac{1}{n} \approx 2,$$

and

$$(l_2)^2 = (1 - \frac{1}{n})^2 + (n-1)(\frac{1}{n})^2 \approx 1.$$

Example 2 *Let* $p = (\frac{2}{n}, ..., \frac{2}{n}, 0, ..., 0)$, $q = (0, ..., 0, \frac{2}{n}, ..., \frac{2}{n})$.
Then,

$$l_1 = n\frac{2}{n} = 2,$$

and

$$(l_2)^2 = n(\frac{2}{n})^2 = \frac{4}{n} \Rightarrow l_2 = \frac{2}{\sqrt{n}}.$$

Algorithm 3 "Plug-in" Estimate Algorithm

1: Take m samples from p
2: Estimate $p(i)$ for each $i \in [n]$: $\widehat{p}(i) = \sharp$ of times i occurs in sample$/m$
3: **if** $\sum_i |\widehat{p}(i) - \frac{1}{n}| > \Delta$ **then**
4: **return** FALSE
5: **else**
6: **return** TRUE

Analysis (better analysis exist):
Pick m s.t. with high probability
Foreach i,

$$|\widehat{p} - p(i)| < \frac{\Delta}{n}$$

by

$$\Delta \neq ||p - u||_1 \leq ||\widehat{p} - u||_1 + \Delta$$

.

Q: How many samples should we take?
A: $\Omega(\frac{n}{\epsilon})$, maybe even worse. For getting a good approximation by chernoff bound, a linear number of samples in needed to estimate P.
Q: Can we do better?
A: Yes. Instead of trying to estimate $p(i)$ for each i in the domain, try to estimate the total probability of collision of P.
We will use the l_2-distance:
The collision probability of distribution P is:

$$Pr_{s,t \in P}[s = t] = \sum_i p_i^2 = (||P||_2)^2$$

.

$$(||P-U||_2)^2 = \sum_{i \in [n]} (p_i - \frac{1}{n})^2 = \sum_i p_i^2 - 2\sum p_i \frac{1}{n} + \sum_i (\frac{1}{n})^2 = \sum_i p_i^2 - \frac{2}{n} + \frac{1}{n} = \sum_i p_i^2 - \frac{1}{n} = (||P||_2)^2 - \frac{1}{n}.$$

Observe that for a uniform distribution U,

$$(||U||_2)^2 = \sum \frac{1}{n^2} = \frac{1}{n},$$

and for any other distribution P,

$$(||P||_2)^2 > \frac{1}{n}.$$

Corollary 7 *We can find the distance between P and U using the known fact that $(||U||_2)^2 = \frac{1}{n}$, and the collision probability of P, known as $(||P||_2)^2 > \frac{1}{n}$, that we can estimate.*