CFGs for Finite Unary Sets

Lecture 11

Binghui Peng

Context

This lecture is based on the paper:

Simulating Finite Automata with Context-Free Grammars
by Domaratzki, Pighizzini, and Shallit
Information Processing Letters, Volume 84, 2002, pp. 339-344.

Modified Chomsky Normal Form

A context-free grammar is in **Modified Chomsky Normal Form (CNF)** if its production rules take one of the following forms:

- $A \to BCD$
- $A \to BC$
- $A \to \sigma$
- $A \to \epsilon$

In this lecture, we focus on **the number of nonterminals (NTs)** required to generate a language.

Allowing rules of the form $A \to BCD$ and $A \to \epsilon$ simplifies the construction for unary sets. Construction for the conventional Chomsky Normal Form will be more complicated and omitted from the lecture.

Objective: Succinct CFGs for Unary Sets

Goal: Construct a small CFG for a set $A \subseteq \{\epsilon, a, a^2, \ldots, a^n\}$.

Plan:

1. Construct a CNF grammar for $\{\epsilon, a, \ldots, a^{124}\}$ with 16 NTs.
2. Show that for any $A \subseteq \{\epsilon, a, \ldots, a^{124}\}$, there exists a CNF with 16 NTs.
3. Generalize: $\forall A \subseteq \{\epsilon, a, \ldots, a^n\}$, there exists a CNF with $O(n^{1/3})$ NTs.
4. Analyze the lower bound: $\exists A \subseteq \{\epsilon, a, \ldots, a^n\}$ requiring $\Omega(n^{1/3})$ NTs.

Construction Strategy for $\{\epsilon, a, \ldots, a^{124}\}$

We define three sets of nonterminals:

- $E_0, \ldots, E_4$ such that $L(E_i) = \{a^i\}$.
- $F_0, \ldots, F_4$ such that $L(F_i) = \{a^{5i}\}$.
- $G_0, \ldots, G_4$ such that $L(G_i) = \{a^{25i}\}$.

Every $m \in \{0, \ldots, 124\}$ can be represented in base 5 as $m = i + 5j + 25k$, where $0 \leq i, j, k \leq 4$. Thus, $a^m = L(E_i)L(F_j)L(G_k)$.

Production Rules for E_i, F_j, G_k

- $E_0 \to \epsilon, E_1 \to a, E_2 \to E_1E_1, E_3 \to E_2E_1, E_4 \to E_3E_1$
- $F_0 \to \epsilon, F_1 \to E_4E_1, F_2 \to F_1F_1, F_3 \to F_2F_1, F_4 \to F_3F_1$
- $G_0 \to \epsilon, G_1 \to F_4F_1, G_2 \to G_1G_1, G_3 \to G_2G_1, G_4 \to G_3G_1$

To generate $\{\epsilon, a, \ldots, a^{124}\}$, include rules:

$$S \to E_iF_jG_k \quad \forall 0 \leq i, j, k \leq 4$$

Property of the Construction

Each string a^m for $0 \leq m \leq 124$ is generated by exactly one rule $S \to E_i F_j G_k$.

Removing the rule $S \to E_i F_j G_k$ specifically omits the string $a^{i+5j+25k}$ from the language.

Subsets of $\{\epsilon, a, \ldots, a^{124}\}$

Theorem: For any $A \subseteq \{\epsilon, a, \ldots, a^{124}\}$, there exists a CNF grammar with 16 nonterminals.

Proof: Starting with the grammar for $\{\epsilon, a, \ldots, a^{124}\}$, remove the rule $S \to E_i F_j G_k$ if $a^{i+5j+25k} \notin A$. The number of nonterminals remains 16.

Generalization to $A \subseteq \{\epsilon, a, \ldots, a^{t^3-1}\}$

Theorem: For $t \in \mathbb{N}$, any $A \subseteq \{\epsilon, a, \ldots, a^{t^3-1}\}$ can be generated by a CNF with $3t + 1$ nonterminals.

Construction:

- $E_0, \ldots, E_{t-1}$ with $L(E_i) = \{a^i\}$ (t NTs)
- $F_0, \ldots, F_{t-1}$ with $L(F_i) = \{a^{ti}\}$ (t NTs)
- $G_0, \ldots, G_{t-1}$ with $L(G_i) = \{a^{t^2 i}\}$ (t NTs)
- Start symbol S with rules $S \to E_i F_j G_k$ (1 NT)

Complexity for Arbitrary n

Theorem: For any $A \subseteq \{\epsilon, a, \ldots, a^n\}$, there exists a CNF grammar with $O(n^{1/3})$ nonterminals.

By choosing $t = \lceil (n+1)^{1/3} \rceil$, we have $t^3 - 1 \geq n$, and the construction yields $3t + 1 = O(n^{1/3})$ nonterminals.

Results for Standard Chomsky Normal Form

Using the standard definition of CNF (rules $A \to BC$ or $A \to \sigma$):

Theorem: For any $A \subseteq \{\epsilon, a, \ldots, a^{t^3-1}\}$, there exists a standard CNF with $4t + 1$ nonterminals.

Corollary: For any $A \subseteq \{\epsilon, a, \ldots, a^n\}$, there exists a standard CNF with $O(n^{1/3})$ nonterminals.

Optimality of $O(n^{1/3})$ Nonterminals

Question: Can we achieve better than $O(n^{1/3})$ nonterminals for all $A \subseteq \{\epsilon, a, \ldots, a^n\}$?

Optimality of $O(n^{1/3})$ Nonterminals

Question: Can we achieve better than $O(n^{1/3})$ nonterminals for all $A \subseteq \{\epsilon, a, \ldots, a^n\}$?

Answer: No. There exists a subset A such that any CNF grammar requires $\Omega(n^{1/3})$ nonterminals.

Lower Bound: Counting Argument

- Number of subsets of $\{\epsilon, a, \ldots, a^{n-1}\}$: 2^n.
- Number of possible CNF grammars with t nonterminals:
  - Rules of form $A \to BC$: t^3 possibilities.
  - Rules of form $A \to \sigma$: $O(t)$ possibilities.
  - Total rules: $t^3 + O(t) \leq 2t^3$.
  - Total grammars: $\leq 2^{2t^3}$.

If $2^{2t^3} < 2^n$, there must exist a subset A that cannot be generated by any CNF with t nonterminals.

Setting $2t^3 < n$ implies $t < (n/2)^{1/3}$.

Thus, for $t = \Omega(n^{1/3})$, there exists a subset A requiring at least t nonterminals.