0368.416701 Sublinear Time Algorithms December 28, 2009

Lecture 11

Lecturer: Ronitt Rubinfeld *Scribe: Netanel Nimni and Nir Muchtar*

1 Lecture Topic

Testing functions for the k-junta property.

1.1 Last Week

We found a property tester for dictator functions.

We defined k-junta functions and introduced a k-junta testing algorithm.

2 Definitions and Lemmas From Last Week

Definition 1. A function f is a k-junta function if its output depends only on k or less input variables.

Definition 2. $X_s \equiv ordered-list(X_i \mid i \in S)$

Definition 3. $X_s Y_{\bar{s}} \equiv Z$ s.t $\forall_{i\in S}\ Z_i = X_i$ and $\forall_{i\in \bar{S}}\ Z_i = Y_i$

Definition 4. $Inf_f(s) \equiv 2Pr_{x,y|x_{\bar{s}}=y_{\bar{s}}}\left[f(x) \neq f(y)\right]$

Last week we showed the following:

$Inf_f(s) \equiv 2Pr_{x,y|x_{\bar{s}}=y_{\bar{s}}}\left[f(x) \neq f(y)\right] = \sum\limits_{T|S\cap T\neq\phi} \hat{f}(T)^2 = \sum\limits_{T} \hat{f}(T)^2 - \sum\limits_{T\subseteq\bar{S}} \hat{f}(T)^2 = 1 - \sum\limits_{T\subseteq\bar{S}} \hat{f}(T)^2$

where the last equality is derived from the boolean Parseval equality.

Lemma 5. If a function f is ϵ-far from being a k-junta then

$\forall_{J:|J|\leq k} Inf_f([n] \setminus J) \geq \epsilon$

3 Main Theorem

3.1 Algorithm For Testing The k-junta Property

given k, ϵ

- randomly partition $1..n$ into s parts $I_1 ... I_s$ where $s = poly(k, \frac{1}{\epsilon})$
- $R \leftarrow \phi$
- repeat up to $r = O(\frac{k}{\epsilon})$ times
  - generate x, y randomly s.t. $x_R = y_R$
  - if $f(x) = f(y)$
    - use binary search to find relevant I_j
    - $R \leftarrow R \cup I_j$
  - if R has more than k relevant parts, REJECT
- ACCEPT

Note: The binary search here is performed on the partitions rather than the input bits themselves. That's why we obtain a running time, which is independent from n

3.2 Main Lemma

If f is $\epsilon - far$ from $k - junta$, and$\mathcal{I}$ is a random partition into $s = \frac{10^{20}k^9}{\epsilon^5}$ parts,
we can get with probability$\geq \frac{5}{6}$:
$\forall_J$ s.t. J is a union of at most k parts of $\mathcal{I}$, $Inf_f([n] \setminus J) \geq \frac{\epsilon}{2}$
.
We'll first see how we can use the lemma to get our desired result.

Claim 6. There exists an Algorithm$\mathcal{T}$ that uses $O(\frac{k}{\epsilon} + klog(k))$ queries such that:

- if a function f is a $k - junta$, $\mathcal{T}$ will always pass on f
- if a function f is $\epsilon - far$ from being a $k - junta$, $Pr[\mathcal{T}\ will\ fail\ f] \geq \frac{2}{3}$

Proof. Let f be a function.

- If f is a $k - junta \Longrightarrow$ any partition of $1..n$ can have at most k relevant partitions (in respect to f),
  and hence the algorithm will ACCEPT.
- If f is $\epsilon - far$ from being a $k - junta \Longrightarrow Pr_{x,y|x_R=y_R}\left[f(x) \neq f(y)\right] = Inf_f([n] \setminus R) \geq \frac{\epsilon}{2}$
  where the first equality is simply the definition and the second inequality is derived from the **main-lemma**.
  $\Longrightarrow E$[time to find more than $k+1$ relevant parts] $\leq (k+1)(\frac{4}{\epsilon})$
  Now, using **Markov's inequality** we get $Pr[don't\ find\ morethenkpartsafter\ 6(\frac{4}{\epsilon})(k+1)] \leq \frac{1}{6}$
  In addition to the error probability of $\frac{1}{6}$ that we get from **Markov's inequality**, we also get an error probability of $\frac{1}{6}$ from the **main-lemma**,
  so altogether we get an error probability of $\frac{1}{3} \Longrightarrow Pr[\mathcal{T}\ will\ fail\ f] \geq \frac{2}{3}$

□

3.3 Proving The Main Lemma

What's left is to prove the following:
for most partitions $\mathcal{T}$ & $\forall_{J|J\ is\ a\ union\ of\ k\ parts\ of\ \mathcal{T}}$ s.t. J is a union of k parts of $\mathcal{T}$:
$Inf_f([n] \setminus R) =_{by\ lemma} \sum_{S \subseteq [n]} \hat{f}(S) - \sum_{S \subseteq J} \hat{f}(S)^2 \overset{?}{\geq} \epsilon' \equiv \frac{\epsilon}{2}$
We **pick** $\theta \equiv \left(\epsilon^2 log(\frac{k}{\epsilon})\right)/10^9 k^4$

Definition 7. $Inf_f^{\leq k}(S) \equiv \sum_{T|T \cap S \neq \phi\ and\ |T| \leq 2k} \hat{f}(T)^2$

Definition 8. $Inf_f^{> k}(S) \equiv \sum_{T|T \cap S \neq \phi\ and\ |T| > 2k} \hat{f}(T)^2$

Definition 9. $Inf_f(i) \equiv Inf_f\left(\{i\}\right)$

3.3.1 Nice Property

$Inf_f(S) \leq Inf_f(S \cup S') \leq Inf_f(S) + Inf_f(S')$

The first inequality derives from the fact that $S \cup S'$is possibly a bigger set then just S, and since $Inf_f(S)$ is a sum of positive elements ($Inf_f(S) = \sum\limits_{T|T\cap S\neq\phi} \hat{f}(T)^2$).

The second inequality is true since:

$$\sum_{T|T\cap(S\cup S')\neq\phi} = \sum_{T|T\cap S\neq\phi \text{and } T\cap S'=\phi} + \sum_{T|T\cap S\neq\phi \text{and } T\cap S'\neq\phi} + \sum_{T|T\cap S=\phi \text{and } T\cap S'\neq\phi}$$

$$=_{\text{merging the first two } \sum s} \sum_{T|T\cap S\neq\phi} + \sum_{T|T\cap S=\phi \text{and } T\cap S'\neq\phi}$$

$$\leq_{\text{adding another } \Sigma} \sum_{T|T\cap S\neq\phi} + \sum_{T|T\cap S=\phi \text{and } T\cap S'\neq\phi} + \sum_{T|T\cap S\neq\phi \text{and } T\cap S'\neq\phi}$$

$$=_{\text{merging the last two } \Sigma s} \sum_{T|T\cap S\neq\phi} + \sum_{T|T\cap S'\neq\phi}$$

and so we get that $Inf_f(S \cup S') \leq Inf_f(S) + Inf_f(S')$.

3.3.2 Lemma

$\sum\limits_{i\in[n]} Inf_f^{\leq 2k}(i) \leq 2k$

Proof. $\sum\limits_{i\in[n]} Inf_f^{\leq 2k}(i) = \sum\limits_{i\in[n]} \sum\limits_{T\ :\ i\in T\ and |\ T\ |\leq 2k} \hat{f}(T)^2$

$= \sum\limits_{T\ :|T|\leq 2k} |\ T\ |\ \hat{f}(T)^2 \leq 2k \sum\limits_{T\ :|T|\leq 2k} \hat{f}(T)^2 \leq 2k \sum\limits_{T} \hat{f}(T)^2 = 2k$

□

3.3.3 "Heavy" Index Groups

Definition 10. we define the group H to be the group of the "heavy" indices, in the sense that they have a large influence on f:

$H = \left\{ i \in [n] \mid Inf_f^{\leq 2k}(i) \geq \Theta \right\}$

Corollary 11. $|\ H\ |\leq \frac{2k}{\Theta}$

Proof. otherwise we get:

$\sum\limits_{i\in[n]} Inf_f^{\leq 2k}(i) \geq \sum\limits_{i\in H} Inf_f^{\leq 2k}(i) \geq \sum\limits_{i\in H} \Theta = |\ H\ |\ \Theta \geq \frac{2k}{\Theta}\Theta = 2k$

in contradiction to the last lemma.

□

The previous **corollary** implies that there aren't too many "heavy" indices.
(less then $\frac{2k}{\Theta}$, which is a lot smaller then s).
This is the reason that with a high probability (more then $\frac{17}{18}$) each of these "heavy" indices is on a different part I_i:

Recall that $s = \frac{10^{20}k^9}{\epsilon^5}$ and that $\theta = \frac{\epsilon^2 log(\frac{k}{\epsilon})}{10^9 k^4} \Longrightarrow \frac{72k^2}{\Theta^2} = \frac{72k^2 \cdot 10^{18} k^8}{\epsilon^4 log^2(\frac{k}{\epsilon})} \leq \frac{10^{20}k^{10}}{\epsilon^5} = s$

since $s \geq \frac{72k^2}{\Theta^2}$,

$Pr\left[\forall_{i\in[n]} \mid H \cap I_i \mid\leq 1\right] \geq 1 - Pr\left[\exists_{i\in[n]} s.t. \mid H\cap I_i \mid> 2\right] \geq 1 - \binom{|H|}{2} \cdot \frac{1}{S} \geq 1 - \left(\frac{2k}{\Theta}\right)^2 \cdot \frac{\Theta^2}{72k^2} \geq \frac{17}{18}$

In other words, we have just shown that with high probability, each set of the partition gets at most one member of H.

3.3.4 Breaking Down The influence

We define Partition subsets of J:

$\mathcal{H} = \{S \subseteq J \mid S \subseteq J \cap H, \ \ | S | \leq 2k\}$
$\mathcal{L} = \{S \subseteq J \mid S \nsubseteq J \cap H, \ \ | S | \leq 2k\}$
$\mathcal{B} = \{S \subseteq J \mid | S | > 2k\}$

We now present the influence in terms of the above subsets:

$$Inf_f([n] \setminus R) = \sum_{S \subseteq [n]} \hat{f}(S) - \sum_{S \subseteq J} \hat{f}(S)^2 = \underbrace{\sum_{S \subseteq [n]} \hat{f}(S)^2 - \sum_{S \in \mathcal{H}} \hat{f}(S)^2}_{t_1} - \underbrace{\sum_{S \in \mathcal{L}} \hat{f}(S)^2}_{t_2} - \underbrace{\sum_{S \in \mathcal{B}} \hat{f}(S)^2}_{t_3}$$

We denote the above marked terms as t_1, t_2, t_3 respectively.
We'll prove that $t1$ is sufficiently large and t_2, t_3 are sufficiently small to prove the main lemma.

1. **Evaluating t_1**

   $t_1 = \sum_{S \subseteq [n]} \hat{f}(S)^2 - \sum_{S \in \mathcal{H}} \hat{f}(S)^2 \geq \sum_{S \subseteq [n]} \hat{f}(S)^2 - \sum_{S \subseteq J \cap H} \hat{f}(S)^2 = Inf_f([n] - J \cap H)$

   From 3.3.3, we get that $|J \cap H| \leq k$, (because $Pr\left[\forall_{i \in [n]} \mid H \cap I_i \mid \leq 1\right] \geq \frac{17}{18}$)
   and by last week's lemma we get:
   $t_1 \geq \epsilon$

2. **Evaluating t_2**

   Claim 12. If $s \geq \frac{16k^2}{\epsilon}$ and $\theta \leq \frac{\epsilon^2}{64} k^2 \log(18s)$, then with probability at least $\frac{17}{18}$
   $Inf_{\hat{f}}^{\leq 2k}(I_i \setminus H) \leq \frac{\epsilon}{4k}$ for every $i \in [s]$

   First we'll use the claim to provide an upper bound on t_2.
   The claim implies that $\forall_J$ s.t. J is a union of at most k parts of $\mathcal{I}$, the following is true:
   $\sum_{S \in \mathcal{L}} \hat{f}(S)^2 \leq Inf_{\hat{f}}^{\leq 2k}(J \setminus H) \leq_{\text{nice property+previous claim}} k \cdot \frac{\epsilon}{4k} = \frac{\epsilon}{4}$
   Explanation of the first inequality:
   By definitions, $Inf_{\hat{f}}^{\leq 2k}(J \setminus H) = \underbrace{\sum_{S \mid S \cap (J \setminus H) \neq \phi, \ \ | S | \leq 2k} \hat{f}(s)^2}_{\mathcal{T}}$

   and$\mathcal{L} = \{S \subseteq J \mid S \nsubseteq J \cap H, \ \ | S | \leq 2k\} = \{S \subseteq J \mid S \cap (J \setminus H) \neq \phi, \ \ | S | \leq 2k\}$
   $\mathcal{L} \subseteq \mathcal{T}$ since $\mathcal{T}$ may contain elements that are not in J.
   $\sum_{S \in \mathcal{L}} \hat{f}(S)^2 \leq \sum_{S \in \mathcal{T}} \hat{f}(S)^2$ and the inequality is derived.
   We'll now prove the claim by using Hoeffding's inequality:
   $Pr[s - E(s) \geq nt] \leq e^{\frac{-2n^2k^2}{\sum (b_i - a_i)^2}}$ where $s = \sum_{i=1}^{n} x_i$ and $a_i \leq x_i \leq b_i$.

   Proof. For each $i \in [s]$ we define: $\forall \ j \in [n] \ X_j \equiv \begin{cases} Inf_{\hat{f}}^{\leq 2k}(j) & j \in I_i \setminus H \\ 0 & otherwise \end{cases}$

   $Inf_{\hat{f}}^{\leq 2k}(I_i \setminus H) \leq_{\text{nice property}} \sum_{j \in I_i \setminus H} Inf_{\hat{f}}^{\leq 2k}(j) \leq \sum_{j \in [n]} X_j$

Thus, $E[\sum_{j\in[n]} X_j] \le \sum_{j\in[n]\setminus H} Inf_f^{\le 2k}(j)\cdot \underbrace{E[1_{j\in I_i}]}_{\frac{1}{s}} = \frac{1}{s}\sum_{j\in[n]\setminus H} Inf_f^{\le 2k}(j) \underset{\text{from } 3.3.2}{\le} \frac{2k}{s} \le \frac{\epsilon}{8k}$.

We'll now use Hoeffding's inequality where $a_i = 0$ and $b_i = Inf_f^{\le 2k}(i)$

$$\sum_{i\in[n]\setminus H}(b_i - a_i)^2 = \sum_{i\in[n]\setminus H} Inf_f^{\le 2k}(i) \le max_{i\in[n]\setminus H}(Inf_f^{\le 2k}(i))\cdot \sum Inf(i) \le \theta\cdot 2k$$

$$Pr[\sum x_j - \underbrace{E[\sum_{j\in[n]} X_j]}_{\frac{\epsilon}{8k}} \ge \frac{\epsilon}{8k}] = Pr[\sum xi \ge \frac{\epsilon}{4k}] \le e^{\frac{(\frac{\epsilon}{2k})^2}{\theta\cdot 2k}} \le \frac{1}{18s}$$

The last inequality applies for each index i, and since we have s indices, summing them up will produce probability of no more than $\frac{1}{18}$. □

3. **Evaluating t_3**

All $S \in \mathcal{B}$ are subsets of J, which is a union of less the k parts

Claim 13. If $S > \frac{72ek}{\epsilon}$, then with probability of at least $\frac{17}{18}$,

$$\sum_{S||S|\ge 2k \wedge s\in less\ than\ k\ parts\ of\ \mathcal{I}} \hat{f}(S)^2 \le \frac{\epsilon}{4}$$

Proof. Since $|S| \ge 2k$ We get,

$$Pr[\text{All elements of } S \text{ sent to no more than k parts in } \mathcal{I}] \le \binom{|S|}{k}\left(\frac{k}{|S|}\right)^{2k+1} \le \left(\frac{e|S|}{k}\right)^k\left(\frac{k}{|S|}\right)^{2k+1} \le \frac{\epsilon}{72}$$

We can now put an upper bound on the term's expectation:

$$E\left[\sum_{S||S|\ge 2k\wedge s\in less\ than\ k\ parts\ of\ \mathcal{I}} \hat{f}(S)^2\right] \le \sum \hat{f}(S)^2 \cdot Pr\left[1_{S\in k\ parts\ of\ \mathcal{L}}\right] \le \frac{\epsilon}{72}\sum \hat{f}(S)^2$$

$$=_{by\ boolean\ parseval} \frac{\epsilon}{72}\cdot 1 \le \frac{k}{72}$$ □

To recap,

$$Inf_f([n]\setminus R) = \underbrace{\sum_{S\subseteq[n]} \hat{f}(S)^2 - \sum_{S\in\mathcal{H}} \hat{f}(S)^2}\ - \underbrace{\sum_{S\in\mathcal{L}} \hat{f}(S)^2}\ - \underbrace{\sum_{S\in\mathcal{B}} \hat{f}(S)^2} \equiv t_1 - t_2 - t_3 \ge \epsilon - \frac{\epsilon}{4} - \frac{\epsilon}{4} \ge \frac{\epsilon}{2}$$

The last inequality proves our main lemma, and in turn, our main theorem.