

New analytic techniques for proving the inherent ambiguity of context-free languages

Florent Koechlin
Inria Grand Est, Loria, Nancy, France

Séminaire Combalgo, LaBRI
January 2023, 24th

Reminder on formal languages

Word: finite sequence of letters: $ab, ba, \varepsilon, \dots$

Formal language: set of words over a finite alphabet Σ .

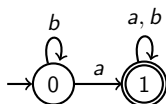
Example 1: $(a + b)^* := \{c_1 \dots c_n : n \in \mathbb{N}, c_i \in \{a, b\}\}$

Example 2: $\{a^n b^n : n \in \mathbb{N}^*\}$

Regular languages

Regular languages are the **simplest** languages in the Chomsky hierarchy. They are exactly the languages recognized by :

- Regular expressions :
 $\Sigma^* a \Sigma^*$, $(a + b)^* b$, $\Sigma^* a \Sigma^{r-1}$, ...

- (Deterministic) finite automata 

Context-free languages

Regular languages \subsetneq Context-free languages

Context-free languages are the **second-level** class of languages in the Chomsky hierarchy. They are exactly the languages recognized by :

- Non-deterministic pushdown automata
- Context-free grammars

$$S \rightarrow aSb \mid \varepsilon, \quad S \rightarrow [S]S \mid \varepsilon, \quad \begin{cases} S \rightarrow aSb \mid C \mid cc \\ C \rightarrow cC \mid c \end{cases}$$

$$S \Rightarrow [S]S \Rightarrow [[S]S]S \Rightarrow [[[S]S]S]S \Rightarrow [[[[S]S]S]S]S \Rightarrow [[[[[S]S]S]S]S]S \Rightarrow [[[[[[S]S]S]S]S]S]S \Rightarrow [[[[[[[S]S]S]S]S]S]S]S$$

$\{a^n b^n \mid n \in \mathbb{N}\}$ is context-free but **not regular**

Context-free grammar

$$S \rightarrow [S]S \mid \varepsilon$$

Derivation

$$\begin{aligned} S &\Rightarrow [S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [S][S] \Rightarrow [S][S] \\ S &\Rightarrow [S][S] \Rightarrow [S][S] \Rightarrow [[S]S] \Rightarrow [[S]S] \Rightarrow [[S]S] \Rightarrow [[S]S] \Rightarrow [S][S] \Rightarrow [S][S] \end{aligned}$$

Context-free grammar

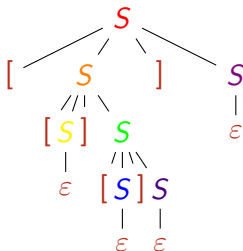
$$S \rightarrow [S]S \mid \varepsilon$$

Derivation

$$S \Rightarrow [S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S$$

$$S \Rightarrow [S]S \Rightarrow [S] \Rightarrow [[S]S] \Rightarrow [[S]S] \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S \Rightarrow [[S]S]S$$

Derivation tree



Unambiguous context-free grammar

Every word in its language has **exactly one** derivation tree.

Unambiguous context-free languages

deterministic CFL \subsetneq unambiguous CFL \subsetneq non det. CFL

$\{a^n b^m c^p \mid n = m \text{ or } m = p\}$ is inherently ambiguous

Relevant intermediate model between deterministic and non-deterministic context-free languages.

Finding inherently ambiguous languages is interesting. However:

- 😞 deciding whether a **grammar** is ambiguous is undecidable [Chomsky-Schützenberger'63]
- 😞 deciding whether a **context-free language** is inherently ambiguous is undecidable [Ginsburg-Ullian'66, Greibach'68]

Standard methods to prove inherent ambiguity

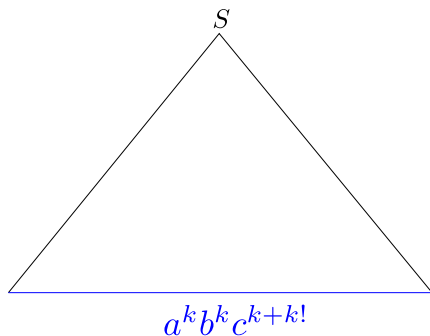
- **Iteration on derivation trees**
By hand or using iteration lemmas (e.g. Ogden's lemma)
- **Iteration on semilinear sets**
- **Generating series**
- **+ closure property** : if R is regular

L unambiguous $\Rightarrow L \cap R$ unambiguous

$L \cap R$ inherently ambiguous $\Rightarrow L$ inherently ambiguous

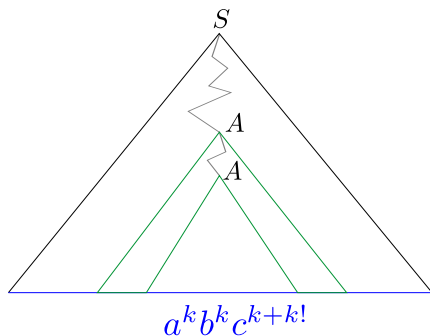
Iteration on trees: $a^n b^m c^p$ with $n = m$ or $m = p$

- Suppose that it is recognized by an unambiguous grammar G
- For k sufficiently big, find an **iterating pair** of a 's and b 's of same length in a derivation of $a^k b^k c^{k+k!}$



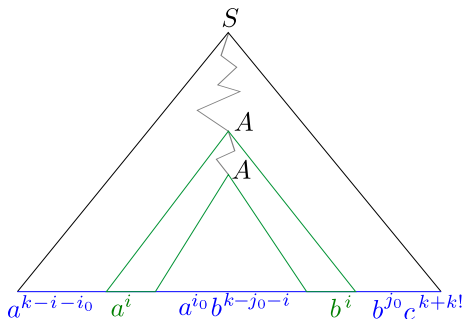
Iteration on trees: $a^n b^m c^p$ with $n = m$ or $m = p$

- Suppose that it is recognized by an unambiguous grammar G
- For k sufficiently big, find an **iterating pair** of a 's and b 's of same length in a derivation of $a^k b^k c^{k+k!}$



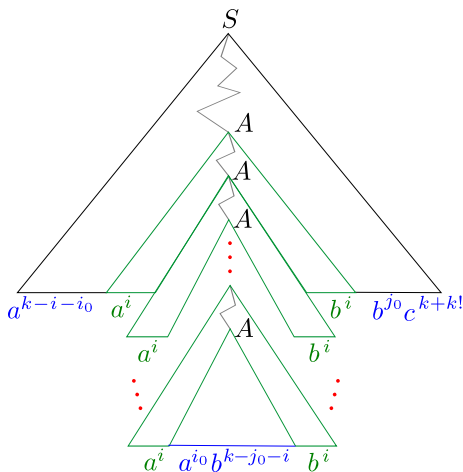
Iteration on trees: $a^n b^m c^p$ with $n = m$ or $m = p$

- Suppose that it is recognized by an unambiguous grammar G
- For k sufficiently big, find an **iterating pair** of a 's and b 's of same length in a derivation of $a^k b^k c^{k+k!}$



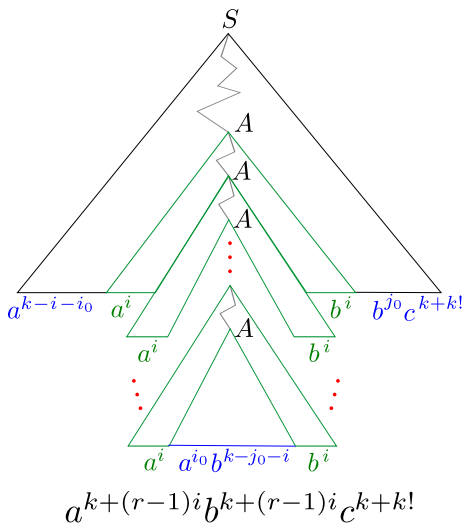
Iteration on trees: $a^n b^m c^p$ with $n = m$ or $m = p$

- Suppose that it is recognized by an unambiguous grammar G
- For k sufficiently big, find an **iterating pair** of a 's and b 's of same length in a derivation of $a^k b^k c^{k+k!}$



Iteration on trees: $a^n b^m c^p$ with $n = m$ or $m = p$

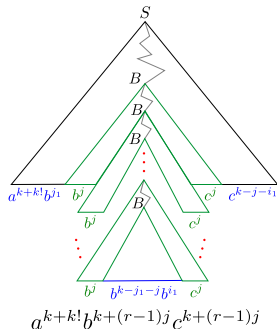
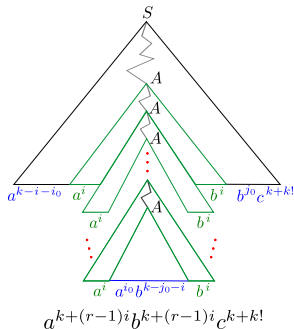
- Suppose that it is recognized by an unambiguous grammar G
- For k sufficiently big, find an **iterating pair** of a 's and b 's of same length in a derivation of $a^k b^k c^{k+k!}$



Iteration on trees: $a^n b^m c^p$ with $n = m$ or $m = p$

Main idea :

- Suppose that it is recognized by an unambiguous grammar G
- For k sufficiently big, find an **iterating pair** of a 's and b 's of same length in a derivation of $a^k b^k c^{k+k!}$
- Derive from it a derivation tree of $a^{k+k!} b^{k+k!} c^{k+k!}$
- Repeat the process from a derivation tree of $a^{k+k!} b^k c^k$ to obtain a different derivation of $a^{k+k!} b^{k+k!} c^{k+k!}$



Methods by iteration

Advantages:

- can handle simple languages that are unreachable with other techniques
- usually bring more information than just inherent ambiguity

Drawbacks:

- are too tedious for complex languages
- are too specific for the studied language
fail on $\{a^n b^m c^p \mid n \neq m \text{ or } m \neq p\}$

Methods based on generating series

Generating series of a language \mathcal{L}

$$L(x) = \sum_{w \in \mathcal{L}} x^{|w|} = \sum_{n \in \mathbb{N}} \ell_n x^n \quad \ell_n : \text{number of words of length } n$$

Example: $(a + b)^*$ $\rightarrow \ell_n = 2^n \rightarrow L(x) = \sum_n 2^n x^n = \frac{1}{1-2x}$

Example: $\{a^n b^n\}$ $\rightarrow \ell_{2n} = 1 \rightarrow L(x) = \frac{1}{1-x^2}$

Theorem [Chomsky-Schützenberger, '63]: The generating series of an **unambiguous** context-free language is **algebraic**.

$$P(x, L(x)) = 0$$

Example: $S \rightarrow [S]S \mid \varepsilon \quad S(x) = x^2 S(x)^2 + 1$

Methods based on generating series

Flajolet's idea: if the series of a context-free language is not algebraic, then it is an **inherently ambiguous context-free language**.

Proposition [Useful criteria, Flajolet '87]:

Let $L(z) = \sum_{n \in \mathbb{N}} \ell_n z^n$ a series.

- If $L(z)$ has **infinitely many singularities**, then $L(z)$ is not algebraic.
- If $\ell_n \sim_{n \rightarrow \infty} \gamma \beta^n n^r$, with $r \notin \mathbb{Q} \setminus \{-1, -2, -3, \dots\}$, then $L(z)$ is not algebraic.
- If ℓ_n does not satisfy a **linear recurrence** with polynomial coefficients in n , then $L(z)$ is not algebraic.

Analytic criteria for inherent ambiguity

Theorem [Flajolet '87]

$\Omega_3 = \{w \in \{a, b, c\}^* : |w|_a \neq |w|_b \text{ or } |w|_b \neq |w|_c\}$ is inherently ambiguous.

Analytic criteria for inherent ambiguity

Theorem [Flajolet '87]

$\Omega_3 = \{w \in \{a, b, c\}^* : |w|_a \neq |w|_b \text{ or } |w|_b \neq |w|_c\}$ is inherently ambiguous.

Analytic proof:

- Suppose that $\Omega_3(x)$ is algebraic
- Let $I = (a + b + c)^* \setminus \Omega_3$
- Then $I(x) = \frac{1}{1-3x} - \Omega_3(x)$ would be algebraic by closure properties
- But $I = \{w \in \{a, b, c\}^* : |w|_a = |w|_b = |w|_c\}$

$$[x^{3n}]I(x) = \binom{3n}{n, n, n} = \frac{(3n)!}{(n!)^3} \sim_{n \rightarrow \infty} 3^{3n} \frac{\sqrt{3}}{2\pi n}$$

Flajolet's analytic method

Advantages :

- is very powerful : P. Flajolet (re)proved the inherent ambiguity of 15 languages, some of which were conjectures, in only one article

$O_3, O_4, \Omega_3, C, S, P_1, P_2, G_{\neq}, G_{<}, G_{>}, G_{=}, H_{\neq}, K_1, K_2, B$

- is robust : it works for both

$\Omega_3 = \{w \in \{a, b, c\}^* : |w|_a \neq |w|_b \text{ or } |w|_b \neq |w|_c\}$ and

$O_3 = \{w \in \{a, b, c\}^* : |w|_a = |w|_b \text{ or } |w|_b = |w|_c\}$.

Remark: $O_3 \cap a^*b^*c^* = \{a^n b^m c^p \text{ with } n = m \text{ or } m = p\}$

Flajolet's analytic method

Advantages :

- is very powerful : P. Flajolet (re)proved the inherent ambiguity of 15 languages, some of which were conjectures, in only one article

$O_3, O_4, \Omega_3, C, S, P_1, P_2, G_{\neq}, G_{<}, G_{>}, G_{=}, H_{\neq}, K_1, K_2, B$

- is robust : it works for both

$\Omega_3 = \{w \in \{a, b, c\}^* : |w|_a \neq |w|_b \text{ or } |w|_b \neq |w|_c\}$ and

$O_3 = \{w \in \{a, b, c\}^* : |w|_a = |w|_b \text{ or } |w|_b = |w|_c\}$.

Remark: $O_3 \cap a^*b^*c^* = \{a^n b^m c^p \text{ with } n = m \text{ or } m = p\}$

Drawbacks:

- does not work on too simple languages, whose series are rational; for instance for:

$$\Omega_3 \cap a^*b^*c^* = \{a^n b^m c^p \text{ with } n \neq m \text{ or } m \neq p\}.$$

$$L(x) = \frac{1}{1-3x} - \frac{1}{1-x^3}$$

In this talk

- rational generating series can still be used to handle the inherent ambiguity of many bounded context-free languages
- I will explain how an old result used to derive iteration on semilinear sets can be derived into two useful criteria on series:
 - The **3-variable criterion**
Re-discover and extension of [Makarov'21]
 - The **interlacing criterion**

Bounded languages

Main question: can we detect the inherent ambiguity of bounded languages using generating series?

A language L is **bounded** with respect to $\langle w \rangle := \langle w_1, \dots, w_d \rangle$ if

$$L \subseteq w_1^* \dots w_d^*$$

where $w^* = \{\varepsilon, w, ww, www, \dots\}$

Example: $\{a^n b^m c^p : \dots\} \subseteq a^* b^* c^*$

Example: $\{(abb)^n (bb)^m c^p : \dots\} \subseteq (abb)^* (bb)^* c^*$

Bounded languages

If L is bounded with respect to $\langle w \rangle$, let us define:

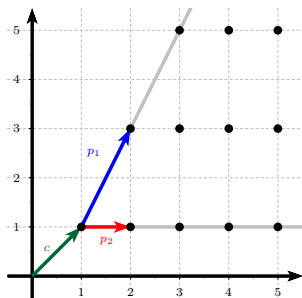
$$\mathcal{S}_{\langle w \rangle}(L) = \{(i_1, \dots, i_d) \in \mathbb{N}^d : w_1^{i_1} \dots w_d^{i_d} \in L\}$$

Example: $\mathcal{S}_{\langle a,b,c \rangle}(\{a^n b^m c^p : n=m \vee m=p\})$
 $= \{(n, m, p) \in \mathbb{N}^3 : n=m \vee m=p\}$

Proposition [Ginsburg and Ullian, '62]: Every bounded context-free language is **semilinear**, i.e., $\mathcal{S}_{\langle w \rangle}(L)$ is semilinear.

Semilinear sets of \mathbb{N}^d (Parikh 61/66)

Linear set: Set of the form $\vec{c} + P^*$ where $P = \{p_1, \dots, p_r\}$ is called a set of **periods**, and $P^* = \{\lambda_1 p_1 + \dots + \lambda_r p_r : \lambda_i \in \mathbb{N}\}$



Semilinear set: Finite union of linear sets : $S = \bigcup_{i=1}^r \vec{c}_i + P_i^*$

Inherent ambiguity of bounded languages

If L is bounded with respect to $\langle w \rangle$:

$$\mathcal{S}_{\langle w \rangle}(L) = \{(p_1, \dots, p_d) \in \mathbb{N}^d : w_1^{p_1} \dots w_d^{p_d} \in L\}$$

Theorem [Ginsburg and Ullian, '66]: A bounded context-free language L is unambiguous **if and only if** $\mathcal{S}_{\langle w \rangle}(L)$ is of the form

$$\mathcal{S}_{\langle w \rangle}(L) = \bigsqcup_{i=1}^r (\vec{c}_i + P_i^*)$$

where:

- the union is disjoint
 - the vectors in each P_i are linearly independent
- } [Eilenberg & Schützenberger, Ito 69]

- each P_i is stratified

Stratified set (Ginsburg and Spanier, '66)

A finite subset $X \subseteq \mathbb{N}^d$ is **stratified** if:

1. every vector in X has **at most two non-zero coordinates**
2. **no two vectors in X have interlacing non-zero coordinates**, i.e. there are no $1 \leq i < j < m < n \leq d$ and two vectors $\vec{x}, \vec{x}' \in X$ such that $x_i x_j x'_m x'_n \neq 0$.

$$\begin{array}{ccc} & \vec{x} & \vec{x}' \\ & \left(\begin{array}{c} \vdots \\ \neq 0 \\ \vdots \\ \vdots \\ \neq 0 \\ \vdots \\ \vdots \end{array} \right) & \left(\begin{array}{c} \vdots \\ \vdots \\ \neq 0 \\ \vdots \\ \vdots \\ \neq 0 \\ \vdots \end{array} \right) \\ i & & j \\ m & & n \end{array}$$

forbidden

Stratified set (Ginsburg and Spanier, '66)

A finite subset $X \subseteq \mathbb{N}^d$ is **stratified** if:

1. every vector in X has **at most two non-zero coordinates**
2. **no two vectors in X have interlacing non-zero coordinates**, i.e. there are no $1 \leq i < j < m < n \leq d$ and two vectors $\vec{x}, \vec{x}' \in X$ such that $x_i x_j x'_m x'_n \neq 0$.

$$\left\{ \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right\}, \left\{ \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}, \left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \\ 1 \end{pmatrix} \right\}$$

Inherent ambiguity of bounded languages

Corollary: If $\mathcal{S}_{\langle w \rangle}(L)$ can not be described as a disjoint union of linear sets with stratified linearly independent periods then L is inherently ambiguous.

For instance, $\{a^n b^m c^p : n = m \text{ or } m = p\}$ is inherently ambiguous if and only if the set $\{(n, m, p) \in \mathbb{N}^3 : n = m \text{ or } m = p\}$ is not a disjoint union of linear sets whose set of periods is a stratified set of linearly independent vectors.

And $\{a^n b^m c^p : n \neq m \text{ or } m \neq p\}$ is inherently ambiguous if and only if the set $\{(n, m, p) \in \mathbb{N}^3 : n \neq m \text{ or } m \neq p\}$ is not ...

Is it useful?

Theoretic advantages :

- it is an equivalence, that leaves the world of derivation trees
- the proof uses very complicated iteration arguments on derivation trees, so we are thankful!

"The proof of the necessity is extremely complicated"

Practical Drawbacks:

- No practical method to prove that $\mathcal{S}_{\langle w \rangle}(L)$ can not be described this way
- it is used in the literature with iteration arguments on semilinear sets: proofs are even trickier than on derivation trees, are too specific, etc.
- hence it has been shadowed by Ogden's Lemma [Ogden '68]

Generating series associated to a semilinear set.

If $S \subseteq \mathbb{N}^d$, then:

$$S(\vec{x}) := \sum_{(v_1, \dots, v_d) \in S} x_1^{v_1} \dots x_d^{v_d} = \sum_{\vec{v} \in S} \vec{x}^{\vec{v}}$$

Example: Linear set $(1, 1) + \{(1, 2), (1, 0)\}^*$

$$S(a, b) = \sum_{n, m} a^{1+1n+1m} b^{1+2n+0m} = \frac{a^1 b^1}{(1 - a^1 b^2)(1 - a^1 b^0)}$$

Example: Linear set $\vec{c} + \{\vec{p}_1, \dots, \vec{p}_r\}^*$ with independent periods

$$S(\vec{x}) = \sum_{\lambda_1, \dots, \lambda_r} \vec{x}^{\vec{c} + \lambda_1 \vec{p}_1 + \dots + \lambda_r \vec{p}_r} = \frac{\vec{x}^{\vec{c}}}{\prod_{j=1}^r (1 - \vec{x}^{\vec{p}_j})}$$

Theorem [Eilenberg & Schützenberger, Ito 69]: If S is semilinear, then $S(\vec{x})$ is rational.

It is useful!

Theorem (3-variable criterion [Koechlin 22])

Let $L \subseteq w_1^* \dots w_d^*$ a *bounded context-free language* with respect to $\langle w \rangle$. Let $S = S_{\langle w \rangle}(L)$ its associated semilinear set.

Let us write

$$S(x_1, \dots, x_d) := \sum_{(i_1, \dots, i_d) \in S} x_1^{i_1} \dots x_d^{i_d} = \frac{P(x_1, \dots, x_d)}{Q(x_1, \dots, x_d)} \in \mathbb{K}(x_1, \dots, x_d)$$

in irreducible form.

Suppose that there exists an *irreducible* polynomial $D \in \mathbb{K}[x_1, \dots, x_d]$ dividing Q , such that D has *three or more variables*.

Then L is *inherently ambiguous*.

It is useful!

Theorem (3-variable criterion [Koechlin 22])

Let $L \subseteq w_1^* \dots w_d^*$ a *bounded context-free language* with respect to $\langle w \rangle$. Let $S = S_{\langle w \rangle}(L)$ its associated semilinear set.

Let us write

$$S(x_1, \dots, x_d) := \sum_{(i_1, \dots, i_d) \in S} x_1^{i_1} \dots x_d^{i_d} = \frac{P(x_1, \dots, x_d)}{Q(x_1, \dots, x_d)} \in \mathbb{K}(x_1, \dots, x_d)$$

in irreducible form.

Suppose that there exists an *irreducible* polynomial $D \in \mathbb{K}[x_1, \dots, x_d]$ dividing Q , such that D has *three or more variables*.

Then L is *inherently ambiguous*.

→ It extends and simplifies the proof of a criterion of [Makarov '21]

Proof

Suppose that L is unambiguous. Then S can be written

$$S = \bigsqcup_{i=1}^r (\vec{c}_i + P_i^*)$$

where the union is disjoint, each P_i is stratified, and the vectors in each P_i are linearly independent.

Proof

Suppose that L is unambiguous. Then S can be written

$$S = \bigsqcup_{i=1}^r (\vec{c}_i + P_i^*)$$

where the union is disjoint, each P_i is stratified, and the vectors in each P_i are linearly independent.

Consequently (with $\vec{x}^{\vec{p}} := x_1^{p_1} \dots x_d^{p_d}$):

$$S(\vec{x}) = \sum_{(i_1, \dots, i_d) \in S} x_1^{i_1} \dots x_d^{i_d} = \sum_{i=1}^r \frac{\vec{x}^{\vec{c}_i}}{\prod_{\vec{p} \in P_i} (1 - \vec{x}^{\vec{p}})} = \frac{P_2(\vec{x})}{Q_2(\vec{x})}$$

with $Q_2(\vec{x}) = \prod_{i=1}^r \prod_{\vec{p} \in P_i} (1 - \vec{x}^{\vec{p}})$. Then D divides Q_2 .

Proof

Suppose that L is unambiguous. Then S can be written

$$S = \bigsqcup_{i=1}^r (\vec{c}_i + P_i^*)$$

where the union is disjoint, each P_i is stratified, and the vectors in each P_i are linearly independent.

Consequently (with $\vec{x}^{\vec{p}} := x_1^{p_1} \dots x_d^{p_d}$):

$$S(\vec{x}) = \sum_{(i_1, \dots, i_d) \in S} x_1^{i_1} \dots x_d^{i_d} = \sum_{i=1}^r \frac{\vec{x}^{\vec{c}_i}}{\prod_{\vec{p} \in P_i} (1 - \vec{x}^{\vec{p}})} = \frac{P_2(\vec{x})}{Q_2(\vec{x})}$$

with $Q_2(\vec{x}) = \prod_{i=1}^r \prod_{\vec{p} \in P_i} (1 - \vec{x}^{\vec{p}})$. Then D divides Q_2 .

As each P_i is stratified, Q_2 is a product of polynomials with at most 2 variables. Hence D cannot divide Q_2 . Contradiction.

Examples [Makarov '21]

1. $\{a^n b^m c^p \text{ with } n \neq m \text{ or } m \neq p\}$ is inherently ambiguous :

- $S = \{(n, m, p) : n \neq m \text{ or } m \neq p\}$

- $S = \mathbb{N}^3 \setminus \{(n, m, p) : n = m = p\}$

- $$\begin{aligned} S(a, b, c) &= \sum_{n,m,p} a^n b^m c^p - \sum_n a^n b^n c^n \\ &= \frac{1}{(1-a)(1-b)(1-c)} - \frac{1}{1-abc} \\ &= \frac{a+b+c-ab-ac-bc}{(1-a)(1-b)(1-c)(1-abc)} \end{aligned}$$

Examples [Makarov '21]

2. $\{a^n b^m c^p \text{ with } n = m \text{ or } m = p\}$ is inherently ambiguous :

$$\frac{1}{(1-ab)(1-c)} + \frac{1}{(1-bc)(1-a)} - \frac{1}{1-abc}$$
$$= \frac{1-3a^2b^2c^2+2a^2b^2c+2ab^2c^2+2a^2bc-ab^2c+2abc^2-a^2b+2abc-bc^2-ac}{(1-a)(1-bc)(1-c)(1-ab)(1-abc)}$$

Examples

3. $\{a^n b^m c^p \text{ with } n = m \text{ or } m \neq p\}$ is inherently ambiguous :

$$\frac{1}{(1-a)(1-b)(1-c)} - \left(\frac{1}{(1-a)(1-bc)} - \frac{1}{1-abc} \right)$$
$$= \frac{3ab^2c^2 - 2ab^2c - 2abc^2 - b^2c^2 + b^2c + bc^2 + ab + ac - 2bc - a + 1}{(1-a)(1-b)(1-c)(1-bc)(1-abc)}$$

Allowing non-distinct symbols: example of primitive words

Primitive words: words that are not the power of a smaller word

$$\mathcal{P} = \{w \in \Sigma^* : \forall u \in \Sigma^*, (w \in u^* \Rightarrow u = w)\}$$

$aaba \in \mathcal{P}$, $abab \notin \mathcal{P}$

Theorem [Petersen '94]: \mathcal{P} is not an unambiguous context-free language.

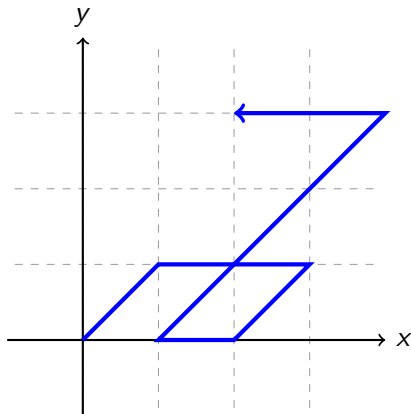
Recall: $\mathcal{P} \cap R$ inherently ambiguous $\Rightarrow \mathcal{P}$ inherently ambiguous

New elementary proof:

- $\mathcal{P} \cap a^*ba^*ba^*b = \{a^nba^m ba^p b : n \neq m \text{ or } m \neq p\}$
- $S_2(a, x, b, y, c, z) = xyz \cdot \frac{a+b+c-ab-ac-bc}{(1-a)(1-b)(1-c)(1-abc)}$

Allowing words : complement of Gessel walks

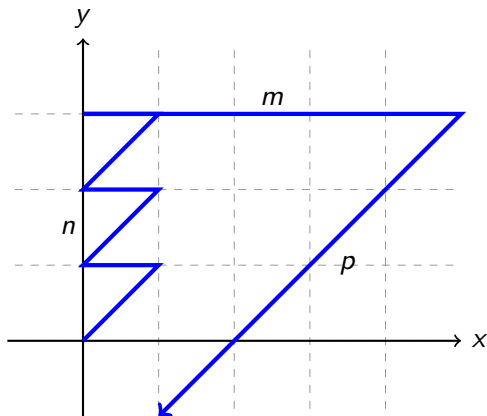
Let \mathcal{G} the set of words in $\{\leftarrow, \rightarrow, \swarrow, \nearrow\}^*$ describing a walk starting at $(0,0)$ and staying in the quarter-plane.



Allowing words : complement of Gessel walks

Then $\bar{\mathcal{G}} = \{\leftarrow, \rightarrow, \swarrow, \nearrow\}^* \setminus \mathcal{G}$ is inherently ambiguous.

Proof: $\bar{\mathcal{G}} \cap (\nearrow\leftarrow)^* \rightarrow^*\swarrow^*$ is inherently ambiguous.



The associated semilinear is $C = \{(n, m, p) \mid n < p \vee m < p\}$.

Allowing words : complement of Gessel walks

Then $\bar{\mathcal{G}} = \{\leftarrow, \rightarrow, \swarrow, \nearrow\}^* \setminus \mathcal{G}$ is inherently ambiguous.

Proof: $\bar{\mathcal{G}} \cap (\nearrow\leftarrow)^* \rightarrow^* \swarrow^*$ is inherently ambiguous.

The associated semilinear is $C = \{(n, m, p) \mid n < p \vee m < p\}$.

$$\begin{aligned} S(a, b, c) &= \frac{1}{(1-a)(1-b)(1-c)} - \frac{1}{(1-abc)(1-ab)(1-b)} - \frac{a}{(1-abc)(1-ab)(1-a)} \\ &= \frac{(1-ab)c}{(1-c)(1-a)(1-b)(1-abc)} \end{aligned}$$

Allowing words : complement of Gessel walks

Then $\bar{\mathcal{G}} = \{\leftarrow, \rightarrow, \swarrow, \nearrow\}^* \setminus \mathcal{G}$ is inherently ambiguous.

Proof: $\bar{\mathcal{G}} \cap (\nearrow\leftarrow)^* \rightarrow^* \swarrow^*$ is inherently ambiguous.

The associated semilinear is $C = \{(n, m, p) \mid n < p \vee m < p\}$.

$$\begin{aligned} S(a, b, c) &= \frac{1}{(1-a)(1-b)(1-c)} - \frac{1}{(1-abc)(1-ab)(1-b)} - \frac{a}{(1-abc)(1-ab)(1-a)} \\ &= \frac{(1-ab)c}{(1-c)(1-a)(1-b)(1-abc)} \end{aligned}$$

Open question: $G(x)$ is algebraic but not \mathbb{N} -algebraic ([Bostan and Kauers, '10], [Banderier and Drmota, '13]). Can we directly prove that $\frac{1}{1-4x} - G(x)$ is not \mathbb{N} -algebraic?

Other examples

- $\{a^n b^m c^p : p \geq n \text{ or } p \geq m\}$
- Product of palindromes
 $C = \{w_1 w_2 : w_1, w_2 \in \{a, b\}^* \text{ are palindromes}\}$
- The complement of every non-singular walk with small steps on the quarter plane is an inherently ambiguous context-free language. [Koe 22]
- And many many other!
- $O_3, O_4, \Omega_3, C, S, P_1, P_2, G_{\neq}, G_{<}, G_{>}, G_{=}, H_{\neq}, K_1, K_2, B$

Limits of this criterion

Advantages :

- is recent
- is robust
- is quick

Drawbacks :

- only for bounded languages *
- this first criterion only deals with the first condition of stratified sets, fails with inherent ambiguity due to interlacing vectors

→ fails on $\{a^n b^m c^p d^q \mid n = p \text{ or } m = q\}$

$$\frac{1}{(1-ac)(1-b)(1-d)} + \frac{1}{(1-bd)(1-a)(1-c)} - \frac{1}{(1-ac)(1-bd)}$$
$$= \frac{1-ab-ac-ad-bc-bd-cd+2abc+2abd+2acd+2bcd-3abcd}{(1-ac)(1-bd)(1-a)(1-b)(1-c)(1-d)}$$

Second criterion

Problem: We need a way to distinguish that $(1 - bd)(1 - ac)$ is bad for $\{a^n b^m c^p d^q \mid n = p \text{ or } m = q\}$

$$\frac{1 - ab - ac - ad - bc - bd - cd + 2abc + 2abd + 2acd + 2bcd - 3abcd}{(1 - ac)(1 - bd)(1 - a)(1 - b)(1 - c)(1 - d)}$$

but is not bad in the series of $\{a^n c^n : n \geq 0\} \cup \{b^n d^n : n \geq 0\}$:

$$\frac{1}{1 - ac} + \frac{1}{1 - bd} - 1 = \frac{1 - abcd}{(1 - ac)(1 - bd)}$$

In other words, find a way to prove that they were in a same set of periods.

Targeting the interlacing condition

Theorem (interlacing criterion [Koechlin 22])

Let $L \subseteq w_1^* \dots w_d^*$ a *bounded context-free language* with respect to $\langle w \rangle$. Let $S = S_{\langle w \rangle}(L)$ its associated semilinear set.

Let us write in irreducible form

$$S(x_1, \dots, x_d) = \frac{P(x_1, \dots, x_d)}{Q(x_1, \dots, x_d)} = \frac{P(\vec{x})}{(1 - x_i^n x_k^m) D(x_j, x_\ell) \tilde{Q}(\vec{x})}$$

Suppose that

- Q is divided by two non-univariate irreducible polynomials $D(x_j, x_\ell)$ and $\pi(x_i, x_k)$ with $j < \ell$ and $i < k$ interlaced (i.e. $i < j < k < \ell$ or $j < i < \ell < k$);
- $\pi(x_i, x_k) = (1 - x_i^n x_k^m)$, with $n, m \geq 1$ and $n \wedge m = 1$;
- finally, $D \nmid P|_{x_i=y^m, x_k=y^{-n}}$ in $\mathbb{Q}(y)[\vec{x}]$ where y is a fresh variable.

Then L is *inherently ambiguous*.

Idea of proof

Suppose that L is unambiguous. Then S can be written

$$S = \bigcup_{i=1}^r (\vec{c}_i + P_i^*)$$

$$S(\vec{x}) = \sum_{s=1}^r \frac{\vec{x}^{\vec{c}_s}}{\prod_{\vec{p} \in P_s} (1 - \vec{x}^{\vec{p}})} = \frac{P(\vec{x})}{(1 - x_i^n x_k^m) D(x_j, x_\ell) \tilde{Q}(\vec{x})}$$

Idea of proof

Suppose that L is unambiguous. Then S can be written

$$S = \bigcup_{i=1}^r (\vec{c}_i + P_i^*)$$

$$S(\vec{x}) = \sum_{s=1}^r \frac{\vec{x}^{\vec{c}_s}}{\prod_{\vec{p} \in P_s} (1 - \vec{x}^{\vec{p}})} = \frac{P(\vec{x})}{(1 - x_i^n x_k^m) D(x_j, x_\ell) \tilde{Q}(\vec{x})}$$

$$\sum_{s \in I_1} \frac{\vec{x}^{\vec{c}_s}}{R_s(\vec{x})} + (1 - x_i^n x_k^m) \sum_{s \in I_2} \frac{\vec{x}^{\vec{c}_s}}{\prod_{\vec{p} \in P_s} (1 - \vec{x}^{\vec{p}})} = \frac{P(\vec{x})}{D(x_j, x_\ell) \tilde{Q}(\vec{x})}$$

with $R_s(\vec{x}) = \frac{\prod_{\vec{p} \in P_s} (1 - \vec{x}^{\vec{p}})}{(1 - x_i^n x_k^m)}$

Idea of proof

Suppose that L is unambiguous. Then S can be written

$$S = \bigsqcup_{i=1}^r (\vec{c}_i + P_i^*)$$

$$S(\vec{x}) = \sum_{s=1}^r \frac{\vec{x}^{\vec{c}_s}}{\prod_{\vec{p} \in P_s} (1 - \vec{x}^{\vec{p}})} = \frac{P(\vec{x})}{(1 - x_i^n x_k^m) D(x_j, x_\ell) \tilde{Q}(\vec{x})}$$

$$\sum_{s \in I_1} \frac{\vec{x}^{\vec{c}_s}}{R_s(\vec{x})} + (1 - x_i^n x_k^m) \sum_{s \in I_2} \frac{\vec{x}^{\vec{c}_s}}{\prod_{\vec{p} \in P_s} (1 - \vec{x}^{\vec{p}})} = \frac{P(\vec{x})}{D(x_j, x_\ell) \tilde{Q}(\vec{x})}$$

with $R_s(\vec{x}) = \frac{\prod_{\vec{p} \in P_s} (1 - \vec{x}^{\vec{p}})}{(1 - x_i^n x_k^m)}$

$$\sum_{s \in I_1} \frac{\vec{x}^{\vec{c}_s} |_{x_i=y^m, x_k=y^{-n}}}{R_s(\vec{x}) |_{x_i=y^m, x_k=y^{-n}}} = \frac{P(\vec{x}) |_{x_i=y^m, x_k=y^{-n}}}{D(x_j, x_\ell) \tilde{Q}(\vec{x}) |_{x_i=y^m, x_k=y^{-n}}}$$

Contradiction.

Examples

$L = \{a^i b^j c^k d^\ell : i \neq k \text{ or } j \neq \ell\}$ is inherently ambiguous.

$$\frac{1}{(1-a)(1-b)(1-c)(1-d)} - \frac{1}{(1-ac)(1-bd)}$$
$$= \frac{abc+abd+acd+bcd-ab-2ac-ad-bc-2bd-cd+a+b+c+d}{(1-ac)(1-bd)(1-a)(1-b)(1-c)(1-d)}$$

- $D(b, d) = (1 - bd)$, $\pi(a, c) = (1 - ac)$
- We need to prove that $(1 - bd) \nmid P|_{a=y, c=1/y}$
- $P|_{a=y, c=1/y} = (y - 2 + \frac{1}{y})(bd - b - d + 1)$
- $(1 - bd) \nmid (bd - b - d + 1)$.

Examples

$L = \{a^i b^j c^k d^\ell : i \neq k \text{ or } j \neq \ell\}$ is inherently ambiguous.

$$\frac{1}{(1-a)(1-b)(1-c)(1-d)} - \frac{1}{(1-ac)(1-bd)}$$
$$= \frac{abc+abd+acd+bcd-ab-2ac-ad-bc-2bd-cd+a+b+c+d}{(1-ac)(1-bd)(1-a)(1-b)(1-c)(1-d)}$$

- $D(b, d) = (1 - bd)$, $\pi(a, c) = (1 - ac)$
- We need to prove that $(1 - bd) \nmid P|_{a=2, c=1/2}$
- $P|_{a=2, c=1/2} = \frac{1}{2}(bd - b - d - 1)$
- $(1 - bd) \nmid \frac{1}{2}(bd - b - d - 1)$.

Examples

$L_2 = \{a^i b^j c^k d^\ell : 3i \neq 5k \text{ ou } 2j \neq 3\ell\}$ is inherently ambiguous.

$$\frac{1}{(1-a)(1-b)(1-c)(1-d)} - \frac{1}{(1-b^3 d^2)(1-a^5 c^3)}$$
$$= \frac{a^5 b^3 c^3 d^2 - a^5 c^3 - b^3 d^2 - abcd + abc + abd + acd + bcd - ab - ac - ad - bc - bd - cd + a + b + c + d}{(1-a)(1-b)(1-c)(1-d)(1-b^3 d^2)(1-a^5 c^3)}$$

- $D(b, d) = (1 - b^3 d^2)$, $\pi(a, c) = (1 - a^5 c^3)$
- $P|_{a=8, c=1/32} = \frac{217}{32}(bd - b - d + 1)$
- $(1 - b^3 d^2) \nmid \frac{217}{32}(bd - b - d + 1)$.

Conclusion

In this talk, we have seen:

- How to use Ginsburg and Ullian criteria with generating series
- We generalized the 3-variable criterion of [Makarov'21] to bounded languages **on words**
- And developed a completely new interlacing criterion

Ideas for further work:

- Develop robust tools for infinite ambiguity
- (Un)Decidability of inherent ambiguity for bounded languages?

Inherent infinite ambiguity

For $K \geq 1$, a grammar is K -ambiguous if every generated word has at most K derivations.

A language is **inherently infinitely** ambiguous if it is not recognized by any finitely ambiguous grammar.

Example: The language of products of palindromes is inherently infinitely ambiguous [Crestin '72].

Idea for further work

1. If L is recognized by a K -ambiguous grammar G , then
 - a. $l_n \leq g_n \leq Kl_n$
 - b. $l_n = \Theta(g_n)$ where g_n is \mathbb{N} -algebraic

Example: Shamir's language is infinitely ambiguous

$$L_k = \{w \in \Sigma \mid w = s\#us^Rv \text{ with } s, u, v \in \{a_1, \dots, a_k\}^*\},$$

[Shamir 70']: proof for $k = 2$ with iteration arguments

New proof: we can prove that $l_n = \Theta(k^{n-1} \log_k(n))$, which is incompatible with algebraicity.

2. Find a way to detect the inherent K -ambiguity of bounded languages.

Idea for further work

1. If L is recognized by a K -ambiguous grammar G , then
 - a. $l_n \leq g_n \leq K l_n$
 - b. $l_n = \Theta(g_n)$ where g_n is \mathbb{N} -algebraic

Example: Shamir's language is infinitely ambiguous

$$L_k = \{w \in \Sigma \mid w = s \# u s^R v \text{ with } s, u, v \in \{a_1, \dots, a_k\}^*\},$$

[Shamir 70']: proof for $k = 2$ with iteration arguments

New proof: we can prove that $l_n = \Theta(k^{n-1} \log_k(n))$, which is incompatible with algebraicity.

2. Find a way to detect the inherent K -ambiguity of bounded languages.

Thank you!