Subsequences in Bounded Ranges: Matching and Analysis Problems

Maria Kosche Tore Koß Florin Manea Viktoriya Pak

Göttingen University

October 18, 2022

Subsequences in Bounded Ranges

w	b	a	c	b	a	a	b	a	d	a

Subsequences in Bounded Ranges

d

ba

bc

w $\boxed{b \quad a \quad c \quad b \quad a \quad a \quad b \quad a \quad d \quad a}$

bb

baab

a

c

b

Subsequences in Bounded Ranges

d

abc

ba

bc

w $\boxed{b \quad a \quad c \quad b \quad a \quad a \quad b \quad a \quad d \quad a}$

bb

baab

a

dd

c

b

Subsequences in Bounded Ranges

d

~~abc~~

ba

bc

$$w \quad \boxed{b \quad a \quad c \quad b \quad a \quad a \quad b \quad a \quad d \quad a}$$

baab

a

bb

~~dd~~

c

b

Subsequences in Bounded Ranges

ababa

d

~~abc~~

bbb

ba

bc

w $\boxed{b \quad a \quad c \quad b \quad a \quad a \quad b \quad a \quad d \quad a}$

bb

baab

a

~~dd~~

c

b

aaaaa

Subsequences in Bounded Ranges

~~ababa~~

d

~~abc~~

~~bbb~~

ba

bc

w $\boxed{b \quad a \quad c \quad b \quad a \quad a \quad b \quad a \quad d \quad a}$

bb

baab

a

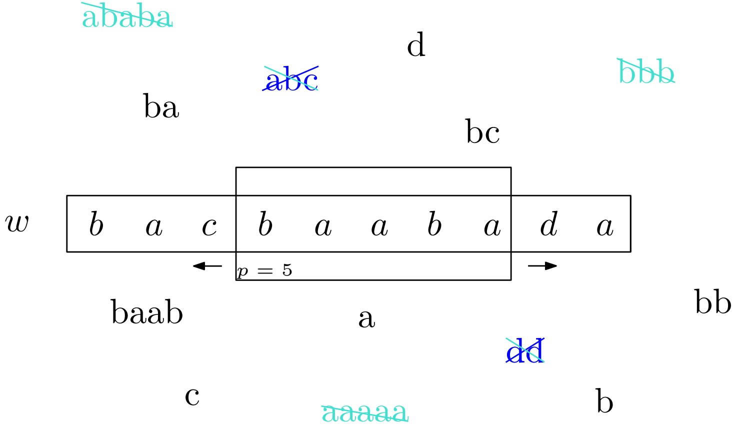
~~dd~~

c

~~aaaaa~~

b

Subsequences in Bounded Ranges



Subsequences in Bounded Ranges

Definition

- We call v a subsequence of w, where $|w| = n$, if there exist positions $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, such that $v = w[i_1]w[i_2] \cdots w[i_k]$.
- We denote the set of all subsequences of length k of w by $\mathrm{Subseq}_k(w)$.
- If v is not a subsequence of w we call it an *absent subsequence*.
- We call an absent subsequence v of w a *minimal absent subsequence* (for short, MAS) of w if every proper subsequence of v is a subsequence of w.
- We call an absent subsequence v of w a *shortest absent subsequence* (for short, SAS) of w if $|v| \leq |v'|$ for any other absent subsequence v' of w.

Motivation: Subsequence Matching in Circular Words

Definition

(i) Two strings $u, w \in \Sigma^*$ are conjugate if and only if there are string x, y such that $u = xy$ and $w = yx$. We call the conjugacy class of w the *circular word* $w_\circ$.

(ii) (u, n) is a (minimal) representative of $w_\circ$ if and only if there is a conjugate w' of w such that $w' = u^{n/|u|}$ (and u is len-lexicographic smaller than every other representative of $w_\circ$).

Open Question (Hegedüs, Nagy 2016)

Can a minimal representative be calculated in linear time?

Motivation: Subsequence Matching in Circular Words

Definition

A string u is a subsequence of the circular word $w_\circ$ if and only if there is a conjugate w' of w such that u is a subsequence of w'

Problem

Given two strings u, w, with $|u| = m$ and $|w| = n$, decide whether u is a subsequence of $w_\circ$.

Idea: The set of conjugates of w equals the set of factors of length $|w|$ of w^2.

Generalization: Why restrict to w^2 and factors of length $|w|$?

Subsequences in Bounded Ranges

Definition

- We call v a p-subsequence of w, where $|w| = n$, if there exist a position $1 \leq i \leq n - p + 1$, such that v is a subsequence of the bounded range $w[i : i + p - 1]$.

Subsequences in Bounded Ranges

Definition

- We call v a p-subsequence of w, where $|w| = n$, if there exist a position $1 \leq i \leq n - p + 1$, such that v is a subsequence of the bounded range $w[i : i + p - 1]$.
- We denote the set of all p-subsequences of length k of w by $p\text{-Subseq}_k(w)$.
- If v is not a p-subsequence of w we call it an *absent p-subsequence*.
- We call an absent p-subsequence v of w a *minimal absent p-subsequence* (for short, p-MAS) of w if every proper subsequence of v is a p-subsequence of w.
- We call an absent p-subsequence v of w a *shortest absent p-subsequence* (for short, p-SAS) of w if $|v| \leq |v'|$ for any other absent subsequence v' of w.

Problems

Given two words $u, w \in \Sigma^*$, with lengths $|u| = m$ and $|w| = n$, and non-negative integers p, k we analyse the following problems.

- p-**Matching:** Is u a p-subsequence of w?
- p-**MAS:** Is u a p-MAS of w?
- (p, k)-**non-Universality:** $p\text{-Subseq}_k(w) \neq \Sigma^k$.
- (p, k)-**non-Equivalence:** $p\text{-Subseq}_k(w) \neq p\text{-Subseq}_k(u)$.
- p-**non-SAS:** Is u not a p-SAS of w?

Matching

Subsequence Matching in Bounded Ranges	
Input:	$u, w \in \Sigma^*$ with $\lvert u\rvert = m, \lvert w\rvert = n$ and an integer p such that $m \leq p \leq n$
Question:	Is u a p-subsequence of w?

Naïve solution in $O(np)$.
We present an $O(mn)$ algorithm, which is optimal unless OVH fails.

If u is not part of the input the presented algorithm requires $O(\log p)$ space, which is also optimal by Ganardi, Hucke, Lohrey (2016).

Matching

We create an array $A[\cdot]$ with m entries. Then read w left-to-right and when reading the t^{th} letter of w update A such that $A[i]$ contains the length of the shortest suffix of $w[t-p+1:t]$ containing $u[1:i]$ (or ∞ if it does not occur):

$$A[i] = \min\{|v| \mid v \text{ is suffix of } w[t{-}p{+}1:t], u[1:i] \in \mathrm{Subseq}_i(v)\}.$$

After any step, if $A[m] \leq p$ holds then u is a p-subsequence of w.

Matching

Given an instance (A, B, d) of OV with $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$ where $v = v[1]v[2] \cdots v[d]$ for all $v \in A \cup B$.

Definition

Let $\psi_A, \psi_B : \{0,1\}^* \to (\{0,1\} \cup \{\$\})^*$ be morphisms such that:

$$\psi_A(x) = \begin{cases} 01\$ & \text{if } x = 0, \\ 00\$ & \text{if } x = 1 \end{cases}$$

$$\psi_B(y) = y\$ \text{ for } y \in \{0,1\}$$

Example:

$$\psi_A(1001) = 00\$01\$01\$00\$$$

$$\psi_A(0100) = 01\$00\$01\$01\$$$

$$\psi_B(1001) = 1\$0\$0\$1\$$$

Matching

Next we construct two words W_A and U_B such that U_B occurs in W_A^2 as $|W_A|$-subsequence if and only if there are orthogonal vectors $a \in A$ and $b \in B$:

$$W_A = [\psi_A(1^d)][\psi_A(0^d)][\psi_A(a_1)][\psi_A(0^d)] \ldots [\psi_A(a_n)][\psi_A(0^d)][\psi_A(1^d)]$$
$$U_B = [\psi_B(1^d)][\psi_B(b_1)][\psi_B(b_2)] \ldots [\psi_B(b_n)][\psi_B(1^d)]$$

Analysis

Minimal Absent Subsequences w.r.t. Bounded Ranges	
Input:	$u, w \in \Sigma^*$ with $\lvert u \rvert = m, \lvert w \rvert = n$ and an integer p such that $m \leq p \leq n$
Question:	Is u a p-MAS of w?

Answer can be computed in time $O(mn)$, which is optimal unless OVH fails.

The algorithm implements the p-subsequence matching algorithm presented above as well as the greedy algorithm to check minimality from Kosche et al. 2021 tailored to bounded ranges.

For the conditional lower bound: reduction from p-subsequence matching.

Analysis

Partial Words non-Universality Problem

Input: A list of partial words $S = \{w_1, \ldots, w_k\}$ over $\{0, 1\}$ of same length L.

Question: Decide whether there is no word $v \in \{0, 1\}^L$ compatible with any of the partial words in S.

Definition

- A partial word u over Σ is a word over $\Sigma \cup \{\Diamond\}$. The positions i with $u[i] \in \Sigma$ are called *defined*.
- A partial word u over Σ and a (full) word w over Σ are compatible if and only if $|u| = |w|$ and $u[i] = w[i]$ for all defined positions i of u.

Theorem (Manea,Tiseanu 2013)

Partial Words non-Universality is NP-complete and cannot be solved in subexponential time $2^{o(L)}\,\mathrm{poly}(L, n)$ *unless* ETH *fails.*

Analysis

k-non-Universality w.r.t. Bounded Range	
Input:	$w \in \Sigma^*$ and integers p, k with $\lvert w \rvert = n$ and $k \leq p \leq n$
Question:	Is $p\text{-Subseq}_k(w) \neq \Sigma^k$?

Naïve solution is already optimal in $O(\lvert\Sigma\rvert^k \operatorname{poly}(n, k))$. For the conditional lower bound: reduction from the Partial Words non-Universality problem.

Analysis

Let $S = \{w_1, \ldots, w_\ell\}$, with $|w_i| = L$ for all $1 \leq i \leq \ell$, be an instance of the Partial Word non-Universality problem. For every $w_i \in S$ let $u_i = u_i[1] \cdots u_i[L]$ where

$$u_i[j] = \begin{cases} 0\$, & \text{if } w_i[j] = 0 \\ 1\$, & \text{if } w_i[j] = 1 \\ 01\$, & \text{if } w_i[j] = \Diamond. \end{cases}$$

We construct an instance of k-non-Universality in Bounded Range with $w = VU$, $k = 2L$ and $p = |V|$ for

$$V = \$^{2L}(001101\$^{2L})^{L-1},$$
$$U = \$^{4L^2} u_1 \$^{4L^2} u_2 \$^{4L^2} \cdots \$^{4L^2} u_\ell \$^{4L^2}.$$

Analysis

k-non-Equivalence w.r.t. Bounded Range

Input:	$v, w \in \Sigma^*$ and integers p, k with $\lvert v \rvert = m, \lvert w \rvert = n$ and $k \leq p \leq m, n$
Question:	Is $p\text{-Subseq}_k(w) \neq p\text{-Subseq}_k(v)$?

Reduction from k-non-Universality in Bounded Range: choose v to be k-universal.

Analysis

Non-Shortest Absenst Subsequence w.r.t. Bounded Range

Input: $u, w \in \Sigma^*$ and an integer p with $|u| = m, |w| = n$ and $m \leq p \leq n$

Question: Is u a p-SAS of w?

Reduction from k-non-Universality in Bounded Range: u is a p-SAS of w if and only if u is p-absent and $p\text{-Subseq}_{k-1}(w) = \Sigma^{k-1}$.

	Unbounded	Bounded Range
k-(non-)Universality	Linear	Exponential
k-(non-)Equivalency	Linear	Exponential
Matching	Linear	Rectangular
MAS	Linear	Rectangular
SAS	Linear	Exponential

	Unbounded	Bounded Range
k-(non-)Universality	Linear	Exponential
k-(non-)Equivalency	Linear	Exponential
Matching	Linear	Rectangular
MAS	Linear	Rectangular
SAS	Linear	Exponential

Thank you for listening!