[Paper Survey]

Profinite Methods in Automata Theory

by Jean-Éric Pin

(Invited Lecture at STACS 2009)

Presentation:

IPL Rinko, Dec. 7, 2010

by Kazuhiro Inaba

The Topic of the Paper

- Investigation on
  (subclasses of) regular languages

  by using
  - Topological method
  - Especially, “profinite metric”

Why I Read This Paper

- I want to have a different point of view on the "Inverse Regularity Preservation" property of str/tree/graph functions
  - A function
    - $f :: string \rightarrow string$
  - is IRP iff
    - For any regular language L, the inverse image $f^{-1}(L) = \{s \mid f(s) \in L\}$ is regular

(Why I Read This Paper)

Application of IRP

- Typechecking $f :: L_{IN} \rightarrow L_{OUT}$?
  - Verify that a transformation always generates valid outputs from valid inputs.

f	L_{IN}	L_{OUT}
XSLT Template for formating bookmarks	XBEL Schema	XHTML Schema
PHP Script	Arbitrary String	String not containing “<script>”

(Why I Read This Paper)

Application of IRP

- Typechecking $f :: L_{IN} \to L_{OUT}$?
  - If f is IRP, we can check this by …

$$\text{f is type-correct}$$

$$\Leftrightarrow f(L_{IN}) \subseteq L_{OUT}$$

$$\Leftrightarrow L_{IN} \subseteq f^{-1}(L_{OUT})$$

$$\Leftrightarrow L_{IN} \cap \overline{f^{-1}(L_{OUT})} = \Phi$$

with counter-example in the unsafe case

(for experts: f is assumed to be deterministic)

(Why I Read This Paper)

Characterization of IRP

- Which function is IRP?
- We know that MTT* is a strict subclass of IRP (as I have presented half a year ago). But how can we characterize the subclass?
- Is there any systematic method to define subclasses of IRP functions?

The paper [Pin 09] looks to provide an *algebraic/topological viewpoint* on regular languages, which I didn't know.

Agenda

- Metrics
- Profinite Metric
- Completion
- Characterization of
  - Regular Languages
  - Inverse Regularity Preservation
  - Subclasses of Regular Languages by "Profinite Equations"
- Summary

Notation

- I use the following notation
  - Σ = finite set of ‘character’s
  - Σ* = the set of finite words (strings)

- e.g.,
  - Σ = {0,1}
    ➔ Σ* = { ε, 0, 1, 00, 01, 10, … }
  - Σ = {a,b,c,…,z,A,B,C,…,Z}
    ➔ Σ* = { ε, a, b, …, HelloWorld, … }

Metrics

- d :: $S \times S \rightarrow R+$
  - is a metric on a set S, if it satisfies:
    - **$d(x, x) = 0$**
    - **$d(x, y) = d(y, x)$**
    - **$d(x, y) \leqq d(x, z) + d(z, y)$**

(triangle inequality)

Example

- $d_R :: R \times R \rightarrow R+$
- $d_R(a, b) = |a-b|$

- $d_2 :: R^2 \times R^2 \rightarrow R+$
- $d_2((a_x,a_y), (b_x,b_y)) = \surd\ (a_x-b_x)^2 + (a_y-b_y)^2$

- $d_1 :: R^2 \times R^2 \rightarrow R+$
- $d_1((a_x,a_y), (b_x,b_y)) = |a_x-b_x| + |a_y-b_y|$
- $d_\infty :: R^2 \times R^2 \rightarrow R+$
- $d_\infty((a_x,a_y), (b_x,b_y)) = \max(|a_x-b_x|, |a_y-b_y|)$

Metrics on Strings : Example

$$d_{cp}(x, y) = 2^{-cp(x,y)}$$

where

- $cp(x,y) = \infty$ if $x=y$
- $cp(x,y) =$ the length of the common prefix of x and y

- $d_{cp}($ "abcabc", "abcdef" $) = 2^{-3} = 0.125$
- $d_{cp}($ "zzz", "zzz" $) = 2^{-\infty} = 0$

Proof : $d_{cp}(x,y)=2^{-cp(x,y)}$ is a metric

- $d_{cp}(x,x) = 0$
- $d_{cp}(x,y) = d_{cp}(y,x)$
  - By definition.
- $d_{cp}(x,y) \leqq d_{cp}(x,z) + d_{cp}(z,y)$
  - Notice that we have either
    - $cp(x,y) \geqq cp(x,z)$ or $cp(x,y) \geqq cp(z,y)$.
  - Thus
    - $d_{cp}(x,y) \leqq d_{cp}(x,z)$ or $d_{cp}(x,y) \leqq d_{cp}(z,y)$.

∎

Profinite Metric on Strings

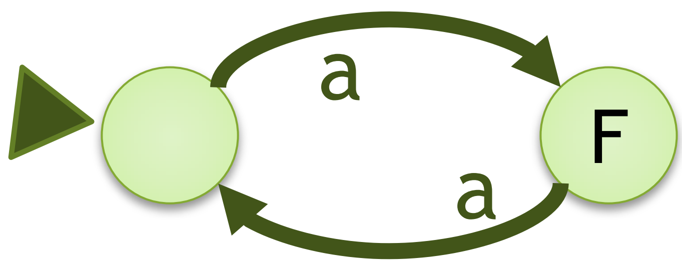
$$d_{mA}(x, y) = 2^{-mA(x,y)}$$

where

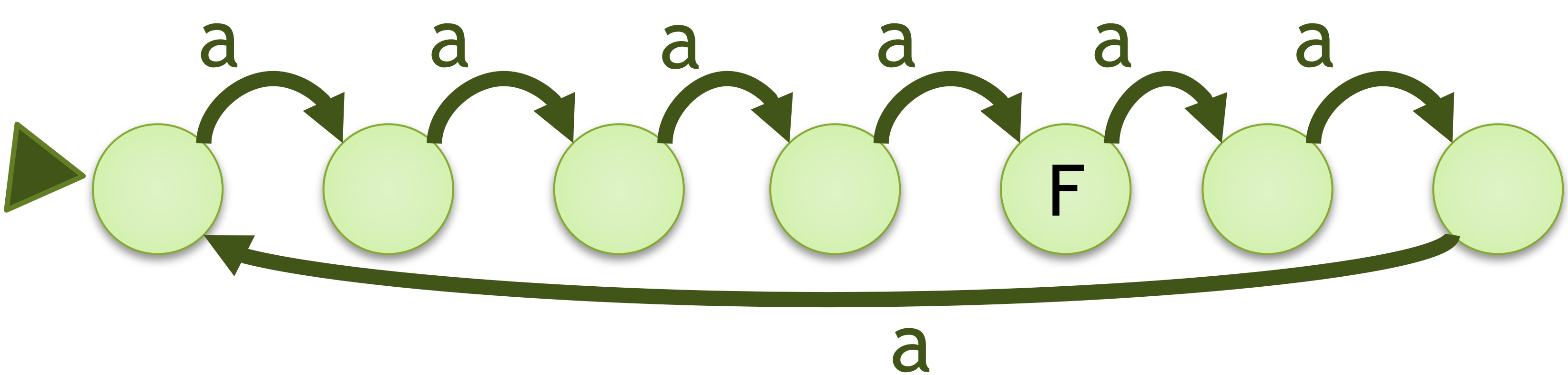
- $mA(x,y) = \infty$ if $x=y$
- $mA(x,y)$ = the size of the minimal DFA (deterministic finite automaton) that distinguishes x and y

Example

- $d_{mA}(\text{“aa”}, \text{“aaa”}) = 2^{-2} = 0.25$
- $d_{mA}(a^{119}, a^{120}) = 2^{-2} = 0.25$

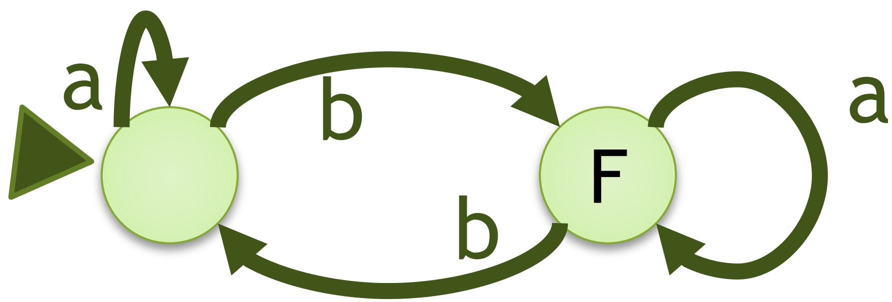


- $d_{mA}(a^{60}, a^{120}) = 2^{-7} = 0.0078125$

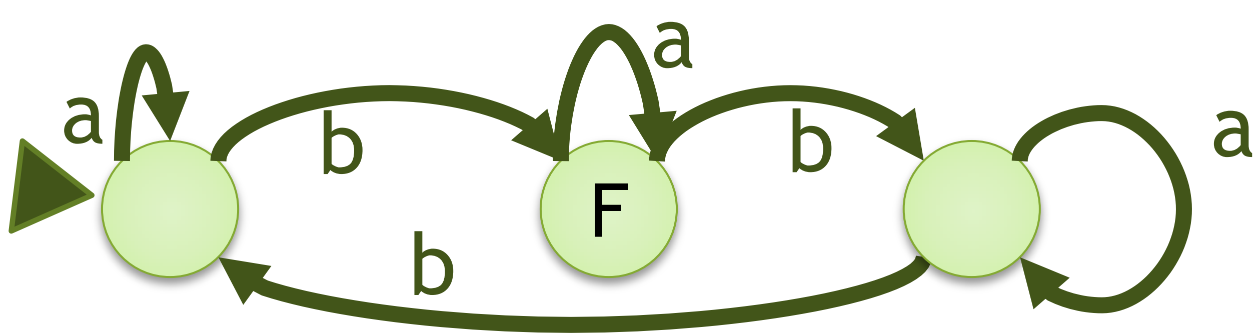


Example

- d_{mA}(“ab”, “abab”) = 2^{-2} = 0.25



- d_{mA}(“abab”, “abababab”) = 2^{-3} = 0.125



Proof : $d_{mA}(x,y)=2^{-mA(x,y)}$ is a metric

- $d_{mA}(x,x) = 0$
- $d_{mA}(x,y) = d_{mA}(y,x)$
  - By definition.
- $d_{mA}(x,y) \leqq d_{mA}(x,z) + d_{mA}(z,y)$
  - Notice that we have either
    - $mA(x,y) \geqq mA(x,z)$ or $mA(x,y) \geqq mA(z,y)$.
  - Thus
    - $d_{mA}(x,y) \leqq d_{mA}(x,z)$ or $d_{mA}(x,y) \leqq d_{mA}(z,y)$.

$\blacksquare$

(Note)

- In the paper another profinite metric is defined, based on the known fact:

  - A set of string L is recognizable by DFA

    if and only if

  - If it is an inverse image of a subset of a finite monoid by a homomorphism

    $L = \psi^{-1}(F)$

    where $\psi :: \Sigma^* \rightarrow M$ is a homomorphism,

    M is a finite monoid, $F \subseteq M$

Completion of Metric Space

- A sequence of elements $x_1, x_2, x_3, \ldots$
  - is Cauchy if

    $\forall \varepsilon > 0, \exists N, \forall i,k > N, d(x_i, x_k) < \varepsilon$
  - is convergent

    $\exists a_\infty, \forall \varepsilon > 0, \exists N, \forall i > N, d(x_i, x_\infty) < \varepsilon$

- Completion of a metric space is the minimum extension of S, whose all Cauchy sequences are convergent.

Example of Completion

- Completion of rational numbers with "normal" distance → Reals
  - Q → R
  - $d_Q(x,y) = |x-y|$ → $d_R(x,y) = |x-y|$

- 1, 1.4, 1.41, 1.41421356, … → √2
- 3, 3.1, 3.14, 3.141592, … → π
- 5, 5, 5, 5, … → 5

Example of Completion

- Completion of finite strings with d_{cp}
  - Σ^*
  - d_{cp} (Common Prefix)

- a, aa, aaa, aaaaaaaa, …
- ab, abab, ababab, …
- zz, zz, zz, zz, …

Example of Completion

- Completion of finite strings with d_{cp}
  ➔ the set of finite and infinite strings
  - Σ^* ➔ Σ^ω
  - d_{cp} (Common Prefix) ➔ d_{cp}

- a, aa, aaa, aaaaaaaa, … ➔ a^ω
- ab, abab, ababab, … ➔ $(ab)^\omega$
- zz, zz, zz, zz, … ➔ zz

Completion of Strings with Profinite Metric

- $d_{mA}(x, y) = 2^{-mA(x,y)}$

- Example of a Cauchy sequence:

  $x_i = w^{i!}$ (for some string **w**)

  w, ww, wwwwww, w^{24}, w^{120}, w^{720}, ...

  (NOTE: w^i is not a Cauchy sequence)

Completion of Strings with Profinite Metric

- Completion of
  - Σ^* with $d_{mA}(x, y) = 2^{-mA(x,y)}$
- yields the set of profinite words $\widehat{\Sigma^*}$

- In the paper, the limit $w^{i!}$ is called

$$x_i = w^{i!} \Rightarrow w^{\omega}$$

with a note:

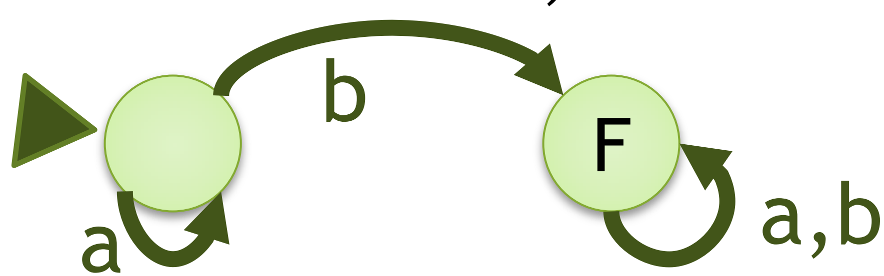
Note that x^{ω} is simply a notation and one should resist the temptation to interpret it as an infinite word.

Difference from Infinite Words

- In the set of infinite words
  - $a^\omega + b = a^\omega$

    (since the length of the common prefix is ω, their distance is 0, hence equal)
- In the set of profinite words
  - $a^\omega + b \neq a^\omega$

    (their distance is 0.25, because of:



p-adic Metric on Q

- Similar concept in the Number Theory

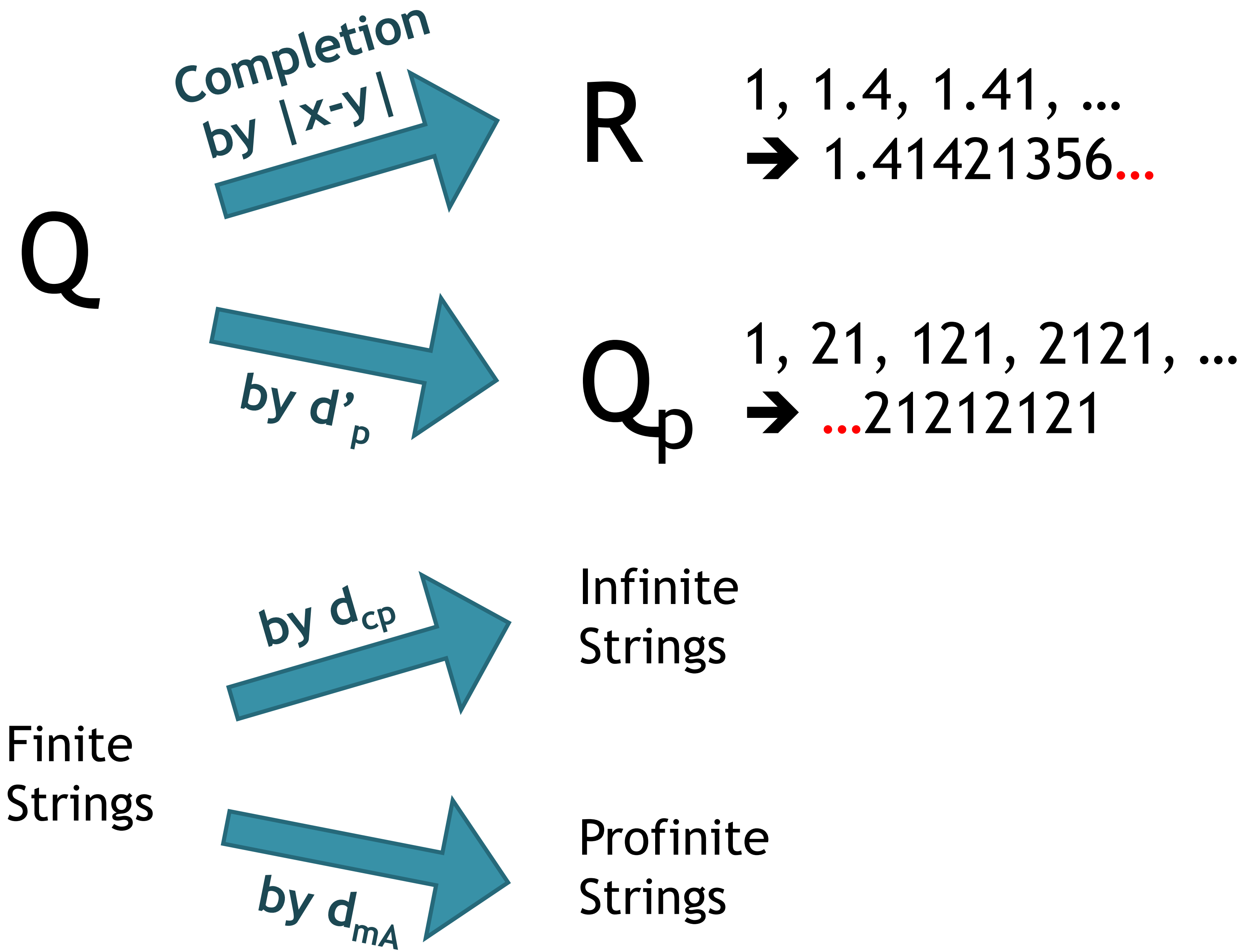
- For each $n \geqq 2$, define **d'_n** as
  - **$d'_n(x,y) = n^{-a}$ if $x-y = b/c\ n^a$**
    - where $a,b,c \in Z$ and b,c is not divisible by n

- When p is a prime, d'_p is called the p-adic metric

Example (p-adic Metric)

- For each $n \geqq 2$, define $\mathbf{d'_n}$ as
  - $\mathbf{d'_n(x,y) = n^{-a}}$ if $\mathbf{x-y = b/c\ n^a}$
    - where $a,b,c \in Z$ and b,c is not divisible by n

- $d'_{10}(12345, 42345) = 10^{-4}$
- $d'_{10}(0.33, 0.43) = 10^{+1}$


Q
Completion by |x-y|
R
1, 1.4, 1.41, …
➔ 1.41421356…
by d'p
Qp
1, 21, 121, 2121, …
➔ …21212121
Finite Strings
by dcp
Infinite Strings
by dmA
Profinite Strings

Theorem [Hunter 1988]

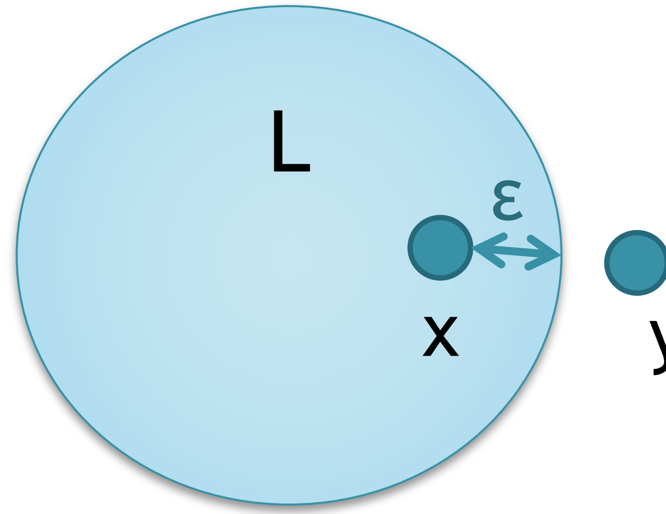
$L \subseteq \Sigma^*$ is regular

if and only if

$cl(L)$ is clopen in $\widehat{\Sigma^*}$

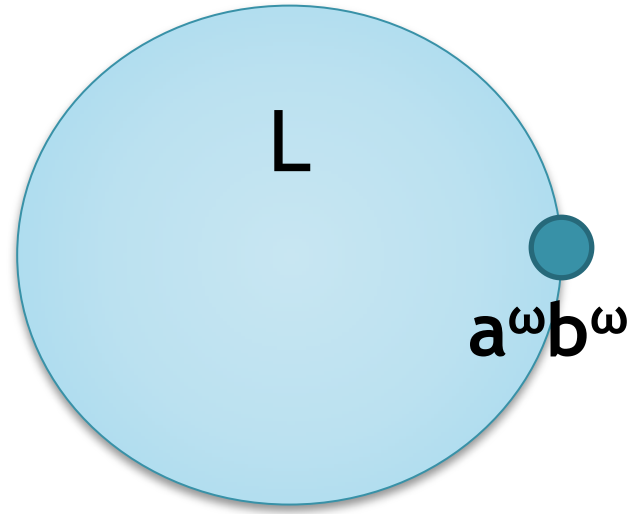
- clopen := closed & open
- closed := complement is open
- S is open := $\forall x \in S, \exists \varepsilon > 0, \{y \mid d(x,y) < \varepsilon\} \subseteq S$
- cl(S) := unique minimum closed set $\supseteq L$

Intuition



- L is regular
  - ⇔
- cl(L) is open
  - ⇔
- $\forall x \in cl(L), \exists \varepsilon, \forall y, d_{mA}(x,y) < \varepsilon \rightarrow y \in cl(L)$
  - ⇔
- If cl(L) contains x, it contains all ‘hard-to-distinguish-from x’ profinite strings

(Non-)example



- $L = \{ a^n b^n \mid n \in nat \}$ is not regular
- Because
  - $\mathbf{a^\omega b^\omega}$ is contained in cl(L)
  - cl(L) do not contain $\mathbf{a^\omega b^{\omega+k!}}$ for each **k**
  - but $\mathbf{d_{mA}(a^\omega b^\omega, a^\omega b^{\omega+k!}) \leqq 2^{-k}}$

Proof Sketch : clopen⇔regular

- L is Regular ⇒ cl(L) is Clopen

  (This direction is less surprising.)
  - It is trivially closed

  - Suppose L is regular but cl(L) is not ppen.
  - Then, $\exists x \in cl(L), \forall \varepsilon, \exists y \notin cl(L), d_{mA}(x,y) < \varepsilon$
  - Then, $\forall n, \exists x \in L, \exists y \notin L, d_{mA}(x,y) < 2^{-n}$
  - Then, ∀size-n DFA, ∃x,y that can't be separated
  - Thus, L is not be a regular language.

(not in the paper: just my thought)

Generalize: Regular ⇒ Clopen

(This direction is less surprising. Why?)

Because it doesn't use any particular property of "regular"

- Let
  - F be a set of predicates string→bool
  - siz be any function F → nat
  - $d_{mF}(x,y) = 2^{-\min\{siz(f) \mid f(x) \neq f(y)\}}$
- L is F-recognizable

  ⇒ cl(L) is clopen with d_{mF}

(not in the paper: just my thought)

Generalize: Regular ⇒ Clopen

(This direction is less surprising. Why?)

Because it doesn't use any particular property of "regular"

- E.g.,
  - $d_{mPA}(x,y) = 2^{-\min\{\#\text{states of PD-NFA separating } x\&y\}}$
- L is context-free

  ⇒ cl(L) is clopen with d_{mPA}

(But this is not at all interesting, because any set is clopen in this metric!!)

Proof Sketch: Clopen ⇒ Regular

- Used lemmas:
  - $\widehat{\Sigma^*}$ is **compact**
    - i.e., if it is covered by an infin union of open sets, then it is covered by their finite subfamily, too.
    - i.e., every infinite seq has convergent subseq
    - The proof relies on the fact: $\mathbf{siz^{-1}(n)}$ is finite
  - Concatenation is **continuous** in this metric
    - i.e., $\forall x\ \forall \varepsilon\ \exists \delta,\ \forall x',\ d(x,x')<\delta \rightarrow d(f(x),f(x'))<\varepsilon$
    - Due to $d_{mA}(wx,wy) \leqq d_{mA}(x,y)$
- *By these lemmas, clopen sets are shown to be covered by finite congruence, and hence regular.* ■

Corollary

$f :: \Sigma^* \rightarrow \Sigma^*$ is IRP

if and only if

$\hat{f} :: \hat{\Sigma}^* \rightarrow \hat{\Sigma}^*$ is continuous

- continuous :=

  $\forall x\ \forall \varepsilon\ \exists \delta,\ \forall x',\ d(x,x')<\delta \rightarrow d(f(x),f(x'))<\varepsilon$
- Known to be equivalent to

  f^{-1}((cl)open) = (cl)open

“Equational Characterization“

- Main interest of the paper
- Many subclasses of regular languages are characterized by
  Equations on Profinite Strings

Example

- A regular language L is **star-free**
  (i.e., in $\{\cup,\cap,\neg,\cdot\}$-closure of fin. langs)
  (or equivalently, **FO-definable**)

  if and only if

  Corollary:
  FO-definability is decidable

- $\mathbf{x^{\omega} \equiv_L x^{\omega+1}}$
  - i.e., $\forall u\ v\ x,\ ux^{\omega}v \in cl(L) \Leftrightarrow ux^{\omega+1}v \in cl(L)$

Example

- A regular language L is **commutative**

  if and only if

> Corollary:
> Commutativity is decidable

- **$xy \equiv_L yx$**
  - i.e., $\forall u\ v\ x,\ u\ xy\ v \in cl(L) \Leftrightarrow u\ yx\ v \in cl(L)$

Example

- A regular language L is **dense**
  ($\forall w,\ \Sigma^* w \Sigma^* \cap L \neq \Phi$)

  if and only if

- **$\{x\rho \equiv_L \rho x \equiv_L \rho,\ \ x \leqq_L \rho\}$**
  - where $\rho = \lim_{n\to\infty} v_n,\ \ v_{n+1}=(v_n\ u_{n+1}\ v_n)^{(n+1)!}$
    $u = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, \ldots\}$
  - i.e., $\forall u\ v\ x, \ldots \ \&\ uxv \in cl(L) \Rightarrow u\rho v \in cl(L)$

Theorem [Reiterman 1982]

- If a family (set of languages) *F* of regular languages is closed under
  - intersection, union, complement,
  - quotient ($q_a(L) = \{x \mid ax \in L\}$), and
  - inverse of homomorphism

if and only if

- It is defined by a set of profinite equations of the form: $u \equiv v$

Other Types of Equations

- [Pin & Gehrke & Grigorieff 2008]

We summarize on a table the various types of equations we have used so far.

Closed under	Equations	Definition
$\cup, \cap$	$u \to v$	$\widehat{\eta}(u) \in \widehat{\eta}(L) \Rightarrow \widehat{\eta}(v) \in \widehat{\eta}(L)$
quotient	$u \leqslant v$	$xvy \to xuy$
complement	$u \leftrightarrow v$	$u \to v$ and $v \to u$
quotient and complement	$u = v$	$xvy \leftrightarrow xuy$
Closed under inverse of morphisms		**Interpretation of variables**
all morphisms		words
nonerasing morphisms		nonempty words
length multiplying morphisms		words of equal length
length preserving morphisms		letters

Summary

- Completion by the Profinite metric

$$d_{mA}(x, y) = 2^{-min\_automaton(x,y)}$$

is used as a tool to characterize (subclasses of) regular languages