

Improved boolean formulas for the Ramsey graphs

PETR SAVICKÝ

Department of logic, Faculty of Philosophy, Charles University,
Prague, Czech Republic

e-mail: petr.savicky@cuni.cz

Abstract

In the theory of the random graphs, there are properties of graphs such that almost all graphs satisfy the property, but there is no effective way to find examples of such graphs. By the well-known results of Razborov, for some of these properties, e.g. some Ramsey property, there are Boolean formulas in ACC representing the graphs satisfying the property and having exponential number of vertices with respect to the number of variables of the formula.

Razborov's proof is based on a probabilistic distribution on formulas of n variables of size approximately $nd^2 \log d$, where d is a polynomial in n , and depth 3 in the basis $\{\wedge, \oplus\}$ with the following property: the restriction of the formula randomly chosen from the distribution to any subset A of the boolean cube $\{0, 1\}^n$ of size at most d has almost uniform distribution on the functions $A \rightarrow \{0, 1\}$.

We show a modified probabilistic distribution on boolean formulas which is defined on formulas of size at most $nd \log^2 d$ and has the same property of the restrictions to sets of size at most d as the original one. This allows us to obtain formulas the complexity of which is a polynomial of a lower degree in n than in Razborov's paper while the represented graphs satisfy the same properties.

1 Introduction

The probabilistic method is a powerful tool allowing to demonstrate the existence of some combinatorial objects without presenting any example. In some situations, this method is even the only known way to show the existence of some objects. Typical examples of such situations in the graph theory may be found in [4]. More recent information on the probabilistic method may be found in [2].

We shall concentrate on two types of graphs the existence of which may be proved at present only by the probabilistic method. Namely, the graphs with nearly maximal number of vertices which do not contain clique and independent set of a specified size and tournaments satisfying the property called $S(k)$ in [4] for nearly maximal possible k . The graphs satisfying the first property will be called the Ramsey graphs. Both these properties of graphs will be discussed in detail in section 3.

A typical probabilistic proof of the existence of a graph satisfying some property consists of showing that a graph chosen at random from the set of all graphs on a fixed set of vertices satisfies the desired property with positive probability. Usually, the distribution of the random graph is such that the edges of the graph are chosen independently. Since every graph may appear

as a result of such a random choice, the proof does not give any information on the structure of the graph. In particular, we cannot expect that a graph obtained in this way may have a small complexity in any reasonable model of computation.

Since no effective way to construct a Ramsey graph is known, one may expect that every such graph is very complex and that it is possible to find some nontrivial lower bounds on the complexity of boolean circuits representing such graphs. Surprisingly, Razborov showed that some of the properties described in [4], including the Ramsey property and the $S(k)$ property for nearly maximal k , may be satisfied by graphs representable by boolean formulas exponentially smaller than the represented graph itself.

Razborov's proof is probabilistic again, although the probabilistic space from which the graphs are chosen is much smaller than the set of all graphs. Moreover, in order to ensure that the resulting graph is representable by a small formula, the probabilistic space consists only of such graphs. In a distribution satisfying this, it is not possible to keep the choice of the edges independent. However, it is possible to define the distribution in such a way that in every group of the edges of a specified size, say d , the edges are chosen almost independently. Then, it is necessary to show that this almost independency of every d edges is sufficient to guarantee the desired property of the graph with positive probability.

To represent a graph we need information consisting of one bit for every pair of distinct vertices to decide whether the corresponding edge is contained in the graph or not. These bits may be represented in different ways. For example, we may take a boolean function of sufficiently large domain and take the values of the function on an arbitrary subset of the domain which is in 1-1 correspondence with the pairs of vertices. For every choice of the subset and its correspondence to the pairs of vertices we obtain a representation of graphs using boolean functions. Consequently, any boolean formula of appropriate number of variables represents a graph, since it represents a boolean function. In Razborov's results, a natural representation of this type is used. We shall describe and use it in section 3.

In Razborov's representation of graphs, a substantial portion (nearly one half of the 2^n elements) of the domain of the boolean function is in 1-1 correspondence with the pairs of vertices. In fact, it is not necessary. Another interesting notion of complexity of graphs was used in [3] and [8]. In spite of the fact that only $\binom{n}{2}$ points of the domain of the functions are used here to code edges, it leads to an interesting concept.

If we take a formula at random, we obtain, using any representation of the graphs described above, a randomly chosen graph. To ensure that every d

edges of this graph are chosen almost independently, it is sufficient to require that the values of the formula on every d points of the domain are almost independent. We shall formulate it more precisely using the terminology of the paper [1]. Let the restriction of any function h to a set D be denoted by $h|_D$.

Definition 1.1 Let a probabilistic distribution on the boolean functions $\tilde{g} : \{0, 1\}^n \rightarrow \{0, 1\}$ be given. We say that the distribution of $\tilde{g}$ is (ε, d) -independent if for every $D \subseteq \{0, 1\}^n$, $|D| = d$ and every $f : D \rightarrow \{0, 1\}$ the following holds

$$\left| \mathbb{P}(\tilde{g}|_D = f) - (1/2)^d \right| \leq \varepsilon.$$

Razborov uses a construction of (ε, d) -independent formulas of size $O(nd \log d \log(1/\varepsilon))$. The main goal of the present paper is to show a slight modification of Razborov's formulas which allows to decrease the size of the formulas to $O(n \log^2 d \log(1/\varepsilon))$. This improvement is presented in section 2. In its application to the graphs we need (ε, d) -independent random formulas, where d is a polynomial in n of a small degree and $\varepsilon = (1/2)^d$. The size of the formulas satisfying this is $O(nd^2 \log d)$ in the original construction and $O(nd \log^2 d)$ in our improved one. The latter estimate may be bounded by a polynomial of a lower degree in n than the first one, if d is unbounded.

The properties of the graphs represented by the random boolean formulas are discussed in section 3. In fact, we reformulate two of Razborov's theorems from [9] using our improved random boolean formulas. The two chosen examples seem to be the most representative. The application of our result to further theorems from [9] is straightforward.

A random formula of n variables may be viewed also as a random boolean vector of the length 2^n , simply by taking the table of the corresponding function. The notion of (ε, d) -independency generalizes to this case in a natural way. In our construction, this random vector is (ε, d) -independent while the number of completely independent random bits needed to construct the vector is at most $n \log^2 d \log(1/\varepsilon)$. Another constructions of long random vectors which are (ε, d) -independent and constructible using only a small number of independent random bits are described in [7] and [1]. The constructions described in the cited papers need at most $2(\log n + \log d + \log(1/\varepsilon))$ of independent random bits. Although these constructions are much better than the construction from the present paper with respect to the number of the random bits, they cannot be used here, since they do not provide a representation of the output vector by a boolean formula of bounded depth and polynomial size.

A particular case of our result may be formulated also in the following way. For every fixed $\varepsilon > 0$ and every n , there is a uniform log-space constructible ACC formula $\phi_n(x, y)$ of depth 3 and size $O(n^3)$ with the following properties. The two input vectors x and y have length n and $O(n^3)$ respectively. If y is taken at random from the uniform distribution, then by taking all possible values of x we get 2^n pseudorandom bits which are ε -biased with respect to any linear test. It means that the sum modulo 2 of any nonempty subset of the bits is 1 with probability $1/2 \pm \varepsilon$. (Hint: in $\tilde{g}[m, r]$ defined later, set $m = n + 1$ and r a sufficiently large constant and use the formula (2)).

2 Random boolean formulas

We shall use the discrete Fourier transform of a probabilistic distribution on boolean functions to compute estimates of the difference between the studied distribution and the uniform one. Let a probabilistic distribution on the functions $\tilde{h} : \{0, 1\}^n \rightarrow \{0, 1\}$ be given. Let $D \subseteq \{0, 1\}^n$. The Fourier transform of the distribution of $\tilde{h}|_D$ is given by the following formula: for every $A \subseteq D$ let

$$\Delta(A) = \mathbb{E}(-1)^{\langle \tilde{h}, X_A \rangle},$$

where X_A is the characteristic function of the set A and $\langle \cdot, \cdot \rangle$ denotes the scalar product of two boolean functions modulo 2.

By standard Fourier analysis, the discrete Fourier transform has an inverse transform. In our situation it can be described by the following formula: for every $f : D \rightarrow \{0, 1\}$

$$\mathbb{P}(\tilde{h}|_D = f) = \frac{1}{2^{|D|}} \sum_{A \subseteq D} \Delta(A) (-1)^{\langle f, X_A \rangle}$$

Using this and the fact that $\Delta(\emptyset) = 0$, we obtain the following consequence:

Lemma 2.1 *For every $f : D \rightarrow \{0, 1\}$*

$$\left| \mathbb{P}(\tilde{h}|_D = f) - \frac{1}{2^{|D|}} \right| \leq \max_{\emptyset \neq A \subseteq D} \left| \mathbb{E}(-1)^{\langle \tilde{h}, X_A \rangle} \right|$$

Let n be a given natural number. We shall denote by $\mathcal{L}$ the set of all the affine functions from $\{0, 1\}^n$ to $\{0, 1\}$, i.e. the functions $c_0 \oplus c_1 x_1 \oplus \dots \oplus c_n x_n$, where $c_i \in \{0, 1\}$ and $\oplus$ denotes the addition modulo 2. Any of these functions may be expressed using a formula of complexity at most n . We shall describe the random construction of the boolean formulas in four steps. In the first step, we take a randomly chosen affine function. In each of the

remaining steps we combine several independently chosen copies of some of the formulas defined in the previous steps. In the following definition, the numbers in square brackets distinguish formulas of different kind and the subscripts are used to distinguish independently chosen copies of the same random formula.

Definition 2.2 Let $\tilde{l}$ be a random affine function of n variables chosen from the uniform distribution on $\mathcal{L}$ and i, m, r be natural numbers. Then, let $\tilde{k}[i]$, $\tilde{h}[m]$ and $\tilde{g}[m, r]$ denote the random formulas described as follows

$$\begin{aligned}\tilde{k}[i] &= \tilde{l}_1 \wedge \dots \wedge \tilde{l}_i \\ \tilde{h}[m] &= \tilde{k}[2] \oplus \dots \oplus \tilde{k}[m] \\ \tilde{g}[m, r] &= \tilde{l} \oplus \tilde{h}_1[m] \oplus \dots \oplus \tilde{h}_r[m]\end{aligned}$$

Clearly, if $r \geq 1$ then the complexity of $\tilde{g}[m, r]$ is at most $n(1 + r(m - 1)(m + 2)/2) \leq nrm(m + 1)/2$.

Notice that for every $a, b \in \{0, 1\}^n$, $a \neq b$, the distribution of the pair $\tilde{l}(a), \tilde{l}(b)$ is uniform on the set $\{0, 1\}^2$. It follows that

$$\mathbb{P}(\tilde{k}[i](a) = 1) = \frac{1}{2^i}$$

$$\mathbb{P}(\tilde{k}[i](a) = 1 \wedge \tilde{k}[i](b) = 1) = \frac{1}{4^i}$$

Lemma 2.3 *Let $|A| \leq 2^i$. Then*

$$\left| \mathbb{E}(-1)^{\langle \tilde{k}[i], X_A \rangle} \right| \leq 1 - 2 \frac{|A|}{2^i} \left(1 - \frac{|A|}{2^i} \right).$$

Proof: For every real numbers $x, y \geq 0$, we have $|x - y| = x + y - 2 \min(x, y)$. Using this for $x = \mathbb{P}(\langle \tilde{k}[i], X_A \rangle = 0)$ and $y = \mathbb{P}(\langle \tilde{k}[i], X_A \rangle = 1)$ we obtain

$$\left| \mathbb{E}(-1)^{\langle \tilde{k}[i], X_A \rangle} \right| = 1 - 2 \min_{\alpha \in \{0, 1\}} \{ \mathbb{P}(\langle \tilde{k}[i], X_A \rangle = \alpha) \}.$$

For the proof of the lemma, we only need to show that for every $\alpha \in \{0, 1\}$

$$\mathbb{P}(\langle \tilde{k}[i], X_A \rangle = \alpha) \geq \frac{|A|}{2^i} \left(1 - \frac{|A|}{2^i} \right). \quad (1)$$

Since

$$\mathbb{P}((\exists a \in A) \tilde{k}[i](a) = 1) \leq \frac{|A|}{2^i},$$

we have

$$P(\langle \tilde{k}[i], X_A \rangle = 0) \geq P((\forall a \in A) \tilde{k}[i] = 0) \geq 1 - \frac{|A|}{2^i} \geq \frac{|A|}{2^i} \left(1 - \frac{|A|}{2^i}\right).$$

Hence, (1) is true for $\alpha = 0$.

Further, for a fixed $a \in A$, we have

$$\begin{aligned} P(\tilde{k}[i](a) = 1 \wedge (\exists b \in A \setminus \{a\}) \tilde{k}[i](b) = 1) &\leq \frac{|A| - 1}{4^i} \\ P(\tilde{k}[i](a) = 1 \wedge (\forall b \in A \setminus \{a\}) \tilde{k}[i](b) = 0) &\geq \frac{1}{2^i} - \frac{|A| - 1}{4^i} \geq \frac{1}{2^i} \left(1 - \frac{|A|}{2^i}\right). \end{aligned}$$

The events considered in the last formula are disjoint for different $a \in A$ and any of these events implies $\langle \tilde{k}[i], X_A \rangle = 1$. Hence, (1) is true also for $\alpha = 1$. $\square$

Theorem 2.4 *Let $m \geq 2$, $r \geq 1$. Then the distribution of $\tilde{g}[m, r]$ is $((1/2)^r, d)$ -independent for every $d \leq 2^{m-1}$.*

Proof: By Lemma 2.1 it is enough to show that for every $\emptyset \neq A \subseteq \{0, 1\}^n$, $|A| \leq 2^{m-1}$ we have

$$\left| E(-1)^{\langle \tilde{g}[m, r], X_A \rangle} \right| \leq \frac{1}{2^r}. \quad (2)$$

Notice that $(-1)^{\langle \tilde{g}[m, r], X_A \rangle}$ is a product of $(-1)^{\langle \tilde{l}, X_A \rangle}$ and of r independent copies of $(-1)^{\langle \tilde{h}[m], X_A \rangle}$. By the well-known theorem on the expected value of the product of several independent variables we obtain:

$$E(-1)^{\langle \tilde{g}[m, r], X_A \rangle} = E(-1)^{\langle \tilde{l}, X_A \rangle} E^r(-1)^{\langle \tilde{h}[m], X_A \rangle}.$$

If $1 \leq |A| \leq 2$, the estimate (2) follows from the fact that $E(-1)^{\langle \tilde{l}, X_A \rangle} = 0$. In the case that $3 \leq |A| \leq 2^{m-1}$ we shall prove (2) by proving that

$$\left| E(-1)^{\langle \tilde{h}[m], X_A \rangle} \right| \leq \frac{1}{2}. \quad (3)$$

Let i be the minimal i such that $|A| \leq 2^i$. Clearly, $2 \leq i \leq m - 1$. Let us denote $t = |A|/2^{i+1}$. Since

$$(-1)^{\langle \tilde{h}[m], X_A \rangle} = (-1)^{\langle \tilde{k}[2], X_A \rangle} \dots (-1)^{\langle \tilde{k}[m], X_A \rangle}$$

and the factors on the right hand side are independent, we obtain that

$$\left| \mathbb{E}(-1)^{\langle \tilde{h}[m], X_A \rangle} \right| \leq \left| \mathbb{E}(-1)^{\langle \tilde{k}[i], X_A \rangle} \right| \cdot \left| \mathbb{E}(-1)^{\langle \tilde{k}[i+1], X_A \rangle} \right|.$$

It follows by Lemma 2.3 that

$$\left| \mathbb{E}(-1)^{\langle \tilde{h}[m], X_A \rangle} \right| \leq (1 - 4t + 8t^2)(1 - 2t + 2t^2)$$

Further, we know that $1/4 \leq t \leq 1/2$. For t from this interval we have $0 \leq 1 - 4t + 8t^2 \leq 2t$ and $0 \leq 1 - 2t + 2t^2 \leq 1 - t$. Since $2t(1 - t) \leq 1/2$, the estimate (3) is proved. $\square$

3 Application to the graphs

Let $G = (V, E)$ be an undirected graph. A nonempty subset $A \subseteq V$ will be called homogeneous if the induced subgraph of G on the set A is either a complete graph or an empty graph.

Definition 3.1 Let $\omega(G)$ be the size of the maximal homogeneous set in a graph G .

The formulation of the following theorem is adopted for the purpose of representation of graphs using boolean functions. The original (and a little stronger) formulation may be found in [4].

Theorem 3.2 (*Erdős, 1947*) *For every sufficiently large even n there is a graph G on $2^{n/2}$ vertices satisfying $\omega(G) \leq n$.*

Despite intensive research, no effective deterministic procedure generating graphs satisfying the property from this theorem is known. The best known effective construction (see. [5]) generates graphs satisfying $\omega(G) \leq 2^{O(\sqrt{n \log n})}$.

For the formulation of our result, we shall use the following property of a probabilistic distribution on graphs.

Definition 3.3 Let $\tilde{G} = (V, \tilde{E})$ be a random graph with a fixed set of vertices V and a randomly chosen set of edges $\tilde{E}$. We say that the distribution of $\tilde{G}$ is (ε, d) -independent, if for every two sets of edges E_0, E_1 , $E_1 \subseteq E_0$, $|E_0| = d$ the following holds:

$$\left| \mathbb{P}(\tilde{E} \cap E_0 = E_1) - \frac{1}{2^d} \right| \leq \varepsilon$$

The probabilistic proof of Theorem 3.2 is based on the following: we take a random graph $\tilde{G}$ on $2^{n/2}$ vertices such that every edge is chosen with probability $1/2$ and independently of the other ones. The probability that the graph $\tilde{G}$ contains a homogeneous set of size n is bounded from above by

$$\sum_{|A|=n} \text{P}(A \text{ is homogeneous in } \tilde{G}) = \binom{2^{n/2}}{n} 2(1/2)^{\binom{n}{2}}.$$

The proof is then completed by showing that this bound tends to zero if n tends to infinity. It is a simple observation that this proof remains valid if instead of the independency of the edges we use the assumption that the distribution of the random graph is (ε, d) -independent, where $d = \binom{n}{2}$ and $\varepsilon = O((1/2)^d)$.

If h is a boolean function of n variables and $u, v \in \{0, 1\}^{n/2}$, we define $h(u, v)$ in the natural way. Let $<_l$ denote the lexicographic ordering of boolean vectors.

Definition 3.4 Let h be a boolean function of n variables. We say that h represents a graph $G = (V, E)$, if $V = \{0, 1\}^{n/2}$ and for every $u, v \in \{0, 1\}^{n/2}$ such that $u <_l v$ we have $\{u, v\} \in E \iff h(u, v) = 1$.

Theorem 3.5 ([9]) *For every sufficiently large even n there is a formula of n variables, of depth 3 in the basis $\{\wedge, \oplus\}$ and of size $O(n^5 \log n)$ representing a graph G_n on $2^{n/2}$ vertices satisfying $\omega(G_n) \leq n$.*

Using our result we obtain the following improvement:

Theorem 3.6 *In Theorem 3.5 the upper bound on complexity of the formulas may be replaced by $O(n^3 \log^2 n)$.*

Proof: Let $d = \binom{n}{2}$, $m = \lceil \log d \rceil + 1$ and $\varepsilon = (1/2)^d$. Then, by Theorem 2.4, the distribution of the graph $\tilde{G}$ represented by the random formula $\tilde{g}[m, d]$ is (ε, d) -independent. Hence, for n large enough, the probability that the graph $\tilde{G}$ satisfies $\omega(\tilde{G}) \leq n$ is positive. The complexity of $\tilde{g}[m, d]$ is at most $nd \log^2 d$. $\square$

Now, we shall turn to the tournaments. A directed graph is a tournament, if for every pair of its vertices there is exactly one directed edge between them.

Definition 3.7 A tournament T satisfies $S(k)$ if for every set A of its vertices of size at most k there is a vertex $v \notin A$ such that for every $u \in A$ there is an edge from v to u .

The formulation of the following theorem is adopted to our situation. The original formulation may be found in [4].

Theorem 3.8 (*Erdős, 1963*) *For every sufficiently large even n there is a tournament on $2^{n/2}$ vertices satisfying $S(\lfloor n/2 - 2 \log n + 2 \rfloor)$.*

So far, no effectively constructed tournament satisfying this property is known. The best known construction (see [6]) yields k nearly $n/4$.

The tournaments may be represented by boolean functions as follows.

Definition 3.9 Let h be a boolean function of n variables. We say that h represents a tournament $T = (V, E)$, if $V = \{0, 1\}^{n/2}$ and for every $u, v \in \{0, 1\}^{n/2}$ such that $u <_l v$ we have $(u, v) \in E \iff h(u, v) = 1$ (and consequently $(v, u) \in E \iff h(u, v) = 0$).

Definition 3.3 generalizes to the tournaments in a straightforward way. To obtain a representation of tournaments satisfying the property from Theorem 3.8 using small boolean formulas, Razborov used the following lemma.

Lemma 3.10 *For every natural number n let a random tournament $\tilde{T}_n$ on $2^{n/2}$ vertices be $((1/2)^{n^3}, d)$ -independent for all $1 \leq d \leq n^3$. Then*

$$\lim_{n \rightarrow \infty} P(\tilde{T}_n \text{ satisfies } S(\lfloor n/2 - 2 \log n + 2 \rfloor)) = 1.$$

Theorem 3.11 ([9]) *For every sufficiently large even n there is a formula of n variables, of depth 3 in the basis $\{\wedge, \oplus\}$ and of size $O(n^7 \log n)$ representing a tournament T_n on $2^{n/2}$ vertices satisfying $S(\lfloor n/2 - 2 \log n + 2 \rfloor)$.*

Using our result we obtain the following improvement:

Theorem 3.12 *In Theorem 3.11 the upper bound on complexity of the formulas may be replaced by $O(n^4 \log^2 n)$.*

Proof: By Theorem 2.4, the assumption of Lemma 3.10 is satisfied by the random tournaments represented by $\tilde{g}[m, r]$, where $r = n^3$ and $m = \lceil \log r \rceil + 1$. The complexity of $\tilde{g}[m, r]$ is then at most $nrm^2 = O(n^4 \log^2 n)$. $\square$

Acknowledgement The author is indebted to Martin Dietzfelbinger and Pavel Pudlák for stimulating discussion on the present paper.

References

- [1] N. Alon, O. Goldreich, J. Hastad, R. Peralta: Simple Constructions of Almost k -wise Independent Random Variables, *Random Structures and Algorithms*, Vol. 3, No. 3 (1992), pp.289–304
- [2] N. Alon, J. H. Spencer, P. Erdős: *The Probabilistic Method*, A Wiley-Interscience Publication,1992
- [3] S. Bublitz: Decomposition of graphs and monotone formula size of homogeneous functions, *Acta Informatica* 23 (1986), pp. 689-696
- [4] P. Erdős, J. H. Spencer: *Probabilistic Methods in Combinatorics*, Akadémiai Kiadó, Budapest,1974
- [5] P. Frankl, R. M. Wilson: Intersection Theorems with Geometric Consequences, *Combinatorica* 1 (1981), pp. 357-368
- [6] R. L. Graham, J. H. Spencer: A constructive solution to a tournament problem, *Canadian Mathematics Boullletin* 14, pp. 45-48
- [7] J. Naor, M. Naor: Small-bias Probability Spaces: Efficient Constructions and Applications, *22nd STOC*, 1990, pp. 213-223
- [8] P. Pudlák, V. Rödl, P. Savický: Graph Complexity, *Acta Informatica* 25 (1988), pp. 515-535
- [9] A. A. Razborov: Bounded-depth formulas over $\{\wedge, \oplus\}$ and some combinatorial problems. In *Complexity of algorithms and applied mathematical logic* (in Russian), Ser. *Voprosy Kibernetiky* (Problems in Cybernetics), ed.: S.I.Adian, Moscow, 1988, pages 149-166