In the last lecture we showed a $n^{3/2}$ lower bound on the formula size of parity. We used the following technical lemma.

**Lemma 1.** $Pr_{a \in R_k}[FSize(f|_a) \leq 4(\frac{k}{n})^{3/2}FSize(f)] \geq \frac{3}{4}$

Today we will exhibit an explicit function whose formula complexity is $\tilde{O}(n^{5/2})$. For this result we need a notion of sub function. Let $f : \Sigma^n \rightarrow \Sigma$ and $g : \Sigma^m \rightarrow \Sigma$, $m \leq n$ be two functions. We say that $g$ is a sub function of $f$ if there exists $a \in R_m$ such that $g = f|_a$. We use the following observation.

**Observation 1.** If $g$ is a sub function of $f$, then $FSize(g) \leq FSize(f)$.

The following theorem is due to Andreev.

**Theorem 1.** There is an explicit function $f$ such that the $FSIZE(f) = \tilde{\Omega}(n^{5/2})$

Fix a function $\psi : \{0, 1\}^{\log n} \rightarrow \{0, 1\}$ such that $FSize(\psi) \geq \frac{2^{\log n}}{10 \log \log n} = \frac{n}{10 \log \log n}$. Define a function $f_\psi(x_1 \cdots x_n)$ as follows: Divide the input $x = x_1 \cdots x_n$ into $b$ blocks each of length $m = n/b$. Let $B_t = x_{tm}x_{tm+1} \cdots x_{tm+b-1}$. Let $\oplus(B_t) = x_{tm} \oplus x_{tm+1} \oplus \cdots \oplus x_{tm+b-1}$. We define $f_\psi(x_1 \cdots x_n) = \psi(\oplus(B_1) \cdots \oplus(B_b))$.

Recall that $R_k$ is the set of all partial assignments that leave $k$ variables unassigned. We will first show that if we randomly pick an assignment $a$ from $R_k$, then with high probability at least one variable from each block will remain unassigned.

Let us first look at the probability that there exists a block in which every variable gets a value. By union bound this probability is at most

$$\sum_{i=1}^{b} Pr[ \text{in block } i \text{ every variable gets a value}].$$

We can view the process of picking $a$ from $R_k$ as randomly picking $k$ variables, and assigning random 0-1 values to remaining variables. Every variable in block $i$ gets a value, if all the $k$ variables picked fall into other blocks. Since each block has $m$ variables, this probability is

$$\left( \frac{n-m}{k} \right) \leq \left( \frac{n-m}{m} \right)^k$$

$$= \left( 1 - \frac{m}{n} \right)^k$$

$$= \left( 1 - \frac{1}{b} \right)^k$$

If we set $k = b \ln(4b)$, then $(1 - \frac{1}{b})^k = (1 - \frac{1}{b})^{b \ln(4b)} = ((1 - \frac{1}{b})^b)^{\ln(4b)} \leq e^{\ln(4b)} = \frac{1}{4b}$.

Thus the probability that there exists a block in which every variable gets a value is at most 1/4. Thus the probability that every block has at least one unassigned variable is at least 3/4. The crucial observation is that if a partial assignment leaves at least one
variable unassigned in each block, then $f_\psi|a$ is a sub function of $\psi$. By Observation 1, if $\text{FSize}(f_\psi|a) \geq \text{FSize}(\psi)$. Thus
\[
\Pr_{a \in R_k}[\text{FSize}(f_\psi|a) \geq \text{FSize}(\psi)] \geq 3/4,
\]
for $k = b \ln(4b)$. By Lemma 1,
\[
\Pr_{a \in R_k}[\text{FSize}(f_\psi|a) \leq 4(\frac{k}{n})^{3/2}\text{FSize}(f_\psi)] \geq \frac{3}{4}
\]
Thus there exist a partial assignment $a$ such that both of the following statements are true.
\[
\text{FSize}(f_\psi|a) \geq \text{FSize}(\psi),
\]
and
\[
\text{FSize}(f_\psi|a) \leq 4(\frac{k}{n})^{3/2}\text{FSize}(f_\psi).
\]
Thus
\[
4(\frac{k}{n})^{3/2}\text{FSize}(f_\psi) \geq \text{FSize}(\psi).
\]
Recall that $\text{FSize}(\psi) \geq n/10 \log \log n$. Since $k = b \ln(4b)$ and $b = \log n$, it follows that $\text{FSize}(f_\psi) = \tilde{\Omega}(n^{2.5})$.

Note that $f_\psi$ is not an explicit function as it is not polynomial-time computable. We now see how we can come with an explicit function. Given a $n$-bit string $y$, view it as a boolean function from $\{0,1\}^{\log n}$ to $\{0,1\}$. Let $g(x,y) = f_y(x)$. Now $f_\psi$ is a subfunction of $g$, and thus $\text{FSize}(g) = \Omega(n^{5/2})$. Now, $g$ is polynomial-time computable.

A few comments on the state of the art. Observe that the above theorem can be improved if the exponent $3/2$ (called shrinkage exponent) from Lemma 1 can be improved. Hastad showed that it can be improved to $2 - o(1)$. This gives an explicit formula with formula complexity $\tilde{\Omega}(n^3)$. This the best known bound for explicit functions.

1. Counting number of subfunctions

We will now exhibit another tool to find lower bounds on the formula size. Recall that a function $g : \Sigma^m \rightarrow \Sigma$ is a subfunction of $f : \Sigma^n \rightarrow \Sigma \ (n \geq m)$, if there is a partial assignment $a \in R_m$ such that $f|a = g$. How many subfunctions can a function have? Consider a function $f$ with many subfunctions. Let $T$ be a formula tree for $f$. For every partial assignment $a$, $T|a$ computes a subfunction of $f$. Intuitively, if $f$ has many subfunctions, then $T$ must be large. Below we make this intuition more precise.

Definition 1. Let $f(x_1, \ldots, x_n)$ be a function. Let $S \subseteq (x_1, \ldots, x_n)$. Let $f_S$: be the set of all subfunctions obtained by assigning values to variables from $\{x_1, \ldots, x_n\} - S$.

Observe that $|f_S| \leq 2^{n-|S|}$.

Definition 2. Let $\phi$ be a formula for $f$ and $S \subseteq (x_1, \ldots, x_n)$. $N(\phi, S) =$ number of times the literals from $S$ appear in $\phi$.

Theorem 2. $\forall f, \forall S \subseteq (x_1, \ldots, x_n) \ N(\phi, S) \geq \frac{1}{4} \log |f_S|$.
Proof. Let $\phi$ be a formula for $f$ and let $T$ be the corresponding binary tree. For each leaf node whose literal is labelled from $S$ consider the path from it to the root. Let $T'$ be the union of all such paths. Note that $T'$ a subtree of $T$ and every node in this $T'$ has indegree 0, 1, or 2.

Let $N(\phi, S) = m$ and let $W$ be the set of nodes from $T'$ whose indegree is 2. We can make the observation that $|W| \leq m - 1$. Partition $T'$ into a disjoint collection of paths $P$ such that each path $p$ in $P$ has the following property: the starting node in $p$ is either a leaf node, or a node in $W$, the end point of $p$ is either a node in $W$ or root, and no internal from $p$ is from $W$. Since every node from $W$ has indegree 2, and every path in $P$ either ends in a node from $W$ or at root, it follows that $|P| \leq 2(|W| + 1).

Let us consider a path $p$ from $P$. Let $p = v_1v_2\cdots v_m$. The node $v_1$ could be a leaf node or a node from $W$. Let us consider the case where $v_1$ is a leaf node and assume that $v_m$ is a node from $W$. Recall that in degree of $v_m$ is 2. Observe that $v_m$ is a leaf node from $S$. Fix an assignment to all leaf nodes from $S$. Let us consider the value that is propagated on the edge $(v_{m-1}, v_m)$? We would like to express this value as a function of (value of) $v_m$. Let us denote the value of $v_m$ as $x$. This depends on the assignment to variables that are not in $S$ and on $v_1$. We would like to express this we fixed an assignment to variables from $S$, this value is either 0, 1, $x$ or $\overline{x}$. For every possible fixing of values to variables from $S$, the value on $(v_{m-1}, v_m)$ is one these four possibilities.

Now build a tree $T''$ as follows: For every path $p \in P$, $p = v_1\cdots v_m$, label the edge $(v_{m-1}, v_m)$ with either 0, 1, $v_1$ or $\overline{v_1}$. The total number of trees that can be built is atmost $4^{|P|} \leq 4^{2|W|+1} \leq 4^{2m}$. Each subfunction from $f_S$ corresponds to one such tree. Thus $|f_S| \leq 4^{2m}$. Thus $m = N(\phi, S) \geq \frac{1}{4} \log |f_S|$.\hfill \Box

We can prove the following theorem due to Nenchporuk using Theorem 2.

**Theorem 3.** Let $f(x_1, \ldots, x_n)$ be a function and let $S_1, \ldots, S_k$ be a partition of $\{x_1, \ldots, x_n\}$. Let $s_i$ be the number of subfunctions obtained by assigning values to $\{x_1, \ldots, x_n\} - S$. Then, $FSIZE(f) \geq \frac{1}{4} \sum_{i=1}^k \log s_i$.

**Proof.** Let $\phi$ be an optimal formula for $f$. Since the sets $S_1, \ldots, S_k$ are mutually disjoint, $FSize(f) \geq \sum_{i=1}^k N(\phi, S_i)$. Thus $FSize(f) \geq \frac{1}{4} \sum_{i=1}^k \log s_i$.\hfill \Box

We can use Nenchporuk's theorem to prove lowerbounds on formula size of certain functions. Let us consider the problem of element distinctness. Given $m$ integers from the set $\{1, \ldots, m^2\}$, $ED(n_1, n_2, \ldots, n_m)$ is 1 if all the numbers are distinct, else the value is zero. Note that this is a function from $2m \log m$ bits to one bit. Let us consider a disjoint partition where $S_i = \{n_i\}$. Let $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_m$ and $b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_m$ be two possible fixings of values to $n_1, \ldots, n_{i-1}, n_{i+1}, n_m$ such that $a_1 \neq a_2 \neq \cdots \neq a_{i-1} \neq a_{i+1} \neq \cdots \neq a_m$ and $b_1 \neq b_2 \neq \cdots \neq b_{i-1} \neq b_{i+1} \neq \cdots \neq b_m$. There must exist an element $a$ that does equal to any $b_k$ ($1 \leq k \leq i-1$, and $i+1 \leq k \leq m$). Now $ED(a_1, \ldots, a_{i-1}, a, a_{i+1}, \ldots, a_m)$ differs from $ED(b_1, \ldots, b_{i-1}, a, b_{i+1}, \ldots, b_m)$. Thus the cardinality of $S_i$ is the total number ways one can pick $m - 1$ distinct integers from $1, \ldots, m^2$. Thus

$$|S_i| \geq \left( \frac{m^2}{m - 1} \right) \simeq 2^{m \log m}.$$
Thus $\text{FSize}(ED) \geq m^2 \log m$. Since $ED$ is a function from $2m \log m$ bits, if we denote $n = 2m \log m$, then $\text{FSize}(ED) = \Omega(n^2 / \log n)$.

It can be shown that for every function $f$, for every partition $S_1, \cdots, S_k$ of variable, the quantity $\sum \log |S_i|$ is atmost $n^2 / \log n$. Thus Nechporuk’s method can not yield bounds better than $n^2 / \log n$. 