TOPIC 1
Boolean Formulas: The Classics

(Related sections: 15.2.2., 20)

Summary: Every formula can be made balanced without increasing its size. Random restrictions can sustantially reduce the formula size. An $\Omega(n^{5/2})$ lower bound on the formula size. Nechiporuk's lower bound for general formulas.

Recall that a *DeMorgan formula* (or just a *formula*) is a circuit with And, Or and Not gates, whose underlying graph is a tree. Such formulas can be defined inductively as follows: (i) every boolean variable x_i and its negation $\overline{x}_i$ is a formula of size 1 (these formulas are called *leaves*); (ii) if F_1 and F_2 are formulas of size l_1 and l_2, then both $F_1 \wedge F_2$ and $F_1 \vee F_2$ are formulas of size $l_1 + l_2$.

Let $|F|$ denote the size of a formula F (i.e. the number of leaves in F). By $L(f)$ we denote the minimal size and by $D(f)$ the minimal depth[1] of a DeMorgan formula computing f.

1.1. Size versus depth

Let F be a DeMorgan formula. Since its underlying graph is a binary tree, we have $|F| \leq 2^{D(F)}$. This implies that, for every function f,

$$D(f) \geq \log L(f).$$

In fact, we have also a converse inequality:

Theorem 1.1 (Spira 1971). *For any f, $D(f) = O(\log L(f))$.*

Proof. We must show how, given a formula F of size ℓ, we can reduce its depth to $O(\log \ell)$. The idea is to separate the leaves of a formula into two formulas each of which has at most $\frac{2}{3}$ of the original size. Then recurse on the subtrees to get log depth.

Let $D(\ell)$ denote the minimal number such that every formula F of size at most ℓ has an equivalent[2] formula F' of depth at most $D(\ell)$. We will argue by induction on ℓ to show that

$$D(\ell) \leq 2 \log_{3/2} \ell = O(\log_2 \ell). \qquad (*)$$

[1]The depth of a formula is, of course, the depth of its underlying tree.
[2]Two formulas are *equivalent* if they compute the same boolean function.

Let F be an arbitrary formula of size ℓ. By Exercise 4.14 from the book, we can find a node v in the underlying graph of F such that

$$\frac{\ell}{3} \leq |F_v| < \frac{2\ell}{3}.$$

Replacing the subformula F_v by contants 0 and 1 we obtain two formulas $F_{v\to 0}$ and $F_{v\to 1}$, each of size

$$|F| - |F_v| \leq \ell - \frac{\ell}{3} = \frac{2\ell}{3}.$$

Consider the following formula

$$F' = \mathrm{sel}(F_v, F_{v\to 0}, F_{v\to 1})$$

where $\mathrm{sel}(x,y,z) = \overline{x}y \vee xz$ is the *selector* function, which outputs y if $x = 0$, and outputs z if $x = 1$. The formula F' computes the same boolean function as F and has depth at most

$$2 + \max\{D(F_v), D(F_{v\to 0}), D(F_{v\to 1})\} \leq 2 + D\left(\frac{2\ell}{3}\right),$$

where by the induction hypothesis (*),

$$D\left(\frac{2\ell}{3}\right) \leq 2\log_{3/2}\left(\frac{2\ell}{3}\right) = 2\left(\log_{3/2}\ell - 1\right).$$

Thus,

$$D(\ell) \leq 2 + D\left(\frac{2\ell}{3}\right) \leq 2\log_{3/2}\ell.$$

□

1.2. The effect of random restrictions

Already in 1961, Subbotovskaya has found an argument to show that some boolean functions require boolean formulas of super-linear size. Her idea was, given a formula F computing some function f, to set randomly some of the variables to constants and show that this restriction reduces the size of F considerably whereas the resulting subfunction of f is not much "easier."

Let us recall some notation. A *partial assignment* (or *restriction*) is a function $\rho : [n] \to \{0, 1, *\}$, where we understand $*$ to mean that the corresponding variable is unassigned. Given a boolean function $f(x_1, \ldots, x_n)$, $f|_\rho$ indicates the resulting *restriction* of f. Let $\mathcal{R}_k$ be the set of all partial assignments which leave exactly k variables unassigned. What we will be interested in is the *random* restrictions $f|_\rho$ that results from choosing a *random* partial assignment from $\mathcal{R}_k$.

The probability distribution of restrictions in $\mathcal{R}_k$ we will use is the following: randomly assign k variables to be $*$, and assign all other variables to be 0 or 1 randomly and independently.

Intuitively, a random restriction should reduce considerably the size of a formula. The following lemma makes this intuition precise.

Lemma 1.2 (Subbotovskaya 1961). *Let f be a boolean function of n variables, and let ρ be a random restriction from $\mathcal{R}_k$. Then, with probability at least $\frac{3}{4}$,*

$$L(f|_\rho) \leq 4 \cdot \left(\frac{k}{n}\right)^{3/2} \cdot L(f).$$

Proof. Let F be an optimal DeMorgan formula for the function f of size $s = L(f)$. Construct the restriction ρ in $n-k$ stages as follows:

- At any stage, choose a variable randomly from the remaining ones, and assign it 0 or 1 randomly.

We analyse the effect of this restriction to the formula F, stage-by-stage.

Suppose the first stage chooses the variable x_i. When this variable is set to a constant, then all the input gates $e \in F$, labeled by the literals x_i and $\overline{x}_i$ will disappear from the formula F. By averaging, the expected number of such literals is s/n.

In fact, the formula is likely to be reduced even further. For each of the input gates e, labeled by x_i or $\overline{x}_i$, consider the gate which e feeds into. For example, suppose the gate is $e \wedge G$ for some subformula G. We may assume w.l.o.g. that G does not contain the literals x_i or $\overline{x}_i$.

Ex. 1.3. Why? *Hint*: Use the minimality of F.

Now, if the variable x_i is assigned 0, then the subformula G will disappear from the formula F, thereby erasing at least one more input gate. Since x_i is assigned 0 or 1 randomly (with probability 1/2), we expect at least $\frac{1}{2} \cdot \frac{s}{n}$ input gates to disappear because of these secondary effects. In total, we thus expect at least

$$\frac{s}{n} + \frac{s}{2n} = \frac{3s}{2n}$$

input gates $e \in F$ to disappear in the first stage, yielding a new formula with expected size at most [3]

$$s - \frac{3s}{2n} = s \cdot \left(1 - \frac{3}{2n}\right) \leq s \cdot \left(1 - \frac{1}{n}\right)^{3/2}.$$

The succeeding stages of the restriction can be analysed in the same way. After each stage the number of variables decrements by one. Hence, the expected size $E[L(f|_\rho)]$ of the final formula is is at most

$$s \cdot \left(1 - \frac{1}{n}\right)^{3/2} \cdot \left(1 - \frac{1}{n-1}\right)^{3/2} \cdot \ldots \cdot \left(1 - \frac{1}{k+1}\right)^{3/2} = s \cdot \left(\frac{k}{n}\right)^{3/2}.$$

By Markov's inequality, the probability that $L(f|_\rho)$ is more than 4 times its expected value is less than $\frac{1}{4}$, which completes the proof. □

Corollary 1.4. *Let f be a boolean function of n variables, and $1 \leq k \leq n$ be an integer. Then there exists a restriction $\rho \in \mathcal{R}_k$ such that*

$$L(f|_\rho) \leq 4 \cdot \left(\frac{k}{n}\right)^{3/2} \cdot L(f).$$

Example 1.5. Let $f = x_1 \oplus x_2 \oplus \cdots \oplus x_n$. Applying Corollary 1.4 with $k = 1$ we get that

$$1 \leq L(f|_\rho) \leq 4 \cdot \left(\frac{1}{n}\right)^{3/2} \cdot L(f),$$

which gives the lower bound $L(f) = \Omega(n^{3/2})$.

[3] Because $\left(1 - \frac{3}{2n}\right)^2 = 1 - \frac{3}{n} + \frac{9}{4n^2}$ whereas $\left(1 - \frac{1}{n}\right)^3 = 1 - \frac{3}{n} + \frac{3}{n^2} - \frac{1}{n^3}$.

1.3. An $\Omega\left(n^{5/2}\right)$ lower bound

Andreev (1987) used Subbotovskaja's argument to prove the first super-quadratic lower bound for formula size.

Arrange the variables $x_1, \ldots, x_n$ into a $b \times m$ matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ & & \cdots & \\ x_{b1} & x_{b2} & \cdots & x_{bm} \end{pmatrix}$$

where $b = \log n$ and $m = n/b$. Given a boolean function $\varphi : \{0,1\}^b \to \{0,1\}$ on b variables, let $f_\varphi(X)$ denote the following boolean function on bm variables X

$$f_\varphi(X) \rightleftharpoons \varphi\left(\bigoplus_{j=1}^{m} x_{1j}, \bigoplus_{j=1}^{m} x_{2j}, \ldots, \bigoplus_{i=j}^{m} x_{bj}\right).$$

Lemma 1.6. *Let ρ be a random restriction from $\mathcal{R}_k$ where $k = \lceil b\ln(4b)\rceil$. Then with probability at least $3/4$, the restriction ρ assigns at least one $*$ to each of the b rows of X.*

Proof. Observe that for every fixed i, $\Pr[\rho(i) = *] = \binom{n-1}{k-1}/\binom{n}{k} = \frac{k}{n}$. Hence,

$$\begin{aligned} & \Pr[\rho \text{ assigns } * \text{ to every row of } X] \\ \geq\ & 1 - \sum_{i=1}^{b} \Pr[\rho(x_{ij}) \neq *, \forall j = 1, \ldots, m] \\ \geq\ & 1 - b \cdot \left(1 - \frac{k}{n}\right)^m \geq 1 - b \cdot e^{-\frac{km}{n}} \\ =\ & 1 - b \cdot e^{-\ln(4b)} = 1 - 1/4 = 3/4. \end{aligned}$$

$\square$

By this lemma, we have that with probability at least $3/4$, φ is a subfunction of $f_\varphi|_\rho$, and hence,

$$\Pr[L(f_\varphi|_\rho) \geq L(\varphi)] \geq \frac{3}{4}.$$

On the other hand, by Lemma 1.4,

$$\Pr\left[L(f_\varphi|_\rho) \leq 4 \cdot \left(\frac{k}{bm}\right)^{3/2} \cdot L(f_\varphi)\right] \geq \frac{3}{4}.$$

Thus, there must be a restriction $\rho \in \mathcal{R}_k$ for which both these events happen, implying that

$$L(f_\varphi) \geq \frac{1}{4} \cdot \left(\frac{bm}{k}\right)^{3/2} \cdot L(f_\varphi|_\rho) \geq \frac{1}{4} \cdot \left(\frac{bm}{k}\right)^{3/2} \cdot L(\varphi). \tag{1.7}$$

Using an easy counting, Shannon has proved that, for almost all boolean functions φ in b variables,

$$L(\varphi) = \Omega\left(\frac{2^b}{\log b}\right).$$

Taking any of these "most complicated" functions φ we get from (1.7) that

$$L(f_\varphi) = \Omega\left(\frac{bm}{k}\right)^{3/2} \cdot \left(\frac{2^b}{\log b}\right) = \Omega\left(\frac{n^{5/2}}{\log n \cdot \log\log n}\right). \tag{1.8}$$

Strictly speaking, this lower bound is not "explicit" since the function φ is not specified. But the construction can be easily adopted to get the same lower bound for the following *explicit* boolean function $A_{b,m}$. This function has $n = 2^b + bm$ variables; the first 2^b variables of $A_{b,m}$ will specify a boolean function φ of b variables by listing its truth table. Then the value of $A_{b,m}$ is defined to be the value of $f_\varphi(X)$ where X is the set of remaining bm variables. From (1.8) we immediately get the following lower bound for $A_{b,m}$.

Theorem 1.9 (Andreev 1987)**.** *Let $b = \lceil \log_2 n - 1\rceil$ and $m = \lceil n/(2b)\rceil$. Then the function $A_{b,m}$ (of n variables) requires DeMorgan formulas of size $\Omega\left(n^{5/2-\varepsilon}\right)$ for every fixed constant $\varepsilon > 0$.*

1.4. Nechiporuk's theorem

We have seen that Subbotovskaja's method of random restrictions can help us to show that some functions require large formulas. In particular - that some special functions (like the Andreev's function above) require size $\Omega\left(n^{5/2-\varepsilon}\right)$. However, for "natural" functions (like the parity function $x_1 \oplus x_2 \oplus \cdots \oplus x_n$) this method yields only $\Omega\left(n^{3/2}\right)$ lower bounds, which are far from the upper bound $O(n^2)$. Fortunately, we have other methods (due to Nechiporuk and Khrapchenko) which can be applied to many explicit functions, and give almost optimal bounds.

Khrapchenko's method is described in Sect. 15.2.2 of the book (Theorem 15.13). In this section we will describe the Nechiporuk's argument and show how it works for explicit functions. Nechiporuk's method gives slightly worse bounds but works also in the case when we allow *all* binary boolean operations (including $\wedge, \vee, \neg, \oplus$, etc.) as gates.

Let $L(f)$ denote the minimum size of a boolean formula (over the basis of all binary boolean operations) computing f.

A *subfunction* of f on $Y \subseteq X$ is a function obtained from f by setting all the variables of $X \setminus Y$ to constants. Let $N_Y(f)$ be the number of different non-constant subfunctions of f on Y. Intuitively, if f has many subfunctions, then it is complicated and hence should require large formulas. This intuition was made precise by Nechiporuk (1966) who proved the following theorem with a weaker constant factor; the version bellow is due to Paterson (unpublished) and Zwick (1991).

Theorem 1.10 (Nechiporuk 1966)**.** *Let f be a boolean function on a variable set X, and let $Y_1, Y_2, \ldots, Y_k$ be disjoint subsets of X. Then*

$$L(f) \geq \sum_{i=1}^{k} \log_5 \Big(2N_{Y_i}(f) + 1\Big).$$

Proof. Let $S_Y(f)$ be the collection of functions g on a variable set Y for which either g or $\neg g$ is a subfunction of f. Clearly, $N_Y(f) \leq |S_Y(f)|$. Let $size_Y(F)$ be the number of occurrences of variables of Y in a boolean formula F. The theorem follows directly from the following claim.

Claim 1.11. *For every binary formula F and every variable set Y, we have*

$$2|S_Y(F)| + 1 \leq 5^{size_Y(F)}.$$

The claim is proved by the induction on the size of F.

The base case $F = x_i$ divides into two subcases $i \in Y$ or $i \notin Y$. If $i \in Y$ then $S_Y(x_i) = \{x_i\}$; if $i \notin Y$ then $S_Y(x_i) = \{0, 1\}$; thus, both subcases satisfy the claim.

Assume now by induction that $F = F_1 * F_2$, where F_1 and F_2 satisfy the claim, and $*$ is a binary operation. For brevity, let $S_1 = S_Y(F_1)$ and $S_2 = S_Y(F_2)$. Consider the following two collections of boolean functions on Y:

$$\begin{aligned} T &\rightleftharpoons \{g_1 * g_2 \,:\, g_1 \in S_1 \text{ and } g_2 \in S_2\}, \\ \widetilde{T} &\rightleftharpoons \{\neg g \,:\, g \in T\}. \end{aligned}$$

It is straightforward to check that

$$S_Y(F) \subseteq T \cup \widetilde{T} \cup S_1 \cup S_2,$$

and therefore,[4]

$$\begin{aligned} 2|S_Y(f)| + 1 &\leq 2 \cdot (|T| + |\widetilde{T}| + |S_1| + |S_2|) + 1 \\ &\leq 4|S_1||S_2| + 2|S_1| + 2|S_2| + 1 \\ &= (2|S_1| + 1) \cdot (2|S_2| + 1) \\ &\leq 5^{size_Y(F_1)} \cdot 5^{size_Y(F_2)} = 5^{size_Y(F)}. \end{aligned}$$

This completes the proof of the claim and hence the proof of the theorem. □

Ex. 1.12. Show that Nechiporuk's theorem cannot yield lower bounds larger than $\Omega(\frac{n^2}{\log n})$.

1.4.1. Applications

To apply Nechiporuk's theorem, the function must have many subfunctions. Below we describe two such functions.

The block equality function. Let $n = (k+1) \cdot b$ where $b = \lceil \log k \rceil + 1$ and $k \geq 3$. Break the variables $x_1, \ldots, x_n$ into $(k+1)$ blocks $Y_0, \ldots, Y_k$ of size b, where

$$Y_i = \{x_{i \cdot b+1}, \ldots, x_{(i+1) \cdot b}\}$$

is the i-th block, $i = 0, 1, \ldots, k$. Define [5]

$$f(x_1, \ldots, x_n) := \begin{cases} 1 & \text{if } \exists i \neq j \colon Y_i = Y_j \\ 0 & \text{otherwise} \end{cases}$$

Claim 1.13. *For every $i = 0, 1, \ldots, k$,*

$$N_f(Y_i) \geq \frac{2^{2k}}{k}.$$

To show this, assume for simplicity that $i = 0$, i.e. let us consider the first block Y_0. Take arbitrary assignments $\alpha_1, \ldots, \alpha_k \in \{0, 1\}^b$ of constants to the remaining blocks $Y_1, \ldots, Y_k$. If some two of these assignments α_i and α_j coincide, then the resulting subfunction $f(Y_0, \alpha_1, \ldots, \alpha_k)$ is constant 1. So, to get non-constant subfunctions we have to restrict ourselves to assignments $\alpha = (\alpha_1, \ldots, \alpha_k)$

---

[4] Because $size_Y(F_1) + size_Y(F_2) = size_Y(F)$.

[5] Here $Y_i = Y_j$ means that these blocks coincide as strings.

where all α_i's are different. There are at least $\binom{2^b}{k}$ such assignments. Moreover, all of them induce different subfunctions.

Ex. 1.14. Verify this last claim.

Thus,

$$N_f(Y_i) \quad \geq \quad \binom{2^b}{k} \quad \geq \quad \binom{2k}{k} = \frac{(2k)!}{(k!)^2} \approx \frac{2^{2k}}{\sqrt{\pi k}} > \frac{2^{2k}}{k}.$$

By Nechiporuk,

$$L(f) \quad \geq \quad \sum_{i=0}^{m} \log_5(N_f(Y_i)) \geq (k+1) \cdot \log_5 \frac{2^{2k}}{k} = \Omega(k^2) = \Omega\left(\frac{n^2}{\log^2 n}\right),$$

because $k = \Omega(n/\log n)$.

Indirect Addressing Function The "indirect addressing function" $f_n(z, Y, x)$ is defined as follows. It has n variables arranged into three groups. Let $n = k + 2m$ where $k = \log b$ and $b = \log m$. The first group z consists of k variables, the second group Y consists of m variables arranged into an $(m/b) \times b$ matrix, and the third group x consists of the remaining m variables.

The value of this function is computed as follows. First we compute the number $i = int(z)$ whose binary code is the input z. Then we take the ith row Y_i of Y, compute the number $j = int(Y_i)$ and let

$$f_n(z, Y, x) = x_j.$$

Ex. 1.15. Prove that $L(f_n) = \Omega(\frac{n^2}{\log n})$. *Hint*: Apply Nechiporuk's theorem with $Y_1, \ldots, Y_{m/b}$ being the rows of Y. For each row Y_i fix the input z so that $int(z) = i$, and observe that after that each assignment of constants to x leads to a different subfunction of f_n on Y_i.

Ex. 1.16. Show that f_n can be computed by a DeMorgan *circuit* of size $O(n)$.