Finite Model Theory

George Voutsadakis[1]

[1]Mathematics and Computer Science
Lake Superior State University

LSSU Math 600

1. Finite Automata and Logic
   - Languages Accepted by Automata
   - Word Models
   - Examples and Applications
   - First-Order Definability

Subsection 1

Languages Accepted by Automata

Languages

- Let $\mathbb{A}$ be a finite alphabet.
- Let $\mathbb{A}^*$ be the set of strings (or words) over $\mathbb{A}$.
- Let $\mathbb{A}^+$ the set of nonempty strings (or words) over $\mathbb{A}$.
- We have
$$\mathbb{A}^* = \mathbb{A}^+ \cup \{\lambda\},$$
where λ is the empty word.
- A **language** over $\mathbb{A}$ is a subset of $\mathbb{A}^+$.
- This is a slight deviation from standard terminology in *automata theory*, where the term *language* signifies a subset of $\mathbb{A}^*$.

Nondeterministic Automata

- A **nondeterministic automaton** M, in short, an **NDA** (over the alphabet $\mathbb{A}$) is given by a tuple

$$M = (S, q_0, \delta, F),$$

where:

  - S is a finite set, the set of **states**;
  - $q_0 \in S$ is the **initial state**;
  - $F \subseteq S$ is the set of (**accepting** or) **final states**;
  - $\delta \subseteq S \times \mathbb{A} \times S$ is the **transition relation**.
    Intuitively, $(q, a, p) \in \delta$ means if M is in state q and reads a, then M can pass into state p.

Extending the Transition to Strings

- This relation induces a function $\widetilde{\delta} : S \times \mathbb{A}^* \to \text{Pow}(S)$, where $\text{Pow}(S)$ denotes the power set of S.
- $\widetilde{\delta}$ is given by

$$\begin{array}{rcl} \widetilde{\delta}(q, \lambda) & := & \{q\}; \\ \widetilde{\delta}(q, wa) & := & \{p : (r, a, p) \in \delta \text{ for some } r \in \widetilde{\delta}(q, w)\}. \end{array}$$

- In particular, $\widetilde{\delta}(q, a) = \{p : (q, a, p) \in \delta\}$, for $a \in \mathbb{A}$.
- If $\widetilde{\delta}(q, a)$ is a singleton for every $a \in \mathbb{A}$, then M is said to be a **deterministic automaton** or an **automaton**.

  In this case, $\widetilde{\delta}(q, w)$ is a singleton, for any $w \in \mathbb{A}^*$.
- If $\widetilde{\delta}(q, w) = \{p\}$, we simply write $\widetilde{\delta}(q, w) = p$.
- Similarly, $\delta(q, a) = p$ stands for $\widetilde{\delta}(q, a) = \{p\}$.

Languages Recognized by NDAs

- The language **recognized** (or **accepted**) by the NDA M is defined by

$$L(M) := \{w \in \mathbb{A}^+ : \widetilde{\delta}(q_0, w) \cap F \neq \varnothing\}.$$

- Hence, in case M is deterministic,

$$L(M) = \{w \in \mathbb{A}^+ : \widetilde{\delta}(q_0, w) \in F\}.$$

- We aim to show that a language is recognized by an automaton if and only if it is definable in monadic second order logic.
- However, we will prove many equivalences which, apart from being useful in the proof, are also interesting in their own.

A Characterization Theorem

- Some of the terms below have not yet been defined.
- They will be in the course of the proof.

Characterization of Regular Languages

For a language $L \subseteq \mathbb{A}^+$, the following are equivalent:

(i) L is the union of equivalence classes of an invariant equivalence relation on $\mathbb{A}^+$ of finite index.

(ii) L is recognized by an automaton.

(iii) L is recognized by an NDA.

(iv) L is regular.

(v) L is definable in monadic second-order logic by a Σ_1^1-sentence.

(vi) L is definable in monadic second-order logic.

- Note that (ii)$\Rightarrow$(iii) and (v)$\Rightarrow$(vi) are trivial.

Invariance and Index

- An equivalence relation $\sim$ on $\mathbb{A}^+$ is called **invariant** if

$$u, v, w \in \mathbb{A}^+ \quad \text{and} \quad u \sim v \quad \text{imply} \quad uw \sim vw.$$

- Denote by $[u]$ the equivalence class of u and by $\mathbb{A}^+/{\sim}$ the set of equivalence classes.
- The **index** of $\sim$ is the cardinality of $\mathbb{A}^+/{\sim}$.

Invariant Equivalence Relations of Finite Index

Proposition

Let $\sim$ be an invariant equivalence relation on $\mathbb{A}^+$ of finite index. Suppose that the language $L \subseteq \mathbb{A}^+$ is the union of equivalence classes,

$$L = [u_1] \cup \cdots \cup [u_r],$$

for some $u_1, \ldots, u_r \in \mathbb{A}^+$. Then L is recognized by an automaton.

- Add $[\lambda]$, "the equivalence class of λ", as a new object to $\mathbb{A}^+/\sim$. Define the automaton

$$M = (S, q_0, \delta, F)$$

as follows:
  - $S := (\mathbb{A}^+/\sim) \cup \{[\lambda]\}$;
  - $q_0 := [\lambda]$;
  - $\delta([u], a) := [ua]$;
  - $F := \{[u_1], \ldots, [u_r]\}$.

Invariant Equivalence Relations of Finite Index (Cont'd)

- By invariance of $\sim$, the transition function δ is well-defined.

  For $u, v \in \mathbb{A}^*$, an induction on the length of v shows that

  $$\widetilde{\delta}([u], v) = [uv].$$

  In particular, $\widetilde{\delta}([\lambda], v) = [v]$.

  Therefore,

  $$\begin{array}{rcl} L(M) & = & \{v \in \mathbb{A}^+ : \widetilde{\delta}(q_0, v) \in F\} \\ & = & \{v \in \mathbb{A}^+ : [v] \in F\} \\ & = & [u_1] \cup \cdots \cup [u_r] \\ & = & L. \end{array}$$

The Pumping Lemma

Lemma (Pumping Lemma)

Let $\sim$ be an invariant equivalence relation on $\mathbb{A}^+$ of finite index. Then there is an $n \geq 0$ such that, for any word $u \in \mathbb{A}^+$, with $|u| \geq n$, there exist $v, w \in \mathbb{A}^+$ and $x \in \mathbb{A}^*$ with

$$u = vwx, \ |vw| \leq n, \text{ and } vw^k \sim vw \text{ for all } k \geq 0.$$

Hence, by invariance, $vw^k y \sim vwy$, for all $k \geq 0$ and $y \in \mathbb{A}^*$.

- Let ℓ be the index of $\sim$ and set $n := \ell + 1$.

  Consider $u \in \mathbb{A}^+$, $u = a_1 \ldots a_s$, where $a_1, \ldots, a_s \in \mathbb{A}$ and $s \geq n$.

  Then, for some i and j with $1 \leq i < j \leq n$, we have $a_1 \ldots a_i \sim a_1 \ldots a_j$.

  Let $v = a_1 \ldots a_i$ and $w = a_{i+1} \ldots a_j$. Thus, $v \sim vw$.

  By invariance of $\sim$, $vw \sim vw^2 \sim vw^3 \sim \cdots$.

Concatenation and Positive Closure

- The **concatenation** of languages L_1 and L_2, denoted by $L_1 L_2$, is the set
$$L_1 L_2 := \{uv : u \text{ is in } L_1 \text{ and } v \text{ is in } L_2\}.$$
- Define:
$$\begin{array}{lll} L^1 & := & L; \\ L^n & := & L^{n-1}L, \quad n > 1. \end{array}$$
- The **plus** (or **positive**) **closure** L^+ of L is the set
$$L^+ := \bigcup_{n \geq 1} L^n.$$

Regular Expressions and Regular Languages

- Regular expressions (over $\mathbb{A}$) are strings over the alphabet

$$\{\varnothing\} \cup \{\boldsymbol{a} : a \in \mathbb{A}\} \cup \{\cup, {}^+,), (\}.$$

- **Regular expressions**, together with the languages they denote, are defined recursively as follows:
  (a) $\varnothing$ is a regular expression and denotes the empty set;
  (b) $\boldsymbol{a}$ is a regular expression and denotes the set $\{a\}$;
  (c) If r and s are regular expressions denoting the languages R and S, respectively, then

$$(r \cup s), \quad (rs), \quad r^+$$

  are regular expressions that denote, respectively, the sets

$$R \cup S, \quad RS, \quad R^+.$$

- A language is **regular** if it is denoted by some regular expression.

Some Conventions

- For convenience, when writing *regular expressions*, we adopt some conventions.
- We omit parentheses when they have no influence on the language they denote.

  E.g., $r_1 \cup \cdots \cup r_k$.
- We assume the following order of operations (in decreasing strength):

  plus closure, concatenation, union.

Languages Recognized by NDA are Regular

Proposition

If L is recognized by an NDA then L is regular.

- Suppose L is recognized by the NDA $M = (S, q_0, \delta, F)$, with $S = \{q_0, \ldots, q_n\}$.

  Let L_k^{ij} be the set of all nonempty strings that M can read starting in q_i and ending in q_j without going through any state numbered $\geq k$,

$$L_k^{ij} := \{b_1 \ldots b_s : s \geq 1, b_1, \ldots, b_s \in \mathbb{A}, \text{ there are } i_0, \ldots, i_s, \text{ such that } i_1, \ldots, i_{s-1} < k, i_0 = i, i_s = j \text{ and } (q_{i_m}, b_{m+1}, q_{i_{m+1}}) \in \delta \text{ for } m < s\}.$$

  Since $L(M) = \bigcup_{q_j \in F} L_{n+1}^{0j}$, it suffices to show that all L_k^{ij} are regular.

  We proceed by induction on k.

Languages Recognized by NDA are Regular (Cont'd)

- Note that $L_0^{ij} = \{a \in \mathbb{A} : (q_i, a, q_j) \in \delta\}$ is a subset of $\mathbb{A}$.
  Suppose $L_0^{ij} = \{a_1, \ldots, a_r\}$.
  Then L_0^{ij} is denoted by $(\boldsymbol{a}_1 \cup \cdots \cup \boldsymbol{a}_r)$ or by $\varnothing$ in case $r = 0$.
  For the induction step, note that a nonempty string is in L_{k+1}^{ij} if it can be read without visiting any state numbered $\geq k+1$.
  Such a string starts in q_i, ends in q_j, and passes through q_k zero times or one or more than one time.
  Hence, we get the expression
  $$L_{k+1}^{ij} = L_k^{ij} \cup L_k^{ik} L_k^{kj} \cup L_k^{ik} (L_k^{kk})^+ L_k^{kj}.$$
  By the induction hypothesis, for all i', j', there is a regular expression $r_k^{i'j'}$ denoting $L_k^{i'j'}$. Therefore, L_{k+1}^{ij} is denoted by the regular expression
  $$r_k^{ij} \cup r_k^{ik} r_k^{kj} \cup r_k^{ik} (r_k^{kk})^+ r_k^{kj}.$$

Subsection 2

Word Models

Word Models

- We fix an alphabet $\mathbb{A}$.
- Let $\tau(\mathbb{A})$ be the vocabulary $\{<\} \cup \{P_a : a \in \mathbb{A}\}$, where:
  - $<$ is binary;
  - The P_a are unary.
- For a given $u \in \mathbb{A}^*$, say $u = a_1 \dots a_n$, we consider structures of the form
$$(B, <, (P_a)_{a \in \mathbb{A}}),$$
where:
  - The cardinality of B equals the length of u;
  - $<$ is an ordering of B;
  - P_a corresponds to the positions in u carrying an a,
$$P_a := \{b \in B : \text{for some } j,\ b \text{ is the } j\text{-th element of } < \text{ and } a_j = a\}.$$
- We call these **word models** for u.
- The class of word models for u is denoted by K_u.

Example

- Suppose $\mathbb{A} = \{a, b\}$.

  Let $u = abbab$.

  Consider the structure

  $$(\{1, \ldots, 5\}, <, P_a, P_b),$$

  where:

  - $<$ is the natural ordering on $\{1, \ldots, 5\}$;
  - $P_a = \{1, 4\}$;
  - $P_b = \{2, 3, 5\}$.

  This structure is a word model for u.

Definability in Monadic Second Order Logic

- Any two word models for u are isomorphic.
- Therefore, we often speak of *the* word model for u, written $\mathcal{B}_u$.
- Note that for $u, v \in \mathbb{A}^+$, a word model for uv is obtained by forming the ordered sum $\mathcal{B}_u \triangleleft \mathcal{B}_v$.
- A language $L \subseteq \mathbb{A}^+$ is **definable in monadic second-order logic**, if there is a sentence φ in $\mathrm{MSO}[\tau(\mathbb{A})]$, such that $\mathrm{Mod}(\varphi) = \bigcup_{u \in L} K_u$, or, more succinctly (but not fully correct), $\mathrm{Mod}(\varphi) = \{\mathcal{B}_u : u \in L\}$.
- A language $L \subseteq \mathbb{A}^+$ is **definable in first-order logic**, if there is a sentence φ in $\mathrm{FO}[\tau(\mathbb{A})]$, such that $\mathrm{Mod}(\varphi) = \bigcup_{u \in L} K_u$, or, more succinctly (but not fully correct), $\mathrm{Mod}(\varphi) = \{\mathcal{B}_u : u \in L\}$.

Definability of the Class of All Word Models

- Let φ_W be the first-order sentence

$$\begin{aligned}\varphi_W \quad := \quad & \text{“} < \text{ is a total ordering”} \wedge \\ & \forall x \bigvee_{a \in \mathbb{A}} P_a x \wedge \bigwedge_{\substack{a,b \in \mathbb{A} \\ a \neq b}} \forall x \neg (P_a x \wedge P_b x).\end{aligned}$$

- Then, $\text{Mod}(\varphi_W)$ is the class of all word models,

$$\text{Mod}(\varphi_W) = \{\mathcal{B}_u : u \in \mathbb{A}^+\}.$$

- So the language $\mathbb{A}^+$ is definable in first-order logic.

Some Notation

- Let $\psi_{\mathsf{min}}(x)$ and $\psi_{\mathsf{max}}(x)$ be first-order formulas defining the first and the last element of the ordering, respectively:

$$\psi_{\mathsf{min}}(x) := \forall y \neg y < x, \qquad \psi_{\mathsf{max}}(x) := \forall y \neg x < y.$$

- For any formula φ of MSO and variables x and y, let $\varphi^{[x,y]}$ be a formula expressing that the closed interval $[x,y]$ satisfies φ.
- Similarly, $\varphi^{]x,y]}$ is a formula expressing that the half-open interval $]x,y]$ satisfies φ.
- Such formulas can be obtained from φ by relativizing the first-order quantifiers to the interval.
- The main clause of an inductive definition is (for a variable $z \neq x, z \neq y$)

$$\begin{array}{rcl}(\exists z\varphi)^{[x,y]} & := & \exists z(x \leq z \wedge z \leq y \wedge \varphi^{[x,y]}); \\ (\exists z\varphi)^{]x,y]} & := & \exists z(x < z \wedge z \leq y \wedge \varphi^{]x,y]}).\end{array}$$

Regular Languages and Monadic Second Order Logic

Proposition

Any regular language is definable in monadic second order logic by a Σ_1^1-sentence.

- We split the proof in two stages.
- In the first stage, we prove by induction on the length of the regular expression r that there is a sentence φ_r of MSO defining the language denoted by r.
- In the second tage, we show that we can replace φ_r by a Σ_1^1-sentence.

Regular Languages and MSO (Stage 1)

- For the base case, we have:

$$\begin{array}{lcl} \varphi_{\varnothing} & := & \exists x \neg x = x; \\ \varphi_{\boldsymbol{a}} & := & \varphi_W \wedge \exists x \forall y (y = x \wedge P_a x). \end{array}$$

For the inductive step, we have:

$$\begin{array}{rcl} \varphi_{(r \cup s)} & := & \varphi_W \wedge (\varphi_r \vee \varphi_s); \\ \varphi_{(rs)} & := & \varphi_W \wedge \text{"the universe is partitioned into two} \\ & & \text{intervals satisfying } \varphi_r \text{ and } \varphi_s\text{, respectively"} \\ & = & \varphi_W \wedge \exists x \exists y \exists z (\psi_{\min}(x) \wedge y < z \wedge \psi_{\max}(z) \wedge \varphi_r^{[x,y]} \wedge \varphi_s^{]y,z]}); \\ \varphi_{r^+} & := & \varphi_W \wedge \text{"there is a set of right endpoints of intervals,} \\ & & \text{which partition the universe, all parts satisfying } \varphi_r\text{"} \\ & = & \varphi_W \wedge \exists X (\exists y (Xy \wedge \psi_{\max}(y)) \wedge \\ & & \exists x \exists y (\psi_{\min}(x) \wedge Xy \wedge \forall z (z < y \to \neg Xz) \wedge \varphi_r^{[x,y]}) \wedge \\ & & \forall x \forall y ((x < y \wedge Xx \wedge Xy \wedge \forall z (x < z < y \to \neg Xz)) \to \varphi_r^{]x,y]})). \end{array}$$

Regular Languages and MSO (Stage 2)

- We obtain a Σ_1^1-sentence by inductively bringing all existential second order quantifiers to the front.

  In general, a monadic second-order formula $\forall\overline{x}\exists Y\chi$, with first-order χ, is not equivalent to a monadic Σ_1^1-formula.
  However, in the case of the formula in the last two lines of φ_{r^+} we can argue as follows:

  Suppose that φ_r is equivalent to $\exists Y_1 \cdots \exists Y_m \chi$.
  In models of φ_W (the only ones of interest), the formula

  $$\forall x \forall y((x < y \land Xx \land Xy \land \forall z(x < z < y \to \neg Xz)) \to \varphi_r^{]x,y]})$$

  is equivalent to

  $$\exists Y_1 \cdots \exists Y_m \forall x \forall y((x < y \land Xx \land Xy \land \forall z(x < z < y \to \neg Xz)) \to \chi^{]x,y]}).$$

  For the nontrivial implication, piece $Y_1, \ldots, Y_m$ together from corresponding subsets chosen in the (disjoint) intervals.

MSO and Invariant Equivalences of Finite Index

Proposition

Let $L \subseteq \mathbb{A}^+$ be definable in monadic second-order logic. Then, there is an invariant equivalence relation on $\mathbb{A}^+$ of finite index, such that L is a union of equivalence classes.

- Assume that there exists a sentence φ of MSO, such that

$$\text{Mod}(\varphi) = \{\mathcal{B}_u : u \in L\}.$$

Let m be the quantifier rank of φ.

Recall that $\mathcal{A} \equiv_m^{\text{MSO}} \mathcal{B}$ means that $\mathcal{A}$ and $\mathcal{B}$ satisfy the same sentences of MSO of quantifier rank $\leq m$.

Define $\sim$ on $\mathbb{A}^+$ by

$$u \sim v \quad \text{iff} \quad \mathcal{B}_u \equiv_m^{\text{MSO}} \mathcal{B}_v.$$

Clearly, $\sim$ is an equivalence relation.

MSO and Invariant Equivalences (Cont'd)

- Now, up to logical equivalence, there are only finitely many sentences of quantifier rank $\leq m$. So the relation $\sim$ is of finite index.

  By definition of m,

  $$\mathcal{B}_u \vDash \varphi \quad \text{and} \quad u \sim v \quad \text{imply} \quad \mathcal{B}_v \vDash \varphi.$$

  Thus,

  $$L = \bigcup \{[u] : u \in \mathbb{A}^+, \mathcal{B}_u \vDash \varphi\}.$$

  Finally, we show that $\sim$ is invariant.

  Assume $u \sim v$ and $w \in \mathbb{A}^+$. Then $\mathcal{B}_u \equiv_m^{\text{MSO}} \mathcal{B}_v$.

  Since $\equiv_m^{\text{MSO}}$ is preserved by ordered sums, we get

  $$\mathcal{B}_{uw} \cong \mathcal{B}_u \lhd \mathcal{B}_w \equiv_m^{\text{MSO}} \mathcal{B}_v \lhd \mathcal{B}_w \cong \mathcal{B}_{vw}.$$

  This shows that $uw \sim vw$.

The Main Theorem Restated

Theorem

For a language $L \subseteq \mathbb{A}^+$ the following are equivalent:

(i) L is the union of equivalence classes of an invariant equivalence relation on $\mathbb{A}^+$ of finite index.

(ii) L is recognized by an automaton.

(iii) L is recognized by an NDA.

(iv) L is regular.

(v) L is definable in monadic second-order logic by a Σ_1^1-sentence.

(vi) L is definable in monadic second-order logic.

- Thus, a language is accepted by an automaton:
  - Exactly in case it is definable in monadic second-order logic;
  - Exactly in case it is specified by means of a regular expression.
- Do both characterizations count as logical descriptions?

Subsection 3

Examples and Applications

Closure Under Boolean Operations and Pumping Lemma

Proposition

(a) The class of languages over $\mathbb{A}$ accepted by automata is closed under the boolean operations (complementation and union).

(b) (**Pumping Lemma**) Let L be accepted by an automaton. Then there is $n \geq 0$, such that for any $u \in \mathbb{A}^+$ with $|u| \geq n$, there exist $v, w \in \mathbb{A}^+$ and $x \in \mathbb{A}^*$ with:

- $u = vwx$;
- $|vw| \leq n$;
- For $k \geq 0$ and $y \in \mathbb{A}^*$,

$$vw^k y \in L \quad \text{iff} \quad vwy \in L.$$

- Part (a) holds, since monadic second-order logic is closed under the boolean connectives $\neg$ and $\vee$.

  Part (b) is a reformulation of the Pumping Lemma.

Example: Ultimately Periodic Subsets of $\mathbb{N}_+$

- Let $\mathbb{A} = \{a\}$.
  Identify $a \dots a$ (of length n) with the natural number n.
  Thus, $\mathbb{A}^+$ is identified with the set $\mathbb{N}_+$ of positive natural numbers.
  A subset L of $\mathbb{N}_+$ is called **ultimately periodic** if there are $p, r \in \mathbb{N}_+$, such that for all $m \geq p$, $m + r \in L$ iff $m \in L$.
  Claim: A subset L of $\mathbb{N}_+$ is accepted by an automaton iff L is ultimately periodic.
  Assume first that L is accepted by an automaton.
  By the Pumping Lemma, there are $n, j, r \in \mathbb{N}_+$ and $\ell \geq 0$, with $n = j + r + \ell$, such that, for all $k \geq 0$ and $s \in \mathbb{N}$,
  $$j + kr + s \in L \quad \text{if} \quad j + r + s \in L.$$
  In particular, if $m \geq p := j + r$, say $m = j + r + s$, then (take $k = 2$)
  $$m + r \in L \quad \text{iff} \quad m \in L.$$

Example (Cont'd)

- Now let L be ultimately periodic.

  Choose corresponding $p, r \in \mathbb{N}_+$, such that, for all $m \geq p$,

  $$m + r \in L \quad \text{iff} \quad m \in L.$$

  Set

  - $L_1 := \{m \in L : m < p\}$;
  - $L_2 := \{m \in L : p \leq m < p + r\}$.

  Then, by periodicity,

  $$L = L_1 \cup L_2 \cup \{m + kr : m \in L_2, k \geq 1\}.$$

  So L is the union of the finite (and hence regular) sets L_1 and L_2 and of the languages denoted by the regular expressions $\boldsymbol{a}^m(\boldsymbol{a}^r)^+$, $m \in L_2$.

  Thus, L is regular.

  So the classes of finite ordered structures of vocabulary $\{<\}$ axiomatizable in MSO coincide with the ultimately periodic ones.

Example

- For $\mathbb{A} = \{a, b\}$ the set

  $$L := \{u \in \mathbb{A}^+ : \text{the number of } a\text{'s in } u \text{ equals the number of } b\text{'s in } u\}$$

  is not accepted by an automaton.

  Choose n according to the Pumping Lemma.

  Consider $a^n b^n$.

  Let its representation, according to the Pumping Lemma, be vwx.

  Since $|vw| \leq n$, we have $w \in \{a\}^+$.

  Hence, the string vw^2x contains more a's than b's.

  Therefore, $vw^2x \notin L$ (while $vwx = a^n b^n \in L$).

  This contradicts the Pumping Lemma.

Bipartite and Balanced (Bipartite) Graphs

- A graph (G, E^G) is **bipartite**, if there is an $X \subseteq G$ such that
$$E^G \subseteq (X \times (G \backslash X)) \cup ((G \backslash X) \times X).$$
- A bipartite graph (G, E^G) is **balanced**, if the set X can be chosen such that, in addition,
$$\|X\| = \|G \backslash X\|.$$
- Denote by BAL the class of finite balanced graphs.
- Denote by $\mathrm{BAL}_<$ the class of finite balanced graphs carrying an arbitrary ordering on their universe,
$$\mathrm{BAL}_< := \{(\mathcal{G}, <) : \mathcal{G} \in \mathrm{BAL}, < \text{ an ordering of } G\}.$$

Non-Axiomatizability of $\mathrm{BAL}_<$ in MSO

Proposition

The class $\mathrm{BAL}_<$, and hence the class BAL, is not axiomatizable in monadic second-order logic.

- Suppose that $\mathrm{BAL}_< = \mathrm{Mod}(\varphi)$ for a sentence φ of MSO.

  Let $\mathbb{A} = \{a, b\}$ and let L be as in the preceding example.

  For $u \in \mathbb{A}^+$, let $\mathcal{B}_u = (B_u, <, P_a, P_b)$ be a word model associated with u, say, with:
  - $B_u = \{1, \ldots, |u|\}$;
  - $<$ the natural ordering.

  Let $\mathcal{G}_u = (B_u, R_u)$ be the bipartite graph given by

$$R_u := \{(i, j) \in B_u \times B_u : P_a i \text{ iff } P_b j\}.$$

  Then, $(\mathcal{G}_u, <) \in \mathrm{BAL}_<$ iff $u \in L$.

Non-Axiomatizability of $\mathrm{BAL}_<$ in MSO (Cont'd)

- Denote by
$$\varphi\frac{(P_a \ldots \leftrightarrow P_b \_\_\_)}{E \ldots \_\_\_}$$
the formula obtained from φ by replacing any subformula of the form Exy by $(P_a x \leftrightarrow P_b y)$.

Then
$$(\mathcal{G}_u, <) \vDash \varphi \quad \text{iff} \quad \mathcal{B}_u \vDash \varphi\frac{(P_a \ldots \leftrightarrow P_b \_\_\_)}{E \ldots, \_\_\_}.$$

Therefore,
$$\mathrm{Mod}\left(\varphi\frac{(P_a \ldots \leftrightarrow P_b \_\_\_)}{E \ldots \_\_\_}\right) = \{\mathcal{B}_u : u \in L\}.$$

A previous theorem now implies that L is accepted by an automaton.

This contradicts the preceding example.

Finite Graphs with a Hamiltonian Circuit

- Let HAM be the class of finite graphs with a Hamiltonian circuit.

Corollary

HAM and $\text{HAM}_<$ are not axiomatizable in MSO.

- Consider a graph of the form $(X \cup Y, E)$ with

$$E = \{(a, b) : (a \in X, b \in Y) \text{ or } (a \in Y, b \in X)\}.$$

Such a graph has a Hamiltonian circuit iff it is balanced.

Asume $\text{HAM}_< = \text{Mod}(\varphi)$ for an MSO-sentence φ.

Then the sentence

$$\exists X \left(\forall x \forall y (Exy \to (Xx \leftrightarrow \neg Xy)) \wedge \varphi \frac{(X \ldots \leftrightarrow \neg X\_\_\_)}{E \ldots \_\_\_} \right)$$

would axiomatize the class $\text{BAL}_<$.

Finite Graphs with a Clique of At Least Half Their Size

- Let CHS be the set of finite graphs which contain a clique of at least half their size.

Corollary

CHS and $CHS_<$ are not axiomatizable in MSO.

- Suppose that $CHS_< = Mod(\varphi)$ for some φ of MSO.

  Then an axiomatization of $BAL_<$ in MSO would be given by

$$\exists X \, (\forall x \forall y (Exy \to (Xx \leftrightarrow \neg Xy))$$
$$\wedge \varphi \frac{X \ldots \wedge X \_\_\_ \wedge \neg \ldots = \_\_\_}{E \ldots \_\_\_} \wedge \varphi \frac{\neg X \ldots \wedge \neg X \_\_\_ \wedge \neg \ldots = \_\_\_}{E \ldots \_\_\_} \Bigg)$$

  Note that the conjunction in the last line implies that both X and its complement have size at least half of the universe.

Subsection 4

First-Order Definability

Plus-Free Regular Languages

- We turn to the problem of characterizing the languages that are accepted by automata and are first-order definable.
- The passage from a regular expression to an MSO formula shows that second-order quantifiers are only needed for the positive closure, i.e., in the transition from a regular expression r to r^+.
- Therefore, if r does not contain the symbol $^+$, the language L denoted by r is first-order definable.
- By induction on the length of such an r, L must then be finite.

Plus-Free Regular Languages and Complementation

Example: Let $\mathbb{A}$ be an alphabet.

For $a \in \mathbb{A}$, the language $\mathbb{A}^+ \backslash \{a\}$ is infinite.

Therefore, it is not definable by a regular expressions without $^+$.

However, it is first-order definable by

$$\varphi_W \wedge (\exists x \neg \psi_{\min}(x) \vee \exists x (\psi_{\min}(x) \wedge \neg P_a x)).$$

- It follows from the example that the class of languages denoted by regular expressions without $^+$ is not closed under complementation.
- On the other hand, the class of first-order definable languages is certainly closed under complementation.

Plus-Free Regular Languages

- We add closure under complementation in the definition of **plus free regular expressions**:
  - $\varnothing, \boldsymbol{a}$ (for $a \in \mathbb{A}$) are plus free regular expressions;
  - If r and s are plus free regular expressions, then so are

$$\sim r, \quad (r \cup s), \quad (rs).$$

- If r denotes the language L, then $\sim r$ denotes $\mathbb{A}^{+} \backslash L$.
- A language is said to be **plus free regular** if it is denoted by a plus free regular expression.

Characterization of Plus-Free Regularity

Theorem

A language is plus free regular iff it is definable in first order logic.

- Suppose a language is plus free regular.

  Then it is defined by a plus free regular expression r.

  Using induction on the structure of r, we construct a first-order sentence defining the same language.

  For the base case:
  - $\varphi_{\varnothing} := \exists x \neg x = x$;
  - $\varphi_{\boldsymbol{a}} := \varphi_W \wedge \exists x \forall y (y = x \wedge P_a(x))$.

  For the induction step:
  - $\varphi_{\sim r} := \varphi_W \wedge \neg \varphi_r$;
  - $\varphi_{(r \cup s)} := \varphi_W \wedge (\varphi_r \vee \varphi_s)$;
  - $\varphi_{(rs)} := \varphi_W \wedge \exists x \exists y \exists z (\psi_{\min}(x) \wedge y < z \wedge \psi_{\max}(z) \wedge \varphi_r^{[x,y]} \wedge \varphi_s^{]y,z]})$.

Characterization of Plus-Free Regularity (Converse)

- Recall that $\tau(\mathbb{A}) = \{<\} \cup \{P_a : a \in A\}$.

  For convenience, we add a constant min to this vocabulary, which henceforth will always denote the first element.

  More precisely, we only look at models of $\varphi_W \wedge \psi_{\text{min}}(\text{min})$.

  We show for a language L that if

  $$\text{Mod}(\varphi_W \wedge \psi_{\text{min}}(\text{min}) \wedge \varphi) = \{(\mathcal{B}_u, \text{min}^{\mathcal{B}_u}) : u \in L\},$$

  then L is plus free regular. We use induction on the quantifier rank of the FO$[\tau(\mathbb{A}) \cup \{\text{min}\}]$-sentence φ.

Characterization of Plus-Free Regularity (Cont'd)

- First assume that φ is atomic.

  Then φ is $\min = \min$ or $P_a \min$ for some $a \in \mathbb{A}$.
  - In the first case, L is $\mathbb{A}^+$.
    Thus, L is denoted by $\sim \varnothing$.
  - Let φ be $P_a \min$. Then $L = \{a\} \cup \{a\}\mathbb{A}^+$.
    Therefore, L is denoted by $\boldsymbol{a} \cup \boldsymbol{a}(\sim \varnothing)$.

  Suppose the languages defined by the sentences φ and ψ are denoted by the plus free expressions r and s, respectively. Then:
  - $\sim r$ corresponds to the sentence $\neg\varphi$;
  - $r \cup s$ corresponds to the sentence $(\varphi \vee \psi)$.

Characterization of Plus-Free Regularity (Cont'd)

- Let $\varphi = \exists x \psi(x)$. Then

$$\begin{aligned}\text{Mod}(\varphi_W \wedge \psi_{\text{min}}(\text{min}) \wedge \exists x \psi(x)) = \\ \text{Mod}(\varphi_W \wedge \psi_{\text{min}}(\text{min}) \wedge \psi(\text{min})) \\ \cup \text{Mod}(\varphi_W \wedge \psi_{\text{min}}(\text{min}) \wedge \exists x(\neg x = \text{min} \wedge \psi(x))).\end{aligned}$$

By the induction hypothesis, the first class of structures on the right corresponds to a plus free regular language.

We turn to the second class.

Let c be a new constant.

Then the finite models of $\varphi_W \wedge \psi_{\text{min}}(\text{min}) \wedge \exists x(\neg x = \text{min} \wedge \psi(x))$ are the $[\tau(\mathbb{A}) \cup \{\text{min}\}]$-reducts of the finite structures $(\mathcal{A}, \text{min}^A, c^A)$ such that

$$(\mathcal{A}, \overset{A}{\text{min}}, c^A) \vDash \varphi_W \wedge \psi_{\text{min}}(\text{min}) \wedge \neg c = \text{min} \wedge \psi(c).$$

Characterization of Plus-Free Regularity (Cont'd)

- Any structure $(\mathcal{A}, \min^A, c^A)$ satisfying

$$\varphi_W \wedge \psi_{\min}(\min) \wedge \neg c = \min \wedge \psi(c)$$

can be written in the form

$$(\mathcal{A}, \min^A, c^A) = (\mathcal{A}_1 \triangleleft \mathcal{A}_2, \min^A, c^A),$$

where:

  - $\triangleleft$ denotes the ordered sum;
  - $(\mathcal{A}_1, \min^A) \vDash (\varphi_W \wedge \psi_{\min}(\min))$;
  - $(\mathcal{A}_2, c^A) \vDash (\varphi_W \wedge \psi_{\min}(c))$.

Let m be the quantifier rank of ψ.

Characterization of Plus-Free Regularity (Cont'd)

- Choose the - up to logical equivalence - finite set $\{(\psi_i(\min), \chi_i(c)) : i \in I\}$ of pairs of FO-sentences of quantifier rank $\leq m$, such that

$$\begin{array}{l}(\mathcal{A}_1, \min^{A_1}) \vDash (\varphi_W \wedge \psi_{\min}(\min) \wedge \psi_i(\min)) \\ \qquad \text{and} \quad (\mathcal{A}_2, c^{A_2}) \vDash (\varphi_W \wedge \psi_{\min}(c) \wedge \chi_i(c)) \\ \text{imply} \quad (\mathcal{A}_1, \min^{A_1}) \lhd (\mathcal{A}_2, C^{A_2}) \vDash \psi(c).\end{array}$$

  By the induction hypothesis there are plus free regular expressions:
  - r_i denoting the language defined by $\varphi_W \wedge \psi_{\min}(\min) \wedge \psi_i(\min)$;
  - s_i denoting the language defined by $\varphi_W \wedge \psi_{\min}(\min) \wedge \chi_i(\min)$.

  Then the plus free regular expression $\bigcup_{i \in I}(r_i s_i)$ denotes the language defined by $(\varphi_W \wedge \psi_{\min}(\min) \wedge \exists x(\neg x = \min \wedge \psi(x)))$.

  Note that, if $(\mathcal{A}_1 \lhd \mathcal{A}_2, \min^{A_1}, c^{A_2}) \vDash \psi(c)$ then, by a previous result, the pair $(\varphi^m_{(\mathcal{A}_1, \min^{A_1})}, \varphi^m_{(\mathcal{A}_2, c^{A_2})})$ of m-isomorphism types belongs (up to logical equivalence) to $\{(\psi_i(\min), \chi_i(c)) : i \in I\}$.

Automata, First Order Logic and Counting Ability

- Let $\mathbb{A} = \{a\}$.
- Identify $\mathbb{A}^+$ with the set $\mathbb{N}_+$ of positive natural numbers.
- Automata do not have the ability to count.

  For instance, they cannot recognize if a given string has prime length. I.e., the set $\{p : p \text{ a prime}\}$ is not accepted by an automaton.
- On the other hand, automata are capable to count modulo a natural number.

  E.g., the set $\{5n : n \geq 1\}$ is accepted by an automaton.
- But first-order logic even lacks this restricted counting ability.

  It is an immediate consequence of a previous result that a subset L of $\mathbb{N}_+$ is first-order definable iff for some $n \geq 1$, $\{m : m \geq n\} \cap L = \varnothing$ or $\{m : m \geq n\} \subseteq L$.

First Order Logic Definability

Theorem

For a language $L \subseteq \mathbb{A}^+$ accepted by an automaton the following are equivalent:

(i) L is definable in first-order logic.

(ii) L is noncounting in the sense that there is an integer $k \geq 1$, such that for every $y \in \mathbb{A}^+$ and $x, z \in \mathbb{A}^*$,

$$xy^k z \in L \quad \text{iff} \quad xy^{k+1} z \in L.$$

- We only prove the implication (i)⇒(ii).
  Suppose $\{\mathcal{B}_u : u \in L\} = \mathrm{Mod}(\varphi)$ for $\varphi \in \mathrm{FO}[\tau(\mathbb{A})]$.
  Let $k := 2^m + 1$, where m is the quantifier rank of φ.

First Order Logic Definability (Cont'd)

- Then, by a previous result, for any $y \in \mathbb{A}^+$, we have

$$\mathcal{B}_{y^k} \cong \lhd^k \mathcal{B}_y \equiv_m \lhd^{k+1} \mathcal{B}_y \cong \mathcal{B}_{y^{k+1}}.$$

Using a previous theorem, we obtain

$$\mathcal{B}_{xy^kz} \cong \mathcal{B}_x \lhd \mathcal{B}_{y^k} \lhd \mathcal{B}_z \equiv_m \mathcal{B}_x \lhd \mathcal{B}_{y^{k+1}} \lhd \mathcal{B}_z \cong \mathcal{B}_{xy^{k+1}z}.$$

In particular,

$$\mathcal{B}_{xy^kz} \vDash \varphi \quad \text{iff} \quad \mathcal{B}_{xy^{k+1}z} \vDash \varphi.$$

So, $xy^kz \in L$ iff $xy^{k+1}z \in L$.

Least Fixed Points: An Appetizer

- The results of this section show that the plus operation cannot be captured in first-order logic.
- An instance of this operation can be viewed as the fixed point of a monotone operation.
- Let $L \subseteq \mathbb{A}^+$ be a language.
- Define $C_L : \text{Pow}(\mathbb{A}^*) \to \text{Pow}(\mathbb{A}^*)$ by

$$C_L(M) := L \cup ML.$$

- Then:
  (a) C_L is monotone, i.e.,

$$M_1 \subseteq M_2 \quad \text{implies} \quad C_L(M_1) \subseteq C_L(M_2).$$

  (b) For $n \geq 1$,

$$\underbrace{C_L(\cdots(C_L}_{n \text{ times}}(\varnothing))\ldots) = L \cup L^2 \cup \cdots \cup L^n.$$

Least Fixed Points: An Appetizer (Cont'd)

- M is a **fixed-point** of C_L if
$$C_L(M) = M.$$
- It can easily be proved that the least - with respect to set-theoretical inclusion - fixed point of C_L is given by
$$C_L(\varnothing) \cup C_L(C_L(\varnothing)) \cup C_L(C_L(C_L(\varnothing))) \cup \cdots.$$
- Hence by Property (b), the least fixed-point of C_L is L^+.