Introduction to Property Testing

Lecture 1

Lecture by Dr. Reut Levi
Typeset by Steven Karas

2020-04-01
Last edited 18:24:40 2020-04-01

Disclaimer These notes are based on the lectures for the course Introduction to Property Testing, taught by Dr. Reut Levi at IDC Herzliyah in the spring semester of 2019/2020. Sections may be based on the lecture slides prepared by Dr. Reut Levi.

1 Course Overview

The classical approach is to use polytime algorithms that read the entire input. Property testing is a subfield of sublinear algorithms that provide probabilistic algorithms that decide on a particular property of the input. Of particular interest to the field is the concept of query complexity.

Approximation algorithms provide a valid solution that has cost within some distance of the optimum.

Approximating a decision problem requires either allowing a probabilistic answer, or by defining an answer as expressed by some distance metric.

2 Course administration

5 homework assignments, each worth 20 except the last which is worth 25. The maximal grade in the course is 100, but the maximum points possible is 105.

3 Agenda

- Binary Sequences

4 Binary Sequences

We will consider today binary sequences $\{0,1\}^*$. We will first use the relative Hamming distance as our distance metric.

Let $x, z \in \{0,1\}^*$ be binary sequences. Let $\delta(x,z) = \begin{cases} |\{i \in [1x1] : x_i \neq z_i\}| & \text{if } |x| = |z| \\ \infty & \text{otherwise} \end{cases}$

Let S be some set of binary sequences. We define the distance of x from S as:

$$\delta_S(x) = \min_{z \in S}\{\delta(x,z)\}$$

We want to distinguish between sequences that are more distant than some parameter ε

Distinguishing for randomized algorithms is defined as if $x \in S$ then accepts with probability $> \frac{1}{2}$, and if x is more than ε far from S then rejects with probability $> \frac{1}{2}$[1].

4.1 Majority

We define the property of majority as the set of all sequences which have strictly more than half of their bits as 1.

[1] Any constant greater than $\frac{1}{2}$ can be used. In the slides $\frac{2}{3}$ is used

$$\text{MAJ} = \left\{ x : \sum_{i=1}^{|x|} x_i > \frac{|x|}{2} \right\}$$

4.1.1 Algorithm

Proposition 1: There exists a $O(\frac{1}{\varepsilon^2})$ time algorithm that distinguishes between $x \in \text{MAJ}$ and x that is ε far from MAJ.

Proof of proposition 1: Popular form of Chernoff's bound:
Given a sequence of random i.i.d. samples $x_1, \ldots, x_n$ in the range $[0, 1]$.

$$\forall \varepsilon \in (0, 1] \ \Pr\left[\left| \frac{1}{n} \sum_{i \in [n]} x_i - p \right| > \varepsilon \right] < 2e^{\frac{-\varepsilon^2 n}{4}}$$

The important takeaway from Chernoff's bound is that the number of queries required is not based on the size of the population, but rather on the desired tolerance.

Algorithm: Query $\theta(\frac{1}{\varepsilon^2})$ random locations and accept iff $\tilde{p}$ = fraction of 1s exceeds $\frac{1-\varepsilon}{2}$.
This algorithm is correct by Chernoff's bound with probability $\geq \frac{2}{3}$ because $|\tilde{p} - p| \leq \frac{\varepsilon}{2}$.
If $x \in \text{MAJ} \Rightarrow p > \frac{1}{2} \Rightarrow \tilde{p} > \frac{1}{2} - \frac{\varepsilon}{2}$ with probability $\geq \frac{2}{3}$. This means we accept with probability $\geq \frac{2}{3}$.
If x is ε-far from MAJ $\Rightarrow p < \frac{1}{2} - \varepsilon \Rightarrow$ with probability $\geq \frac{2}{3}$. $\tilde{p} < \frac{1}{2} - \varepsilon + \frac{1}{2}$.

4.1.2 Lower bound on number of queries

Any randomized algorithm that exactly decides MAJ must make $\Omega(n)$ queries.

Proof: Let D_1 be a distribution over No-instances: uniform[2] over strings with Hamming weight $\lfloor \frac{n}{2} \rfloor$.
Let D_2 be a distribution over Yes-instances[3]: uniform over strings with Hamming weight $\lfloor \frac{n}{2} \rfloor + 1$.
We claim that if $X_n \sim D_1$ and $Y_n \sim D_2$ then

$$|\Pr[A(X_n) = 1] - \Pr[A(Y_n) = 1]| \leq \frac{q}{n}$$

where A is an algorithm that makes q queries.
For any A that makes $< \frac{n}{3}$ queries:

$$|\Pr[A(X_n) = 1] - \Pr[A(Y_n) = 1]| < \frac{1}{3}$$

which implies that either $\Pr[A(X_n) = 1] > \frac{1}{3}$ or that $\Pr[A(Y_n) = 1] < \frac{2}{3}$. In the first case the algorithm does not sufficiently distinguish instances that do not have a majority. In the second case the algorithm does not sufficiently distinguish instances that do have a majority.

4.1.3 Limitations of deterministic algorithms

Any deterministic algorithm that distinguishes between MAJ and 0.5-far from MAJ must make $\frac{n}{2}$ queries.

Example:

0000 0000 0 - 0.5 far
0000 1111 1 - MAJ

[2] A uniform distribution is defined as a distribution over a set (e.g. $\{0, \ldots, n\}$) with equal probability to choose any element in the set.
[3] As mentioned, a typical approach for proving lower bounds is to define two subsets and show that it is difficult to distinguish between the two subsets

5 For next time

Due to passover, the next lecture will take place in three weeks. Ensure we have access to the book[1], read the appendix A (labeled as Probabilistic Preliminaries).

References

[1] Oded Goldreich. *Introduction to property testing.* Cambridge University Press, 2017.