Theory of Computation

Thomas Zeugmann

Hokkaido University
Laboratory for Algorithmics

http://www-alg.ist.hokudai.ac.jp/~thomas/ToC/

Lecture 8: $\mathcal{CF}$ and Homomorphisms

Substitutions I

In this lecture we continue with further useful properties and characterizations of context-free languages. First, we look at substitutions.

Substitutions I

In this lecture we continue with further useful properties and characterizations of context-free languages. First, we look at substitutions.

Definition 1

Let Σ and Δ be any two finite alphabets. A mapping $\tau\colon \Sigma \longrightarrow \wp(\Delta^*)$ is said to be a ***substitution***. We extend τ to be a mapping $\tau\colon \Sigma^* \longrightarrow \wp(\Delta^*)$ (i.e., to strings) by defining

(1) $\tau(\lambda) = \lambda$,

(2) $\tau(wx) = \tau(w)\tau(x)$ for all $w \in \Sigma^*$ and $x \in \Sigma$.

The mapping τ is generalized to languages $L \subseteq \Sigma^*$ by setting

$$\tau(L) = \bigcup_{w \in L} \tau(w)\ .$$

Substitutions II

So, a substitution maps every symbol of Σ to a language over Δ. The language a symbol is mapped to can be finite or infinite.

Substitutions II

So, a substitution maps every symbol of Σ to a language over Δ. The language a symbol is mapped to can be finite or infinite.

Example 1

Let $\Sigma = \{0, 1\}$ and let $\Delta = \{a, b\}$. Then, the mapping τ defined by $\tau(\lambda) = \lambda$, $\tau(0) = \{a\}$ and $\tau(1) = \{b\}^*$ is a substitution.

Let us calculate $\tau(010)$. By definition,

$$\tau(010) = \tau(01)\tau(0) = \tau(0)\tau(1)\tau(0) = \{a\}\{b\}^*\{a\} = \underline{a}\langle\underline{b}\rangle\underline{a}\ ,$$

where the latter equality is by the definition of regular expressions.

Substitutions III

Next, we want to define what is meant by closure of a language family $\mathcal{L}$ under substitutions. Here special care is necessary. At first glance, we may be tempted to require that for every substitution τ the condition $\tau(L) \in \mathcal{L}$ has to be satisfied. But this is a too strong demand. Why?

Substitutions III

Next, we want to define what is meant by closure of a language family $\mathcal{L}$ under substitutions. Here special care is necessary. At first glance, we may be tempted to require that for every substitution τ the condition $\tau(L) \in \mathcal{L}$ has to be satisfied. But this is a too strong demand. Why?

Consider $\Sigma = \{0, 1\}$, $\Delta = \{a, b\}$ and $\mathcal{L} = \mathcal{REG}$. Furthermore, suppose that $\tau(0) = L$, where L is any recursively enumerable but non-recursive language over Δ.

Substitutions III

Next, we want to define what is meant by closure of a language family $\mathcal{L}$ under substitutions. Here special care is necessary. At first glance, we may be tempted to require that for every substitution τ the condition $\tau(L) \in \mathcal{L}$ has to be satisfied. But this is a too strong demand. Why?

Consider $\Sigma = \{0, 1\}$, $\Delta = \{a, b\}$ and $\mathcal{L} = \mathcal{REG}$. Furthermore, suppose that $\tau(0) = L$, where L is any recursively enumerable but non-recursive language over Δ.

Then we obviously have $\tau(\{0\}) = L$, too. Consequently, $\tau(\{0\}) \notin \mathcal{REG}$. On the other hand, $\{0\} \in \mathcal{REG}$, and thus we would conclude that $\mathcal{REG}$ is not closed under substitution. Also, the same argument would prove that $\mathcal{CF}$ is not closed under substitution.

Substitutions IV

The point to be made here is that we have to restrict the set of allowed substitutions to those ones that map the elements of Σ to languages belonging to $\mathcal{L}$. Therefore, we arrive at the following definition.

Substitutions IV

The point to be made here is that we have to restrict the set of allowed substitutions to those ones that map the elements of Σ to languages belonging to $\mathcal{L}$. Therefore, we arrive at the following definition.

Definition 2

Let Σ be any alphabet, and let $\mathcal{L}$ be any language family over Σ. We say that $\mathcal{L}$ is ***closed under substitutions*** if for every substitution $\tau\colon \Sigma \longrightarrow \mathcal{L}$ and every $L \in \mathcal{L}$ we have $\tau(L) \in \mathcal{L}$.

Homomorphisms I

Definition 3

Let Σ and Δ be any two finite alphabets. A mapping $\varphi\colon \Sigma^* \longrightarrow \Delta^*$ is said to be a ***homomorphism*** if

$$\varphi(vw) = \varphi(v)\varphi(w) \quad \text{for all } v, w \in \Sigma^* .$$

φ is said to be a ***λ-free homomorphism***, if additionally

$$\varphi(w) = \lambda \quad \text{implies} \quad w = \lambda \quad \text{for all } w \in \Sigma^* .$$

Moreover, if $\varphi\colon \Sigma^* \longrightarrow \Delta^*$ is a homomorphism then we define the ***inverse of the homomorphism*** φ to be the mapping $\varphi^{-1}\colon \Delta^* \longrightarrow \wp(\Sigma^*)$ by setting for each $s \in \Delta^*$

$$\varphi^{-1}(s) = \{w \mid w \in \Sigma^* \text{ and } \varphi(w) = s\} .$$

Homomorphisms II

So, a homomorphism is a special case of a substitution.

Homomorphisms II

So, a homomorphism is a special case of a substitution.

That is, a homomorphism is a substitution that maps every symbols of Σ to a *singleton* set. Clearly, by the definition of homomorphism, it already suffices to declare the mapping φ for the symbols in Σ. Note that, when dealing with homomorphisms we usually identify the language containing exactly one string by the string itself, i.e., instead of $\{s\}$ we shortly write s.

Homomorphisms II

So, a homomorphism is a special case of a substitution.

That is, a homomorphism is a substitution that maps every symbols of Σ to a *singleton* set. Clearly, by the definition of homomorphism, it already suffices to declare the mapping φ for the symbols in Σ. Note that, when dealing with homomorphisms we usually identify the language containing exactly one string by the string itself, i.e., instead of $\{s\}$ we shortly write s.

Example 2

Let $\Sigma = \{0, 1\}$ and let $\Delta = \{a, b\}$. Then, the mapping $\varphi\colon \Sigma^* \longrightarrow \Delta^*$ defined by $\varphi(0) = ab$ and $\varphi(1) = \lambda$ is a homomorphisms but *not a λ-free homomorphism*. Applying φ to 1100 yields $\varphi(1100) = \varphi(1)\varphi(1)\varphi(0)\varphi(0) = \lambda\lambda abab = abab$ and to the language $\underline{1}\langle\underline{0}\rangle\underline{1}$ gives $\varphi(\underline{1}\langle\underline{0}\rangle\underline{1}) = \langle\underline{ab}\rangle$.

Remarks

For seeing the importance of the notions just introduced consider the language $L = \{a^n b^n \mid n \in \mathbb{N}\}$. This language is context-free. Thus, we intuitively know that $\{0^n 1^n \mid n \in \mathbb{N}\}$ is also context-free, since we could go through the grammar and replace all occurrences of a by 0 and all occurrences of b by 1.

Remarks

For seeing the importance of the notions just introduced consider the language $L = \{a^n b^n \mid n \in \mathbb{N}\}$. This language is context-free. Thus, we intuitively know that $\{0^n 1^n \mid n \in \mathbb{N}\}$ is also context-free, since we could go through the grammar and replace all occurrences of a by 0 and all occurrences of b by 1.

This observation would suggest that if we replace all occurrences of a and b by strings v and w, respectively, we also get a context-free language. However, it is much less intuitive that we also obtain a context-free language if all occurrences of a and b are replaced by context-free sets of strings V and W, respectively.

Remarks

For seeing the importance of the notions just introduced consider the language $L = \{a^n b^n \mid n \in \mathbb{N}\}$. This language is context-free. Thus, we intuitively know that $\{0^n 1^n \mid n \in \mathbb{N}\}$ is also context-free, since we could go through the grammar and replace all occurrences of a by 0 and all occurrences of b by 1.

This observation would suggest that if we replace all occurrences of a and b by strings v and w, respectively, we also get a context-free language. However, it is much less intuitive that we also obtain a context-free language if all occurrences of a and b are replaced by context-free sets of strings V and W, respectively.
Nevertheless, we just aim to prove this *closure* property. In the following we always assume two finite alphabets Σ and Δ as in the definition of substitution.

Closure under Substitutions I

Theorem 1

$\mathcal{CF}$ is closed under substitutions.

Closure under Substitutions I

Theorem 1

$\mathcal{CF}$ is closed under substitutions.

Proof. Let $L \in \mathcal{CF}$ be arbitrarily fixed and let τ be a substitution such that $\tau(a)$ is a context-free language for all $a \in \Sigma$. We have to show that $\tau(L)$ is context-free. We shall do this by providing a context-free grammar $\overline{\mathcal{G}} = [\overline{T}, \overline{N}, \overline{\sigma}, \overline{P}]$ such that $L(\overline{\mathcal{G}}) = \tau(L)$.

Closure under Substitutions I

Theorem 1

$\mathcal{CF}$ is closed under substitutions.

Proof. Let $L \in \mathcal{CF}$ be arbitrarily fixed and let τ be a substitution such that $\tau(a)$ is a context-free language for all $a \in \Sigma$. We have to show that $\tau(L)$ is context-free. We shall do this by providing a context-free grammar $\overline{\mathcal{G}} = [\overline{T}, \overline{N}, \overline{\sigma}, \overline{P}]$ such that $L(\overline{\mathcal{G}}) = \tau(L)$.

Since $L \in \mathcal{CF}$, there exists a context-free grammar $\mathcal{G} = [\Sigma, N, \sigma, P]$ in Chomsky normal form such that $L = L(\mathcal{G})$. Next, let $\Sigma = \{a_1, \ldots, a_n\}$ and consider $\tau(a)$ for all $a \in \Sigma$. By assumption, $\tau(a) \in \mathcal{CF}$ for all $a \in \Sigma$. Thus, there are context-free grammars $\mathcal{G}_a = [T_a, N_a, \sigma_a, P_a]$ such that $\tau(a) = L(\mathcal{G}_a)$ for all $a \in \Sigma$. Without loss of generality, we can assume the sets $N, N_{a_1}, \ldots, N_{a_n}$ to be pairwise disjoint and disjoint to all terminal alphabets considered.

Closure under Substitutions II

At this point we need an *idea how to proceed*. To get this idea, we look at possible derivations in $\mathcal{G}$. Suppose we have a derivation

$$\sigma \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} x_1 x_2 \cdots x_m \ ,$$

where all $x_i \in \Sigma$ for $i = 1, \ldots, m$.

Closure under Substitutions II

At this point we need an *idea how to proceed*. To get this idea, we look at possible derivations in $\mathcal{G}$. Suppose we have a derivation

$$\sigma \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} x_1 x_2 \cdots x_m ,$$

where all $x_i \in \Sigma$ for $i = 1, \dots, m$. Then, since $\mathcal{G}$ is in Chomsky normal form, we can conclude that there must be productions $(h_{x_i} \rightarrow x_i) \in P$, $i = 1, \dots, m$, and hence achieve the following.

$$\sigma \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} h_{x_1} h_{x_2} \cdots h_{x_m} \overset{m}{\underset{\mathcal{G}}{\Longrightarrow}} x_1 x_2 \cdots x_m , \tag{1}$$

where all $h_{x_i} \in N$.

Closure under Substitutions III

Taking into account that the image $\tau(x_1 \cdots x_m)$ is obtained by calculating

$$\tau(x_1)\tau(x_2) \cdots \tau(x_m) ,$$

we see that for every string $w_1 w_2 \cdots w_m$ in this image there must be a derivation

$$\sigma_{x_i} \xRightarrow[\mathcal{G}_{x_i}]{*} w_i \quad i = 1, \ldots, m .$$

This directly yields the idea for constructing $\overline{\mathcal{G}}$.

Closure under Substitutions IV

We aim to cut the derivation in (1) when having obtained $h_{x_1}h_{x_2}\cdots h_{x_m}$. Instead of deriving $x_1x_2\cdots x_m$, all we need is to generate $\sigma_{x_1}\cdots\sigma_{x_m}$, and thus, we have to replace the productions $(h_{x_i} \rightarrow x_i) \in P$ by $(h_{x_i} \rightarrow \sigma_{x_i}) \in \overline{P}$, $i = 1, \ldots, m$. So we define:

Closure under Substitutions IV

We aim to cut the derivation in (1) when having obtained $h_{x_1} h_{x_2} \cdots h_{x_m}$. Instead of deriving $x_1 x_2 \cdots x_m$, all we need is to generate $\sigma_{x_1} \cdots \sigma_{x_m}$, and thus, we have to replace the productions $(h_{x_i} \rightarrow x_i) \in P$ by $(h_{x_i} \rightarrow \sigma_{x_i}) \in \overline{P}$, $i = 1, \ldots, m$. So we define:

$$\begin{aligned}
\overline{T} &= \bigcup_{a \in \Sigma} T_a \\
\overline{N} &= N \cup \left(\bigcup_{a \in \Sigma} N_a \right) \\
\overline{\sigma} &= \sigma \\
\overline{P} &= \left(\bigcup_{a \in \Sigma} P_a \right) \cup P[a /\!/ \sigma_a] \,.
\end{aligned}$$

We set $\overline{\mathcal{G}} = [\overline{T}, \overline{N}, \overline{\sigma}, \overline{P}]$.

Closure under Substitutions V

It remains to show that $\tau(L) = L(\overline{\mathcal{G}})$.

Claim 1. $\tau(L) \subseteq L(\overline{\mathcal{G}})$.

Closure under Substitutions V

It remains to show that $\tau(L) = L(\overline{\mathcal{G}})$.

Claim 1. $\tau(L) \subseteq L(\overline{\mathcal{G}})$.

If $\sigma \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} x_1 \cdots x_m$, where $x_i \in \Sigma$ and if $\sigma_{x_i} \overset{*}{\underset{\mathcal{G}_{x_i}}{\Longrightarrow}} w_i$, where $w_i \in T^*_{x_i}$, $i = 1, \ldots, m$, then we derive $x_1 \cdots x_m$ as follows:

$$\sigma \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} h_{x_1} \cdots h_{x_m} \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} x_1 \cdots x_m \ ,$$

where all $h_{x_i} \in N$. By construction, we can thus generate

$$\sigma \overset{*}{\underset{\overline{\mathcal{G}}}{\Longrightarrow}} h_{x_1} \cdots h_{x_m} \overset{*}{\underset{\overline{\mathcal{G}}}{\Longrightarrow}} \sigma_{x_1} \cdots \sigma_{x_m} \overset{*}{\underset{\overline{\mathcal{G}}}{\Longrightarrow}} w_1 \cdots w_m \ .$$

Hence, Claim 1 follows.

Closure under Substitutions VI

Claim 2. $L(\overline{\mathcal{G}}) \subseteq \tau(L)$.

Now, we start from $\sigma \overset{*}{\Rightarrow} w$, where $w \in \overline{T}^*$. If $w = \lambda$, then also $\sigma \ \rightarrow \ \lambda$ in P, and we are done.

Closure under Substitutions VI

Claim 2. $L(\overline{\mathcal{G}}) \subseteq \tau(L)$.

Now, we start from $\sigma \stackrel{*}{\Rightarrow} w$, where $w \in \overline{T}^*$. If $w = \lambda$, then also $\sigma \rightarrow \lambda$ in P, and we are done. Otherwise, the construction of $\overline{\mathcal{G}}$ ensures that the derivation of w must look as follows.

$$\sigma \underset{\overline{\mathcal{G}}}{\overset{*}{\Longrightarrow}} \sigma_{x_1} \cdots \sigma_{x_m} \underset{\overline{\mathcal{G}}}{\overset{*}{\Longrightarrow}} w\,.$$

Closure under Substitutions VI

Claim 2. $L(\overline{\mathcal{G}}) \subseteq \tau(L)$.

Now, we start from $\sigma \overset{*}{\Rightarrow} w$, where $w \in \overline{T}^*$. If $w = \lambda$, then also $\sigma \to \lambda$ in P, and we are done. Otherwise, the construction of $\overline{\mathcal{G}}$ ensures that the derivation of w must look as follows.

$$\sigma \underset{\overline{\mathcal{G}}}{\overset{*}{\Longrightarrow}} \sigma_{x_1} \cdots \sigma_{x_m} \underset{\overline{\mathcal{G}}}{\overset{*}{\Longrightarrow}} w .$$

By our construction we then know that $\sigma \underset{\mathcal{G}}{\overset{*}{\Longrightarrow}} x_1 \cdots x_m$ as we have shown in (1). Also, there are strings $w_1, \ldots, w_m \in \overline{T}^*$ such that $w = w_1 \cdots w_m$ and $\sigma_{x_i} \underset{\mathcal{G}_{x_i}}{\overset{*}{\Longrightarrow}} w_i$ for all $i = 1, \ldots, m$.

Closure under Substitutions VII

Consequently, $w_i \in \tau(x_i)$. Therefore, $w \in \tau(L)$ and we are done.

Closure under Substitutions VII

Consequently, $w_i \in \tau(x_i)$. Therefore, $w \in \tau(L)$ and we are done.

Finally, putting Claim 1 and 2 together, we see that

$$\tau(L) = L(\overline{\mathcal{G}}) \; .$$

$\blacksquare$

Closure under Substitutions VII

Consequently, $w_i \in \tau(x_i)$. Therefore, $w \in \tau(L)$ and we are done.

Finally, putting Claim 1 and 2 together, we see that

$$\tau(L) = L(\overline{\mathcal{G}}) \, .$$

$\blacksquare$

Our Theorem allows the following nice corollary.

Corollary 2

$\mathcal{CF}$ is closed under homomorphisms.

Proof. Since homomorphisms are a special type of substitution, it suffices to argue that every singleton subset is context-free. But this is obvious, because we have already shown that every finite language belongs to $\mathcal{REG}$ and that $\mathcal{REG} \subseteq \mathcal{CF}$. Thus, the corollary follows. $\blacksquare$

Dyck Languages I

When we started to study context-free languages, we emphasized that many programming languages use balanced brackets of different kinds. Therefore, we continue with a closer look at bracket languages. Such languages are called *Dyck languages*.

Dyck Languages I

When we started to study context-free languages, we emphasized that many programming languages use balanced brackets of different kinds. Therefore, we continue with a closer look at bracket languages. Such languages are called *Dyck languages*.

In order to define Dyck languages, we need the following notations. Let $n \in \mathbb{N}^+$ and let

$$X_n = \{a_1, \overline{a}_1, a_2, \overline{a}_2, \ldots, a_n, \overline{a}_n\} .$$

We consider the set X_n as a set of different bracket symbols, where a_i is an opening bracket and $\overline{a}_i$ is the corresponding closing bracket. Thus, it is justified to speak of X_n as a set of n different bracket symbols.

Dyck Languages II

Now we are ready to define Dyck languages.

Definition 4

A language L is said to be a ***Dyck language*** with n bracket symbols if L is isomorphic to the language D_n generated by the following grammar $\mathcal{G}_n = [X_n, \{\sigma\}, \sigma, P_n]$, where P_n is given by

$$P_n = \{\sigma \rightarrow \lambda,\ \sigma \rightarrow \sigma\sigma,\ \sigma \rightarrow a_1\sigma\bar{a}_1,\ \ldots,\ \sigma \rightarrow a_n\sigma\bar{a}_n\}\,.$$

Dyck Languages II

Now we are ready to define Dyck languages.

Definition 4

A language L is said to be a ***Dyck language*** with n bracket symbols if L is isomorphic to the language D_n generated by the following grammar $\mathcal{G}_n = [X_n, \{\sigma\}, \sigma, P_n]$, where P_n is given by

$$P_n = \{\sigma \rightarrow \lambda,\ \sigma \rightarrow \sigma\sigma,\ \sigma \rightarrow a_1\sigma\bar{a}_1,\ \ldots,\ \sigma \rightarrow a_n\sigma\bar{a}_n\}\,.$$

The importance of Dyck languages will become immediately transparent, since we are going to prove a beautiful characterization theorem for context-free languages by using them.

Chomsky-Schützenberger Theorem I

Theorem 3 (Chomsky-Schützenberger Theorem)

For every context-free language L *there are* $n \in \mathbb{N}^+$*, a homomorphism* h *and a regular language* R_L *such that*

$$L = h(D_n \cap R_L) \ .$$

Chomsky-Schützenberger Theorem II

Proof. Consider any arbitrarily fixed context-free language L. Without loss of generality we can assume that $\lambda \notin L$. Furthermore, let $\mathcal{G} = [T, N, \sigma, P]$ be a context-free grammar in Chomsky normal form such that $L = L(\mathcal{G})$. Let $T = \{x_1, \ldots, x_m\}$ and consider all productions in P. Since $\mathcal{G}$ is in Chomsky normal form, all productions have the form $h_i \rightarrow h_i' h_i''$ or $h_j \rightarrow x$. Let t be the number of all nonterminal productions, i.e., of all productions $h_i \rightarrow h_i' h_i''$. Note that for any two such productions it is well possible that some but not all nonterminal symbols coincide.

Chomsky-Schützenberger Theorem III

In all we have m terminal symbols and t nonterminal productions. Thus, we try the Dyck language D_{m+t} over

$$\begin{aligned} X_{m+t} \;&=\; \{\overline{x}_1, \ldots, \overline{x}_m, \overline{x}_{m+1}, \ldots, \overline{x}_{m+t}, \\ &\quad x_{m+1}, \ldots, x_{m+t}, x_1, \ldots, x_m\}\,. \end{aligned}$$

Chomsky-Schützenberger Theorem III

In all we have m terminal symbols and t nonterminal productions. Thus, we try the Dyck language D_{m+t} over

$$\begin{aligned} X_{m+t} &= \{\overline{x}_1, \ldots, \overline{x}_m, \overline{x}_{m+1}, \ldots, \overline{x}_{m+t}, \\ &\quad x_{m+1}, \ldots, x_{m+t}, x_1, \ldots, x_m\}. \end{aligned}$$

Next, we consider the mapping $\chi_{m+t} \colon X_{m+t} \longrightarrow T^*$ defined as follows.

$$\chi_{m+t}(x_j) = \begin{cases} x_j, & \text{if } 1 \leqslant j \leqslant m\,; \\ \lambda, & \text{if } m+1 \leqslant j \leqslant m+t\,; \end{cases}$$

and $\chi_{m+t}(\overline{x}_j) = \lambda$ for all $j = 1, \ldots, m+t$. We leave it as an exercise to show that χ_{m+t} is a homomorphism.

Chomsky-Schützenberger Theorem IV

Now we are ready to define the following grammar $\mathcal{G}_L = [X_{m+t}, N, \sigma, P_L]$, where

$$
\begin{aligned}
P_L &= \{h \rightarrow x_i\bar{x}_i \mid 1 \leqslant i \leqslant m \text{ and } (h \rightarrow x_i) \in P\} \\
&\cup \{h \rightarrow x_i\bar{x}_i\bar{x}_{m+j}h''_j \mid 1 \leqslant i \leqslant m,\ (h \rightarrow x_i) \in P,\ 1 \leqslant j \leqslant t\} \\
&\cup \{h_j \rightarrow x_{m+j}h'_j \mid 1 \leqslant j \leqslant t\}\,.
\end{aligned}
$$

Clearly, $\mathcal{G}_L$ is a regular grammar. We set $R_L = L(\mathcal{G}_L)$, and aim to prove that

$$L = \chi_{m+t}(D_{m+t} \cap R_L)\,.$$

This is done via the following claims and lemmata.

Chomsky-Schützenberger Theorem V

Claim 1. $L \subseteq \chi_{m+t}(D_{m+t} \cap R_L)$.

The proof of Claim 1 is mainly based on the following lemma.

Lemma 4

Let $\mathcal{G}$ *be the grammar for* L *fixed above, let* $\mathcal{G}_L$ *be the grammar for* R_L *and let* $h \in N$. *If*

$$h \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_1 \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_2 \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} \cdots \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_{n-1} \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_n \in T^*$$

then there exists a $q \in D_{m+t}$ *such that* $h \overset{*}{\underset{\mathcal{G}_L}{\Longrightarrow}} q$ *and*

$\chi_{m+t}(q) = w_n$.

Chomsky-Schützenberger Theorem VI

The lemma is shown by induction on the length n of the derivation. For the induction basis let $n = 1$. Thus, our assumption is that

$$h \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_1 \in T^* \ .$$

Since $\mathcal{G}$ is in Chomsky normal form, we can conclude that $(h \ \rightarrow \ w_1) \in P$. So, by the definition of Chomsky normal form, we must have $w_1 = x$ for some $x \in T$.

Chomsky-Schützenberger Theorem VII

We have to show that there is a $q \in D_{m+t}$ such that $h \overset{*}{\underset{\mathcal{G}_L}{\Longrightarrow}} q$ and $\chi_{m+t}(q) = x$. By construction, the production $h \rightarrow x\overline{x}$ belongs to P_L (cf. the first set of the definition of P_L). Thus, we can simply set $q = x\overline{x}$. Now, the induction basis follows, since the definition of χ_{m+t} directly yields

$$\chi_{m+t}(q) = \chi_{m+t}(x\overline{x}) = \chi_{m+t}(x)\chi_{m+t}(\overline{x}) = x\lambda = x\ .$$

Chomsky-Schützenberger Theorem VIII

Assuming the induction hypothesis for $n \geqslant 1$, we are going to perform the induction step to $n + 1$. So, let

$$h \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_1 \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} \cdots \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_n \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_{n+1} \in T^*$$

be a derivation of length $n + 1$.

Chomsky-Schützenberger Theorem VIII

Assuming the induction hypothesis for $n \geqslant 1$, we are going to perform the induction step to $n + 1$. So, let

$$h \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_1 \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} \cdots \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_n \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} w_{n+1} \in T^*$$

be a derivation of length $n + 1$.

Because of $n \geqslant 1$, and since the derivation has length at least 2, we can conclude that the production used to derive w_1 must be of the form $h \rightarrow h'h''$, where $h, h', h'' \in N$. Therefore, there must be a j such that $1 \leqslant j \leqslant t$ and $h = h_j$ as well as $w_1 = h_j'h_j''$.

Chomsky-Schützenberger Theorem IX

The latter observation implies that there must be v_1, v_2 such that $w_{n+1} = v_1 v_2$ and

$$h'_j \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} v_1 \quad \text{and} \quad h''_j \overset{*}{\underset{\mathcal{G}}{\Longrightarrow}} v_2 \ .$$

Since the length of the complete derivation is $n + 1$, both the generation of v_1 and of v_2 must have a length smaller than or equal to n.
Hence, we can apply the induction hypothesis. That is, there are strings q_1 and q_2 such that $q_1, q_2 \in D_{m+t}$ and $\chi_{m+t}(q_1) = v_1$ as well as $\chi_{m+t}(q_2) = v_2$.

Chomsky-Schützenberger Theorem X

Furthermore, by the induction hypothesis we additionally know that

$$h'_j \xRightarrow[\mathcal{G}_L]{*} q_1 \quad \text{and} \quad h''_j \xRightarrow[\mathcal{G}_L]{*} q_2 \,.$$

Taking into account that $(h_j \to h'_j h''_j) \in P$ we know by construction that $h_j \to x_{m+j} h'_j$ is a production in P_L. Thus,

$$h = h_j \xRightarrow[\mathcal{G}_L]{1} x_{m+j} h'_j \xRightarrow[\mathcal{G}_L]{*} x_{m+j} q_1$$

is a regular derivation. Moreover, the last step of this derivation must look as follows:

$$x_{m+j} q'_1 h_k \xRightarrow[\mathcal{G}_L]{1} x_{m+j} q'_1 x\overline{x} \,.$$

where $h_k \to x\overline{x}$ is the rule applied and where x is determined by the condition $q_1 = q'_1 x\overline{x}$.

Chomsky-Schützenberger Theorem XI

Now, we replace this step by using the production $h_k \rightarrow x\overline{x}x_{m+j}h_j''$ which also belongs to P_L. Thus, we obtain

$$h = h_j \overset{1}{\underset{\mathcal{G}_L}{\Longrightarrow}} x_{m+j}h_j' \overset{*}{\underset{\mathcal{G}_L}{\Longrightarrow}} x_{m+j}q_1\overline{x}_{m+j}h_j''$$

$$\overset{*}{\underset{\mathcal{G}_L}{\Longrightarrow}} x_{m+j}q_1\overline{x}_{m+j}q_2 =: q \in D_{m+t} \,.$$

The containment in D_{m+t} is due to the correct usage of the brackets x_{m+j} and $\overline{x}_{m+j}$ around q_1 and the fact that $q_2 \in D_{m+t}$ as well as by the definition of the Dyck language. Finally, the definition of χ_{m+t} ensures that $\chi_{m+t}(x_{m+j}q_1\overline{x}_{m+j}q_2) = v_1v_2$. This proves the lemma and Claim 1 immediately follows for $h = \sigma$.

Chomsky-Schützenberger Theorem XII

Claim 2. $L \supseteq \chi_{m+t}(D_{m+t} \cap R_L)$.

Again, the proof of the claim is mainly based on a lemma which we state next.

Lemma 5

Let $\mathcal{G}$ be the grammar for L fixed above, let $\mathcal{G}_L$ be the grammar for R_L and let $h \in N$. If

$$h \underset{\mathcal{G}_L}{\overset{1}{\Longrightarrow}} w_1 \underset{\mathcal{G}_L}{\overset{1}{\Longrightarrow}} \cdots \underset{\mathcal{G}_L}{\overset{1}{\Longrightarrow}} w_n \in D_{m+t}$$

then $h \underset{\mathcal{G}}{\overset{*}{\Longrightarrow}} \chi_{m+t}(w_n)$.

Chomsky-Schützenberger Theorem XIII

The lemma is shown by induction on the length of the derivation. We perform the induction basis for $n = 1$. Consider

$$h \overset{1}{\underset{\mathcal{G}_L}{\Longrightarrow}} w_1 \in D_{m+t} \; .$$

Hence, we must conclude that $(h \; \to \; w_1) \in P_L$. So, there must exist $x_i\overline{x}_i$ such that $w_1 = x_i\overline{x}_i$, $1 \leqslant i \leqslant m$ and $(h \; \to \; x_i\overline{x}_i) \in P_L$. By the definition of P_L we conclude that $(h \; \to \; x_i) \in P$. Hence

$$h \overset{1}{\underset{\mathcal{G}}{\Longrightarrow}} x_i = \chi_{m+t}(x_i\overline{x}_i) = \chi_{m+t}(w_1) \; .$$

This proves the induction basis.

Chomsky-Schützenberger Theorem XIV

The induction step is provided in the book.

Again, Claim 2 is a direct consequence of the latter lemma for $h = \sigma$.

Claim 1 and Claim 2 together imply the theorem. $\blacksquare$

Thank you!