Scribed By: Eunjung KIM
Büchi's Theorem on finite strings
Week 1: 25, 27 February 2025

Contents

1 Terminology: quantifier rank, type

We introduce some important terminology, which will be also used for proving Theorem 7.

Definition 1 (Quantifier rank)**.** *A* quantifier rank *of a formula ψ is the maximum depth of its nested quantifiers. That is,*

- *an atomic formula has quantifier rank 0,*
- *the quantifier rank of a boolean combination $(\wedge, \vee, \neg)$ of formulas is the maximum quantifier rank over the formulas,*
- *one existential / universal quantification increases the quantifier rank by exactly 1.*

The set of all MSO*-formulas of quantifier rank at most k is denoted by* $\text{MSO}[k]$

One can consider all MSO-sentences over τ satisfied by some τ-structure. There are infinitely many MSO-sentences and such a set consisting of all true MSO-sentences in some τ-structure could be infinitely large. However, when you restrict to sentences of quantifier rank up to k, the set becomes finite, size bounded by a function of k.

Definition 2 (Rank-k (ℓ, m)-type)**.** *For a relational structure $\mathbb{A}$ over τ, ℓ-tuple $\vec{v} = (v_1, \ldots, v_\ell) \in A^\ell$ of elements of A and m-tuple $\vec{V} = (V_1, \ldots, V_m) \in (2^A)^m$ of subsets of A, we define the* MSO rank-k ℓ, m-type *of $(\mathbb{A}, \vec{v}, \vec{V})$ as the set of all* MSO*-formulas with ℓ free individual variables and m free set variables satisfied by $(\mathbb{A}, \vec{v}, \vec{V})$. That is,*

$$\textsf{mso-type}_k(\mathbb{A}, \vec{v}, \vec{V}) = \{\psi \in \text{MSO}[k] \mid \mathbb{A} \models \psi(\vec{v}, \vec{V})\}.$$

When $\ell = 0, m = 0$, the MSO *rank-k ℓ, m-type of a structure $\mathbb{A}$ is simply called the* rank-k type *of $\mathbb{A}$. Notice that* $\textsf{mso-type}_k(\mathbb{A})$ *is the set of all* MSO*-sentences of quantifier rank at most k which holds on $\mathbb{A}$.*

An MSO rank-k ℓ, m-type *(when a τ-structure $\mathbb{A}$, ℓ-tuple of elements and m-tuple of subsets of A is not specified) is the set S of* MSO*-formulas in* $\text{MSO}[k]$ *with ℓ free individual variables and m free set variables such that*

- CONSISTENCY*: there exist a τ-structure $\mathbb{A}$ and an ℓ-tuple $\vec{v}$ and an m-tuple of sets $\vec{V}$ over A which satisfies all the formulas in S, and*

- COMPLETENESS*: for any* MSO*-formula ψ with ℓ free individual variables and m free set variables, exactly one of ψ or $\neg\psi$ is included in the set*

It is not difficult to see that for each fixed k, l, m, there are finitely many (bounded by a function of k) MSO-formulas of quantifier rank up to k with ℓ free individual variables and m free set variables. This is particularly because in the base case, i.e. when a formula is quantifier-free, the formula is a boolean combination of atomic formulas on $\ell + m + k$ free variables. The number of atomic formulas is bounded by a function of ℓ, m and τ, and the number of their boolean combinations (up to logical equivalence!) are again bounded. We state this observation without proof.

Observation 3. *For any fixed k, l, m, the number of* MSO *rank-k ℓ, m-types is finite, determined solely by k, ℓ, m and the vocabulary τ.*

Note that for τ-structure $\mathbb{A}$, $\textsf{mso-type}_k(\mathbb{A})$ is an MSO rank-k type; the consistency of the set $\textsf{mso-type}_k(\mathbb{A})$ is witnessed by the very structure $\mathbb{A}$. The completeness of $\textsf{mso-type}_k(\mathbb{A})$ is clear from the fact that $\textsf{qr}(\psi) = \textsf{qr}(\neg\psi)$ for any formula ψ and exactly one of $\mathbb{A} \models \psi$ and $\mathbb{A} \models \neg\psi$ holds.

The consistency and completeness of MSO rank-k type implies that the converse also holds. That is, for any MSO rank-k type q there exists a τ-structure $\mathbb{A}$ such that $q = \textsf{mso-type}_k(\mathbb{A})$.

Lemma 4. *Let $\mathcal{Q}$ be the set of all* MSO *rank-k types in* MSO$[k]$*. Then for every $Q \in \mathcal{Q}$, there exists a τ-structures $\mathbb{A}$ such that* $\textsf{mso-type}_k(\mathbb{A}) = Q$.

Proof: Choose an arbitrary $Q \in \mathcal{Q}$. By consistency of MSO rank-k type, there exists a τ-structure $\mathbb{A}$ which satisfies all sentences in Q. This implies that $Q \subseteq \textsf{mso-type}_k(\mathbb{A})$. We want to show $Q = \textsf{mso-type}_k(\mathbb{A})$. Suppose $\psi \in \textsf{mso-type}_k(\mathbb{A}) \setminus Q$. Because Q, as an MSO rank-k type, is complete and Q contains $\neg\psi$. Then $\textsf{mso-type}_k(\mathbb{A})$ contains both $\neg\psi$ and ψ, which is impossible. $\square$

Definition 5 (Disjoint union on τ-structures)**.** *When the vocabulary τ contains only the predicates and no constant symbols*[1]*, the* disjoint union $\mathbb{A} \cup \mathbb{B}$ *of τ-structures $\mathbb{A}$ and $\mathbb{B}$ with disjoint universe is defined as:*

- *the universe of $\mathbb{A} \cup \mathbb{B}$ is $A \cup B$*
- *the interpretation $R^{\mathbb{A}\cup\mathbb{B}}$ of R is $R^{\mathbb{A}} \cup R^{\mathbb{B}}$ for each predicate $R \in \tau$.*

The so-called *compositionality* of MSO logic is of central importance. It is rather loosely defined and needs to be appropriately formulated in relevant settings. Informally speaking, it says that whether a given MSO-sentence (or formula) holds on a relational structure is determined by whether MSO-sentences hold on relational *sub*structures, when the original structure is formed from the substructures with well-regulated combination rule. The next lemma observes the simplest case of MSO compositionality. We postpone the proof till we learn *Ehrenfeucht-Fraïssé game.*

Lemma 6. *Let $\mathbb{A}, \mathbb{A}', \mathbb{B}, \mathbb{B}'$ be τ-structures such that* $\textsf{mso-type}_k(\mathbb{A}) = \textsf{mso-type}_k(A')$ *and* $\textsf{mso-type}_k(\mathbb{B}) = \textsf{mso-type}_k(\mathbb{B}')$*. Then it holds that* $\textsf{mso-type}_k(\mathbb{A} \cup \mathbb{B}) = \textsf{mso-type}_k(\mathbb{A}' \cup \mathbb{B}')$.

2 Büchi's theorem on strings

We explore the surprising connection between MSO logic on strings and regular languages.

[1]Why do we need this restriction?

Theorem 7 (Büchi'60, Elgot'61,Trakhtenbrot'62)**.** *[1, 2] A language is regular if and only if it is definable in* MSO.

Theorem 7 crucially relies on the compositionality of MSO logic on strings under concatenation[2]. We defer the proof of Lemma 8 for now.

Lemma 8 (MSO is compositional under concatenation)**.** *Let s_i, s_i' for $i = 1, 2$ be two strings over the alphabet Σ. If*

$$\mathsf{mso\text{-}type}_k(s_1) = \mathsf{mso\text{-}type}_k(s_2) \quad \textit{and} \quad \mathsf{mso\text{-}type}_k(s_1') = \mathsf{mso\text{-}type}_k(s_2'),$$

then it holds that $\mathsf{mso\text{-}type}_k(s_1 \cdot s_1') = \mathsf{mso\text{-}type}_k(s_2 \cdot s_2')$

Proof of Theorem 7:
$\diamond$ Forward implication. We use the fact that any regular language has a regular expression. To establish that a regular language is MSO-definable, it suffices to prove that a regular language L are MSO-definable for each of the following cases by induction on the length of the regular expression R generating L:

- $R = a$ for some letter $a \in \Sigma$:
$$\varphi_a := \exists x (P_a(x) \wedge \forall z (x = z)).$$
- $R = \epsilon$:
$$\varphi_a := \neg \exists x (x = x).$$
- $R = \emptyset$:
$$\varphi_\emptyset := \exists x (x \neq x)$$
- $R = R_1 \cup R_2$: by induction hypothesis, there exists MSO-sentences φ_1 and φ_2 such that $s \models \varphi_i$ if and only if $s \in L(R_i)$ for $i = 1, 2$. Now $\varphi := \varphi_1 \vee \varphi_2$ is the desired MSO-sentence.
- $R = \bar{R}'$: by induction hypothesis, there exists MSO-sentences φ' such that $s \models \varphi'$ if and only if $s \in L(R')$. Now $\varphi := \neg\varphi'$ is the desired MSO-sentence.

• $R = R_1 \cdot R_2$: Again by induction hypothesis, there exists MSO-sentences φ_1 and φ_2 such that $s \models \varphi_i$ if and only if $s \in L(R_i)$ for $i = 1, 2$.

If we simply take the conjunction $\varphi_1 \wedge \varphi_2$ to define $L(R)$, then for the evaluation $\varphi_1 \wedge \varphi_2$ on a given string s we consider an interpretation of a variable of φ_1 in s. But what we actually want is to evaluate φ_1 on the substring $s[1 : z]$, i.e. *up to some position* z. Likewise, we want to evaluate φ_2 on the substring $s[z + 1, n]$. For this, we need to modify the original sentence φ_1 defining $L(R_1)$ so that, even when the variables are interpreted in s, in practice its interpretation is confined to the prefix (likewise φ_2 for the suffix of s). We can achieve this effect by replacing every occurrence

- $\exists x\ \psi$ by $\exists x (x \leq z) \wedge \psi(x)$, and
- $\exists X\ \psi$ by $\exists X (\forall x (x \in X \to x \leq z)) \wedge \psi(X)$

in φ_1 using z as a free individual variable. A similar relativization can be applied to the universal quantifiers in φ_1.

- $\forall x\ \psi(x)$ by $\forall x (x \leq z \to \psi)$, and
- $\forall X\ \psi$(X) by $\forall X (\forall x (x \in X \to x \leq z)) \to \psi(X)$ (read as: "if $X \leq z$ then ψ holds").

[2]The compositionality of MSO logic on strings under concatenationa is a special case of Feferman-Vaught Theorem.

A symmetric modification applies to φ_2. Apparently, the free variable z points at the last position of the prefix so that it matches a string from $L(R_1)$. Let $\varphi_1^{pf}(z)$ and $\varphi_2^{sf}(z)$ be the respective formulas obtained as above. It is not difficult to see that

$$\varphi := \exists z\ \varphi_1^{pf}(z) \wedge \varphi_2^{sf}(z)$$

is an MSO-sentence defining $L(R)$.

• $R = (R')^*$: by induction hypothesis, there exists MSO-sentences ψ such that $s \models \psi$ if and only if $s \in L(R')$. The trouble of using *delimiters* as in the case of concatenation using individual variables does not work as the operation $*$ may require arbitrarily many delimiters. However, using set variables, we can designate the delimiters *simultaneously* no matter how many substrings, each of which is generated by R'. Let us introduce a free set variable Z, which shall be interpreted as the set of last positions of substrings s_i when s is written as $s_1 \cdot s_2 \cdot \dots \cdot s_n$, each $s_i \in L(R')$. Cleary, s is a string generated by $R = (R')^*$ if and only if s can be written in this way for some $n \geq 1$ or $s = \epsilon$.

Next, we want to talk about the *interval between the delimiters*. Specifically, we want to define the substring s_i for each i. For this, we need a formula with free set variable Z and I which tests if (i) I is indeed an interval, i.e. contiguous, (ii) the *maximum* element in I is a delimiter, i.e. belongs to Z, and (iii) there is a unique element in I that belongs to Z. Let $\varphi_{good}(I, Z)$ be such a formula.

We define a formula $\varphi_{\max}(z, I)$ which says that z is the maximum element in the set I. The following formula serves this purpose:

$$\varphi_{\max}(z, I) := \forall x\ (x \in I \rightarrow x \leq z).$$

It is not difficult to write (left to the readers) to write $\varphi_{good}(I, Z)$ using $\varphi_{\max}(z, I)$.

We also define a formula which says that an interval I *does not start in the middle*; for this we can use the following formula[3]:

$$\varphi_{maximal}(I, Z) := \forall x\ (x < I \rightarrow \exists z\ (z < I \wedge z \in Z \wedge x \leq z)).$$

Using $\varphi_{good}(I, Z)$ and $\varphi_{maximal}(I, Z)$, it is an easy exercise to write an MSO-formula $\varphi_{dlm}(Z)$ which evaluates to TRUE for delimiters Z if and only if each *maximal interval* I *w.r.t* Z satisfies the formula ψ :

$$\varphi_{dlm}(Z) := \forall I\ (\varphi_{good}(I, Z) \wedge \varphi_{maximal}(I, Z) \rightarrow \psi^{int}(I))$$

and here, $\psi^{int}(I)$ is a *relativization of* ψ with respect to I. The idea is the same as in the case of concatenation. We want to *activate* an interpretation of a variable in φ only when the interpretation is confined to the interval I. The implementation of $\psi^{int}(I)$ with an actual MSO-sentence is an easy exercise.

Finally, we write a sentence which defines the language generated by the regular expression $(R')^*$:

$$\varphi := \nexists x\ (x = x) \vee (\exists Z\ (\varphi_{dlm}(Z) \wedge \forall z\ \varphi_{\max}(z) \rightarrow z \in Z))$$

where $\varphi_{\max}(z)$ is a formula which checks if z is the maximum element in the entire universe; one can obtain the formula as a simple modification of $\varphi_{\max}(z, I)$.

$\diamond$ Backward implication. Suppose that ψ is an MSO-sentence which defines a language $L \subseteq \Sigma^*$ and let $k := \mathsf{qr}(\psi)$.

[3] Remark on $\varphi_{maximal}(I, Z)$: If I contains the first element of the string, then for every x (interpretation of x) the formula $(x < I)$ is FALSE and the formula after the quantifier $\forall x$ is satisfied for every x. Therefore $\varphi_{maximal}(I, Z)$ holds for any such I.

Claim 1. *Let ψ is an* MSO*-sentence with $k := \mathsf{qr}(\psi)$ which defines $L \subseteq \Sigma^*$. Then, for any string $w \in \Sigma$, it holds that $\mathsf{mso\text{-}type}_k(w) \in \{\mathsf{mso\text{-}type}_k(s) \mid s \in L\}$ if and only if $w \models \psi$.*

PROOF OF THE CLAIM: Recall that $w \in L$ if and only if $w \models \psi$, which implies the backward implication. To see the forward direction, suppose that $\mathsf{mso\text{-}type}_k(w) = \mathsf{mso\text{-}type}_k(s)$ for some $s \in L$. As ψ has quantifier rank at most k, it holds that $s \models \psi$, or equivalently the sentence ψ is contained in $\mathsf{mso\text{-}type}_k(s)$. Therefore, $\psi \in \mathsf{mso\text{-}type}_k(w)$, finishing the proof. $\Diamond$

Consider the 5-tuple $M = (Q, \Sigma, \delta, q_0, F)$, where

- Q is the set of all MSO rank-k types $(0, 0$-types).
- q_0 is $\mathsf{mso\text{-}type}_k(\epsilon)$.
- F is the set of $\{\mathsf{mso\text{-}type}_k(s) \mid s \in L\}$.
- δ is a function from $Q \times \Sigma \to Q$ defined as

$$\delta(q, a) = \mathsf{mso\text{-}type}_k(sa) \quad \text{if there is a string } s \in \Sigma^* \text{ s.t. } \mathsf{mso\text{-}type}_k(s) = q$$

First, we claim that M defines a deterministic finite automaton. For this, it suffices to show that (i) for every $q \in Q$ and for every $a \in \Sigma$, $\delta(q, a)$ has a unique value and it belongs to Q, and (ii) the number of states Q is finite. Indeed, if $\mathsf{mso\text{-}type}_k(s) = \mathsf{mso\text{-}type}_k(s')$, then we have $\mathsf{mso\text{-}type}_k(sa) = \mathsf{mso\text{-}type}_k(s'a)$ for any $a \in \Sigma$ by Lemma 8. This means that, if $q = \mathsf{mso\text{-}type}_k(s)$ for some string s, then $\delta(q, a)$ is uniquely defined for every $a \in \Sigma$. By Lemma 4, there exists a string s such that $q = \mathsf{mso\text{-}type}_k(s)$ for every $q \in Q$. Therefore, $\delta(q, a)$ has a unique value for every $(q, a) \in Q \times \Sigma$. To see that $\delta(q, a) \in Q$, notice that the string sa certifies $\mathsf{mso\text{-}type}_k(sa) \in Q$, thus $\delta(q, a) \in Q$, whenever $\mathsf{mso\text{-}type}_k(s) = q$. This proves (i). That (ii) holds is immediate from Observation 3.

Secondly, we want to show that $L(M) = L$. For this we use the following claim.

Claim 2. *The run of M on a string $s \in \Sigma^*$ ends in $\mathsf{mso\text{-}type}_k(s)$.*

PROOF OF THE CLAIM: We prove by induction on $|s|$. When $s = \epsilon$, then the claim trivially holds by definition of q_0, the start state. Let $a \in \Sigma$ be the symbol such that $s = s'a$. By induction hypothesis, $\mathsf{mso\text{-}type}_k(s')$ is the last state of the run of M on s'. Now after reading the symbol a, the transition function δ updates the state from $\mathsf{mso\text{-}type}_k(s')$ to $\mathsf{mso\text{-}type}(s'a)$ by construction of δ. $\Diamond$

It remains to observe that Claim 2 and Claim 1 establish $L(M) = L$. $\square$

An alternative proof of the forward implication in Theorem 7 which constructs an MSO-sentence *simulating* an accepting run of the automaton M recognizing L is presented in [3].

References

[1] J. Richard Büchi. Weak second-order arithmetic and finite automata. *Mathematical Logic Quarterly*, 6(1-6):66–92, 1960.

[2] Calvin C. Elgot. Decision problems of finite automata design and related arithmetics. *Transactions of the American Mathematical Society*, 98(1):21–51, 1961.

[3] Leonid Libkin. *Elements of Finite Model Theory*. Springer, 2004.