Periodicity of Degenerate Strings

Estéban Gabory, Eric Rivals, Michelle Sweering, Hilde Verbeek & Pengfei Wang
Centrum Wiskunde & Informatica, Université Montpellier and CNRS

Degenerate Strings

A **degenerate string** permits a choice of several different characters at each position, as to encode several classical strings. For example $\hat{w} = \{\substack{a\\b}\}\{\substack{a\\c}\}b$ is a degenerate string; its **language** is $\mathcal{L}(\hat{w}) = \{aab, acb, bab, bcb\}$.

Two degenerate strings of the same length **match** if at every position there is a non-empty intersection. For example, $\{\substack{a\\b}\}\{\substack{a\\c}\}b$ matches $\{\substack{a\\c}\}c\{\substack{b\\c}\}$ as both have acb in their language.

Types of periods

Periodicity is a way of modeling repetitions in strings. A degenerate string $\hat{w}$ of length n has

- **weak period** p when $\hat{w}$'s prefix and suffix of length $n-p$ match;
- **medium period** p when all multiples of p are a weak period;
- **strong period** p when a string $w \in \mathcal{L}(\hat{w})$ has period p.

For a degenerate string $\hat{w}$, the sets of weak, medium and strong periods are denoted by $P_w(\hat{w})$, $P_m(\hat{w})$ and $P_s(\hat{w})$ respectively.

Autocorrelations

An **autocorrelation** is a 0-indexed vector representing the periods of a degenerate string. Its p-th element is 3 if $p \in P_s$, 2 if $p \in P_m \setminus P_s$, 1 if $p \in P_w \setminus P_m$ and 0 if $p \notin P_w$.

Characterizing period sets

P_w, P_m and P_s are the period sets of *some* length-n degenerate string iff.:

i. $\{0\} \subseteq P_s \subseteq P_m \subseteq P_w$;
ii. $p \geq \frac{n}{2} \wedge p \in P_w \Longrightarrow p \in P_s$;
iii. $p \in P_m$ iff. for all $k \in \mathbb{N}$ with $kp \in \{0, \ldots, n-1\}, kp \in P_w$;
iv. $p \in P_s$ iff. for all $k \in \mathbb{N}$ with $kp \in \{0, \ldots, n-1\}, kp \in P_s$.

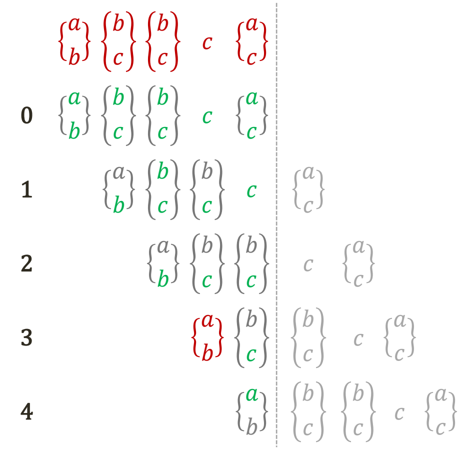
These conditions are sufficient for alphabets of size 3 and greater.

Counting period sets

Let Ω_n^w, Ω_n^m and Ω_n^s be the families of period sets of length-n degenerate strings. We show $|\Omega_n^w| = 2^{n-1}$ and based on prior work $|\Omega_n^m| = |\Omega_n^s| = \alpha^{(n-1)(1+O(\exp((-1+\epsilon)\sqrt{\log n \log\log n})))}$ with $1.5729 < \alpha < 1.5745$.

Example

Let $\hat{w} = \{\substack{a\\b}\}\{\substack{b\\c}\}\{\substack{b\\c}\}c\{\substack{a\\c}\}$. Its periods are $P_w(\hat{w}) = \{0,1,2,4\}$, $P_m(\hat{w}) = \{0,2,4\}$ and $P_s(\hat{w}) = \{0,4\}$. The autocorrelation of $\hat{w}$ is 31203.



Other contributions

- We characterize the structure of autocorrelations.
- We show the sets of auto-correlations form lattices.
- We give formulae for the number of degenerate strings with a given autocorrelation, by counting independent sets in graphs.

This project was funded by NWO Gravitation-grant NETWORKS-024.002.003, EU Horizon 2020 Marie Skłodowska-Curie grant agreements No 872539 and 956229 and CWI's Constance van Eeden PhD Fellowship.
We also thank Solon Pissis for helpful insights and feedback.