Progetto di Ricerca GNCS 2019

Studio di proprietà combinatoriche di linguaggi formali ispirate dalla biologia e da strutture bidimensionali

Responsabile: Maria Madonia - UNICT

Partecipanti strutturati:

E. Barcucci, A. Bernini, L. Ferrari, A. Frosini, E. Pergola
UNIFI

S. Rinaldi- UNISI

D. Giammarresi – UNIROMA2

M. Anselmo, R. Zizza, R. Zaccagnino - UNISA

Partecipanti non strutturati:

R. Pinzani - UNIFI

G. Cerbai - UNIFI

R. Zaccagnino - UNISA

Le tematiche del progetto si inquadrano nell'ambito della teoria dei linguaggi formali

Studio di modelli di evoluzione del genoma

DNA computing

Analisi di particolari cammini discreti nel piano

Generalizzazione alle due dimensioni del concetto (unidimensionale) di codice di stringhe

Two-dimensional codes

Marcella Anselmo
Salerno

Dora Giammarresi
Roma Tor Vergata

Maria Madonia
Catania

Convegno 2020
GNCS-INDAM

11-13 febbraio 2020 - MONTECATINI

1dim versus 2dim

Strings

s =

b	b	a	b	a	b	a

- Σ^* the set of all strings over Σ
- $L \subseteq \Sigma^*$ is a language over Σ

- length(s) =# of symbols

Pictures

p =

a	b	a	a	a	b	b
b	b	a	b	a	b	a
a	b	a	b	b	b	a
a	b	a	b	a	b	a

- Σ^{**} the set of all pictures over Σ
- $L \subseteq \Sigma^{**}$ is a 2D-language over Σ

- size(p) =(# of rows, # of cols)

Strings and 1dim-languages

- **Problems and Results**
  - **theoretical** : recognizability , automata, grammars, codes, combinatorics on words, periodicity, avoiding regularities (squares, palindroms, overlaps)….
  - **algorithmic** : string matching, encoding, compression, code syncronization….

Pictures and 2dim-languages

- **General Main Goal**
  - Do all the same things by exploiting intrinsic two-dimensional properties …
- **General Main Troubles**
  - …deleting a picture from a bigger picture does not give a picture! (not a rectangle)
  - …⊕, ⊖ concatenations are partial operations!

Pictures and 2dim-languages

- **Finite Automata** (4way, OTA, TS, Sgraffito, Restarting)

  Blum Hewitt (1967), Inoue, Takanami (1993), G, Restivo (1997), Prusa, Frantisek, Otto (1999-2017), Anselmo, G., Madonia (2007-19)

- **Grammars** (matrix-, image-, array-, TRG- grammars)

  Cherubini, Crespi-Reghizzi, Lonati, Pradella, San Pietro (2003-11), Bozapalidis (2006), Anselmo, G., Madonia (2013-17)

- **Picture Codes** (undecidability, prefix- finite-delay codes)

  Beauquier, Nivat (2003), Moczurad, Moczurad (2004), Bozapalidis (2006), Anselmo, G., Madonia (2013-19)

- **Combinatorial properties** (periods, repetitions, overlaps)

  Amir, Benson (1992-98), Gamard, Richomme (2013-17), Barcucci, Bernini, Bilotta, Pinzani (2015-17), Anselmo, G., Madonia (2015-19)

- **Pattern matching**

  Crochemore, Iliopoulos, Galil, Giancarlo, Park ... (1992-2017)

- **Compression**

  Lempel, Ziv (1986), Brittain, El-Sakka (2005)

One-dimensional world & codes

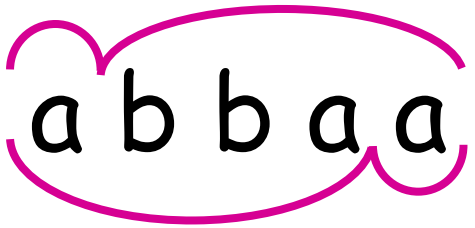
- **1d**: Uniquely decipherable codes of strings

well-established theory

$S \subseteq \Sigma^*$ is a **code** iff any $w \in \Sigma^*$ has at most **one** decomposition with strings of S

Examples: $S = \{a, abb\}$ is a code

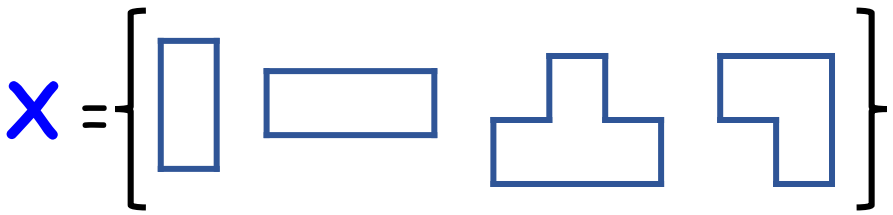
$S_1 = \{a, abba, bbaa\}$ is not a code

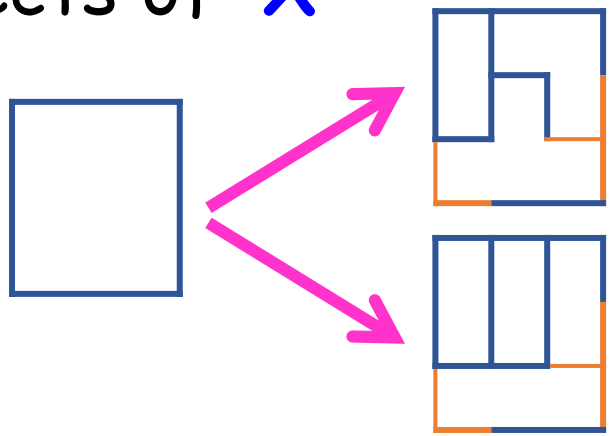


Two-dimensional world & codes

- **2d**: The notion of **code** is related to the problem of decomposing\tiling:

A set **X** of two-dimensional objects is a **code** iff any two-dimensional object has at most **one** decomposition\tiling with objects of **X**

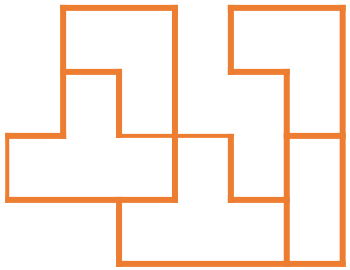




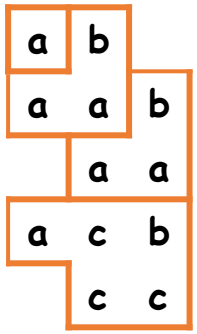
X is **not** a code.

Two-dimensional codes: previous works

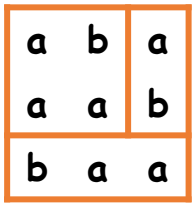
- Beauquier&Nivat (2003)



- M.Moczurad&W.Moczurad (2004)



- Bozapalidis&Grammaticopoulou (2006)



Unfortunately: always **undecidability** results already for finite sets

A different approach

First: the two-dimensional objects

- **Pictures**: rectangular arrays of symbols taken from a finite alphabet

  Σ the finite alphabet

  Σ^{**} the set of all pictures over Σ

Second: the operation

- **Tiling star**

Tiling star of X

- In **1d**

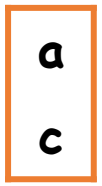
The **Kleene star** of a set **S**, denoted **S***, is the set of strings obtained by concatenating strings of **S**

- In **2d**

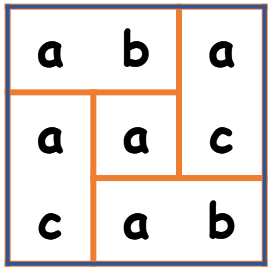
Definition [Simplot'91]: The **tiling star** of **X**, denoted **X****, is the set of pictures obtained by composing pictures of **X** in a way to cover a rectangular area.

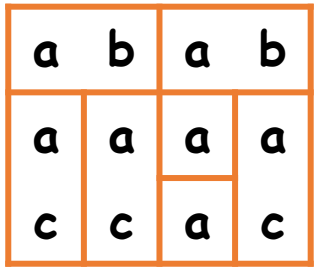
Tiling star of X (ctd.)

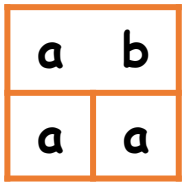
Example: Let X = { 



}







$\in X^{**}$

Definition: If a picture $p \in X^{**}$ then p is tilable in X and the way to obtain p by composing pictures of X is a tiling decomposition of p on X

Two-dimensional codes

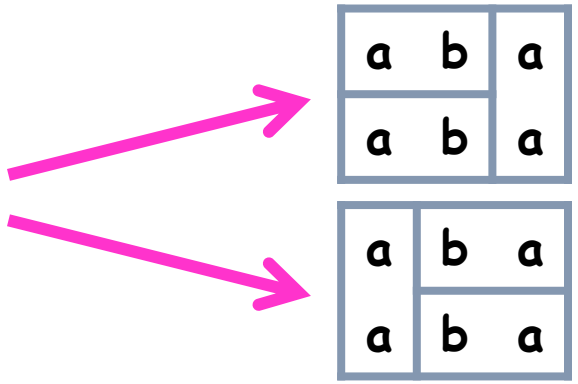
Definition [AGM'13]: Let X be a set of pictures.
X is a code iff any $p \in \Sigma^{**}$ has at most one tiling decomposition on X

Examples: Let $X_1 = \left\{ \begin{array}{|c|c|}\hline a & b \\ \hline \end{array} \quad \begin{array}{|c|}\hline a \\ \hline b \\ \hline \end{array} \quad \begin{array}{|c|c|}\hline a & a \\ \hline a & a \\ \hline \end{array} \right\}$ X_1 is a code.

Let $X_2 = \left\{ \begin{array}{|c|c|}\hline a & b \\ \hline \end{array} \quad \begin{array}{|c|c|}\hline b & a \\ \hline \end{array} \quad \begin{array}{|c|}\hline a \\ \hline a \\ \hline \end{array} \right\}$ X_2 is not a code:

consider $p =$

a	b	a
a	b	a



Undecidability

Proposition [AGM'13]: Let $X \subseteq \Sigma^{**}$.

It is **undecidable** whether X is a code.

Also a good news

a **decidable** (sub-)family of codes:
the **strong prefix codes**!

One-dimensional prefix set

- In **1d**, starting from the classical definition of string prefix of another string

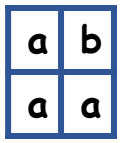
A set of strings **S** is **prefix** if **no** string in **S** is prefix of another string in **S**

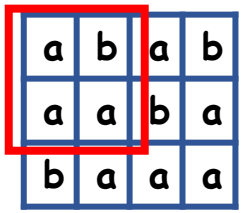
Example: **S** = {abaa, aba, bb} is not prefix. a b a a

Fact: **S** prefix implies **S** code

Two-dimensional prefix set: first attempt

Definition: A picture x is prefix of a picture p if x corresponds to the top-left portion of p.

x = 

is a prefix of p = 

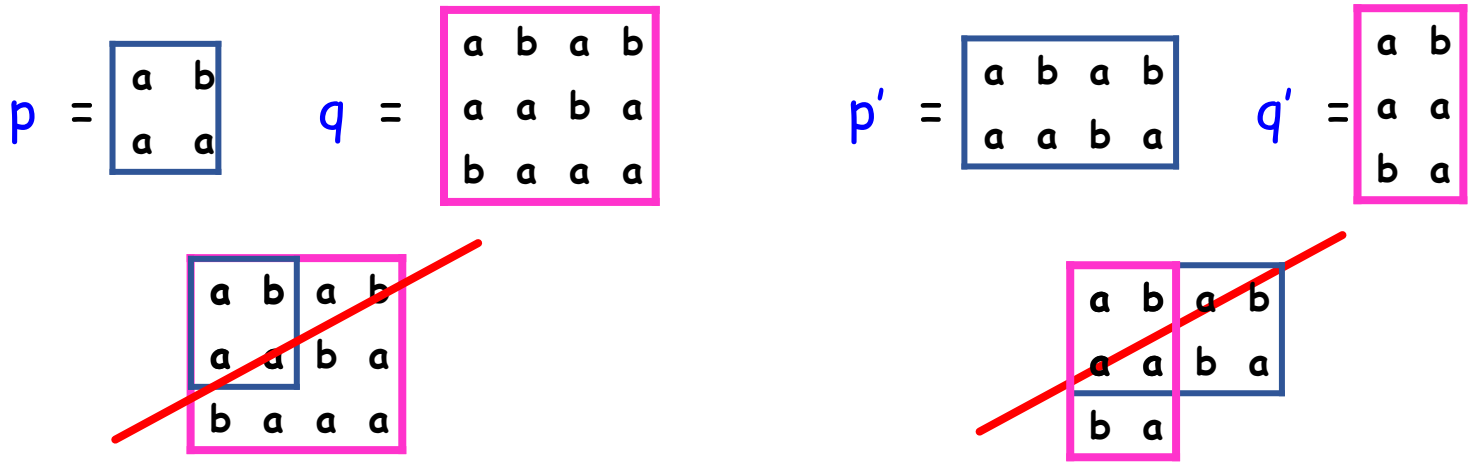
Trivial generalization: Prefix set of pictures

BUT … X prefix does not imply X code

Strong prefix set: definition

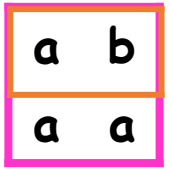
Definition: Let **X** be a set of pictures. **X** is **strong prefix** if

- no picture in **X** is prefix of another picture in **X**
- for any p, q ∈ **X**, p and q cannot be "overlapped" starting from their top-left corner

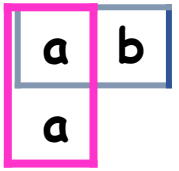


Strong prefix sets: examples

Examples: $X_1 = \{$ a b , a b / a a $\}$ is not strong prefix



$X_2 = \{$ a b , a / a $\}$ X_2 is not strong prefix



$X_3 = \{$ a b a , a b b , b / b , a a / a a , a a / a b , b a / a a , b a / a b , b b / a a , b b / a b $\}$

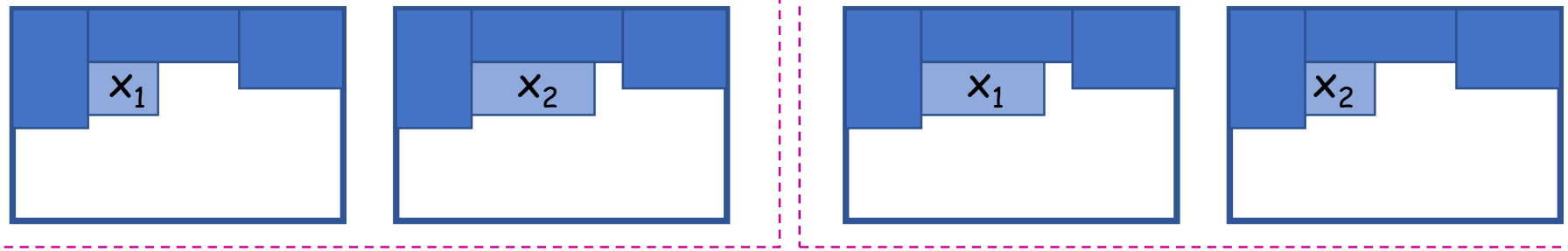
is strong prefix

Strong prefix sets are codes

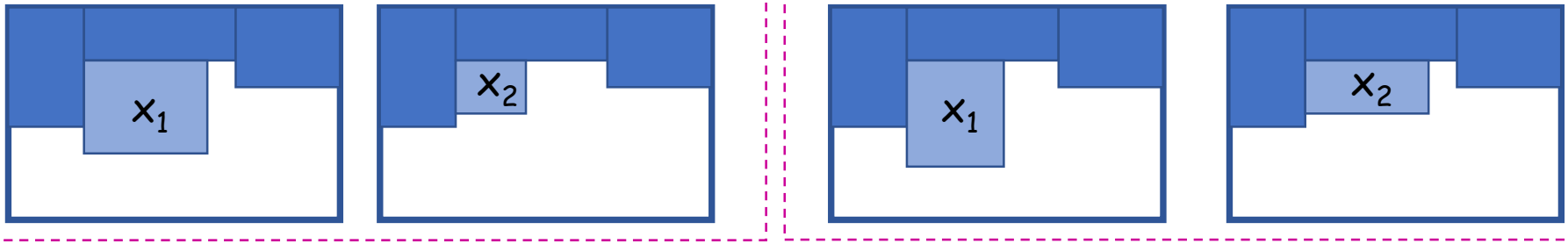
Proposition: Let X be a set of pictures. If X is strong prefix then X is a code

Idea of proof: by contradiction

First case : $|x_1|_{row} = |x_2|_{row}$



Second case : $|x_1|_{row} > |x_2|_{row}$



Maximality

Definition: A strong prefix code $X \subseteq \Sigma^{**}$ is **maximal strong prefix** over Σ if it is **not** properly contained in any other strong prefix code over Σ.

Decidability of maximality

Proposition: It is **decidable** whether a **finite** **strong prefix** code of pictures **X** is **maximal** **strong prefix**

Sketch of proof:

The test whether a new picture can be added to **X** (and **X** remains strong prefix) can be limited to a finite number of "small" pictures. A "big" picture can/cannot be added iff some of its prefixes can/cannot.

Let $\max \{|x|_{row} \text{ with } x \in \mathbf{X}\} = \mathbf{r_X}$ and $\max \{|x|_{col} \text{ with } x \in \mathbf{X}\} = \mathbf{c_X}$.

- **Test** all pictures p with $|p|_{row} \leq \mathbf{r_X}$ and $|p|_{col} \leq \mathbf{c_X}$.

  If no such picture can be added, then no picture with $|p|_{row} > \mathbf{r_X}$ and $|p|_{col} > \mathbf{c_X}$ can be added.

Embedding

Proposition: Let **X** be a finite strong prefix code of pictures. It is possible to construct a finite **Y** **maximal** strong prefix containing **X**.

Idea of proof. **Y** can be incrementally obtained, starting from **X** and adding "small" pictures that do not overlap any picture of the current **Y**

Remark: The construction can output different sets depending on the order in which it processes the candidate picture to be added

Embedding: an example

Example Let $X = \{ aba,\ abb,\ \begin{matrix} b \\ b \end{matrix} \}$

Different sets can be obtained by the construction

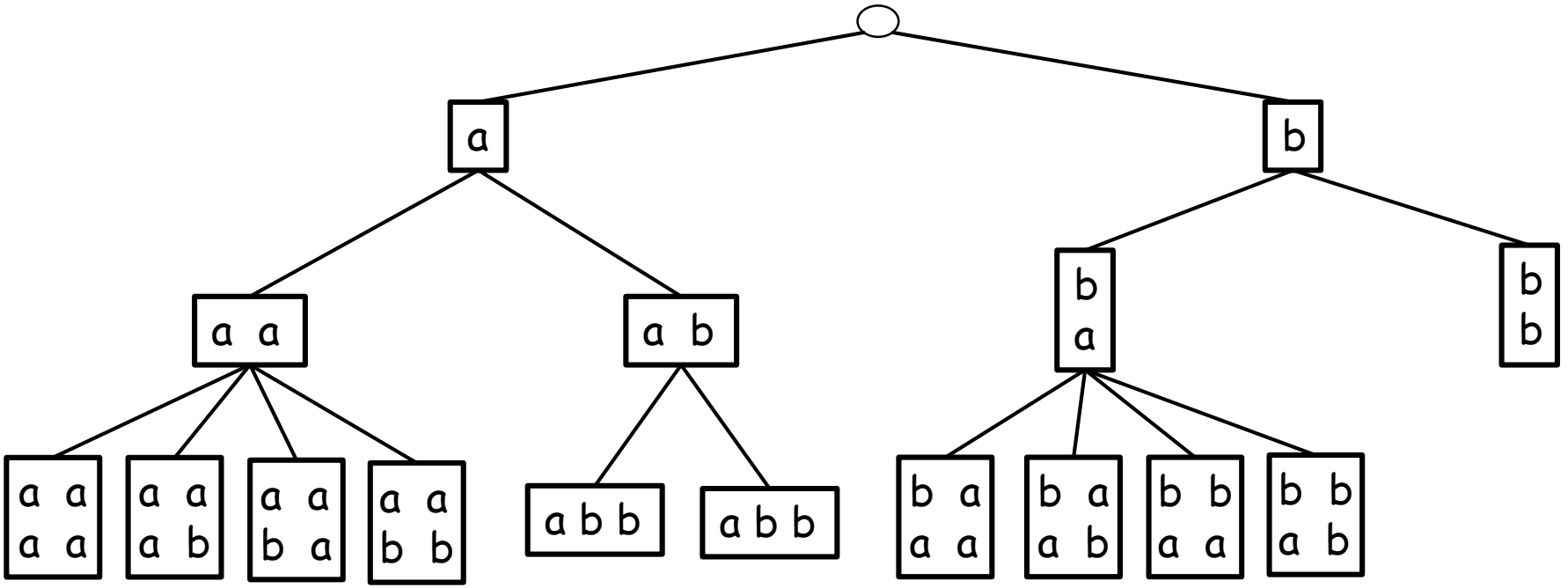
$$Y = \{ aba,\ abb,\ aa,\ \begin{matrix} b \\ b \end{matrix},\ \begin{matrix} b \\ a \end{matrix} \}$$

$$Y' = \{ aba,\ abb,\ \begin{matrix} b \\ b \end{matrix},\ \begin{matrix} a & a \\ a & a \end{matrix},\ \begin{matrix} a & a \\ a & b \end{matrix},\ \begin{matrix} a & a \\ b & b \end{matrix},\ \begin{matrix} a & a \\ b & a \end{matrix},\ \begin{matrix} b & a \\ a & a \end{matrix},\ \begin{matrix} b & a \\ a & b \end{matrix},\ \begin{matrix} b & b \\ a & a \end{matrix},\ \begin{matrix} b & b \\ a & b \end{matrix} \}$$

Example of a finite MSP

X = { a b a , a b b , b b , a a a a , a a a b , a a b a , a a b b , b a a a , b a a b , b b a a , b b a b }

X is a finite maximal strong prefix code.



Measure of 1d languages and codes

Definition: Let Σ be an alphabet and π be a probability distribution on Σ.

The probability of a string $w=a_1a_2\ldots a_n \in \Sigma^*$ is defined as

$$\pi(w) = \prod_{1 \le i \le n} \pi(a_i)$$

The measure of a language $S \subseteq \Sigma^*$ relative to π is defined as

$$\mu_\pi(S) = \sum_{w \in S} \pi(w)$$

Theorem: Let $S \subseteq \Sigma^*$ be a code and π be probability distribution on Σ. Then $\mu_\pi(S) \le 1$.

Moreover, if $S \subseteq \Sigma^*$ is a finite code, then

$\mu_\pi(S) = 1$ iff S is a maximal code

Measure of 2d languages and codes

Definition: Let Σ be an alphabet and π be a probability distribution on Σ.

The probability of a picture $\mathbf{p} \in \Sigma^{**}$ is defined as

$$\pi(\mathbf{p}) = \prod_{\substack{1 \le i \le row(\mathbf{p}) \\ 1 \le j \le col(\mathbf{p})}} \pi(\mathbf{p}(i,j))$$

The measure of a language $\mathbf{X} \subseteq \Sigma^{**}$ relative to π is defined as

$$\mu_\pi(\mathbf{X}) = \sum_{\mathbf{p} \in \mathbf{X}} \pi(\mathbf{p})$$

Proposition: There exist 2d codes of measure greater than 1.

Measure of 2d strong prefix codes

Theorem 1: Let $X \subseteq \Sigma^{**}$ be a finite strong prefix code and π be a probability distribution on Σ. Then $\mu_\pi(X) \leq 1$.

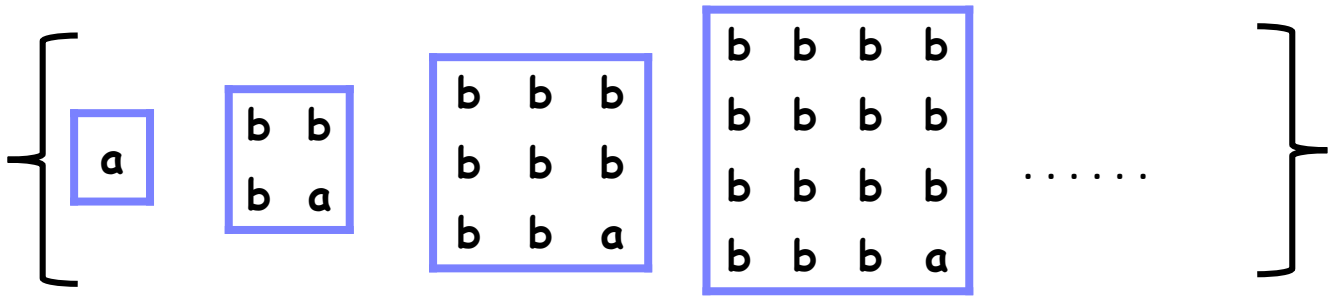
Moreover $\mu_\pi(X) = 1$ iff X is a finite **maximal** strong prefix code.

Remark: the theorem provides a simple algorithm to test whether a finite strong prefix code is a finite **maximal** strong prefix code.

Further results

- Remove the **finiteness** hypothesis
- Study codes related to recognizable picture languages
- Generalization in 2D of circular codes of strings

Examples

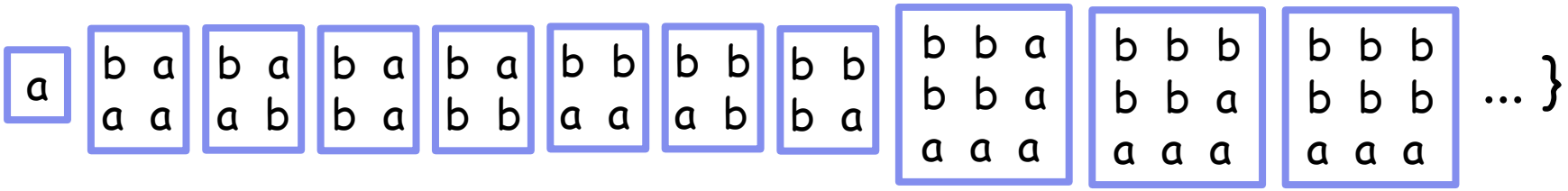
$X =$ 

is an infinite strong prefix code

X = the set of all square pictures, over Σ = {a, b}, that contain symbol b in all positions apart for the bottom right corner where symbol a occurs.

Example

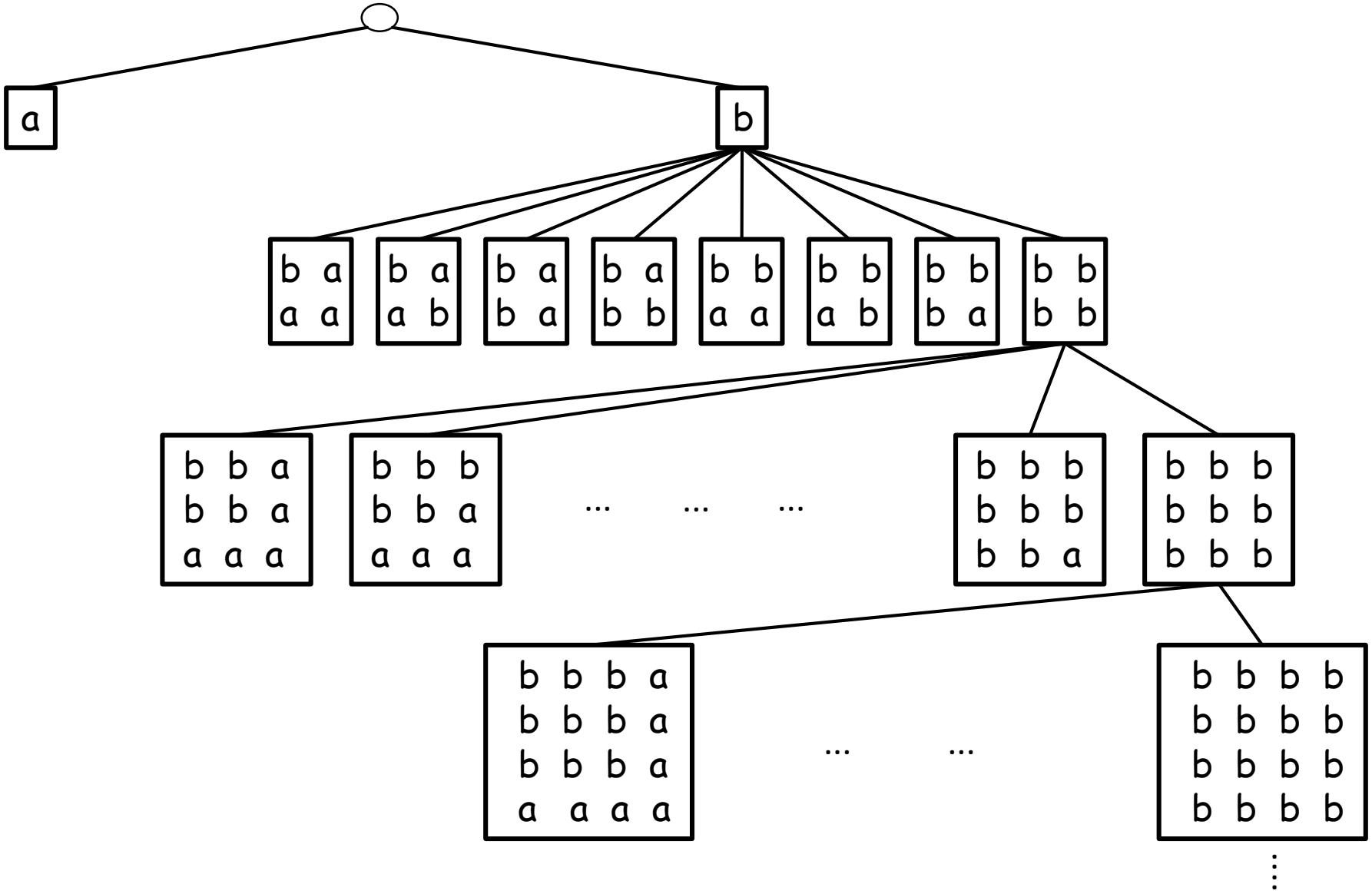
Consider the set X_∞ of square pictures, over $\Sigma = \{a, b\}$, with b **everywhere unless** for the last column or row where there is at least one occurrence of a

$X_\infty = \{$  $\ldots\}$

X_∞ is an **infinite** maximal strong prefix code.

- X_∞ is strong prefix. No picture in X_∞ is prefix of another picture in X_∞
- X_∞ is **maximal** strong prefix. If one adds a picture p, then three cases hold:
  - the top left corner of p is *a*
  - the top left corner of p is *b* and p does not contain an *a*
  - the top left corner of p is *b* and p contains an *a*

A tree for X_∞



Grazie!