Computational Logic and Applications, CLA '05 DMTCS proc. **AF**, 2006, 69–76

Undecidable problems concerning densities of languages

Jakub Kozik

Jagiellonian University, Faculty of Mathematics and Computer Science, Institute of Computer Science, ul. Nawojki 11, 30-072 Kraków, Poland

In this paper we prove that the question whether a language presented by a context free grammar has density, is undecidable. Moreover we show that there is no algorithm which, given two unambiguous context free grammars on input, decides whether the language defined by the first grammar has a relative density in the language defined by the second one. Our techniques can be extended to show that this problem is undecidable even for languages given by grammars from $LL(k)$ (for sufficiently large fixed $k \in \mathbb{N}$).

Keywords: asymptotic density, context-free languages, decidability

In the theory of formal languages the notion of *density* was introduced by Berstel in [Ber72]; density for regular languages was studied extensively (comp. [SS78]). There theory of formal power series proved to be the main tool in this development.

A second approach was presented in [BGvOS04]. The authors used theory of Markov chains to obtain some results concerning computational complexity of the problem of density for regular languages.

A natural extension of the notion of density is a *relative density* (also introduced in [Ber72]). The relative density of a language S in L is the limit probability that a uniformly chosen word of a bounded length from L belongs to S. Using this definition many problems concerning densities (comp. [Zai05], [CFGG04]) can be reformulated in the theory of languages. The problem whether, for a given pair of languages, the first language has a relative density in the second one is decidable for the regular languages ([Koz05]).

In the first part of this paper we show that the existence of the density for a language given by a context free language, is undecidable.

In the second part we focus on unambiguous context free languages. The well-known theorem of Chomsky and Shützenberger states that generating functions of unambiguous context free languages are algebraic. It has been shown in [Ber72] that if there exists a density of some unambiguous context free language it is an algebraic number. This observation (with several refinements) has been used in [Kem80] and [Fla87] to prove the inherent ambiguity of several context free languages.

In the second part we address the following question: given a pair of unambiguous context free grammars, decide whether the language defined by the first one has a density in the language defined by the second one. We show that there is no algorithm answering this question. The problem whether a given context free grammar is unambiguous is undecidable, and thus our question is defined on a non-recursive

set. An easy modification of our proof shows that the problem of existence of a relative density is undecidable even for grammars belonging to $LL(k)$ (for sufficiently large fixed $k \in \mathbb{N}$). Note that for any fixed $k \in \mathbb{N}$ the set of $LL(k)$-grammars is recursive.

It is shown in [Koz06] that the question is decidable if we assume that the language defined by the first grammar is a subset of the language defined by the second one.

1 Relative densities

We denote the empty word by ε. For a language L over some finite alphabet we let $L(n)$ denote the number of words of length n in L.

The density of a language can be defined in a number of ways. Some authors ([Koz05], [BGvOS04]) define the density of a language L over an alphabet Σ as

$$d(L) = \lim_{n\to\infty} \frac{\mathrm{Card}(L \cap \Sigma^n)}{\mathrm{Card}(\Sigma)^n},$$

which is an asymptotic probability that an uniformly chosen word of length n belongs to L. We call the number $d(L)$ the density or the ordinary density of the language L. This definition is usually used for structures like trees ([Zai05], [CFGG04]) and graphs.

On the other hand, J. Berstel ([Ber72], [SS78]) uses the following definition:

$$d_c(L) = \lim_{n\to\infty} \frac{\sum_{i=0}^{n} \mathrm{Card}(L \cap \Sigma^i)}{\sum_{i=0}^{n} \mathrm{Card}(\Sigma)^i}.$$

We refer to the number $d_c(L)$ as to the cumulative density of a language L.

It is an easy exercise to show that a language L over a non unary alphabet has density if and only if it has cumulative density (and in such a case $d(L) = d_c(L)$). Thus the decidability of the problem does not depend on the definition we use.

Throughout the paper we use the notion of a *step language*. For a fixed alphabet Σ we will say that the language $L \subset \Sigma^*$ is a *step language* if there exists $q \in \mathbb{Q}$ and $N \in \mathbb{N}$ such that:

$$\frac{L(n)}{\mathrm{Card}(\Sigma)^n} = \begin{cases} 0, & n < N; \\ q, & n \geq N. \end{cases}$$

Clearly every step language has the density q.

Although the function of density is not countably additive on disjoint languages, for disjoint step languages it is.

Lemma 1.1 *If $\{L_i\}_{i\in\mathbb{N}}$ is a family of pairwise disjoint nonempty step languages then:*

$$d\Big(\bigcup_{i\in\mathbb{N}} L_i\Big) = \sum_{i\in\mathbb{N}} d(L_i)$$

Proof: Let $L = \bigcup_{i\in\mathbb{N}} L_i$. Since d is finitely additive (on disjoint sets) we get $\sum_{i=0}^{N} d(L_i) < 1$ for any $N \in \mathbb{N}$. Therefore the series $\sum_{i\in\mathbb{N}} d(L_i)$ is convergent.

Let us assume that the ordering of the languages L_i is such that for every $i \in \mathbb{N}$ the shortest word from L_i is no longer then the shortest word from L_{i+1}. Note that for every $n \in \mathbb{N}$ there is finitely many languages L_i with words of length n. We define $\nu : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$ a language L_i contains words of length n if and only if $i \leq \nu(n)$. Then

$$\frac{L(n)}{\mathrm{Card}(\Sigma)^n} = \sum_{i=0}^{\nu(n)} \frac{L_i(n)}{\mathrm{Card}(\Sigma)^n} = \sum_{i=0}^{\nu(n)} d(L_i).$$

The last equality follows from the fact that each language L_i is a step language and contains words of length n. It shows that:

$$d(L) = \lim_{n\to\infty} \frac{L(n)}{\mathrm{Card}(\Sigma)^n} = \sum_{i=0}^{\infty} d(L_i).$$

□

We define the relative density of a language S in a language L to be

$$d(S|L) = \lim_{n\to\infty} \frac{\mathrm{Card}(S \cap L \cap \Sigma^n)}{\mathrm{Card}(L \cap \Sigma^n)}, \tag{1}$$

and its cumulative version as

$$d_c(S|L) = \lim_{n\to\infty} \frac{\sum_{i=0}^{n} \mathrm{Card}(S \cap L \cap \Sigma^i)}{\sum_{i=0}^{n} \mathrm{Card}(L \cap \Sigma^i)}.$$

Note that the fraction in (1) is well-defined if $L \cap \Sigma^n \neq \emptyset$. By an abuse of notation, we say that a limit is well-defined if there are words of almost every length in L. By the same token the relative cumulative density requires $L \neq \emptyset$. It is easy to see that for any two languages L, S the existence of $d(S|L)$ implies the existence of $d_c(S|L)$ and that the inverse implication does not hold.

In the second part of the paper we show that there is no algorithm which, given a pair of unambiguous context free grammars (G_1, G_2) on input, decides whether $d(L_1|L_2)$ exists (where L_1, L_2 are the languages defined by G_1, G_2 respectively). The proofs we present can be applied directly to prove the same theorems for relative cumulative density.

Finally, we note that the definition of relative density given by Berstel in [Ber72] assumes that $S \subset L$. For regular languages, considered by Berstel, the difference is insignificant (since regular languages are closed with respect to intersections). Unambiguous context free languages are not closed with respect to the intersection and, moreover, the problem of inclusion of two languages given by unambiguous context free grammars is undecidable. If we restrict our problem to pairs of grammars such that the inclusion holds for the corresponding languages, then it becomes decidable (as was shown in [Koz06]).

2 Density of a context free language

The problem of existence of density in full language is undecidable even for context free languages. To show this we encode Post Correspondence Problem into the problem of density. Our proof is a modification of the standard argument for the undecidability of the problem of emptiness of intersection of two languages given by context free grammars.

Theorem 2.1 *The question of existence of ordinary density for a language given by a context free grammar is undecidable.*

Proof: Let $\{(\alpha_i, \beta_i)\}_{i\in\{1,\dots,n\}}$ be an instance of Post Correspondence Problem over the alphabet $\Sigma = \{a, b\}$. Let $\Gamma = \{a, b, c, d\}$ and L_1 be a language generated by the grammar:

$$L_1 \to \alpha_1 L_1 \overleftarrow{\beta_1} | \dots | \alpha_n L_1 \overleftarrow{\beta_n} \tag{2}$$

$$L_1 \to cFc \tag{3}$$

$$F \to aF|bF|cF|dF|\varepsilon$$

where $\overleftarrow{\beta_i}$ denotes reverted word β_i. We use the same symbol to denote the starting symbol of the grammar and the whole language it generates. We assign numbers to productions from (2) in such a way that the production $L_1 \to \alpha_i L_1 \overleftarrow{\beta_i}$ has number i.

Let W be the set of all pairs of words that factorize into pairs of words from the instance of the PCP. Formally $(w, v) \in W$ if and only if there exist $k \in \mathbb{N}$ and $\nu \in \mathbb{N}^k$ such that $w = \alpha_{\nu_1} \dots \alpha_{\nu_k}$ and $v = \beta_{\nu_1} \dots \beta_{\nu_k}$. We can rewrite the definition of L_1 as follows:

$$L_1 = \bigcup_{(w,v)\in W} wc\Gamma^* c\overleftarrow{v}.$$

Note that for two different pairs of words $(w_1, v_1), (w_2, v_2)$ languages $w_1c\Gamma^*cv_1$ and $w_2c\Gamma^*cv_2$ are disjoint. For every pair (w, v) the language $wc\Gamma^*cv$ is a step language with its density equal to $\frac{1}{\mathrm{Card}(\Gamma)^{|w|+|v|+2}}$. Hence L_1 is a disjoint union of step languages with positive densities and as a result of Lemma 1.1 it has positive density as well.

Let L_2 be the language generated by

$$L_2 \to aL_2a|bL_2b|cFc,$$

$$F \to aF|bF|cF|dF|\varepsilon.$$

It consists of words of a form $wcvc\overleftarrow{w}$ for some $w \in \Sigma^*$ and $v \in \Gamma^*$. Therefore

$$L_2 = \bigcup_{w\in\Sigma^*} wc\Gamma^* c\overleftarrow{w}.$$

Therefore L_2 is a disjoint union of step languages with positive density and, by Lemma 1.1, positive density for L_2 exists.

The language $L_1 \cap L_2$ is not empty if and only if a given instance of PCP has a solution. Suppose there is a solution of PCP. Let $(i_1, \dots, i_k)$ be the sequence of numbers of the pairs in that solution. By definition it implies that $\alpha_{i_1} \dots \alpha_{i_k} = \beta_{i_1} \dots \beta_{i_k}$. Using the same sequence of productions in the first grammar, production (3), and $F \to \varepsilon$ we obtain a word $\alpha_{i_1} \dots \alpha_{i_k} cc\overleftarrow{\beta_{i_k}} \dots \overleftarrow{\beta_{i_1}} = wcc\overleftarrow{w}$ which belongs to L_2, and hence the intersection is not empty. On the other hand whenever the intersection is not empty it contains some word $wcvc\overleftarrow{w}$ for $w \in \Sigma^*$. Let us consider a derivation of this word in the first grammar. From the construction of the grammar the first and the last occurrence of c must be generated by production (3). Therefore words w and $\overleftarrow{w}$ are generated by productions from (2). Let $(i_1, \dots, i_k)$ be the sequence of

numbers of these productions. Then we have $w = \alpha_{i_1} \dots \alpha_{i_k}$ and $\overleftarrow{w} = \overleftarrow{\beta_{i_k}} \dots \overleftarrow{\beta_{i_1}}$. Hence the sequence $(i_1, \dots, i_k)$ is a solution of the given instance of PCP.

Moreover whenever the intersection contains a word $wcvc\overleftarrow{w}$ for some $w \in \Sigma^*$, it also includes the whole language $wc\Gamma^* c\overleftarrow{w}$. If this intersection contains $w_1 c v_1 c\overleftarrow{w_1}$ and $w_2 c v_2 c\overleftarrow{w_2}$ for different $w_1, w_2 \in \Sigma^*$, the languages $w_1 c\Gamma^* c\overleftarrow{w_1}$ and $w_2 c\Gamma^* c\overleftarrow{w_2}$ are disjoint.

Let $P \in \Sigma^*$ be the language of words w of the form $w = \alpha_{i_1} \dots \alpha_{i_k} = \beta_{i_1} \dots \beta_{i_k}$ for some $(i_1, \dots, i_k) \in \mathbb{N}$. Trivially

$$L_1 \cap L_2 = \bigcup_{w \in P} wc\Gamma^* c\overleftarrow{w}$$

and by previous considerations we know that all languages which are summed are disjoint step languages with positive densities. By Lemma 1.1 $L_1 \cap L_2$ has positive density as well.

We conclude by saying that $L_1 \cap L_2$ has a positive density if and only if corresponding instance of PCP has a solution, and density 0 otherwise.

The last step is to modify the second grammar. Let L_2' be a language generated by the grammar below:

$$L_2' \to aL_2'a|bL_2'b|cF_1c|dF_2d$$

$$F_2 \to aF_1|bF_1|cF_1|dF_1|\varepsilon$$

$$F_1 \to aF_2|bF_2|cF_2|dF_2.$$

It is easily seen that

$$L_2' = \bigcup_{w \in \Sigma^*} (wc\Gamma(\Gamma\Gamma)^* c\overleftarrow{w} \cup wd(\Gamma\Gamma)^* d\overleftarrow{w}).$$

Since language L_2' has exactly the same number of words of every length than language L_2, it has the same density.

Analogous considerations show that the instance of PCP has a solution if and only if $L_1 \cap L_2'$ is not empty.

Let us assume that the given PCP has a solution. Let $P \in \Sigma^*$ be the language of words w of the form $w = \alpha_{i_1} \dots \alpha_{i_k} = \beta_{i_1} \dots \beta_{i_k}$ for some $(i_1, \dots, i_k) \in \mathbb{N}$.

Then $I = L_1 \cap L_2'$ is an union of distinct languages $wc\Gamma(\Gamma\Gamma)^* c\overleftarrow{w}$ for $w \in P$. Let $S = wc\Gamma(\Gamma\Gamma)^* c\overleftarrow{w}$ for some $w \in P$. Then S has no density since

$$\lim_{k\to\infty} \frac{S(2k)}{\mathrm{Card}(\Gamma)^{2k}} = \lim_{k\to\infty} \frac{0}{\mathrm{Card}(\Gamma)^{2k}} = 0$$

and

$$\lim_{k\to\infty} \frac{S(2k+1)}{\mathrm{Card}(\Gamma)^{2k+1}} = \lim_{k\to\infty} \frac{\mathrm{Card}(\Gamma)^{2k+1-2(|w|+1)}}{\mathrm{Card}(\Gamma)^{2k+1}} = \frac{1}{\mathrm{Card}(\Gamma)^{2(|w|+1)}}.$$

Thus $I(2n) = 0$ and $I(2n+1) \geq S(2n+1)$. We proved that the given instance of PCP has a solution if and only if $L_1 \cap L_2'$ has no density.

Unfortunately, context free languages are not closed under intersections so we cannon use $L_1 \cap L_2'$ directly to prove undecidability. Let us consider $L = L_1 \cup L_2'$ instead. If the intersection $L_1 \cap L_2'$ is empty then L is a disjoint union of step languages with positive densities and hence it has positive density. On the other hand we proceed to show that if $L_1 \cap L_2'$ is not empty then $L_1 \cup L_2'$ has no density. Let us denote

the densities of languages L_1 and L_2 by d_1 and d_2 (respectively). Since the intersection $L_1 \cap L_2'$ contains only words of odd length we have $L(2n) = L_1(2n) + L_2'(2n)$ which implies:

$$\lim_{k\to\infty} \frac{L(2k)}{\mathrm{Card}(\Gamma)^{2k}} = \lim_{k\to\infty} \frac{L_1(2k) + L_2'(2k)}{\mathrm{Card}(\Gamma)^{2k}} = d_1 + d_2.$$

When intersection is not empty we got $L_1 \cap L_2 \supset S = wc\Gamma(\Gamma\Gamma)^* c\overleftarrow{w}$ for some $w \in P$. Let $\delta = \frac{1}{\mathrm{Card}(\Gamma)^{2|w|+2}}$. For sufficiently large n we have $\delta = \frac{S(2n+1)}{\mathrm{Card}(\Gamma)^{2n+1}}$ and since $L(2n+1) \leq L_1(2n+1) + L_2'(2n+1) - S(2n+1)$, so we obtain

$$\limsup_{k\to\infty} \frac{L(2k+1)}{\mathrm{Card}(\Gamma)^{2k+1}} \leq \lim_{k\to\infty} \frac{L_1(2k+1) + L_2'(2k+1) - S(2k+1)}{\mathrm{Card}(\Gamma)^{2k+1}} = d_1 + d_2 - \delta.$$

Therefore the sequence $(\frac{L(n)}{\mathrm{Card}(\Gamma)^n})_{n\in\mathbb{N}}$ diverges and as a result $L = L_1 \cup L_2'$ has no density.

We have shown that a given instance of PCP has a solution if and only if the language $L_1 \cup L_2'$ has no density. The language $L_1 \cup L_2'$ is context free, and its grammar can be effectively constructed by adding a new production $L \to L_1 | L_2'$ to the productions of L_1 and L_2'. Note that the grammar for L_1 can be constructed directly from any given instance of PCP, and the grammar for L_2' is always the same. It shows that the problem of existence of density in a full language for a language given by a context free grammar is undecidable. □

3 Unambiguous context free languages

Between regular and context free languages there exists an interesting class of unambiguous context free languages. A context free grammar is unambiguous if every word has at most one derivation in it. The language is unambiguous context free if it is generated by some unambiguous context free grammar. Unfortunately the problem of existence of relative density is undecidable for that class of grammars. To prove this fact we need a modified version of the Post Correspondence Problem.

The language $L \subset \Sigma^*$ is a prefix code if there are no words $w, v \in L$ such that w is a prefix of v. In such a case every word from Σ^* has at most one factorization into elements of L.

A new problem abbreviated by UPCP is defined as follows:

Given finite set of pairs of words $\{(\alpha_i, \beta_i)\}_{i\in\{1,\ldots,n\}}$ on input return true if $\alpha_i \neq \alpha_j$ for $i \neq j$ and $\{\alpha_i\}_{i\in\{1,\ldots,n\}}$ is a prefix code and $\{(\alpha_i, \beta_i)\}_{i\in\{1,\ldots,n\}}$ has a solution as an instance of PCP. Return false otherwise.

Such modification does not change the undecidability of the problem. The encoding of the halting problem of the Post Machine into Post Correspondence Problem from [Man74] constructs systems in which $\{\alpha_i\}_{i\in\{1,\ldots,n\}}$ is a prefix code and $\alpha_i \neq \alpha_j$ for $i \neq j$. Moreover the property of being a prefix code can be easily verified algorithmically for finite languages.

Theorem 3.1 *There is no algorithm which for two unambiguous context free grammars decides whether language defined by the first one has a relative density in the language defined by the second one.*

Proof: Let $\{(\alpha_i, \beta_i)\}_{i\in\{1,\ldots,m\}}$ be an instance of UPCP over alphabet Σ. We construct two unambiguous context free languages such that a given instance of UPCP has a solution if and only if the first language has no density in the second one.

It is easy to decide whether $\{\alpha_i\}_{i=1\ldots m}$ is a prefix code, if it is not or if $\alpha_i = \alpha_j$ for $i \neq j$ we put $L_1 = L_2 = \Sigma^*$. In the remaining case each pair of words (w_1, w_2) has at most one factorization into $\{(\alpha_i, \beta_i)\}_{i\in\{1,\ldots,n\}}$ (since w_1 has at most one). Therefore by construction from the proof of 2.1 we obtain unambiguous languages L_1, L_2'(the second one was unambiguous before).

Let us consider $d(L_1|L_2')$. If the instance of UPCP has no solution, the intersection of languages is empty and $d(L_1|L_2') = 0$.

If, on the other hand, the intersection $I = L_1 \cap L_2'$ is not empty, it contains words of odd length only. Therefore we have

$$\lim_{k\to\infty} \frac{I(2k)}{L_2'(2k)} = 0. \tag{4}$$

Let $S = wc\Gamma(\Gamma\Gamma)^*c\overleftarrow{w} \subset L_1 \cap L_2'$. Let $\delta = \frac{1}{\mathrm{Card}(\Gamma)^{2|w|+2}}$. For sufficiently large n we have $\delta = \frac{S(2n+1)}{\mathrm{Card}(\Gamma)^{2n+1}}$. Then

$$\lim_{k\to\infty} \frac{I(2k+1)}{L_2'(2k+1)} = \lim_{k\to\infty} \frac{I(2k+1)}{\mathrm{Card}(\Gamma)^{2k+1}} \cdot \frac{\mathrm{Card}(\Gamma)^{2k+1}}{L_2'(2k+1)} \geq \frac{\delta}{d_2} \tag{5}$$

where d_2 is the density of L_2' in the full language Γ^*. It follows from (4) and (5) that there is no relative density of L_1 in L_2'.

We proved that the construction used in the proof of Theorem 2.1 gives languages L_1, L_2' such that an instance of $UPCP$ has a solution if and only if there is no relative density of L_1 in L_2'. □

Grammars used in the last proof belong to simpler classes. The grammar constructed for the language L_1 is in $LL(k)$ for $k = \max\{|\alpha_i| : i = 1, \ldots, n\}$ and L_2' is in $LL(1)$. In the encoding of the halting problem of the Post Machine into PCP (from [Man74]) all pairs in a constructed system are short, except the pair which describes the initial configuration. The same encoding used to the halting problem of Post Machine with empty initial word constructs instances of UPCP consisting of short pairs only. Even in such restricted setting the problem remains undecidable. In this way k can be always reduced to 4.

Corollary 3.2 *The problem of relative density for languages presented by context free grammars belonging to $LL(4)$ is undecidable.*

4 Summary

By further modification of the Post Correspondence Problem we can obtain a situation when the constructed language $L_1 \cap L_2'$ is regular. Nevertheless it is still undecidable whether is has density (or equivalently in that case: whether it is empty or not).

On the other hand the following theorem is proved in [Koz06]:

Theorem 3.1' *There exists an algorithm which for two unambiguous context free grammars decides whether the language defined by the first one has relative density in the language defined by the second one,* provided the first language is a subset of the second one.

References

[Ber72] J. Berstel. Sur la densité asymptotique de langages formels. *ICALP*, 1972.

[BGvOS04] M. Bodirsky, T. Gärtner, T. von Oertzen, and J. Schwinghammer. Effciently computing the density of regular languages. *LATIN*, 2004.

[CFGG04] B. Chauvin, P. Flajolet, D. Gardy, and B. Gittenberger. And/or trees revisited. *Combinatorics, Probability & Computing*, 13, 2004.

[Fla87] P. Flajolet. Analytic models and ambiguity of context-free languages. *Theoretical Computer Science*, 49, 1987.

[Kem80] R. Kemp. A note on the density of inherently ambiguous context-free languages. *Acta Inf.*, 14, 1980.

[Koz05] J. Kozik. Conditional densities of regular languages. *Electronic Notes in Theoretical Computer Science*, 14, 2005.

[Koz06] J. Kozik. Decidability of the problem of relative density of languages. in prepartion, 2006.

[Man74] Z. Manna. *Mathematical Theory of Computation.* Mcgraw-Hill Computer Science Series. Mcgraw-Hill College, 1974.

[SS78] A. Salomaa and M. Soittola. *Automata-Theoretic Aspects of Formal Power Series.* Texts and Monographs in Computer Science. Springer-Verlag, 1978.

[Zai05] M. Zaionc. On the asymptotic density of tautologies in logic of implication and negation. *Reports on Mathematical Logic*, 39, 2005.