Combinatorics on words

An *alphabet* is a non-empty set of symbols Σ whose elements are called *letters* or *symbols*. We usually denote letters by the beginning of the english alphabet $a, b, c, \ldots$.

A *string* is a finite or infinite ordered list of letters in Σ, i.e., sequences of the form

$$w = (a_1, a_2, \ldots, a_n), \quad w = (a_1, a_2, \ldots).$$

We write strings w/o parenthesses and separating commas; so $w = (a_1, \ldots, a_n)$ is usually written as

$$w = a_1 a_2 \cdots a_n$$

where $a_i \in \Sigma$.

The *empty* string $()$ is denoted by $\varepsilon = ()$.

The *length* of a string is denoted by $|w|$, so if $\Sigma = \{a, b\}$, then

$$|\varepsilon| = 0, \quad |abba| = 4.$$

Also, $|w|_a$ denotes the number of occurrences

A string x is a <u>subword</u> of a string w if there are strings y, z such that

$$w = yxz .$$

Also, x is a <u>prefix</u> of w if there is a string y s.t. $w = xy$, and x is a <u>suffix</u> if there is a string z such that

$$w = zx .$$

The string x is a <u>subsequence</u> of w if x is obtained from w by eliminating one or more occurrences of various letters.

Thus abba is a subsequence of

aababba, but not of bbbaaaabaa.

If $w = a_1 a_2 \cdots a_n$, we let for $1 \le i \le j \le n$

$$w[i] = a_i$$

and

$$w[i..j] = a_i a_{i+1} \cdots a_j .$$

Finally, if Σ is an alphabet, let

Σ^* = set of all finite strings over Σ,

Σ^+ = set of all non-empty finite strings over Σ.

A <u>language</u> over Σ is any subset $L \subseteq \Sigma^*$

Given strings $w = a_1 a_2 \cdots a_n$ and $x = b_1 b_2 \cdots b_m$ we define their <u>concatenation</u> $w \circ x = wx$ by

$$wx = a_1 a_2 \cdots a_n b_1 b_2 \cdots b_m .$$

So $\varepsilon w = w = w\varepsilon$ and $(wx)y = w(xy) =: xyz$.

Note that in this way, Σ^* becomes a <u>monoid</u>, i.e., a semigroup with identity under the concatenation operation.

Note that the length function l is a morphism from $(\Sigma^*, \circ)$ to $(\mathbb{N}, +)$.

For simplicity, we also write

$$x^d = \underbrace{xxx \cdots x}_{d \text{ times}} .$$

So $x^0 = \varepsilon$.

We say that Σ^* is the <u>free</u> monoid on the set Σ.

Theorems of Lyndon - Schützenberger

Lemma Suppose $u, v, x, y \in \Sigma^*$ and $uv = xy$. Then there is some $t \in \Sigma^*$ such that

$$ut = x \quad \text{and} \quad ty = v \qquad \text{if } |u| \leq |x|$$

and

$$xt = u \quad \text{and} \quad tv = y \qquad \text{if } |x| \leq |u|.$$

The first theorem of Lyndon - Schützenberger characterizes when a string has identical non-trivial proper pre- and suffixes.

Thm Let $x, y, z \in \Sigma^+$. Then $xy = yz$ if and only if there are $u \in \Sigma^+$, $v \in \Sigma^*$ and $d \geq 0$ such that

$$x = uv, \quad z = vu,$$

$$y = (uv)^d u = u(vu)^d = \overbrace{uvuvuv \ldots uvu}^{d}$$

Proof The implication from right to left is obvious, in this case

$$xy = uv(uv)^d u = (uv)^{d+1} u = (uv)^d uvu = yz.$$

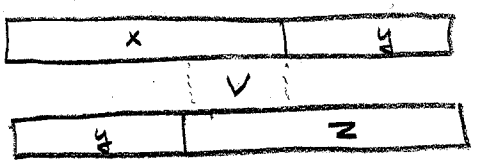
The converse direction is by induction on $|y| \geq 1$.

$|y| = 1$: Then $y = a$ for some $a \in \Sigma$ and so

$xa = az$. Then, x begins with a and z ends with a. So $x = ax'$ and $z = z'a$, whence $ax'a = az'a$, i.e., $x' = z'$. So letting $u = a$, $v = x'$ and $d = 0$ we have the result.

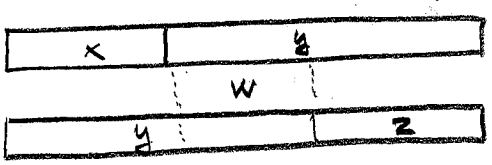
$|y| > 1$:

Case $|x| \geq |y|$:



We can find $v \in \Sigma^*$ such that $x = yv$ and $z = vy$. Letting $u = y$ and $d = 0$, we are done.

Case $|x| < |y|$:



Again take $w \in \Sigma^*$ s.t. $xw = y = wz$.

Since $|w| < |y|$, we have by induction some $u \in \Sigma^+$, $v \in \Sigma^*$ and $d \geq 0$ such that

$x = uv$, $z = vu$, $w = (uv)^d u$.

But then also $y = xw = uv(uv)^d u = (uv)^{d+1}$.

Corollary If w has a common prefix and suffix, i.e. for $x, y, z \in \Sigma^+$, $w = xy = yz$, then w has a common prefix and suffix that do not overlap in w, i.e., for some $u \in \Sigma^+$ and $v \in \Sigma^*$,

$$w = uvu.$$

Now, consider problems such as

- when can we have $x^2 = y^3$?
- when can we have $xy = yx$?

Theorem Let $x, y \in \Sigma^+$. TFAE

(i) $xy = yx$

(ii) there is $z \in \Sigma^+$ and $k, l \geq 1$ such that $x = z^k$, $y = z^l$

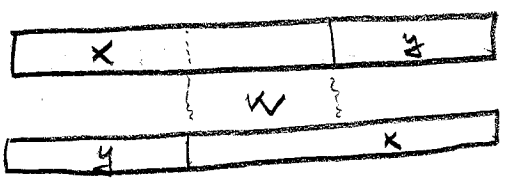
(iii) there are $i, j \geq 1$ such that $x^i = y^j$.

Proof

(i) $\Rightarrow$ (ii): By induction on $|xy|$.

$|xy| = 2$: Then there must be some $a \in \Sigma$ st. $x = y = a$, whence (ii) follows.

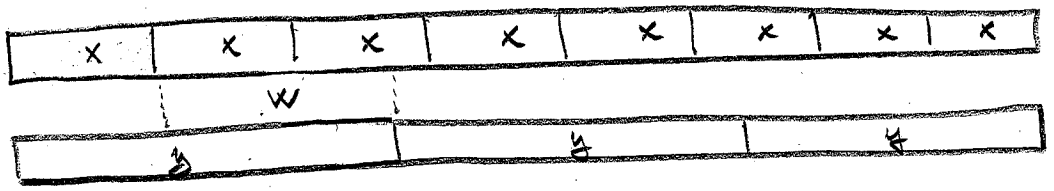
$|xy| = n$: WLOG, assume that $|x| \geq |y|$ and find w st. $yw = x$.



So $wy = yw$. If $|w| = 0$, then $x = y$ and we can take $z = x = y$. If $|w| \geq 1$, then by induction there is some z and $k, l \geq 1$ s.t. $w = z^k$, $y = z^l$, whence $x = yw = z^{l+k}$, showing (ii).

(ii) $\Rightarrow$ (iii): Suppose $z \in \Sigma^+$, $k, l \geq 1$ and $x = z^k$, $y = z^l$. Then clearly $x^l = z^{kl} = y^k$.

(iii) $\Rightarrow$ (i): Assume that for some $i, j \geq 1$, $x^i = y^j$.



WLOG, assume $|x| \leq |y|$. If $|x| = |y|$, the result is trivial, so suppose instead that $|x| < |y|$ and find $w \in \Sigma^+$ s.t. $xw = y$. Then,

$$y^j = (xw)^j = x^i = x(wx)^{j-1}w,$$ whence

$$x^{i-1} = (wx)^{j-1}w \text{ and so } x^i = (wx)^{j-1}wx = (wx)^j.$$

Then $(xw)^j = (wx)^j$ and so $xw = wx$.

It follows that

$$xy = xxw = xwx = yx.$$

□

This result leads to the following definition

Definition A string $w \in \Sigma^+$ is a **power** if there is some $z \in \Sigma^+$ and $k \geq 2$ s.t. $w = z^k$.

Non-powers are called **primitive** strings.

So "beef" is primitive, while "tartar" is a power.

Corollary Any string $w \in \Sigma^+$ can be expressed uniquely as $w = z^k$, where $k \geq 1$ and z is primitive.

Definition If L and K are two languages, let $LK = \{wx \mid w \in L \text{ \& } x \in K\}$. Similarly, let $L^0 = \{\varepsilon\}$, $L^{n+1} = LL^n$ and $L^\omega = \{w_1 w_2 w_3 \ldots \mid w_i \in L\}$.

For simplicity write

$$xL = \{xw \mid w \in L\} = \{x\}L.$$

Thus, $\emptyset L = L\emptyset = \emptyset$, while $\varepsilon L = L\varepsilon = L$.

Also, for $w \in \Sigma^+$, set $w^\omega = www\ldots$ and $w^+ = \{\ldots w^2 w^3\}$

Theorem Let $x, w \in \Sigma^+$ and $\alpha \in w\{w,x\}^\omega$, $\beta \in x\{w,x\}^\omega$. TFAE

(i) α and β agree on a prefix of length $\geq |w| + |x| - \gcd(|w|, |x|)$

(ii) $xw = wx$

(iii) $\alpha = \beta$.

Pf (iii) $\Rightarrow$ (i) is trivial.

(i) $\Rightarrow$ (ii) : Suppose $xw \neq wx$. We prove that for some $i < |w| + |x| - \gcd(|w|, |x|)$, we have $\alpha[i] \neq \beta[i]$. The proof is by induction on $|x| + |w|$.

$|x| + |w| = 2$: Then there are $a \neq b \in \Sigma$ st. $x = a$ and $w = b$, whence $\alpha[1] \neq \beta[1]$.

$|x| + |w| = n$: If $|x| = |w|$, then $x \neq w$ and so for some $i \leq |x|$, $\alpha[i] \neq \beta[i]$.

Otherwise, suppose wlog that $|x| < |w|$. If x is not a prefix of w there is $i \leq |x|$ st. $\alpha[i] \neq \beta[i]$, and if x is a prefix of w, we find some $y \in \Sigma^+$ such that $xy = w$. In the latter

$|y|$, so $\gcd(|w|, |x|) = \gcd(|y|, |w|)$.

Also, since $xxy = xw \neq wx = xyx$, we have $xy \neq yx$.

Now $\alpha = x\gamma$, $\beta = x\delta$, where

$$\gamma \in y\{x,y\}^\omega \text{ and } \delta \in x\{x,y\}^\omega,$$

so by induction there is some

$i \leq |x| + |y| - \gcd(|w|, |x|)$ so that $\gamma[i] \neq \delta[i]$,

where $j = |x| + i \leq |x| + |w| - \gcd(|w|, |x|)$ and

$\alpha[j] \neq \beta[j]$.

(ii) $\Rightarrow$ (iii): If $xw = wx$, then by the preceding theorem there is some $z \in \Sigma^+$ of which both x and w are powers. But then $\alpha = \beta$. □

Definition If $p, q \geq 0$ and $x \in \Sigma^+$ with $q \mid |x|$, we set $x^{p/q} = x^a r$, where $a = \left\lfloor \frac{p}{q} \right\rfloor$ and r is the prefix of x of length $\text{frac}\left(\frac{p}{q}\right)$.

Theorem Suppose $x, y \in \Sigma^+$ and $p, q \geq 2$ are rational numbers with $x^p = y^q$. Then $x, y \in w^+$ for some $w \in \Sigma^+$.

Proof Wlog, $|x| \geq |y|$. So x^ω and y^ω have a common prefix of length $\geq p|x| = q|y| \geq 2|x|$
$\geq |x| + |y| - \gcd(|x|, |y|)$. It follows that

$\in +$ □

Conjugates and borders

Two strings x and y are <u>conjugate</u>, written $x \sim y$, if we can write

$$x = uv \quad , \quad y = vu \quad , \quad u, v \in \Sigma^*.$$

Thus, conjugacy is an equivalence relation on Σ^* with finite equivalence classes.

<u>Theorem</u> Suppose $x \sim y$. Then x is a power if and only if y is a power, and, moreover, if $x = z^k$, then $y = w^k$ where $z \sim w$.

<u>Proof</u> Suppose $x = z^k = uv$ and $y = vu$, $k \geq 2$.

If $|z|$ divides $|u|$, then clearly $u = z^i$, $v = z^j$ for some i, j and thus $y = z^{j+i} = z^k$.

If $|z|$ does not divide $|u|$, then we can write

$u = z^i r$, $v = s z^j$, where $rs = z$ and $r, s \in \Sigma^+$. Thus,

$$y = vu = s z^j z^i r = s(rs)^{j+i} r = (sr)^{j+i+1},$$

so y is a power of sr which is a conjugate of $rs = z$. □

Definition A string w is said to be **bordered** if it can be written as $w = xyx$ where $x \in \Sigma^+$ and $y \in \Sigma^*$. Thus, w is bordered if and only if it has common nontrivial pre- and suffixes.

Theorem Let $w \in \Sigma^+$. Then w is primitive if and only if its has an unbordered conjugate.

Pf Note that by the preceding theorem, being primitive is conjugacy invariant. Also, any power is clearly bordered, showing the implication $\Leftarrow$. Conversely, suppose w is primitive.

Fix an ordering $<$ of Σ and let $<_{lex}$ be the induced lexicographical ordering of Σ^*, i.e,

$x <_{lex} y \iff$ x is a proper prefix of y

or

there are $z \in \Sigma^*$, $a, b \in \Sigma$ with

$a < b$ and

za is a prefix of x,

zb is a prefix of y.

$x \leq_{lex} y \iff x <_{lex} y$ or $x = y$.

Let now x be the lexicographically least conjugate of w and suppose towards a contradiction that x is bordered, i.e., $x = uvu$ for $u \in \Sigma^+$, $v \in \Sigma^*$.

Note now that by the choice of x,

$uvu \leq_{lex} uuv$, i.e., $vu \leq_{lex} uv$, and

$uvu \leq_{lex} vuu$, i.e., $uv \leq_{lex} vu$.

So $uv = vu$ and thus by a preceding theorem, $u = z^k$, $v = z^l$ for some $z \in \Sigma^+$.

It follows that x and thus also w is a power contradicting the assumption on w. □

Lemma Let x be a power, $x = z^k$, $k \geq 2$, $z \in \Sigma^+$, and let w be a subword of x of length $> |z|$. Then w is bordered.

Proof Exercise. □

Theorem (Lyndon - Schützenberger)

Let $x, y, z \in \Sigma^+$. Then there are $i, l, k \geq 2$ s.t.

$$x^i = y^l z^k$$

if and only if $x, y, z \in w^+$ for some $w \in \Sigma^+$,

Proof Suppose towards a contradiction that the theorem fails and let $x \in \Sigma^+$ be of minimal length such that for some $y, z \in \Sigma^+$ and $i, j, k \geq 2$ we have

$$(*) \qquad x^i = y^j z^k,$$

but for no $w \in \Sigma^+$ do we have $x, y, z \in w^+$.

By minimality, x is primitive.

Suppose first that $|y| \geq |x|$. Then there is a rational number $p \geq 2$ such that $y^p = x^p$, where $x, y \in v^+$ for some $v \in \Sigma^+$. By eliminating copies of v on both sides of $(*)$, we find that $v^n = z^k$ for some $n \geq 1$. Thus, $v, z \in w^+$ for some $w \in \Sigma^+$, whence $x, y, z \in w^+$. ↯ A similar argument applies if $|z| \geq |x|$.

So suppose instead that $|y|, |z| < |x|$.

Assume first that $i > 2$. Since x is primitive, it has an unbordered conjugate f, say $x = uv$, $f = vu$. Then $x^i = (uv)^i = u(vu)^{i-1}v = uf^{i-1}v = y^j z^k$. Since $i - 1 \geq 2$, either y^j or z^k contains f as a subword and $|f| > |y|, |z|$. By the preced-

Now, suppose instead that $i = 2$.

Then if $|y^j| = |z^k|$, we would have $x = y^j = z^k$ and thus also $x, y, z \in w^+$ for some w ↯.

So wlog suppose $|y^j| > |z^k|$. Then $y^j = xu^n$ for some primitive u and $n \geq 1$. Similarly, $u^n z^k = x$, whence $u^{2n} z^k = u^n x \sim y^j$.

It follows that $u^{2n} z^k = v^j$ for some $v \sim y$.

Since $|v| = |y| < |x|$, the minimality of x implies that for some w, $u, z, v \in w^+$.

It follows that also $x \in w^+$ and so, as before, $y \in w^+$.

□