6.889 Sublinear Time Algorithms April, 29 2019

Lecture 21

Lecturer: Ronitt Rubinfeld *Scribe: Malvika Joshi*

1 Topics Covered

We continued lower bounds with techniques from communication complexity and also covered an overview of a closeness testing lower bound. These were the topics covered in class.

- Another $\tilde{\Omega}(\sqrt{n})$ samples lower bound on uniformity testing. This is looser than the previous lower bound we have showed on the pset which was $\Omega(\sqrt{n})$, but will illustrate useful techniques.
- Sketch of $\Omega(n^{2/3})$ samples lower bound for testing closeness of two distributions.

2 Simultaneous Message Passing Model

This communication model which will be used today is different from last lecture. The participants in the model are *Alice*, *Bob* and a *referee*. Alice and Bob have private coins and they each get inputs x and y resp. Alice will compute message m_a and Bob will compute m_b independently using their own coins and send them to the referee. The goal of the referee is to compute $f(x, y)$ while knowing m_a and m_b. There is only one round of message passing and complexity of the protocol is the number of bits in m_a and m_b.

2.1 Known lower bound

Suppose Alice and Bob want to compute f where:

$$f(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

then the following theorem is known.

Theorem 1 *Alice and Bob need to communication $\Omega(\sqrt{k})$ bits to the referee to compute $f(x, y)$, where $|x| = |y| = k$.*

3 Error Correcting Codes

We will use some known facts about the existence of error correcting codes C with the following properties:
$\exists\ C : \{0, 1\}^k \to \{0, 1\}^n$ such that,

1. $n = O(k)$ (so $k/n \geq$ constant)
2. The fraction of bits different in two different code numbers, $\delta = \Omega(1)$.
3. Each code number, $y \in \text{Range}(C)$ has an equal number of 0 and 1 bits.

4 Uniformity testing lower bound

We use the error correcting code C and the uniformity tester T that uses s samples, with error parameter δ to give the following protocol for computing $f(x, y)$.

4.1 The protocol

Alice: given x

- Compute $C(x)$
- Compute $A = \{i \mid C(x)_i = 1\}$
- Send s uniformly distributed samples from A (distribution U_A) to *referee.*

Bob: given y

- Compute $C(y)$
- Compute $B = \{i \mid C(y)_i = 0\}$
- Send s uniformly distributed samples from B (distribution U_B) to *referee.*

The referee does the following:

1. Receives samples from Alice and Bob, m_A and m_B.
2. Construct s samples of the distribution $q = \frac{U_A+U_B}{2}$. The ith sample is obtained by tossing a coin and if heads, getting the next sample from m_A, otherwise getting the next sample from m_B.
3. Run the uniformity tester T on these samples and if it accepts then output 1, otherwise output 0.

The referee will get s samples from each Alice and Bob, to make sure that even in the case when all the coins output head (or tail), there will be enough samples in m_A (or m_B).

4.2 Correctness

If $x = y$, then $C(x) = C(y)$, and so A and B are equal sized partitions of $[n]$, so $U_{[n]} = \frac{U_A+U_B}{2} = q$. Hence, T accepts q and the protocol will correctly output 1.

If $x \neq y$, then, $C(x) \neq C(y)$ on at least δ fraction of the domain and have equal number of 1s and 0s. Hence, $|A \cap B| = \delta/2$ and $|[n] \setminus (A \cup B)| = \delta/2$. So, q does not have $\delta/2$ fraction of the domain. Therefore, $||q - U_{[n]}||_1 \geq \delta$ and q will be rejected by T and the protocol will correctly output 0.

4.3 Message complexity

Each sample that Alice and Bob send will be encoded in $\log n$ bits, so they send $s \log n$ total bits, and since the communication complexity of calculating $f(x, y)$ has to be $\Omega(\sqrt{k})$, where $|x| = |y| = k$,

$$s = \Omega\left(\frac{\sqrt{k}}{\log n}\right) = \Omega\left(\frac{\sqrt{n}}{\log n}\right) = \tilde{\Omega}(\sqrt{n})$$

.

Using the fact from Section 3 that $n = O(k)$. Hence, T should use $\tilde{\Omega}(\sqrt{n})$ samples.

5 Sketch of Closeness testing Lower Bound

We are given samples from two distributions p and q and the tester should pass if $p = q$ and fail if $||p - q||_1 > \epsilon$. Here is a sketch of obtaining a $O(n^{2/3})$ samples lower bound for the tester.

Consider two distributions p_0 and q_0. Each contain $n^{2/3}$ "heavy" elements, each with weight $\frac{1}{2n^{2/3}}$ and $n/4$ "light" elements, each with weight $2/n$. The light elements in the two distributions are disjoint and the heavy elements are the same.

Then for all possible relabellings of the domain π, $\pi(p_0), \pi(p_0)$ is a positive pair (should pass by tester) and $(\pi(p_0), \pi(q_0))$ is a negative pair, because the L_1 distance of the two, since the light elements are disjoint is $2 \cdot n/4 \cdot 2/n = 1$.

Note that only the statistics of collisions are matter to the tester. This idea is similar to the one in the pset, of tester operating only on the *fingerprint* of the sample. In positive pairs we should have collisions in both heavy and light elements. In negative pairs we should only have collisions (across the two distributions) in heavy pairs. After $o(n^{2/3})$ samples, the probability of seeing any heavy element 4 times or a light element 3 times in the same distribution is very small and can be ignored.

Let $2m$ be the number of samples, for simplicity assume that m are heavy elements and m are light elements. Let H be the number of collisions of heavy elements and L be the number of collisions of light elements. Then, H has the same distribution for positive and negative pairs, but $L = 0$ for negative pairs.

Calculating the expected number of collisions and variance:

$$\text{Exp}[H] = \frac{m^2}{2n^{2/3}} \text{ and } \text{Var}[H] \approx \frac{m^2}{n^{2/3}}$$

$$\text{Exp}[L] \approx \text{Var}[L] \approx \frac{m^2}{n}$$

We need to show that we cannot distinguish the distribution of $H + L$ from the distribution of L. This can be achieved by showing that their L_1 distance is small. If the two distributions were Gaussian then is done by showing $\sqrt{\text{Var}[H]} \geq \text{Exp}[L]$, so intuitively it's hard to detect adding such a small change to H. This is true when $m < n^{2/3}$, because then,

$$\text{Exp}[L] \approx \frac{m^2}{n} < \frac{mn^{2/3}}{n^{1/3}} \leq \frac{m}{n^{1/3}} \approx \sqrt{\text{Var}[H]}$$

However they are not exactly Gaussian, so it requires a bit more work.