What is this course about?

- Examining the power of an abstract machine

What can this box of tricks do?

What is this course about?

- Examining the power of an abstract machine
- Domains of discourse: automata and formal languages

Automaton is the box of tricks, language recognition is what it can do.

What is this course about?

- Examining the power of an abstract machine
- Domains of discourse: automata and formal languages
- Formalisms to describe languages and automata

Very useful for future courses.

What is this course about?

- Examining the power of an abstract machine
- Domains of discourse: automata and formal languages
- Formalisms to describe languages and automata
- Proving a particular case: relationship between regular languages and finite automata

Perhaps the simplest result about power of a machine. Finite Automata are simply a formalisation of finite state machines you looked at in Digital Electronics.

A word about formalisms to describe languages

- Classically (i.e. when I was young) this would be done using formal grammars.

e.g. $S \rightarrow NV$
e.g. $I \rightarrow ID$, $I \rightarrow D$, $I \rightarrow -D$

A word about formalisms to describe languages

- Classically (i.e. when I was young) this would be done using formal grammars.
- Here will we use rule induction

Excuse to introduce now, useful in other things

Syllabus for this part of the course

- Inductive definitions using rules and proofs by rule induction.
- Abstract syntax trees.
- Regular expressions and pattern matching.
- Finite automata and regular languages: Kleene's theorem.
- The Pumping Lemma.

mathematics needed for computer science

Common theme: mathematical techniques for defining formal languages and reasoning about their properties.

Key concepts: inductive definitions, automata

Relevant to:

Part IB	Compiler Construction, Computation Theory, Complexity Theory, Semantics of Programming Languages
Part II	Natural Language Processing, Optimising Compilers, Denotational Semantics, Temporal Logic and Model Checking

N.B. we do not cover the important topic of context-free grammars, which prior to 2013/14 was part of the CST IA course *Regular Languages and Finite Automata* that has been subsumed into this course.

see course web page for relevant Tripos questions

Formal Languages

Alphabets

An **alphabet** is specified by giving a finite set, Σ, whose elements are called **symbols**. For us, any set qualifies as a possible alphabet, so long as it is finite.

Examples:

- $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $\mathbf{10}$-element set of decimal digits.
- $\{a, b, c, \ldots, x, y, z\}$, $\mathbf{26}$-element set of lower-case characters of the English language.
- $\{S \mid S \subseteq \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}\}$, $\mathbf{2^{10}}$-element set of all subsets of the alphabet of decimal digits.

Non-example:

- $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, set of all non-negative whole numbers is not an alphabet, because it is infinite.

Strings over an alphabet

A **string of length** n (for $n = 0, 1, 2, \ldots$) over an alphabet Σ is just an ordered n-tuple of elements of Σ, written without punctuation.

Σ^* denotes set of all strings over Σ of any finite length.

Examples:

notation for the string of length 0

- If $\Sigma = \{a, b, c\}$, then ε, a, ab, aac, and $bbac$ are strings over Σ of lengths zero, one, two, three and four respectively.
- If $\Sigma = \{a\}$, then Σ^* contains ε, a, aa, aaa, $aaaa$, etc.

In general, a^n denotes the string of length n just containing a symbols

Strings over an alphabet

A **string of length** n (for $n = 0, 1, 2, \ldots$) over an alphabet Σ is just an ordered n-tuple of elements of Σ, written without punctuation.

Σ^* denotes set of all strings over Σ of any finite length.

Examples:

- If $\Sigma = \{a, b, c\}$, then ε, a, ab, aac, and $bbac$ are strings over Σ of lengths zero, one, two, three and four respectively.
- If $\Sigma = \{a\}$, then Σ^* contains ε, a, aa, aaa, $aaaa$, etc.
- If $\Sigma = \emptyset$ (the empty set), then what is Σ^*?

Strings over an alphabet

A **string of length** n (for $n = 0, 1, 2, \ldots$) over an alphabet Σ is just an ordered n-tuple of elements of Σ, written without punctuation.

Σ^* denotes set of all strings over Σ of any finite length.

Examples:

- If $\Sigma = \{a, b, c\}$, then ε, a, ab, aac, and $bbac$ are strings over Σ of lengths zero, one, two, three and four respectively.
- If $\Sigma = \{a\}$, then Σ^* contains ε, a, aa, aaa, $aaaa$, etc.
- If $\Sigma = \emptyset$ (the empty set), then $\Sigma^* = \{\varepsilon\}$.

Concatenation of strings

The **concatenation** of two strings u and v is the string uv obtained by joining the strings end-to-end. This generalises to the concatenation of three or more strings.

Examples:

If $\Sigma = \{a, b, c, \ldots, z\}$ and $u, v, w \in \Sigma^*$ are $u = ab$, $v = ra$ and $w = cad$, then

$$vu = raab$$
$$uu = abab$$
$$wv = cadra$$
$$uvwuv = abracadabra$$

Concatenation of strings

The **concatenation** of two strings u and v is the string uv obtained by joining the strings end-to-end. This generalises to the concatenation of three or more strings.

Examples:

If $\Sigma = \{a, b, c, \ldots, z\}$ and $u, v, w \in \Sigma^*$ are $u = ab$, $v = ra$ and $w = cad$, then

$$vu = raab$$
$$uu = abab$$
$$wv = cadra$$
$$uvwuv = abracadabra$$

N.B. $(uv)w = uvw = u(vw)$ (any u, v, w)

$u\epsilon = u = \epsilon u$

The length of a string $u \in \Sigma^*$ is denoted $|u|$.

Formal languages

An extensional view of what constitutes a formal language is that it is completely determined by the set of 'words in the dictionary':

Given an alphabet Σ, we call any subset of Σ^* a (formal) **language** over the alphabet Σ.

We will use inductive definitions to describe languages in terms of grammatical rules for generating subsets of Σ^*.

Inductive Definitions

Axioms and rules

for inductively defining a subset of a given set U

- **axioms** $$\frac{}{a}$$ are specified by giving an element a of U

- **rules** $$\frac{h_1 \; h_2 \; \cdots \; h_n}{c}$$ are specified by giving a finite subset $\{h_1, h_2, \ldots, h_n\}$ of U (the **hypotheses** of the rule) and an element c of U (the **conclusion** of the rule)

Axioms and rules

for inductively defining a subset of a given set U

- **axioms** $$\frac{}{a}$$ are specified by giving an element a of U

which means that a is in the subset we are defining

- **rules** $$\frac{h_1 \; h_2 \; \cdots \; h_n}{c}$$

are specified by giving a finite subset $\{h_1, h_2, \ldots, h_n\}$ of U (the **hypotheses** of the rule) and an element c of U (the **conclusion** of the rule)

which means that c is in the subset we are defining if all of $h_1, h_2, \ldots, h_n$ are

Derivations

Given a set of axioms and rules for inductively defining a subset of a given set U, a **derivation** (or proof) that a particular element $u \in U$ is in the subset is by definition

> a finite rooted tree with vertexes labelled by elements of U and such that:
>
> - the root of the tree is u (the conclusion of the whole derivation),
> - each vertex of the tree is the conclusion of a rule whose hypotheses are the children of the node,
> - each leaf of the tree is an axiom.

usually draw with leaves at top, root at bottom

Example

$U = \{a, b\}^*$

axiom: $\dfrac{}{\varepsilon}$

rules: $\dfrac{u}{aub}$ $\quad \dfrac{u}{bua}$ $\quad \dfrac{u \quad v}{uv}$ (for all $u, v \in U$)

Example derivations:

$$\dfrac{\dfrac{\varepsilon}{ab} \qquad \dfrac{\dfrac{\varepsilon}{ab}}{aabb}}{abaabb} \qquad\qquad \dfrac{\dfrac{\dfrac{\varepsilon}{ba} \qquad \dfrac{\varepsilon}{ab}}{baab}}{abaabb}$$

Example

$U = \{a, b\}^*$ The universal set from which we are specifying a subset.

axiom: $\dfrac{}{\varepsilon}$

rules: $\dfrac{u}{aub}$ $\qquad \dfrac{u}{bua}$ $\qquad \dfrac{u \quad v}{uv}$ (for all $u, v \in U$)

Example derivations:

$$\dfrac{\dfrac{\varepsilon}{ab} \qquad \dfrac{\dfrac{\varepsilon}{ab}}{aabb}}{abaabb} \qquad\qquad \dfrac{\dfrac{\dfrac{\varepsilon}{ba} \qquad \dfrac{\varepsilon}{ab}}{baab}}{abaabb}$$

Example

$U = \{a,b\}^*$ It is the set of all finite strings containing a's & b's.

axiom: $\dfrac{}{\varepsilon}$

rules: $\dfrac{u}{aub}$ $\quad \dfrac{u}{bua}$ $\quad \dfrac{u \quad v}{uv}$ (for all $u, v \in U$)

Example derivations:

$$\dfrac{\dfrac{\varepsilon}{ab} \qquad \dfrac{\dfrac{\varepsilon}{ab}}{aabb}}{abaabb} \qquad\qquad \dfrac{\dfrac{\dfrac{\varepsilon}{ba} \qquad \dfrac{\varepsilon}{ab}}{baab}}{abaabb}$$

Example

$U = \{a, b\}^*$

Now the axioms and rules to define the subset :

axiom: $\dfrac{}{\varepsilon}$

rules: $\dfrac{u}{aub}$ $\qquad \dfrac{u}{bua}$ $\qquad \dfrac{u \quad v}{uv}$ $\qquad$ (for all $u, v \in U$)

Example derivations:

$$\dfrac{\dfrac{\varepsilon}{ab} \qquad \dfrac{\dfrac{\varepsilon}{ab}}{aabb}}{abaabb} \qquad\qquad \dfrac{\dfrac{\dfrac{\varepsilon}{ba} \qquad \dfrac{\varepsilon}{ab}}{baab}}{abaabb}$$

Inductively defined subsets

Given a set of axioms and rules over a set U, the subset of U **inductively defined** by the axioms and rules consists of all and only the elements $u \in U$ for which there is a derivation with conclusion u.

For example, for the axioms and rules on Slide 13

- $abaabb$ is in the subset they inductively define (as witnessed by either derivation on that slide)
- $abaab$ is not in that subset (there is no derivation with that conclusion – why?)

(In fact $u \in \{a,b\}^*$ is in the subset iff it contains the same number of a and b symbols.)

rules or templates?

$$\frac{u \qquad v}{uv} \qquad \text{(for all } u, v \in U\text{)}$$

is really a template for a (potentially) infinite set of rules

Example: transitive closure

Given a binary relation $R \subseteq X \times X$ on a set X, its **transitive closure** R^+ is the smallest (for subset inclusion) binary relation on X which contains R and which is **transitive** ($\forall x, y, z \in X.\ (x, y) \in R^+ \ \&\ (y, z) \in R^+ \Rightarrow (x, z) \in R^+$).

R^+ is equal to the subset of $X \times X$ inductively defined by

axioms $\dfrac{}{(x, y)}$ (for all $(x, y) \in R$)

rules $\dfrac{(x, y) \qquad (y, z)}{(x, z)}$ (for all $x, y, z \in X$)

Example: reflexive-transitive closure

Given a binary relation $R \subseteq X \times X$ on a set X, its **reflexive-transitive closure** R^* is defined to be the smallest binary relation on X which contains R, is both transitive and **reflexive** ($\forall x \in X.\ (x, x) \in R^*$).

R^* is equal to the subset of $X \times X$ inductively defined by

axioms $\dfrac{}{(x, y)}$ (for all $(x, y) \in R$) $\dfrac{}{(x, x)}$ (for all $x \in X$)

rules $\dfrac{(x, y) \qquad (y, z)}{(x, z)}$ (for all $x, y, z \in X$)

Example: reflexive-transitive closure

Given a binary relation $R \subseteq X \times X$ on a set X, its **reflexive-transitive closure** R^* is defined to be the smallest binary relation on X which contains R, is both transitive and **reflexive** ($\forall x \in X.\ (x, x) \in R^*$).

R^* is equal to the subset of $X \times X$ inductively defined by

axioms $\dfrac{}{(x, y)}$ (for all $(x, y) \in R$) $\qquad \dfrac{}{(x, x)}$ (for all $x \in X$)

rules $\dfrac{(x, y) \qquad (y, z)}{(x, z)}$ (for all $x, y, z \in X$)

we can use Rule Induction to prove this

Example: reflexive-transitive closure

Given a binary relation $R \subseteq X \times X$ on a set X, its **reflexive-transitive closure** R^* is defined to be the smallest binary relation on X which contains R, is both transitive and **reflexive** ($\forall x \in X.\ (x, x) \in R^*$).

R^* is equal to the subset of $X \times X$ inductively defined by

axioms $\dfrac{}{(x, y)}$ (for all $(x, y) \in R$) $\dfrac{}{(x, x)}$ (for all $x \in X$)

rules $\dfrac{(x, y) \qquad (y, z)}{(x, z)}$ (for all $x, y, z \in X$)

we can use Rule Induction to prove this, since $S \subseteq X \times X$ being closed under the axioms & rules is the same as it containing R, being reflexive and being transitive.

Inductively defined subsets

Given a set of axioms and rules over a set U, the subset of U **inductively defined** by the axioms and rules consists of all and only the elements $u \in U$ for which there is a derivation with conclusion u.

Derivation is a finite (labelled) tree with u at root, axiom at leaves and each vertex the conclusion of a rule whose hypotheses are the children of the vertex.

(We usually draw the trees with the root at the bottom.)

Rule Induction

Theorem. The subset $I \subseteq U$ inductively defined by a collection of axioms and rules is closed under them and is the least such subset: if $S \subseteq U$ is also closed under the axioms and rules, then $I \subseteq S$.

Given axioms and rules for inductively defining a subset of a set U, we say that a subset $S \subseteq U$ is **closed under the axioms and rules** if

- for every axiom $\overline{\;a\;}$, it is the case that $a \in S$
- for every rule $\dfrac{h_1 \; h_2 \; \cdots \; h_n}{c}$, if $h_1, h_2, \ldots, h_n \in S$, then $c \in S$.

E.G. for the axiom & rules

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u\ v}{uv} \quad \text{for all } u,v \in \{a,b\}^*$$

the subset

$$\{u \in \{a,b\}^* \mid \#_a(u) = \#_b(u)\}$$

(where $\#_a(u)$ is the number of 'a's in the string u)

E.G. for the axiom & rules

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u \quad v}{uv} \qquad \text{for all } u,v \in \{a,b\}^*$$

the subset

$$\{u \in \{a,b\}^* \mid \#_a(u) = \#_b(u)\}$$

is closed under the axiom & rules.

N.B. for a given set $\mathcal{R}$ of axioms & rules

$$\{u \in U \mid \forall S \subseteq U.(S \text{ closed under } \mathcal{R}) \implies u \in S\}$$

is closed under $\mathcal{R}$ (Why?) and so is the smallest such (with respect to subset inclusion, $\subseteq$)

N.B. for a given set $\mathcal{R}$ of axioms & rules

$$\{u \in U \mid \forall S \subseteq U.(S \text{ closed under } \mathcal{R}) \Longrightarrow u \in S\}$$

is closed under $\mathcal{R}$ (Why?) and so is the smallest such (with respect to subset inclusion, $\subseteq$)

This set contains all items that are in every set that is closed under $\mathcal{R}$

Theorem. The subset $I \subseteq U$ inductively defined by a collection of axioms and rules is closed under them and is the least such subset: if $S \subseteq U$ is also closed under the axioms and rules, then $I \subseteq S$.

"the least subset closed under the axioms & rules"

is sometimes take as the definition of

"inductively defined subset"

Proof of the Theorem [Page 23 of notes]

Closure part

- I is closed under each axiom $\dfrac{}{a}$

  Because we can construct a derivation witnessing $a \in I$ ...

  ... which is simply a tree with one node containing a

Closure part (2)

- I is closed under each rule $r = \dfrac{h_1\ h_2\ \ldots h_n}{a}$

  Because if $h_1\ h_2 \ldots h_n \in I$ ...

  we have n derivations from axioms to each h_i and so ...

  we can just make these the n children to our rule r to form a big tree ...

  which is a derivation witnessing $c \in I$

Proof of the Theorem

so we have closure under rules & axioms

Now the "least such subset" part

We need to show, for every $S \subseteq U$

$$(S \text{ closed under axioms and rules }) \Rightarrow \; I \subseteq S$$

That is, I is the least subset, in that any other subset that is closed under the axioms & rules contains I.

Least Subset

So we need to show that every element of I is contained in any set $S \subseteq U$ which is closed under the rules & axioms

Q: How can we characterise an element of I?
A: For each element of I there is a derivation that witnesses its membership

So let's do induction on the height of the derivation (i.e. the height of the tree)

Least Subset - Proof By Induction

$P(n) \triangleq$ "all derivations of height n have their conclusion in S"

Need to show:

- $P(0)$ (consider these to be single (axiom) node derivations)
- $\forall (k \leq n)\ P(k) \Rightarrow P(n+1)$

since if $P(n)$ is true for all n, then all derivations have their conclusion in S, and thus every element of I is in S.

Least Subset – Proof By Induction

$P(n) \triangleq$ "all derivations of height n have their conclusion in S"

- $P(0)$:
  trivially true since conclusion is an axiom and S is closed under axioms

Least Subset - Proof By Induction

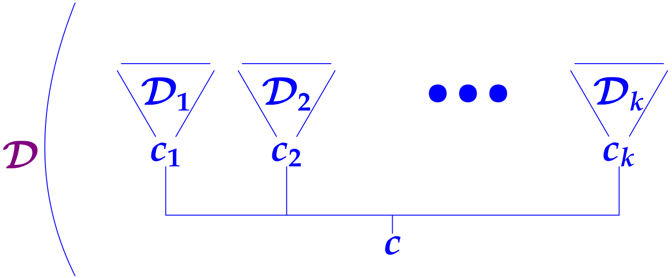
$P(n) \triangleq$ "all derivations of height n have their conclusion in S"

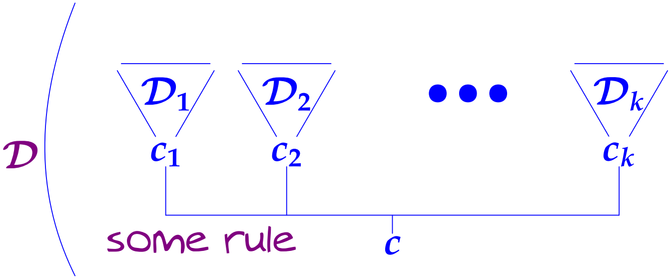
- $P(0)$:
  trivially true since conclusion is an axiom and S is closed under axioms
- $\forall (k \leq n)\ P(k) \Rightarrow P(n+1)$:

Least Subset - Proof By Induction

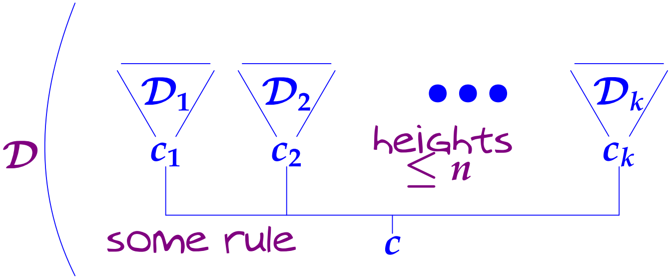
$P(n) \triangleq$ "all derivations of height n have their conclusion in S"

- $P(0)$:
  trivially true since conclusion is an axiom and S is closed under axioms
- $\forall(k \leq n)\ P(k) \Rightarrow P(n+1)$:
  Suppose $\forall(k \leq n)\ P(k)$ and that $\mathcal{D}$ is a derivation of height $n+1$ with, say, conclusion c


$\mathcal{D}$
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_k$
c_1
c_2
c_k
c



c is the result of applying some rule to a set of conclusions $c_1\ c_2 \ldots c_k$



But the derivations for the c_i all have height $\leq n$. So the c_i are all in S by assumption

and since S is closed under all axioms & rules, $c \in S$

so $\forall(k \leq n)\ P(k) \Rightarrow P(n+1)$

Thus every element in I is in any S that is closed under the axioms & rules that inductively defined I.

Thus I is the least subset that is closed under those axioms & rules.

Rule Induction

Theorem. The subset $I \subseteq U$ inductively defined by a collection of axioms and rules is closed under them and is the least such subset: if $S \subseteq U$ is also closed under the axioms and rules, then $I \subseteq S$.

We use the theorem as method of proof: given a property $P(u)$ of elements of U, to prove $\forall u \in I.\, P(u)$ it suffices to show

- **base cases:** $P(a)$ holds for each axiom $\dfrac{}{a}$
- **induction steps:** $P(h_1)\ \&\ P(h_2)\ \&\ \cdots\ \&\ P(h_n) \Rightarrow P(c)$ holds for each rule $\dfrac{h_1\ h_2\ \cdots\ h_n}{c}$

Example using rule induction

Let I be the subset of $\{a, b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u \quad v}{uv}$$

Example using rule induction

Let I be the subset of $\{a,b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u \ v}{uv}$$

Associated Rule Induction:

Example using rule induction

Let I be the subset of $\{a, b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u \quad v}{uv}$$

Associated Rule Induction:

- $P(\epsilon)$

Example using rule induction

Let I be the subset of $\{a,b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u\ v}{uv}$$

Associated Rule Induction:

- $P(\epsilon)$
- $\forall u \in I \,.\, P(u) \Rightarrow P(aub)$

Example using rule induction

Let I be the subset of $\{a, b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u \; v}{uv}$$

Associated Rule Induction:

- $P(\epsilon)$
- $\forall u \in I \,.\, P(u) \Rightarrow P(aub)$
- $\forall u \in I \,.\, P(u) \Rightarrow P(bua)$

Example using rule induction

Let I be the subset of $\{a,b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

$$\frac{}{\epsilon} \qquad \frac{u}{aub} \qquad \frac{u}{bua} \qquad \frac{u \quad v}{uv}$$

Associated Rule Induction:

- $P(\epsilon)$
- $\forall u \in I \,.\, P(u) \Rightarrow P(aub)$
- $\forall u \in I \,.\, P(u) \Rightarrow P(bua)$
- $\forall u, v \in I \,.\, P(u) \land P(v) \Rightarrow P(uv)$

Example using rule induction

Let I be the subset of $\{a,b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

For $u \in \{a,b\}^*$, let $P(u)$ be the property

u contains the same number of a and b symbols

We can prove $\forall u \in I.\ P(u)$ by rule induction:

- **base case:** $P(\varepsilon)$ is true (the number of as and bs is zero!)

Example using rule induction

Let I be the subset of $\{a, b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

For $u \in \{a, b\}^*$, let $P(u)$ be the property

u contains the same number of a and b symbols

We can prove $\forall u \in I.\, P(u)$ by rule induction:

- **base case:** $P(\varepsilon)$ is true (the number of as and bs is zero!)
- **induction steps:** if $P(u)$ and $P(v)$ hold, then clearly so do $P(aub)$, $P(bua)$ and $P(uv)$.

Example using rule induction

Let I be the subset of $\{a, b\}^*$ inductively defined by the axioms and rules on Slide 17 of the notes.

For $u \in \{a, b\}^*$, let $P(u)$ be the property

u contains the same number of a and b symbols

We can prove $\forall u \in I.\, P(u)$ by rule induction:

- **base case:** $P(\varepsilon)$ is true (the number of as and bs is zero!)
- **induction steps:** if $P(u)$ and $P(v)$ hold, then clearly so do $P(aub)$, $P(bua)$ and $P(uv)$.

(It's not so easy to show $\forall u \in \{a, b\}^*.\, P(u) \Rightarrow u \in I$ – rule induction for I is not much help for that.)