The *rational* construction of a (Wheeler) DFA

Giovanni Manzini, Alberto Policriti, Nicola Prezza, and Brian Riccardi

Order is important ...

Order is important ...

Order is important ...

... and some orders are more important than others (e.g. the order of $\mathbb{Q}$)

Automata and (Co-lexicographic) Order

$\mathcal{A}$ *is input-consistent*: $(\forall u, v \in Q)(\delta(u, a_1) = \delta(v, a_2) \rightarrow a_1 = a_2)$.

$$\delta(q) = q' \text{ stands for } \delta(q, \lambda(q')) = q'.$$

Definition

A Wheeler DFA (WDFA) $\mathcal{A} = (Q, s, \delta, F, \prec)$ is such that $(Q, \prec)$ is a total order with s as minimum, and letting $v_1 = \delta(u_1)$, and $v_2 = \delta(u_2)$:

i $v_1 \prec v_2 \;\Rightarrow\; \lambda(v_1) \leqslant \lambda(v_2)$;

ii $(\lambda(v_1) = \lambda(v_2) \wedge v_1 \prec v_2) \Rightarrow u_1 \prec u_2$.

(Gagie-Manzini-Sirén. TCS 2017) (Alanko-D'Agostino-P.-Prezza. Inf. and Comp. 2021)

(Cotumaccio-D'Agostino-P.-Prezza. JACM. 2023)

Automata and (Co-lexicographic) Order

$\mathcal{A}$ *is input-consistent*: $(\forall u, v \in Q)(\delta(u, a_1) = \delta(v, a_2) \rightarrow a_1 = a_2)$.

$$\delta(q) = q' \text{ stands for } \delta(q, \lambda(q')) = q'.$$

Definition

A Wheeler DFA (WDFA) $\mathcal{A} = (Q, s, \delta, F, \prec)$ is such that $(Q, \prec)$ is a total order with s as minimum, and letting $v_1 = \delta(u_1)$, and $v_2 = \delta(u_2)$:

- i $v_1 \prec v_2 \;\Rightarrow\; \lambda(v_1) \leqslant \lambda(v_2)$;
- ii $(\lambda(v_1) = \lambda(v_2) \wedge v_1 \prec v_2) \Rightarrow u_1 \prec u_2$.

(Gagie-Manzini-Sirén. TCS 2017) (Alanko-D'Agostino-P.-Prezza. Inf. and Comp. 2021)

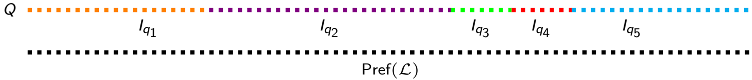
(Cotumaccio-D'Agostino-P.-Prezza. JACM. 2023)

$\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ is ordered and W(i)-W(ii) extend the ordering to strings reaching states

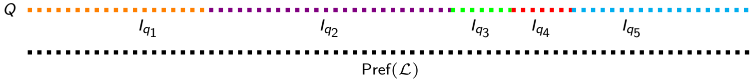
co-lexicographically

align to the right and compare right-to-left (cbaabba $<$ aaacba)

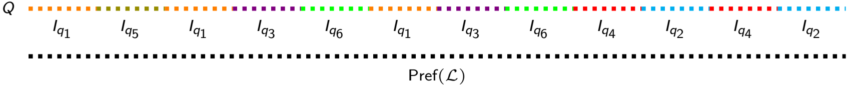
WDFA: states are intervals of strings



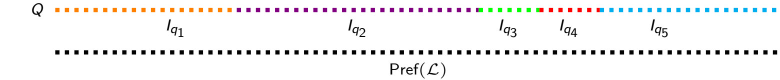
WDFA: states are intervals of strings



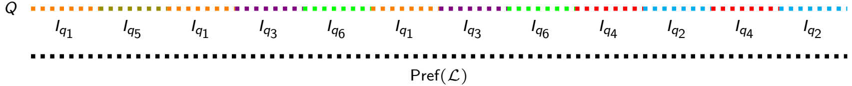
DFA

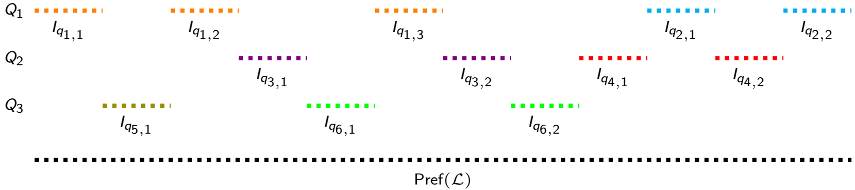


WDFA: states are intervals of strings

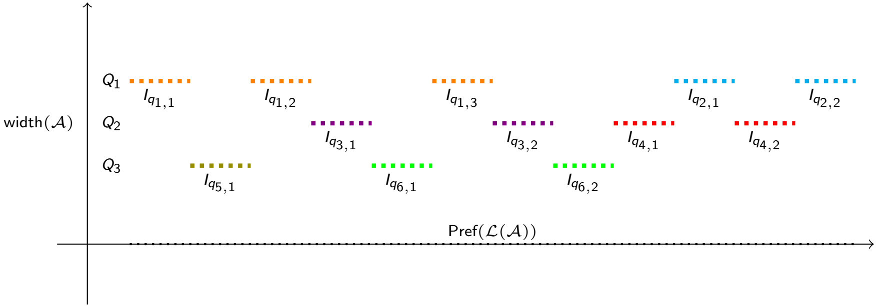


DFA

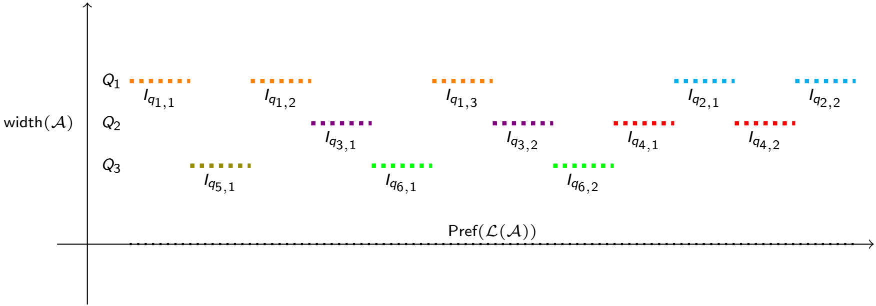




… coordinates



... coordinates



- (substring closure) membership in $\mathcal{L}(\mathcal{A})$ in $O(\text{width}(\mathcal{A})^2)$ per matched character;
- any NFA $\mathcal{N}$ is equivalent to a DFA $\mathcal{D}$ with at most $2^{\text{width}(\mathcal{N})}(|\mathcal{N}| - \text{width}(\mathcal{N}) + 1)$ states;
- any automaton of width p can be encoded in $O(\log p + \log|\Sigma|)$ bits per transition.

(Cotumaccio-D'Agostino-P.-Prezza. JACM. 2023)

Intermezzo: Path coherence

Definition (Path coherence)

Given a NFA $\mathcal{A}$ and an order $\prec$ over the set of its states, we say that $\mathcal{A}$ is *path coherent* iff strings send intervals (of states) into intervals:

$$\forall q \preceq q' \; \forall \alpha \; \exists p \preceq p' \; \big([q, q'] \overset{\alpha}{\rightsquigarrow} [p, p']\big).$$

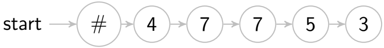
The rational embedding of strings

Definition (**The Rational Embedding of Σ^***)

The *Rational Embedding* of Σ^* is the map $\mathbb{q} : \Sigma^* \to \mathbb{Q}[0,1)$ such that, for any $\alpha = \alpha_1 \ldots \alpha_m \in \Sigma^*$:

$$\mathbb{q}(\alpha) = \sum_{i=1}^{m} \alpha_i \cdot (\sigma + 2)^{-(m-i+1)}.$$

Example



$\mathbb{q}(\alpha) = \mathbb{q}(47753) = 0.35774$

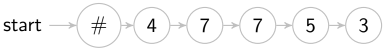
The rational embedding of strings

Definition (**The Rational Embedding of Σ^***)

The *Rational Embedding* of Σ^* is the map $\mathbb{q} : \Sigma^* \to \mathbb{Q}[0,1)$ such that, for any $\alpha = \alpha_1 \dots \alpha_m \in \Sigma^*$:

$$\mathbb{q}(\alpha) = \sum_{i=1}^{m} \alpha_i \cdot (\sigma + 2)^{-(m-i+1)}.$$

Example



$\mathbb{q}(\alpha) = \mathbb{q}(47753) = 0.35774$

property:
$\boldsymbol{\alpha} < \boldsymbol{\beta}$ (in co-lex order) $\Leftrightarrow \mathbb{q}(\boldsymbol{\alpha}) < \mathbb{q}(\boldsymbol{\beta})$ (as rational numbers).

The rational embedding of DFAs

Definition

$I_{\mathbb{Q}[0,1)}$ the collection of convex sets of rationals in $\mathbb{Q}[0,1)$:

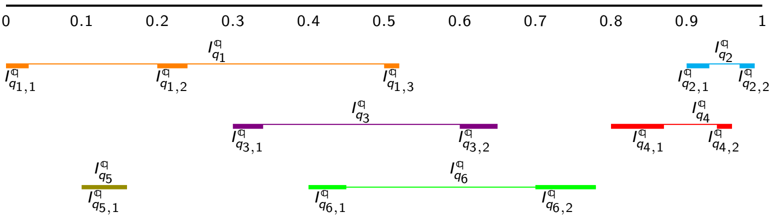
Definition (**The Rational Embedding of a DFA)**

The *Rational Embedding* of $\mathcal{A} = (Q, s, \delta, F)$ is the map $I^{\mathfrak{q}} : Q \to I_{\mathbb{Q}[0,1)}$ defined as follows: for any $q \in Q$,

$$I^{\mathfrak{q}}(q) = \bigcap \{J \in I_{\mathbb{Q}[0,1)} \mid (\forall \alpha \in I_q)(\mathfrak{q}(\alpha) \in J)\}.$$

$I^{\mathfrak{q}}(q)$ is the convex closure (hull) of I_q. (Notation: $I_q^{\mathfrak{q}} = I^{\mathfrak{q}}(q)$)

The rational embedding of DFAs



Determinism: $q \neq q' \Rightarrow I_q \cap I_{q'} = \emptyset$.
But it might be that $q \neq q' \wedge I^{\mathbb{Q}}_q \cap I^{\mathbb{Q}}_{q'} \neq \emptyset$.

From (Alanko-D'Agostino-P.-Prezza. Inf. and Comp. 2021)[Theorem 4.3]:

$\mathcal{A}$ is Wheeler iff $q \neq q' \Rightarrow I^{\mathbb{Q}}_q \cap I^{\mathbb{Q}}_{q'} = \emptyset$.

The rational embedding of DFAs

Example

There are *"solid"* collections of rational embeddings of strings accepted by a DFA: Σ^* (non-denumerable set of accumulation points)

The rational embedding of DFAs

Example

There are *"solid"* collections of rational embeddings of strings accepted by a DFA: Σ^* (non-denumerable set of accumulation points)

Example

Missing largest and smallest digit $\Rightarrow$ (in general) there is a successor but no predecessor.

$$0.7348 < \cdots < 0.73488 < \cdots < 0.734888 < \cdots 0.7345 < 0.73451$$

The rational embedding of DFAs

Example

There are *"solid"* collections of rational embeddings of strings accepted by a DFA: Σ^* (non-denumerable set of accumulation points)

Example

Missing largest and smallest digit $\Rightarrow$ (in general) there is a successor but no predecessor.

$$0.7348 < \cdots < 0.73488 < \cdots < 0.734888 < \cdots 0.7345 < 0.73451$$

Question: are $I_q^{\mathbb{q}}$'s left and right limits (ℓ_q, r_q) always in $\mathbb{Q}$?

Ordering states: Entanglement

Definition

$Q' \subseteq Q$ is entangled if there exists a monotone sequence $(\alpha_i)_{i \in \mathbb{N}}$ in $Pref(\mathcal{L}(\mathcal{D}))$ such that:

$$\forall u' \in Q' \; \delta(s, \alpha_i) = u' \text{ for infinitely many } i's$$

Ordering states: Entanglement

Definition

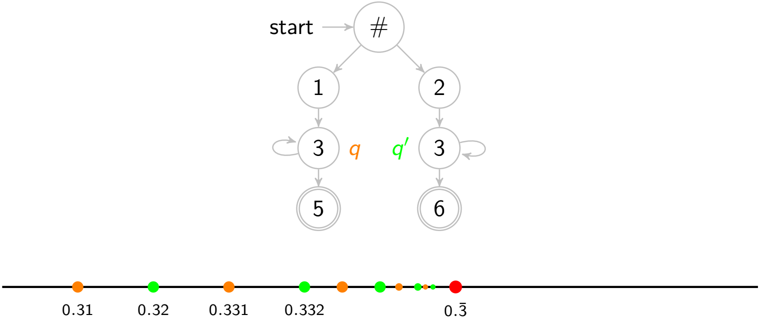
$Q' \subseteq Q$ is entangled if there exists a monotone sequence $(\alpha_i)_{i \in \mathbb{N}}$ in $Pref(\mathcal{L}(\mathcal{D}))$ such that:

$$\forall u' \in Q' \; \delta(s, \alpha_i) = u' \text{ for infinitely many } i's$$

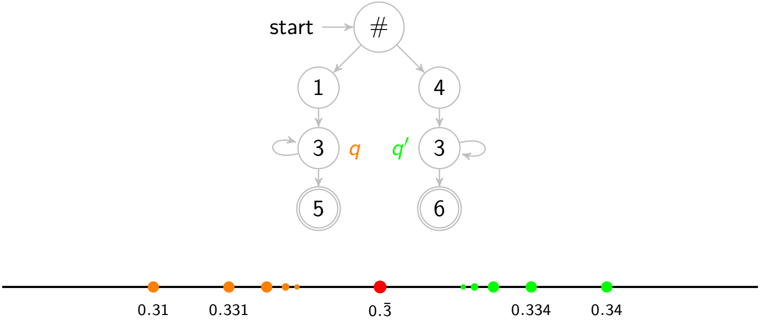
Lemma

If a value x is a left-accumulation point (resp. right-accumulation point) for both the sets $I_q^{\mathbb{Q}}$ and $I_{q'}^{\mathbb{Q}}$ then q and q' are entangled.

Examples



Examples



Finding Left and Right Limits

Lemma

Let $\mathcal{L} = \mathcal{L}(\mathcal{D})$, with $\mathcal{L}$ Wheeler and $\mathcal{D}$ either minimum or Wheeler. For any $q \in Q$ we have:

$$\ell_q = 0.a_{q,1} \cdots a_{q,h}\overline{a_{q,h+1} \cdots a_{q,h+j}},$$

with $h + j \leqslant |Q|$, and $j > 0$ if and only if $\ell_q \notin I_q^{\mathfrak{q}}$.

Finding Left and Right Limits

Lemma

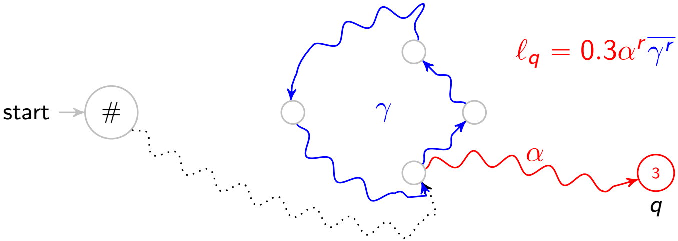
Let $\mathcal{L} = \mathcal{L}(\mathcal{D})$, with $\mathcal{L}$ Wheeler and $\mathcal{D}$ either minimum or Wheeler. For any $q \in Q$ we have:

$$\ell_q = 0.a_{q,1} \cdots a_{q,h}\overline{a_{q,h+1} \cdots a_{q,h+j}},$$

with $h + j \leqslant |Q|$, and $j > 0$ if and only if $\ell_q \notin I_q^{\mathbb{q}}$.

Proof's idea

walk backward from q and find α whose rational embedding $\mathbb{q}(\alpha)$ is the smallest among those reaching q.



Finding Left and Right Limits

Lemma

Let $\mathcal{L} = \mathcal{L}(\mathcal{D})$, with $\mathcal{L}$ Wheeler and $\mathcal{D}$ either minimum or Wheeler. For any $q \in Q$ we have:

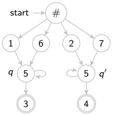
$$\ell_q = 0.a_{q,1} \cdots a_{q,h}\overline{a_{q,h+1} \cdots a_{q,h+j}},$$

with $h + j \leqslant |Q|$, and $j > 0$ if and only if $\ell_q \notin I_q^{\mathbb{Q}}$.

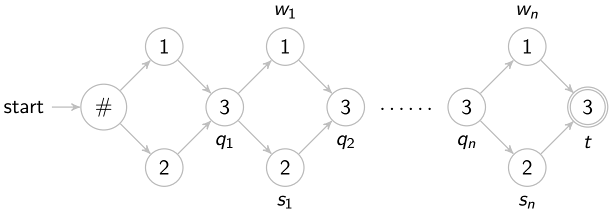
Theorem

If $\mathcal{L} = \mathcal{L}(\mathcal{D}) = \mathcal{L}(\mathcal{D}_w)$, with $\mathcal{L}$ Wheeler and $\mathcal{D}$ either minimum or Wheeler, then for all $q \in Q$, we have $\ell_q, r_q \in \mathbb{Q}$.

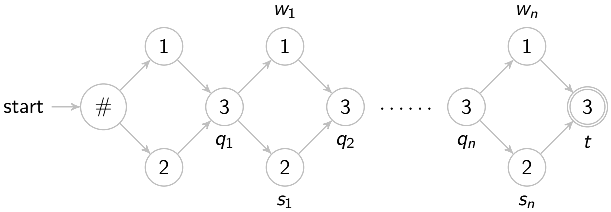
More on the algorithmic side on (Becker-Cenzato-Kim-Kodric-P.-Prezza. SPIRE 2023.)
More on the entanglement:

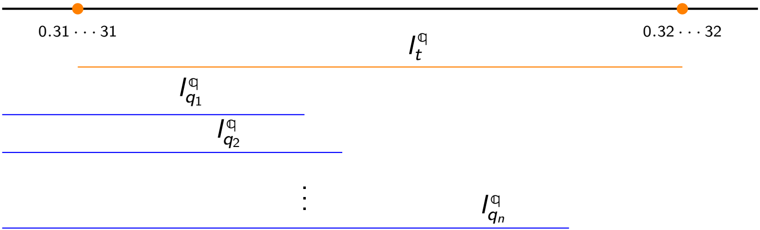


Minimum DFA vs. Minimum Wheeler-DFA

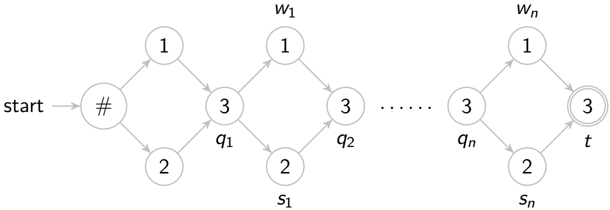


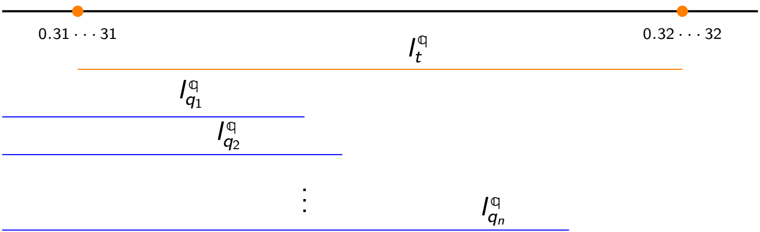
Minimum DFA vs. Minimum Wheeler-DFA





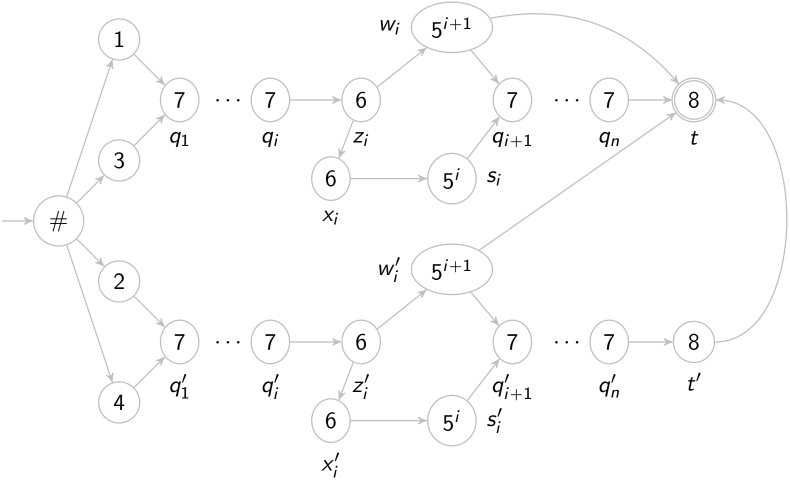
Minimum DFA vs. Minimum Wheeler-DFA

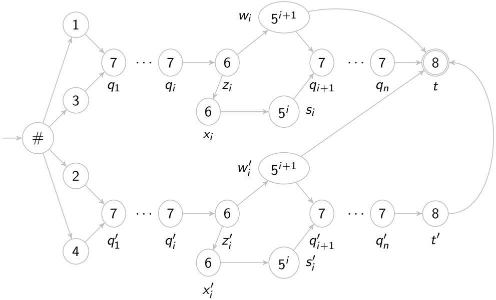




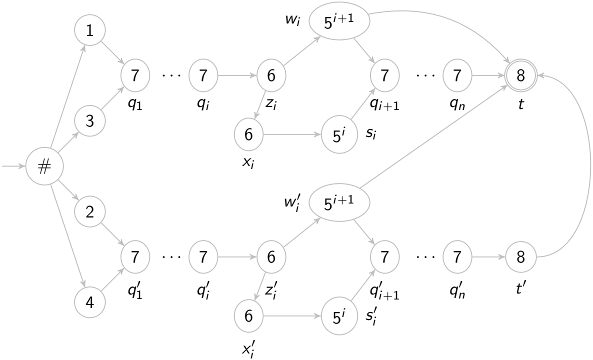
Question: is the $|Q_w|/|Q|$ related to width$(\mathcal{D})$?

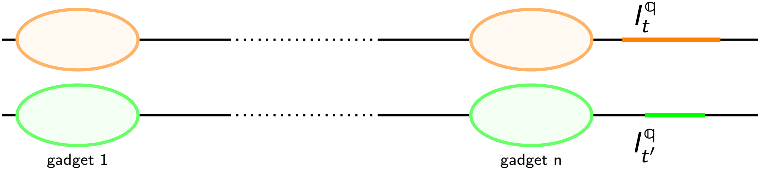
Lower bound


1
3
#
2
4
7
7
6
6
w_i
5^{i+1}
5^i
s_i
7
7
8
q_1
q_i
z_i
x_i
q_{i+1}
q_n
t
w'_i
5^{i+1}
5^i
s'_i
q'_1
q'_i
z'_i
x'_i
q'_{i+1}
q'_n
t'
8



State type	Left limit	Right limit
$s_{i,j}$	$0.5^j6675^i67\ldots$	$0.5^j6675^{i-1}667\ldots$
$w_{i,j}$	$0.5^j675^i67\ldots$	$0.5^j675^{i-1}667\ldots$
x_i	$0.6675^i67\ldots$	$0.6675^{i-1}667\ldots$
z_i	$0.675^i67\ldots$	$0.675^{i-1}667\ldots$
q_i	$0.75^i67\ldots$	$0.75^{i-1}667\ldots$
t	$0.85^n67\ldots$	$0.8875^{n-1}667\ldots$
t'	$0.875^n67\ldots$	$0.875^{n-1}667\ldots$


#
1
2
3
4
7
6
8
5^{i+1}
5^i
w_i
z_i
x_i
s_i
q_1
q_i
q_{i+1}
q_n
t
w'_i
z'_i
x'_i
s'_i
q'_1
q'_i
q'_{i+1}
q'_n
t'


$I_t^{\mathfrak{q}}$
$I_{t'}^{\mathfrak{q}}$
gadget 1
gadget n

Theorem

Let $\mathcal{L} = \mathcal{L}(\mathcal{D}) = \mathcal{L}(\mathcal{D}_w)$*, with* $\mathcal{L}$ *Wheeler,* $\mathcal{D}$ *minimum,* $\mathcal{D}_w$ *minimum Wheeler, and let* $f(\cdot,\cdot)$ *be such that* $|\mathcal{D}_w| = O(\, f(|\mathcal{D}|, width(D))\,)$*. Then, for any* $k, p \in \mathbb{N}$*,*

$$f(n, p) \notin O(n^k + 2^p).$$

The arithmetic way

Formally, for the *left* case, we consider the problem of finding the set of all real-valued vectors $x \in \mathbb{R}^Q$ that satisfy the following constraint satisfaction program, that we name $\mathcal{P}_{Left}$:

$$
\begin{aligned}
&(1) \quad x_s = 0, \\
&(2) \quad 0 < x_q < 1, && (\forall q \in Q \backslash \{s\}) \\
&(3) \quad (\sigma + 2) \cdot x_q - \lambda(q) = \min \left\{ x_{q'} \mid \delta(q') = q \right\}, && (\forall q \in Q \backslash \{s\})
\end{aligned}
$$

The arithmetic way

Formally, for the *left* case, we consider the problem of finding the set of all real-valued vectors $x \in \mathbb{R}^Q$ that satisfy the following constraint satisfaction program, that we name $\mathcal{P}_{Left}$:

$$\begin{aligned} &(1) \quad x_s = 0, \\ &(2) \quad 0 < x_q < 1, && (\forall q \in Q \setminus \{s\}) \\ &(3) \quad (\sigma + 2) \cdot x_q - \lambda(q) = \min \left\{ x_{q'} \mid \delta(q') = q \right\}, && (\forall q \in Q \setminus \{s\}) \end{aligned}$$

Lemma

Let $\mathcal{L}$ be a Wheeler language, and $\mathcal{D} = (Q, s, \delta, F)$ be either minimum or Wheeler accepting $\mathcal{L}$, and let $\ell \in \mathbb{Q}^Q$ be the vector of left limits. Then, ℓ is a solution of $\mathcal{P}_{Left}$.

Theorem

Let $\mathcal{D} = (Q, s, \delta, F)$ be either minimum or Wheeler accepting $\mathcal{L}$ Wheeler, and $\ell \in \mathbb{Q}^Q$ be the vector of left limits. Then, $\mathcal{P}_{Left}$ always admits ℓ as its unique solution.

The arithmetic way

Consider the following linear program $\mathcal{P}^*_{Left}$:

$$
\begin{array}{lll}
\text{maximize:} & \sum_{q \in Q} x_q, & \\
\text{subject to:} & x_s = 0, & \\
 & 0 < x_q < 1, & \forall q \in Q \backslash \{s\}, \\
 & (\sigma + 2) \cdot x_q - \lambda(q) \leqslant x_{q'}, & \forall q, q' \in Q \text{ s.t. } \delta(q') = q,
\end{array}
$$

Theorem

Let $\mathcal{D} = (Q, s, \delta, F)$ be either minimum or Wheeler accepting $\mathcal{L}$ Wheeler, and $\ell \in \mathbb{Q}^Q$ be the vector of left limits. Then $\mathcal{P}^*_{Left}$ always admits ℓ as its unique solution.

Conclusions and Open problems

- A parameter measuring the DFA vs. WDFA growth (*finite* entanglement)
- On-line splitting of minimum DFA (self-adjusting splitting)
- Optimal *disentanglement*

Conclusions and Open problems

- A parameter measuring the DFA vs. WDFA growth (*finite* entanglement)
- On-line splitting of minimum DFA (self-adjusting splitting)
- Optimal *disentanglement*

Optimization of the Hasse automaton $\mathcal{H}$ construction

$$\text{width}(\mathcal{L}(\mathcal{H})) = \text{width}(\mathcal{H}) = \textit{ent}(\mathcal{H})$$

Thank you for your attention.