Boolean Circuit Complexity of Regular Languages

Maris Valdats
University of Latvia
Faculty of Computing
Riga, Raiņa Bulv. 19, Latvia
d20416@lanet.lv

In this paper we define a new descriptional complexity measure for Deterministic Finite Automata, BC-complexity, as an alternative to the state complexity. We prove that for two DFAs with the same number of states BC-complexity can differ exponentially. In some cases minimization of DFA can lead to an exponential increase in BC-complexity, on the other hand BC-complexity of DFAs with a large state space which are obtained by some standard constructions (determinization of NFA, language operations), is reasonably small. But our main result is the analogue of the "Shannon effect" for finite automata: almost all DFAs with a fixed number of states have BC-complexity that is close to the maximum.

State complexity of deterministic finite automata (DFA) [1][5] has been analyzed for more than 50 years and all this time has been the main measure to estimate the descriptional complexity of finite automata. Minimization algorithm [6] for it was developed as well as methods to prove upper and lower bounds for various languages.

It is hard to find any evidence of another complexity measure for finite automata. Transition complexity [3] could be one, it counts the number of transitions, but there is not much use of it for DFAs (it is proportional to the state complexity), it is used in the nondeterministic case.

But intuitively not all DFAs with the same number of states have the same complexity. We try to illustrate it with the following example.

Consider a DFA that recognizes a language in the binary alphabet which consists of words in which there is an even number of ones among the last 1000 input letters. One can easily prove that it needs 2^{1000} states, however such a DFA can easily be implemented by keeping its state space in a 1000 bit register which remembers the last 1000 input letters.

On the other hand, consider a "random" DFA with a binary input tape and 2^{1000} states. There is essentially no better way to describe it as with its state transition table which consists of 2^{1001} lines which (as it is widely assumed) is more than particles in our universe.

It is easy to represent a large number of states in a compact form: 2^n states fit into n state bits of the state register. This is true for the "random" DFA as well. But the computation performed by the transition function on this register can be very easy in some cases and hard in some other. Therefore it seems natural to introduce a complexity measure for DFAs which measures the complexity of the transition function.

Automata with a large state space which is kept in a state register have been used before, but not in the widest sense. One example of such a usage is FAPKC [8] (Finite Automata Public Key Cryptosystem), a public key cryptosystem developed in the 80's by Renji Tao. In FAPKC the state space of an automaton is considered to be a vector space and the transition function is expressed as a polynomial over a finite field.

In this paper we consider the following model: we arbitrarily encode the state space into a bit vector (state register) and express the transition function as a Boolean circuit. The BC-complexity of the DFA

Z. Ésik and Z. Fülöp (Eds.): Automata and Formal Languages 2014 (AFL 2014)
EPTCS 151, 2014, pp. 342–354, doi:10.4204/EPTCS.151.24

is (approximately) the complexity of this circuit and this notion extends to regular languages in a natural way.

BC-complexity was first analyzed in [9] where it was considered for transducers. Here we define it for DFAs what allows to extend the definition to regular languages.

The main result of this paper is the Shannon effect for the BC-complexity of regular languages: it turns out that most of the languages have BC-complexity that is close to the maximum. To obtain it we first estimate upper (and lower) bounds for BC-complexity compared to state complexity (Theorem. 3.3), afterwards by counting argument we show that the complexity of most of the languages is around this upper bound.

Influence of state minimization to BC-complexity were analyzed already in [9] for transducers and for DFAs it is essentially the same: it turns out that for some regular languages BC-complexity of their minimal automaton is much (superpolynomially) larger than BC-complexity for some other (non-minimal) DFA that recognizes it. Finally we look how BC-complexity behaves if we do some standard constructions on automata (determinization of an NFA, language operations).

1 Preliminaries

1.1 Finite Automata and Regular Languages

We use a standard notion of DFA [4], it is a tuple $(Q,\Sigma,\delta,q_0,\tilde{Q})$, where Q is the state space, Σ is the input alphabet, $\delta:\Sigma\times Q\to Q$ is the transition function, $q_0\in Q$ is the start state and $\tilde{Q}\subseteq Q$ is the set of accepting states.

DFA starts computation in the state q_0 and in each step it reads an input letter $x\in\Sigma$ and changes its state. If the current state of a DFA is $q\in Q$ and it reads an input letter $x\in\Sigma$ then it moves to state $\delta(x,q)$. If after reading the input word DFA is in a state $q\in\tilde{Q}$ then this word is accepted, otherwise it is rejected. DFA A recognizes language L iff it accepts all words from this language and rejects all words not in the language. Two DFAs that recognize the same language are called equivalent.

The state complexity of a DFA is the number of states in its state space $C_s(A)=|Q|$. For each DFA A there is a unique minimal DFA $M(A)$ which is equivalent to A and has minimal state complexity. There is an effective minimization algorithm for finding it [1].

We will need the estimation of the number of DFAs with s states. Denote $\mathfrak{A}_s$ to be the number of pairwise non-equivalent minimal DFAs with s states over k-letter alphabet. In [2] it is estimated to be larger than $2^{s-1}(s-1)s^{(k-1)s}$, we will use the following reduced estimation (true for $s\geq 3$) which will be sufficient for us:

Theorem 1.1 ([2]) *$\mathfrak{A}_s\geq 2^s s^{(k-1)s}$ for $s\geq 3$.*

1.2 Boolean circuits

We will use the standard notion of a Boolean circuit and restrict our attention to circuits in the standard base (&, $\vee$, $\neg$). The size of the circuit $C(F)$ is the number of gates plus the number of outputs of the circuit F. Boolean circuit F with n inputs and m outputs represents a Boolean function $(y_1,\ldots,y_m)=F(x_1,\ldots,x_n)$ in a natural way.

Each function $f:\{0,1\}^n\to\{0,1\}^m$ can be represented by a Boolean circuit in (infinitely many) different ways. The complexity of this function $C(f)$ is the size of the smallest circuit that represents this function.

We will also need a formula for the upper bound of the number of different Boolean circuits with a given complexity C. Denote $N(n,m,C)$ to be the number of circuits with n input variables, m output variables and no more than C gates, that correspond to different Boolean functions. Then:

Theorem 1.2

$$N(n,m,C) \leq 9^{C+n}(C+n)^{C+m}$$

Proof Assign to inputs numbers from 1 to n, and numbers from $n+1$ to $n+C$ to the gates. Each gate is characterised with its two inputs (at most $(n+C)^2$ possibilities) and type (AND, OR, NOT, 3 possibilities). There are no more than $(n+C)^m$ ways how to assign outputs of the circuit and each circuit is counted $C!$ times, one for each numbering of gates. Therefore the total number of circuits can be estimated as:

$$N(n,m,C) < \frac{(3(C+n)^2)^C \cdot (C+n)^m}{C!} < 9^C(1+\frac{n}{C})^C(C+n)^{C+m},$$

here we have used, that $C! > C^C/3^C$ for all C.

Further, as $(1+1/x)^x < e < 9$ for arbitrary $x > 0$, then

$$(1+\frac{n}{C})^C = ((1+\frac{n}{C})^{\frac{C}{n}})^n < 9^n$$

from where the result follows. □

A classical result about Boolean functions states that most of functions $f : \{0,1\}^n \to \{0,1\}^m$ have approximately the same circuit complexity which is close to maximum. This property is called Shannon effect.

Theorem 1.3 ([7]) *For any Boolean function* $f : \{0,1\}^n \to \{0,1\}^m$

$$C(f) \lesssim \frac{m2^n}{n+\log m},$$

For almost all Boolean functions $f : \{0,1\}^n \to \{0,1\}^m$

$$C(f) \gtrsim \frac{m2^n}{n+\log m}.$$

Here and further $\log = \log_2$ and we use the notation

$$f(n) \lesssim g(n) \Leftrightarrow \lim_{n\to\infty} \frac{f(n)}{g(n)} \leq 1.$$

2 Encodings and Representations of a DFA

Classical representations of automata are table forms or state transition diagrams. They are essentially the same, a state diagram can be thought of as a visualization of a table form. Table form lists the transition function of an automaton as a table where each line corresponds to a pair of state and input letter. In state transition diagram each state is denoted by a circle and for each transition $(q,x) \to q'$ an arrow is drawn from state q to state q' above which letter x is written.

Both of these representations show each state of an automaton separately, therefore with these methods it is not possible to effectively describe an automaton with a large number of states.

One can encode s states into $\lceil \log(s) \rceil$ (or more) state bits which can be kept in a *state register*. Also, input letters can be encoded as a bit vectors. Every automaton has infinitely many such encodings.

The transition function in this case will take as an input a state register and an encoded input letter, and produce a (next) state register. It is thus a Boolean function and it is natural to represent it with a Boolean circuit.

Another question is how to represent the set of accepting states $\tilde{Q}$. We represent it by a Boolean circuit implementing its characteristic function. Therefore a representation of a DFA will consist of an encoding of its state space and input alphabet and two circuits: one for its transition function and one for the characteristic function of the set of accepting states. We call these circuits *transition circuit* and *acceptance circuit*, respectively.

An encoding $E(X)$ of a set X onto a binary string is an injective mapping $f_X : X \to \{0,1\}^{b_X}$ where b_X is the length of the encoding. As the mapping is injective then $b_X \geq \lceil \log |X| \rceil$.

An encoding of a DFA consists of an encoding of its input alphabet f_Σ and an encoding of the state space f_Q which we call input encoding and state encoding, respectively. Additionally for the state encoding we ask that the start state is encoded as a string of all zeros $f_Q(q_0) = 0^{b_Q}$.

Definition Let $A = (Q, \Sigma, \delta, q_0, \tilde{Q})$ be a given DFA and (f_Σ, f_Q) be its encoding. A pair of Boolean circuits (F, G) is a representation of A under encoding (f_Σ, f_Q) iff

- F has $b_\Sigma + b_Q$ input variables and b_Q output variables,
- G has b_Q input variables and one output variable,
- for all $x \in \Sigma$ and $q \in Q$ if $q' = \delta(x, q)$, then $f_Q(q') = F(f_\Sigma(x), f_Q(q))$,
- $G(f_Q(q)) = 1 \iff q \in \tilde{Q}$ for all $q \in Q$.

In other words, transition circuit F reads encoded input $f_\Sigma(x)$ as its first b_Σ input bits, encoded state $f_Q(q)$ as following b_Q input bits and has encoded next state $f_Q(q')$ as its b_Q output bits. Acceptance circuit G reads encoded state $f_Q(q)$ and outputs 1 as its only output bit iff $q \in \tilde{Q}$.

As noted before minimal values for b_Σ and b_Q are $\lceil \log(|\Sigma|) \rceil$ and $\lceil \log(|Q|) \rceil$ respectively, but they can be larger as well. Whether allowing them to be larger gives a possibility to construct smaller representations of DFAs, is an interesting open question.

It is natural to encode the state space Q with $|Q|$ lexicographically first bit strings of length $\lceil \log |Q| \rceil$, in such a case we will say that the state encoding is minimal. The notion of minimal input encoding is introduced similarly. We call an encoding of a DFA *minimal encoding* if both encodings: state and input are minimal.

3 BC-complexity

In this section we define the main concept of this paper, BC-complexity of a DFA. We start from the bottom:

Definition BC-complexity of a representation of a DFA (F, G) is the sum of complexities of its transition circuit and acceptance circuit and the number of state bits:

$$C_{\mathrm{BC}}((F, G)) = C(F) + C(G) + b_Q.$$

The number of state bits b_Q is included in the definition to avoid situation that an automaton has a large number of states but zero BC-complexity. It is natural to assume that it costs something to create

a circuit even if it has no gates and this is one of the possibilities how to reflect this in the definition. Another possibility would be to use the complexity of "wires" instead of the complexity of gates for the underlying circuits, but we prefer to use the standard complexity for the circuits.

Definition BC-complexity of a DFA A, $C_{\mathrm{BC}}(A)$, is the minimal BC-complexity of its representations:

$$C_{\mathrm{BC}}(A) = \min\{C_{\mathrm{BC}}((F,G)) : (F,G) \text{ represents } A\}.$$

Although the name "circuit complexity" also sounds reasonable, we use the abbreviation "BC-complexity" to avoid confusion with the circuit complexity of regular languages.

Definition BC-complexity of a regular language L is the minimal BC-complexity of all DFAs that recognize L:

$$C_{\mathrm{BC}}(L) = \min\{C_{\mathrm{BC}}(A) : A \text{ recognizes } L\}.$$

First we observe that we can optimize our acceptance circuit by rearranging states. If we encode states in such a way that all accepting states have smaller index than rejecting states (or vice versa) then the acceptance circuit can be reduced to a comparison operation whose complexity is not greater than $4n$ where n is the number of state bits.

But this is not the best optimization that can be achieved by rearranging states. For the upper bound in the following Theorem 3.1 different arrangement is used.

Theorem 3.1 *If $|\Sigma| = k \geq 2$ then for any DFA A with s states,*

$$\lceil \log(s) \rceil \leq C_{\mathrm{BC}}(A) \lesssim (k-1)s.$$

If $|\Sigma| = 1$ then for any DFA A with s states,

$$\lceil \log(s) \rceil \leq C_{\mathrm{BC}}(A) \lesssim \frac{s}{\log s}.$$

Proof Lower bound. For any representation (F,G) there are $b_Q \geq \lceil \log s \rceil$ state bits, therefore BC-complexity cannot be smaller than $\lceil \log s \rceil$.

For upper bound if we just construct an optimal representation (F,G) under some arbitrary minimal encoding (with $\lceil \log s \rceil$ state bits) then the BC-complexity of this representation according to Theorem 1.3 can be estimated as

$$C_{\mathrm{BC}}((F,G)) \leq C(F) + C(G) + \lceil \log s \rceil \lesssim \frac{ks\lceil \log s \rceil}{\log(ks\lceil \log s \rceil)} + \frac{s}{\log s} + \lceil \log s \rceil \lesssim ks$$

To improve the result to $(k-1)s$ we will choose a specific minimal encoding where states are ordered in a way that for one input letter the corresponding transition function is simple. Denote q to be the encoding of the current state, q' to be the encoding of the next state and x to be the encoding of the input letter. We split the transition circuit F in two parts F_1 and F_2 where part F_1 computes the next state for one specific input letter a and part F_2 does it for other $k-1$ input letters.

If we look at the state transition graph for the letter a then it splits into connected components each of which has the form

$$q_1 \to q_2 \to \ldots q_{m-1} \to q_m \to q_j$$

where $1 \leq j \leq m$. Each such component is uniquely defined with two numbers m (the number of states in it) and j (the length of the "tail"), we call m to be the length of a component. We order all these

components by m and j lexicographically what naturally leads to the ordering of states. Consider all components with parameters (m, j) and denote by $M = M(m, j)$ the index (encoding) of the first state of the first such component and by $N = N(m, j)$ the index of the last state of the last such component.

The transition function is $q' = q + 1$ except for the last state q_m of each component for which it is $q' = q - (m - j)$. As each of these components have m states then q corresponds to the last state of some component iff $q + 1 = M \mod m$.

The circuit F_1 should compute the following function $q' = F_1(q)$:

```
q' = q+1
for all pairs (m, j)
  if M(m, j)<=q<=N(m, j) and q+1 == M(m, j) mod m:
    q' = q-(m-j)
```

Here M and N are the boundaries within which all components with parameters (m, j) are placed. It is easy to check that circuits for subtraction $q' = q - (m - j)$ and comparison $(M \leq q \leq N)$ are of size $O(\log s)$, for modulo comparison $q + 1 = M \mod m$ it is of size $O(\log s^2)$. Therefore the total size of the circuit F_1 is $K * c \log s^2$ where K is the number of different pairs (m, j) that correspond to some components that are present in the transition graph and c is some constant. We need to estimate the maximum value of K.

One can easily see that maximum value of K is obtained when each component with parameters (m, j) appears exactly once and all the smallest components are used. Let u be the maximum length of a component (maximal value of m) under the condition that all the possible smallest components are used. For each m there are m possible different types of components ($1 \leq j \leq m$) therefore $K \leq u(u+1)/2$.

On the other hand the total length of all components up to the length $u - 1$ should be less than s:

$$\sum_{m=1}^{u-1} m^2 = \frac{(u-1)u(2u-1)}{6} \leq s$$

whence it follows that $u \leq 2\sqrt[3]{s}$. Therefore

$$K \leq \frac{u(u+1)}{2} \leq \frac{2\sqrt[3]{s}(2\sqrt[3]{s}+1)}{2} \leq 4s^{\frac{2}{3}}$$

.

The size of the transition circuit F_2 for the other $k-1$ input letters can be estimated from Theorem 1.3:

$$C(F_2) \lesssim \frac{(k-1)s\lceil \log s \rceil}{\log((k-1)s\lceil \log s \rceil)} \lesssim (k-1)s.$$

The size of the acceptance circuit can be estimated (from Theorem 1.3) as $C(G) \lesssim \frac{s}{\log s}$.

After reordering of states we also have to ensure that the start state is 0. This can increase the complexity of both circuits by no more than $3 \log s$ (it is necessary to add at most n negations to the input of the transition circuit F, the output of the transition circuit F, and the input of the acceptance circuit G). There are also $\log s$ state bits which are included in the computation of BC-complexity. We omit these terms of logarithmic order in the computation of BC-complexity because asymptotically they are negligible.

The BC-complexity of the automaton therefore can be estimated as

$$C_{\mathrm{BC}}(A) \leq C(F) + C(G) + b_Q \lesssim 4c(\log s)^2 s^{2/3} + \frac{s}{\log s} + (k-1)s$$

If $k \geq 2$ then the dominant term of this expression is $(k-1)s$ and $C_{\mathrm{BC}}(A) \lesssim (k-1)s$. For one letter alphabet the dominant term is $s/\log s$, therefore $C_{\mathrm{BC}}(A) \lesssim \frac{s}{\log s}$. □

Consider language L_n in binary alphabet $\Sigma = \{0,1\}$ such that $x \in L_n$ iff $|x| = k$ and $x_{k-n+1} = 1$ (the n-th letter from the end is "1"). The state complexity of this language is 2^n, one has to remember in a state register the last n input letters. But the BC-complexity of it is n. Circuits F, G that represent the natural encoding of a DFA A_n that recognizes L_n have no gates, they are shown in figure 1. Therefore the BC-complexity of (the representation (F,G) of) A_n is the number of state bits which is n. This example shows that the lower bound of Theorem 3.1 is strictly reachable.

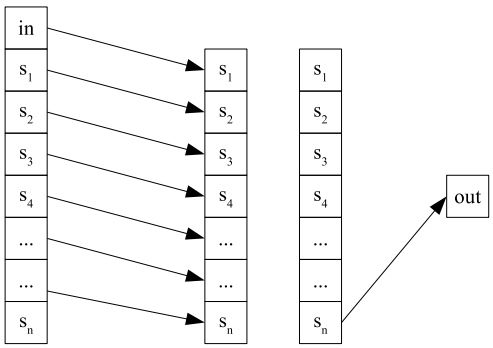

Figure 1: Representation (F,G) of the DFA A_n

Further we try to reach the upper bound. First we find a language (based on the Shannon function) for which the BC-complexity is at least $s/\log^2(s)$, afterwards by counting argument we show that BC-complexity for most languages is close to $(k-1)s$. That matches the upper bound of Theorem 3.1 and can be thought of as the Shannon effect for BC-complexity.

Denote by Sh_n the Shannon function on n bits: lexicographically first Boolean function with n input bits and one output bit with maximal complexity of its minimal circuit. Consider a language L_n^{Sh} that consists of all words $x_1x_2\ldots x_k$ in binary alphabet such that $Sh_n(x_{k-n+1}, x_{k-n+2}, \ldots, x_k) = 1$. State complexity of this language is not larger than 2^n: it is enough to remember the last n input letters. But its BC-complexity is at least $2^n/n^2$.

Theorem 3.2 *BC-complexity of L_n^{Sh} is at least $2^n/n^2$.*

Proof Let (F,G) be a pair of Boolean circuits that represents some DFA A_n recognizing L_n^{Sh}. Assume F has one input bit that represents input letter from the tape and $m = b_Q$ state bits. By concatenating n circuits F together with one circuit G as in figure 2. (state bit output of j-th circuit is passed as state bit input of $j+1$-st) one can obtain a circuit whose size is not larger than $nC(F)+C(G)$ and which computes Shannon function Sh_n on its n input bits. From Theorem 1.3 the complexity of this circuit is at least $2^n/n$. From $nC(F)+C(G) > 2^n/n$ we get that $C(F)+C(G) > 2^n/n^2$. □

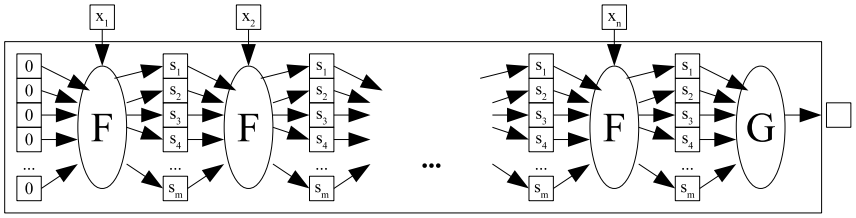

Figure 2: Circuit construction for the Shannon function Sh_n

Theorem 3.2 shows that for some language L_n^{Sh} with s states its BC-complexity is at least $s/(\log s)^2$. Next theorem is an extension of this result. With the use of nonconstructive methods (counting argument) one can show that this value can be raised up to $(k-1)s$. But in the beginning we will need a formula to estimate the number of automata with a given BC-complexity.

Theorem 3.3 *Fix Σ and denote $\mathfrak{A}(c)$ to be the class of those minimal DFAs whose BC-complexity is less than c. If $|\Sigma| = k \geq 2$ then for any $\varepsilon > 0$*

$$\lim_{s\to\infty} \frac{|\mathfrak{A}((1-\varepsilon)(k-1)s)|}{|\mathfrak{A}_s|} = 0$$

If $|\Sigma| = 1$ then for any $\varepsilon > 0$

$$\lim_{s\to\infty} \frac{|\mathfrak{A}((1-\varepsilon)\frac{s}{\log s})|}{|\mathfrak{A}_s|} = 0$$

Proof By Theorem 1.1 $\mathfrak{A}_s \geq 2^s s^{(k-1)s}$. Denote $l = 2^k$, it is clear that no more than l input bits for data input will be used for the representation for which BC-complexity is minimal. If more bits are used, then some of them will be equal as there are only 2^k functions that maps k inputs letters to $\{0,1\}(bits)$.

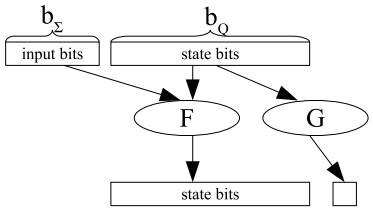

Figure 3: Merged acceptance and transition circuits

Consider a representation (F,G) of some encoding $E(A)$ of A. Merge these two circuits F and G and obtain one circuit H with $b_Q + b_\Sigma$ inputs and $b_Q + 1$ output bits, the first b_Q of which correspond to the output of the transition circuit F, but the last output bit corresponds to the output of the acceptance circuit G (Figure 3). The complexity of this circuit H is $C(F) + C(G)$, for any two minimal automata these "merged" circuits will be different.

Now we want to estimate the number of representations with BC-complexity less that c. Such representations have at least $\lceil \log s \rceil$ and no more than c state bits. The complexity of the "merged" circuit H for a representation with b_Q state bits cannot be more than $c - b_Q$, the number of such circuits H from theorem 1.2 is not larger than

$$N(b_Q + b_\Sigma, b_Q + 1, c - b_Q) < N(b_Q + l, b_Q + 1, c - b_Q) < 9^{c+l}(c+l)^{c+1}.$$

Therefore the number of representations with complexity c is not larger than

$$\sum_{b_Q=1}^{c} N(b_Q + l, b_Q + 1, c - b_Q) < c9^{c+l}(c+l)^{c+1} < 9^{c+l}(c+l)^{c+2}.$$

To prove the theorem we have to show that

$$\lim_{s\to\infty} \frac{9^{c+l}(c+l)^{c+2}}{2^s s^{(k-1)s}} = 0$$

or what is equivalent to that

$$\lim_{s\to\infty} \log\left(9^{c+l}(c+l)^{c+2}\right) - \log\left(2^s s^{(k-1)s}\right) = -\infty$$

for the stated values of c.

For $k \geq 2$ if we substitute $c = (1-\varepsilon)(k-1)s$ then after simplification we obtain an equation of the form

$$\lim_{s\to\infty} -\varepsilon s \log s + O(s) = -\infty$$

which is true. The same happens in the case $k = 1$ if we substitute $c = (1-\varepsilon)s/(\log s)$. □

We have shown in Theorem 3.1 that BC-complexity for any regular language with state complexity s and input alphabet of size $k \geq 2$ is not "much larger" than $(k-1)s$. Theorem 3.3 states that for minimal encodings recognition of almost all such languages would require circuits of size around $(k-1)s$. This can be thought of as the "Shannon effect" for the BC-complexity of automata: for almost all automata its value is close to the maximum.

4 Minimization of BC-complexity

For the state complexity of DFA an efficient minimization algorithm [6] is well known which, given a DFA, finds the state complexity of it as well as the the minimal DFA itself. This is in a big contrast with complexity measures of general programs (Turing machines) for which their complexity (space or time) cannot be determined by any means in the general case.

It is easy to notice that finding the BC-complexity of a DFA is NP-hard.

Theorem 4.1 *Finding the BC-complexity of a DFA given its arbitrary representation is NP-hard.*

Proof We will reduce SAT problem to finding minimal BC-complexity of a DFA. Given a SAT problem instance that contains n variables, consider a DFA with n state bits (2^n states), that works in one letter alphabet, its state transition function is a "circle", that goes through all the states, and accepting states are those, for which this SAT instance gives positive output.

Now assume that this SAT instance is not satisfiable — then this DFA never accepts and therefore its minimal DFA has 1 state (0 state bits) and its BC-complexity is 0. If this SAT instance is satisfiable, then any representation of it will have some state bits and therefore its BC-complexity will be at least 1. Therefore if one could efficiently find BC-complexity of a given DFA, he could also solve any SAT problem. □

Further we show one interesting property of BC-complexity — that for some DFAs BC-complexity is significantly smaller than for their equivalent minimal DFAs. The theorem is based on the conjecture that $PSPACE \not\subseteq P/Poly$. The proof of this theorem for transducers can be found in [9], for DFAs it is almost the same and is omitted here. Denote by $M(L)$ the minimal DFA recognizing language L.

Theorem 4.2 *If there is a polynomial $p(x)$ such that $C_{\mathrm{BC}}(M(L)) < p(C_{\mathrm{BC}}(L))$ for all regular languages L then $PSPACE \subseteq P/Poly$.*

It means that in some cases by minimizing the number of states (minimizing state complexity) BC-complexity of the transition function can increase superpolynomially. And on the other hand, sometimes allowing equivalent states in the automaton helps to keep BC-complexity small.

5 BC-complexity applications

5.1 Nondeterministic automata

Theorems 3.3 and 3.1 suggest that for most DFAs in k-letter alphabet with s states BC-complexity is around $(k-1)s$. But in many cases when DFAs with a large state space are constructed by some standard method, it turns out that their BC-complexity is exponentially smaller than this maximal expected value — it is of order $Polylog(s)$. Further we look at some of these standard constructions starting with the determinization of an NFA.

Theorem 5.1 *If a language R over alphabet Σ, $|\Sigma| = k$ can be recognized by an NFA N with n states and t transitions, then it can also be recognized by a DFA A for which $C_{BC}(A) \leq t + (k+1)n + k\log k$.*

Proof Consider a DFA A that is obtained by a standard construction from NFA N. Its set of states is the powerset of the set of states of N. The state space of A will consist of 2^n states (may be some of them will not be reachable), which can be encoded in n state bits. Each state bit of an encoding of A will correspond to one state of N. For input letters we choose arbitrary minimal input encoding into $\log k$ bits.

The transition circuit of A can be obtained from the transition function of N. NFA N after reading input letter $x \in \Sigma$ will be in state q_i, if there is a state q_j, in which it was before (NFA can be in many states simultaneously) and from which reading input letter x leads to state q_i. Denote by Q_a^i subset of states of N from which reading letter a leads to state q_i. Denote by Q_t a subset of states in which N is after reading t letters. If N reads input letter a in step t then:

$$q_i \in Q_{t+1} \leftrightarrow (Q_t \cap Q_a^i) \neq \emptyset.$$

In the circuit it means that if x denotes the encoded input letter then

$$q_i' = \bigvee_{a \in \Sigma} ((x = a) \& \bigvee_{q \in Q_a^i} q).$$

To construct all k subcircuits $x = a$ we need $\log k$ negations and $k(\log k - 1)$ conjunctions.

The size of the block $\& \bigvee_{q \in Q_a^i} q$ is the number of transitions entering state q on input a therefore the total number of these inner disjunctions and conjunctions for all output bits q_i' is t. There are also $(k-1)n$ outer disjunctions $\bigvee_{a \in \Sigma}$. In total the complexity of the transition circuit is not larger than $k(\log k - 1) + \log k + (k-1)n + t \leq t + (k-1)n + k\log k$.

Acceptance circuit G is a disjunction of all the final states of N, the complexity of this it is not larger than $n-1$. Also $b_Q = n$ have to be added to the BC-complexity. Therefore the total BC-complexity of A is not larger than $t + (k+1)n + k\log k$. □

As the number of transitions is not larger than kn^2 then

Corollary 5.2 *If a language R in alphabet Σ, $|\Sigma| = k$ can be recognized with an NFA N with n states, then it can also be recognized with a DFA A for which $C_{BC}(A) \leq kn^2 + (k+1)n + k\log k$.*

5.2 Language operations

State complexity of language operations has been studied long ago, e.g. in [10]. The result of some of the operations (e.g. reversing) can lead to exponentially larger automata than the original one. Here we analyze how BC-complexity changes with languages operations and observe that in those cases when

the state complexity increases exponentially it leads to automata whose state transition function is very structured therefore its BC-complexity is exponentially smaller than state complexity.

For all operations we assume that we are given two languages L_1 and L_2 and $m = C_s(L_1)$, $n = C_s(L_2)$, $a = C_{\mathrm{BC}}(L_1)$, $b = C_{\mathrm{BC}}(L_2)$, $k = |\Sigma|$. We start with the union and intersection.

Theorem 5.3 *If $L_3 = L_1 \cup L_2$ or $L_3 = L_1 \cap L_2$ then $C_{\mathrm{BC}}(L_3) \leq a + b + 1$.*

Proof Assume circuits (F_1, G_1) represent a DFA recognizing L_1 and (F_2, G_2) represent a DFA recognizing L_2. The transition function for a DFA recognizing L_3 would consist of circuits F_1 and F_2 working in parallel. The acceptance circuit consists of circuits G_1 and G_2 working on corresponding parts of bit vector followed by a disjunction (for union) or conjunction (for intersection) gate. The number of state bits is the sum of state bits for representations (F_1, G_1) and (F_2, G_2). The complexity of such a representation is $C(F_1) + C(F_2) + C(G_1) + C(G_2) + 1 + b_Q = a + b + 1$. □

The complement of the language can be computed by the same pair of circuits as the language itself with negation added at the end of the acceptance circuit.

Theorem 5.4 *If $L_3 = \Sigma^* \setminus L_1$ then $C_{\mathrm{BC}}(L_3) \leq a + 1$.*

A word $x_1 x_2 \ldots x_n$ belongs to the reverse language L_1^R iff $x_n \ldots x_2 x_1$ belongs to L_1. NFA N recognizing L_1^R can be obtained from the DFA A recognizing L_1 by setting the start state of N to be any accepting state of A, setting q_0 of A to be the only accepting state of N and reversing all the arrows. DFA recognizing L_1^R can be obtained from N by running the standard process of determinization.

Theorem 5.5 $C_{BC}(L_1^R) \leq (2k+1)m + k \log k$

Proof This follows directly from Theorem 5.1 and the fact, that NFA obtained by reversing all the transitions has exactly km transitions. □

Language $L_1 L_2$ which is the concatenation of languages L_1 and L_2 consists of all words uw such that $u \in L_1$ and $w \in L_2$.

Theorem 5.6 $C_{\mathrm{BC}}(L_1 L_2) \leq a + (2k+1)n + k \log k$

Proof Assume DFA A_1 recognizes L_1, DFA A_2 recognizes L_2. NFA that recognizes $L_1 L_2$ can be obtained from A_1 and A_2 by adding ε-transitions from all the accepting states of A_1 to the start state of A_2. The standard construction of DFA from this NFA can be optimized — it will consist of circuits F_1 and G_1 representing A_1 together with a circuit $N(A_2)$ constructed from A_2 as from NFA as in Theorem 5.1. Circuit G_1 sets state bit corresponding to state q_0 of A_2 to "1" iff A_1 is in accepting state.

By Theorem 5.1 $C(N(A_2)) \leq t + (k+1)n + k \log k$ and, since A_2 is a deterministic automaton, $t = kn$. Together it gives that $C_{\mathrm{BC}}(L_1 L_2) \leq C(F_1) + C(G_1) + C(N(A_2)) \leq a + kn + (k+1)n + k \log k = a + (2k+1)n + k \log k$. □

Theorem 5.7 $C_{\mathrm{BC}}(L_1^*) \leq km^2 + (k+1)m + k \log k$.

Proof NFA recognizing L_1^* can be obtained from DFA recognizing L_1 by adding ε-transitions from all the accepting states to the start state. The resulting NFA therefore also has m states and the result follows from Corollary 5.2. □

Operation	State complexity	BC-complexity
$L_1 \cup L_2$	mn	$a+b+1$
$L_1 \cap L_2$	mn	$a+b+1$
$\Sigma^* - L_1$	m	$a+1$
L^R	2^m	$(2k+1)m + k\log k$
$L_1 L_2$	$(2m-1)2^{n-1}$	$a+(2k+1)n + k\log k$
L_1^*	$2^{m-1}+2^{m-2}$	$km^2+(k+1)m + k\log k$

Table 1: State complexity and BC-complexity of language operations

6 Conclusions and open problems

In this paper a new measure of complexity, BC-complexity of DFAs and regular languages, was considered. Transition function of a DFA as well as the characteristic function of the set of accepting states are expressed as Boolean circuits and their circuit complexity is taken as a complexity measure (BC-complexity) of this DFA. It turns out that BC-complexity can vary exponentially for DFA with the same number of states (Theorem 3.1). Theorem 3.3 states that almost all DFAs BC-complexity is close to maximum (”Shannon effect”).

In all asymptotic constructions minimal encodings for state and input alphabet where used, but it is not known if minimal encodings are always optimal. We think that sometimes they are not, but showing an example where other encoding than minimal would be more efficient (in the sense of minimizing BC-complexity) is an interesting open question.

In section 4 it was shown that BC-complexity of a regular language can be much smaller than the BC-complexity of the minimal DFA that recognizes it. On the other hand, DFAs with a large state space that are obtained in many standard operations (determinization of NFA, language operations), have a ”good” structure so that their BC-complexity can be relatively small.

References

[1] Trakhtenbrot B. Barzdins J.: *Finite Automata: Behavior and synthesis*. Science, Moscow.

[2] Michael Domaratzki, Derek Kisman & Jeffrey Shallit (2002): *On the number of distinct languages accepted by finite automata with n states. Journal of Automata, Languages and Combinatorics* 7(4), pp. 469–486.

[3] Gregor Gramlich & Georg Schnitger (2007): *Minimizing nfa's and regular expressions. Journal of Computer and System Sciences* 73(6), pp. 908 – 923, doi:10.1016/j.jcss.2006.11.002.

[4] Mealy George H. (1955): *A method for synthesizing sequential circuits. Bell System Technical Journal* 34(5), pp. 1045–1079, doi:10.1002/j.1538-7305.1955.tb03788.x.

[5] J.E. Hopcroft & J.D. Ullman (1979): *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Cambridge.

[6] John Hopcroft (1971): *An $n\log n$ Algorithm for Minimizing States in a Finite Automaton. Theory of Machines and Computations*, pp. 189–196.

[7] Lupanov O.B. (1984): *Asymptotic Estimates of Complexity of Control Systems*. Moscow University Press.

[8] Tao R. (2009): *Finite Automata and Application to Cryptography*. Springer.

[9] Maris Valdats (2011): *Transition Function Complexity of Finite Automata*. In Markus Holzer, Martin Kutrib & Giovanni Pighizzini, editors: *Proc. of DCFS, Lecture Notes in Computer Science* 6808, Springer, pp. 301–313, doi:10.1007/978-3-642-22600-7.

[10] Sheng Yu (2000): *State Complexity of Regular Languages*. *Journal of Automata, Languages and Combinatorics* 6, pp. 221–234.