Carathéodory Extensions of Subclasses of Regular Languages

Ryoma Sin’ya (Akita University)
DLT 2021 Aug 17.

秋田大学
Akita University

Outline

1. **Background: density and measurability**
2. Carathéodory extensions of local varieties
3. Conclusion

Density of formal languages

The density of a language L over A is defined as

$$\delta_A(L) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#(L \cap A^i)}{\#(A^i)}.$$

Example 1: $\delta_A((AA)^*) = \dfrac{1}{2}$.

Example 2: $\delta_A(aA^*) = \dfrac{1}{\#(A)}$.

Example 3: $L_\perp = \{w \in A^* \mid 3^n \leq |w| < 3^{n+1}$ for some even $n\}$ does **not** have a density.

→ The value $\dfrac{1}{n} \displaystyle\sum_{i=0}^{n-1} \frac{\#(L_\perp \cap A^i)}{\#(A^i)}$ can be larger than 2/3 and smaller than 1/3 *infinitely many times*, hence $\delta_A(L)$ diverges.

Density of formal languages

The density of a language L over A is defined as

$$\delta_A(L) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#(L \cap A^i)}{\#(A^i)}.$$

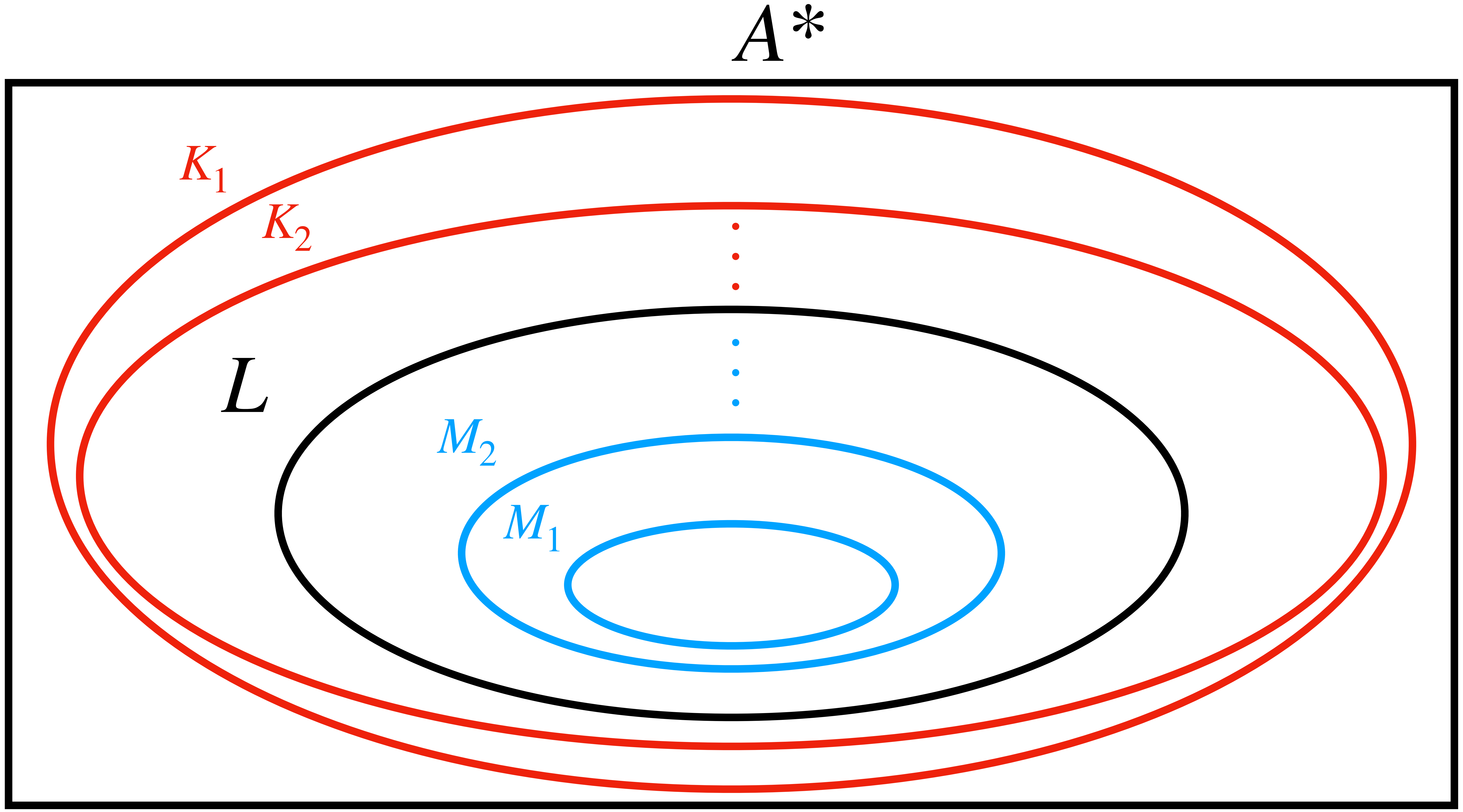
Example 1: $\delta_A((AA)^*) = \dfrac{1}{2}$.

Example 2: $\delta_A(aA^*) = \dfrac{1}{\#(A)}$.

Example 3: $L_\perp = \{w \in A^* \mid 3^n \leq |w| < 3^{n+1} \text{ for some even } n\}$ does **not** have a density.

Theorem (cf. [Berstel 1973]):
Every regular language *do have a rational density*.

$\mathscr{C}$-measurability [SOFSEM 2021]



L is said to be $\mathscr{C}$**-measurable** if there exists an *infinite sequence of pairs of languages* $(M_n, K_n)_{n \in \mathbb{N}}$ in $\mathscr{C}$ such that $M_n \subseteq L \subseteq K_n$ and $\lim_{n \to \infty} \delta_A(K_n \setminus M_n) = 0$.

Example of a regular measurable language

Theorem [SOFSEM2021]:

The semi-Dyck language $\mathrm{D} = \{\varepsilon, ab, aabb, abab, \ldots\}$ over $A = \{a, b\}$ is regular measurable.

Proof: Let $L_k = \{w \in A^* \mid |w|_a = |w|_b \mod k\}$ for each $k \geq 1$.

the # of occurrences of a in w

Then, for each $k \geq 1$, $\mathrm{D} \subseteq L_k$ and $\delta_A(L_k) = \dfrac{1}{k} \to 0$ (if $k \to \infty$) .

Thus the infinite sequence $(\emptyset, L_k)_{k\geq 1}$ converges to D .

Note: there is no regular language L such that $\mathrm{D} \subseteq L$ and $\delta_A(L) = 0$.

Known results [SOFSEM 2021]

All regular measurable languages

The set of all primitive words

Q

Many complex CFLs

There are uncountably many regular measurable languages

$\mathrm{M}_2 = \{w \in \{a,b\}^* \mid |w|_a > 2|w|_b\}$

$L_\perp = \{w \in A^* \mid 3^n \leq |w| < 3^{n+1} \text{ for some even } n\}$

Measurability *à la* Carathéodory

The outer-$\mathscr{C}$-measure (over A) of $L \subseteq A^*$ is defined as

$$\bar{\mu}_{\mathscr{C}}(L) = \inf\{\delta_A(K) \mid L \subseteq K \in \mathscr{C}_A\}$$

where $\mathscr{C}_A$ is the class of all languages in $\mathscr{C}$ over A.

Theorem [S]:
For any language $L \subseteq A^*$ (with a certain density condition),
L is $\mathscr{C}$-measurable if and only if it satisfies the *Carathéodory condition*:

$$\bar{\mu}_{\mathscr{C}}(M) = \bar{\mu}_{\mathscr{C}}(M \cap L) + \bar{\mu}_{\mathscr{C}}(M \cap \overline{L}) \text{ for any language } M \subseteq A^*.$$

We call $\mathrm{Ext}_A(\mathscr{C}) = \{L \subseteq A^* \mid L \text{ is } \mathscr{C}\text{-measurable}\}$ the **Carathéodory extension** of a class $\mathscr{C}$.

Motivation of this work

The class $\mathrm{Ext}_A(\mathrm{REG}) = \{L \subseteq A^* \mid L$ is REG-measurable$\}$ and the class $\mathrm{Ext}_A(\mathrm{REG}) \cap \mathrm{CFL}$ are somewhat hard to analyse.

What about "***miniatures***" of $\mathrm{Ext}_A(\mathrm{REG}) \cap \mathrm{CFL}$?

e.g., $\mathrm{RExt}_A(\mathscr{L}) = \{L \in \mathrm{REG}_A \mid L$ is $\mathscr{L}$-measurable$\}$

a ***regular extension*** of some *subclass* $\mathscr{L}$ *of regular languages.*

Outline

1. Background: density and measurability
2. **Carathéodory extensions of local varieties**
3. Conclusion

Closure properties

Theorem [S]:
If language classes $\mathscr{C} \subseteq \mathscr{D}$ satisfies the following conditions:
(1) every language in $\mathscr{D}$ has the density,
(2) $\mathscr{C}$ and $\mathscr{D}$ are closed under left and right quotients,
then $\mathrm{Ext}_A(\mathscr{C}) \cap \mathscr{D}_A$ is closed under left and right quotients.

Proof idea: Let $L \in \mathscr{D}_A$ be a $\mathscr{C}$-measurable language.

There exists a convergent sequence (K_n, M_n) of L in $\mathscr{C}_A$.

We can show that $(w^{-1}K_n, w^{-1}M_n)$ converges to $w^{-1}L$
(albeit that some non-trivial calculation on densities is required).

Closure properties

Theorem [S]:

If language classes $\mathscr{C} \subseteq \mathscr{D}$ satisfies the following conditions:

(1) every language in $\mathscr{D}$ has the density,

(2) $\mathscr{C}$ and $\mathscr{D}$ are closed under left and right quotients,

then $\mathrm{Ext}_A(\mathscr{C}) \cap \mathscr{D}_A$ is closed under left and right quotients.

Theorem [S]:

If language classes $\mathscr{C} \subseteq \mathscr{D}$ satisfies the following conditions:

(1) every language in $\mathscr{D}$ has the density,

(2) $\mathscr{C}$ and $\mathscr{D}$ are closed under Boolean operations,

then $\mathrm{Ext}_A(\mathscr{C}) \cap \mathscr{D}_A$ is closed under Boolean operations.

Local varieties and Eilenberg theorem

A family of regular languages $\mathscr{L}$ over A is called ***local variety*** [Adámek et al. 2014] if it is closed under left-and-right quotients and Boolean operations

There is a corresponding notion for a *family of finite monoids* generated by A, called ***local pseudovariety***.

And there is an *Eilenberg theorem for local varieties* [Gehrke et al. 2008] roughly stating “there is a *natural bijection* between local varieties and local pseudovarieties”.

Star-free languages

A language L is said to be ***star-free*** if it can be represented as a finite applications of Boolean operations and concatenations to finite languages.
The class SF of all star-free languages over A is a local variety.

Theorem [Schutzenberger 1965]:
L is star-free if and only if the syntactic monoid M_L of L is *aperiodic*, i.e.,
M_L has no non-trivial subgroup.

Regular extension of local varieties

Theorem [S]:
For a local variety $\mathscr{L}$ over A, $\mathrm{RExt}_A(\mathscr{L}) = \{L \in \mathrm{REG}_A \mid L$ is $\mathscr{L}$-measurable$\}$
is also a local variety. Moreover, RExt_A is a **closure operator** on local varieties.

(**extensive**) $\mathscr{L} \subseteq \mathrm{RExt}_A(\mathscr{L})$ (**monotone**) $\mathscr{L} \subseteq \mathscr{L}' \Rightarrow \mathrm{RExt}_A(\mathscr{L}) \subseteq \mathrm{RExt}_A(\mathscr{L}')$
(**idempotent**) $\mathrm{RExt}_A(\mathrm{RExt}_A(\mathscr{L})) = \mathrm{RExt}_A(\mathscr{L})$

Theorem [S]:
SF $\subsetneq \mathrm{RExt}_A(\mathrm{SF}) \subsetneq \mathrm{REG}_A$ if A contains at least two letters,
i.e., RExt_A extends SF *non-trivially*.

Note: SF $\subsetneq \mathrm{RExt}_A(\mathrm{SF})$ is clear because $(aa)^* \notin \mathrm{SF}$, but
$(aa)^* \subseteq \overline{A^*(A\backslash\{a\})A^*} \in \mathrm{SF}$ and hence $(aa)^* \in \mathrm{RExt}_A(\mathrm{SF})$.

Regular extension of SF

Lemma [S]:
If the density of a star-free language $L \subseteq A^*$ is positive, then L contains words of even and odd length.

Theorem [S]:
$\mathrm{SF} \subsetneq \mathrm{RExt}_A(\mathrm{SF}) \subsetneq \mathrm{REG}_A$ if A contains at least two letters,
i.e., RExt_A extends SF *non-trivially*.

Note: by Lemma above, we can deduce that any star-free subset of $(AA)^*$ is of density zero and hence $(AA)^* \notin \mathrm{RExt}_A(\mathrm{SF})$.

Regular extension of $\mathbf{SF}$

Lemma [S]:

If the density of a star-free language $L \subseteq A^*$ is positive, then L contains words of even and odd length.

Proof sketch:

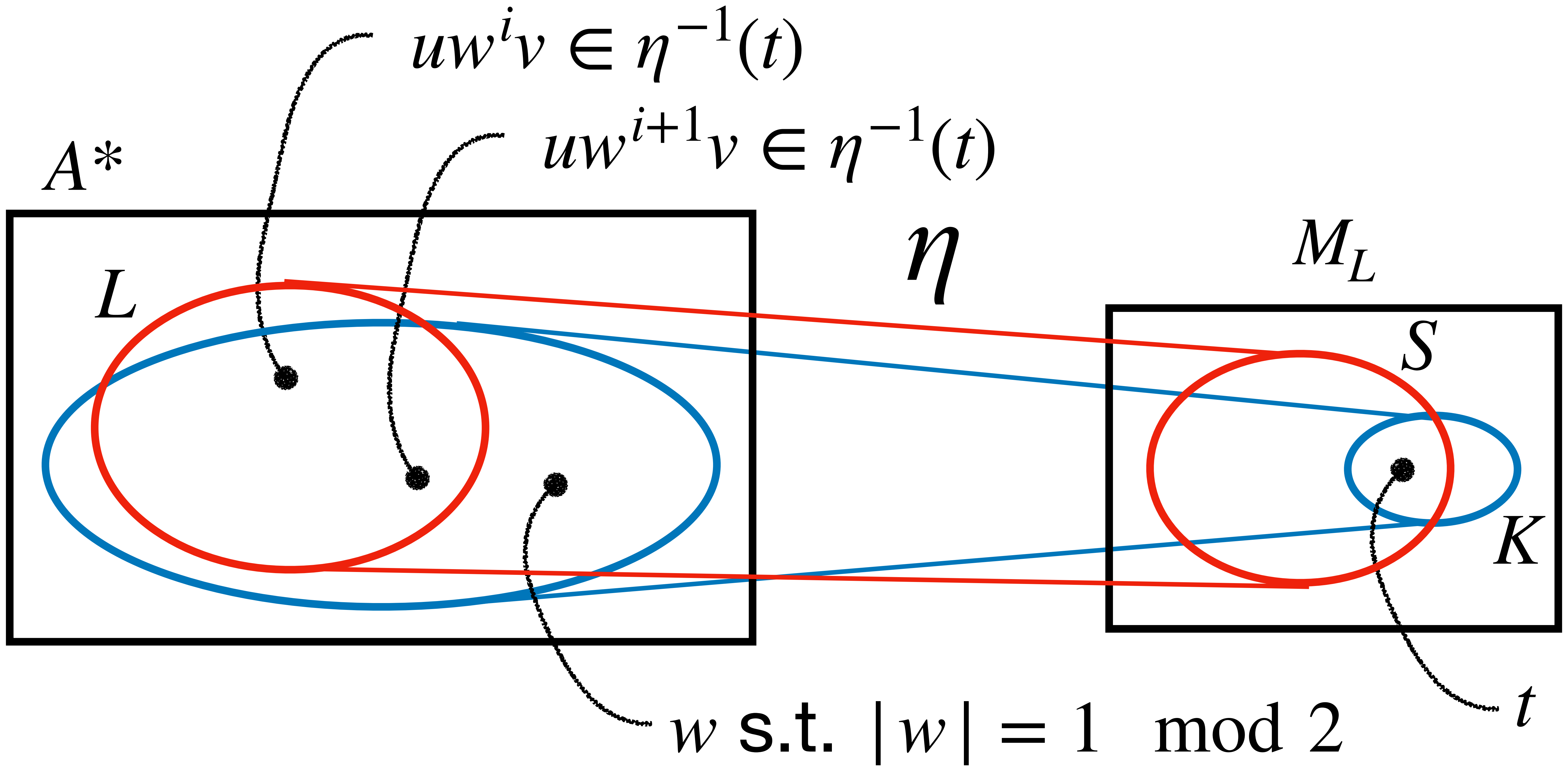
Claim 1

"$\delta^*_A(L) > 0$" and "M_L is finite" imply

S intersects with **the minimal ideal** K of M_L.

Claim 2

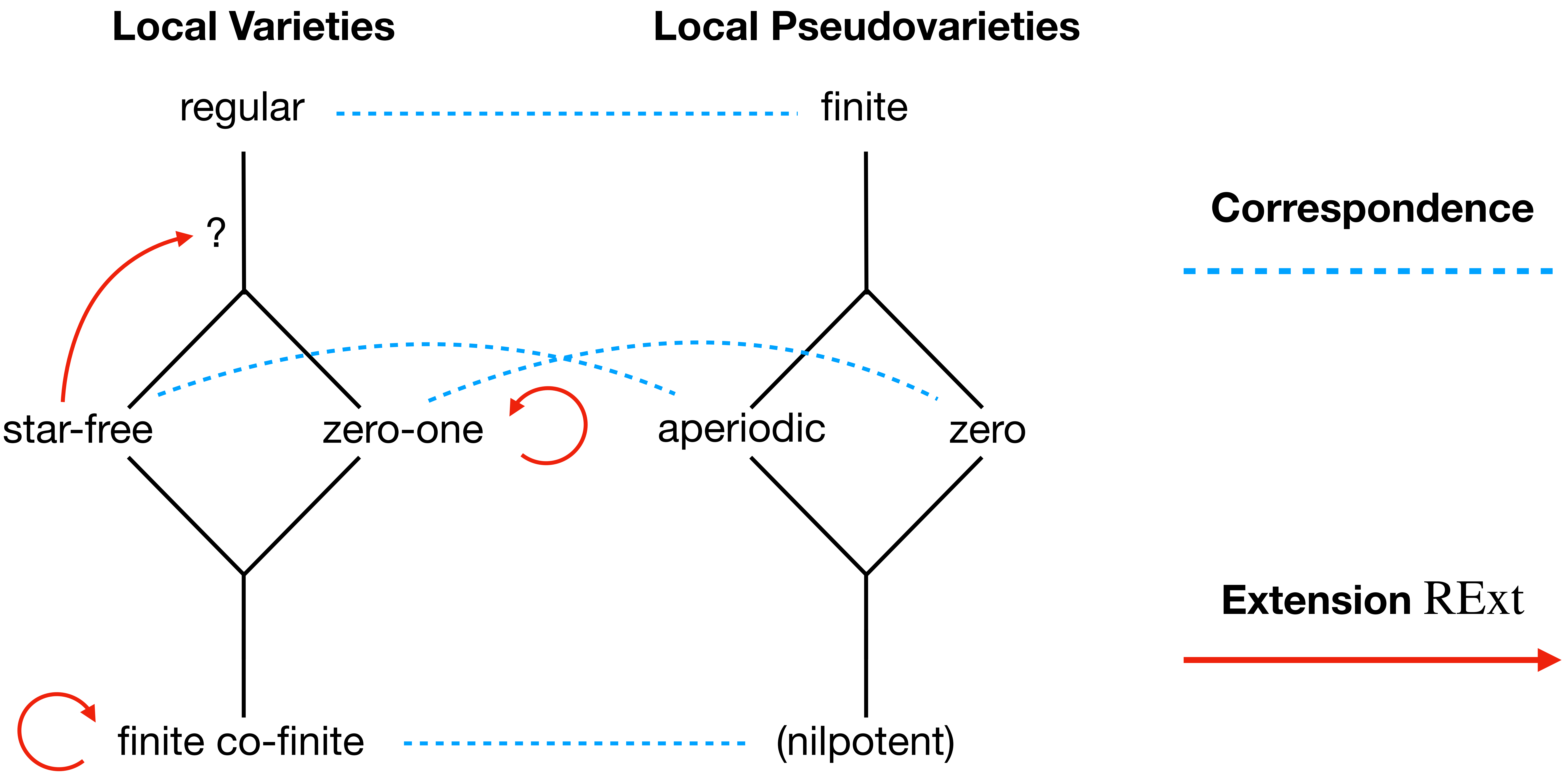
The inverse image of K is of density 1 (**infinite monkey theorem**), thus it contains word of odd length, say, w. Let $\eta(w) = m$.



Claim 3

By Schutzenberger's theorem there exists $i \geq 1$ such that $x^i = x^{i+1}$ for any $x \in M_L$. By the minimality of K, there exist u and v such that $\eta(uw^iv) = \eta(u)m^i\eta(v) = t = \eta(u)m^{i+1}\eta(v) = \eta(uw^{i+1}v)$.

Summary



*zero-one = the class of all regular languages with density zero or one

Conclusion and future work

- We consider Carathéodory extensions of local varieties of regular languages. These extensions could be considered as a "miniature" of the class of regular measurable context-free languages: a difficult object.

- We showed that the extension operator RExt_A is a closure operator on local varieties, and it extends SF non-trivially, and it does not extend some other local varieties (FIN or ZO).

- By Eilenberg theorem for local varieties, there exists a closure operator

$$\mathrm{MExt}_A(\mathscr{M}) = F^{-1}(\mathrm{RExt}_A(F(\mathscr{M})))$$

  on local pseudovarieties of finite monoids

  (where F is the natural bijection between local varieties and local pseudovarieties).

  Can we characterize this operator in purely algebraic way?

Thanks!

(Akita-Inu)

秋田大学
Akita University