*Theorie des Algorithmischen Lernens*
*Sommersemester 2007*

# Teil 2: Lernen formaler Sprachen

Version 1.0

# Gliederung der LV

**Teil 1: Motivation**
1. Was ist Lernen
2. Das Szenario der Induktiven Inf erenz
3. Natürlichkeitsanforderungen

**Teil 2: Lernen formaler Sprachen**
1. Grundlegende Begriffe und Erkennungstypen
2. Die Rolle des Hypothesenraums
3. Lernen von Patternsprachen
4. Inkrementelles Lernen

**Teil 3: Lernen endlicher Automaten**

**Teil 4: Lernen berechenbarer Funktionen**
1. Grundlegende Begriffe und Erkennungstypen
2. Reflexion

**Teil 5: Informationsextraktion**
1. Island Wrappers
2. Query Scenarios

# 7 Parameters of Inductive Inference

1. objects to be learned
2. examples (syntax)
3. examples (semantics, i.e. connection to object to be learnt)
4. learning device
5. hypothesis space (syntax of hypotheses)
6. semantics of hypotheses
7. success criteria

# A few examples to start

- set of all finite languages on $\Sigma = \{a, b, c\}$
- set of all regular languages on $\Sigma = \{a, b, c\}$
- set of all decidable languages on $\Sigma = \{a, b, c\}$
- set of all enumerable languages on $\Sigma = \{a, b, c\}$
- set of all formal languages on $\Sigma = \{a, b, c\}$
- $L_0 = \{a^n \mid n \in \mathbb{N}\}$,
  $L_{i+1} = \{a, \ldots, a^{i+1}\}$ (d.h. $L_1 = \{a\}, L_2 = \{a, aa\}, L_3 = \{a, aa, aaa\}, \ldots$)
- $L_i = \Sigma^* \setminus \{a^i\}$

# Identification by Enumeration

**Theorem 2.1**:
The set of all context-free languages is learnable from complete information.

*Proof.*

Let $(G_j)_{j \in \mathbb{N}}$ be an enumeration of all context-sensitive grammars. Define learning machine $M$ as follows:

On input $i_n$ do: search the least $j \in \mathbb{N}$ such that the language described by the grammar $G_j$ is consistent with $i_n$. Output the language described by $G_j$.

# Analysis

Now, let $L$ be a context-free language and $i$ be a complete presentation of $L$.

Observations:

- if $M$ outputs a hypothesis on $i_n$, it is consistent with $i_n$
- $M$ only changes its hypothesis if necessary
- $M$ always outputs the smallest hypothesis (w.r.t. the enumeration of grammars) that is consistent with $i_n$
- Let $m$ be the least index such that $L(G_m) = L$. Then, $M$ *never* outputs a hypothesis with an index larger than $m$.

  $\longrightarrow$ $M$ converges in the limit!!!
- Assume to the contrary that $M$ converges to a wrong hypothesis, i.e. to $L'$ with $L' \neq L$.

  $\longrightarrow$ There is a $w$ in the difference of $L$ and $L'$.

  $\longrightarrow$ $w$ sometimes occurs in $i$

  $\longrightarrow$ $L'$ is refused eventually which contradicts our assumption.

$\longrightarrow$ $M$ converges to a correct hypothesis, i.e. learns $L$!

# Hypothesis Space

- $M$ sometimes outputs hypotheses for languages that are not context-free
  - hypothesis space contains unnecessary elements
- What happens if we use all regular grammars as hypothesis space?
- What happens if we use all context-free grammars as hypothesis space?
- What happens if we use all chomsky-languages (i.e. all languages that have a finite grammar) as hypothesis space?
- What about all Java programs working as acceptors?
  - What about generators?

$\longrightarrow$ **hypothesis space must at least contain all languages to be learned**

# Identification by Enumeration: In General

***Identification by Enumeration*** works correctly if

- all target concepts can be *enumerated*
- *consistency* can be effectively *decided* in this enumeration
- the information about the target concept is *correct* and *complete in the limit*

$\longrightarrow$ works for arbitrary enumerations!

# Identification by Enumeration: A nice idea?

- consistent working manner
- only change hypothesis if necessary
- semantic finite (i.e. once a correct hypothesis is output, it is never changed)

Drawbacks

- does not work in all cases
- efficiency???

# Identification by Enumeration: A stupid idea?

**Lemma 2.2**:
There is no learning algorithm which outperforms (w.r.t. convergence speed) identification by enumeration (IBE).

*Proof.*

Let $(L_j)_{j\in\mathbb{N}}$ be an enumeration of languages.

Let $L = L_m$ be an arbitrary language out of it and $i$ be an arbitrary complete presentation for $L$.

Now assume that IBE needs $k$ examples until it converges, but some other algorithm $M$ needs only $k'$ with $k' < k$.

Consider the language $L'$ output by IBE on $k'$ examples and a complete presentation $i'$ for it which starts with the first $k'$ examples as above. IBE has reached its point of convergence already after $k'$ examples while $M$ needs at least one more mind change (i.e. converges slower on $i'$).

# Does this also work with positive examples?

What happens with identification by enumeration when only positive examples are available?

Consider the following class $\mathcal{L}_{sf}$:

- $L_0 = \{a^n \mid n \in \mathbb{N}\}$
- $L_{i+1} = \{a, \ldots, a^{i+1}\}$ for all $i \in \mathbb{N}$

How to enumerate it?

Idea: First put $\{a\}$, then $\{a, aa\}$, then $\{a, aa, aaa\}$, ...

Problem: Where to put $L_0$???

Terminus technicus: ***Overgeneralisation***

# Limits of learning from positive examples

**<u>Lemma 2.3</u>**: (Gold 1967)
The set $\mathcal{L}_{\textit{sf}}$ is not learnable from positive examples only.

*Proof.*

Assume the contrary, i.e. some learning device $M$ identifying $\mathcal{L}_{\textit{sf}}$ from positive examples. We now construct a sequence $t$ of positive examples for $L_0$:

```
t:= empty sequence;
i:= 1;
do forever:
        repeat until M(t) describes the language {a, ..., a^i}:
                append a^i to t
        i++;
```

Since $M$ learns each $L_i$ (i.e. eventually outputs a hypothesis for $L_i$), in the limit a sequence $a, \ldots a, aa, \ldots aa, aaa, \ldots, aaa, \ldots$ is constructed:

- which contains all positive examples for $L_0$
- on which $M$ infinitely often changes ist hypothesis, i.e. does not converge