Move-to-Front, Distance Coding and Inversion Frequencies Revisited

Travis Gagie
and
Giovanni Manzini

CPM '07
http://www.mfn.unipmn.it/~manzini/cpm07

~~Move-to-Front,~~ Distance Coding ~~and Inversion Frequencies~~ Revisited

Travis Gagie
and
Giovanni Manzini

CPM '07
http://www.mfn.unipmn.it/~manzini/cpm07

Outline

Outline

Computing the BWT

abracadabra

Computing the BWT

abracadabra$

Computing the BWT

abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$
abracadabra$

Computing the BWT

abracadabra$		abracadabra$
abracadabra$		bracadabra$a
abracadabra$		racadabra$ab
abracadabra$		acadabra$abr
abracadabra$		cadabra$abra
abracadabra$	$\Rightarrow$	adabra$abrac
abracadabra$		dabra$abraca
abracadabra$		abra$abracad
abracadabra$		bra$abracada
abracadabra$		ra$abracadab
abracadabra$		a$abracadabr
abracadabra$		$abracadabra

Computing the BWT

abracadabra$		abracadabra$		bracadabra$ a
abracadabra$		bracadabra$a		racadabra$a b
abracadabra$		racadabra$ab		abra$abraca d
abracadabra$		acadabra$abr		ra$abracada b
abracadabra$		cadabra$abra		adabra$abra c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $
abracadabra$		dabra$abraca		acadabra$ab r
abracadabra$		abra$abracad		a$abracadab r
abracadabra$		bra$abracada		dabra$abrac a
abracadabra$		ra$abracadab		bra$abracad a
abracadabra$		a$abracadabr		cadabra$abr a
abracadabra$		$abracadabra		$abracadabr a

Computing the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Rightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Computing the BWT

abracadabra$		abracadabra$		bracadabra$ a		**a**
abracadabra$		bracadabra$a		racadabra$a b		**b**
abracadabra$		racadabra$ab		abra$abraca d		**d**
abracadabra$		acadabra$abr		ra$abracada b		**b**
abracadabra$		cadabra$abra		adabra$abra c		**c**
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Rightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		**r**
abracadabra$		abra$abracad		a$abracadab r		**r**
abracadabra$		bra$abracada		dabra$abrac a		**a**
abracadabra$		ra$abracadab		bra$abracad a		**a**
abracadabra$		a$abracadabr		cadabra$abr a		**a**
abracadabra$		$abracadabra		$abracadabr a		**a**

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Rightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a			a
abracadabra$		bracadabra$a		racadabra$a b			b
abracadabra$		racadabra$ab		abra$abraca d			d
abracadabra$		acadabra$abr		ra$abracada b			b
abracadabra$		cadabra$abra		adabra$abra c			c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Rightarrow$		$
abracadabra$		dabra$abraca		acadabra$ab r			r
abracadabra$		abra$abracad		a$abracadab r			r
abracadabra$		bra$abracada		dabra$abrac a			a
abracadabra$		ra$abracadab		bra$abracad a			a
abracadabra$		a$abracadabr		cadabra$abr a			a
abracadabra$		$abracadabra		$abracadabr a			a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Inverting the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		**$abracadabra**		a

A useful property of the BWT

abracadabra$		abracadabra$		bracadabra$ a		a
abracadabra$		bracadabra$a		racadabra$a b		b
abracadabra$		racadabra$ab		abra$abraca d		d
abracadabra$		acadabra$abr		ra$abracada b		b
abracadabra$		cadabra$abra		adabra$abra c		c
abracadabra$	$\Rightarrow$	adabra$abrac	$\Rightarrow$	abracadabra $	$\Leftrightarrow$	$
abracadabra$		dabra$abraca		acadabra$ab r		r
abracadabra$		abra$abracad		a$abracadab r		r
abracadabra$		bra$abracada		dabra$abrac a		a
abracadabra$		ra$abracadab		bra$abracad a		a
abracadabra$		a$abracadabr		cadabra$abr a		a
abracadabra$		$abracadabra		$abracadabr a		a

Outline

Computing DC

abdbcrraaaa

Computing DC

abdbcrraaaa

$\Downarrow$

first character: 1

Computing DC

abdbcrraaaa

$\Downarrow$

first character: 1

other characters: 2, 3, 5, 6

Computing DC

~~abdber~~raaaa

$\Downarrow$

first character: 1

other characters: 2, 3, 5, 6

distances: 7

Computing DC

abdbcrraaaa

⇓

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2

Computing DC

abdbcrraaaa

$\Downarrow$

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

Computing DC

abdbcrraaaa

$\Downarrow$

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

Computing DC

abdbcrraaaa

⇓

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

⇓

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Inverting DC

abdbcrraaaa

$\Downarrow$

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

$\Downarrow$

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Inverting DC

abdbcrraaaa

first character: 1
other characters: 2, 3, 5, 6
distances: 7, 2, 1, 1, 1, 1, 1
last run: 4

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Inverting DC

abdbcrraaaa

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Inverting DC

abdbcrraaaa

⇕

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

⇕

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Inverting DC

abdbcrraaaa

⇕

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

⇕

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Inverting DC

abdbcrraaaa

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Inverting DC

abdbcrraaaa

⇕

first character: 1

other characters: 2, 3, 5, 6

distances: 7, 2, 1, 1, 1, 1, 1

last run: 4

⇕

1, 2, 3, 5, 6, 7, 2, 1, 1, 1, 1, 1, 4

Two useful properties of DC

1. Suppose Pfx is a prefix-free code for the positive integers such that $|\text{Pfx}(i)| \leq a \log i + b$ for some constants a and b; then

$$|\text{Pfx}(\text{DC}(s))| \leq a\left(H_0(s)n + \log n\right) + b|\text{DC}(s)| + g$$

   for some g depending only on the alphabet size.
2. $|\text{DC}(s)|$ is the number of maximal runs in s plus a term depending only on the alphabet size.

Outline

Manzini, 2001

Theorem (Man01)

For any string s of length n and $k \geq 0$,

$$|\mathrm{BW0}(s)| \leq 8H_k(s)n + (0.08 + \mu)n + (1 + o(1)) \log n + g_k$$

for some g_k depending only on the alphabet size and k.

Manzini, 2001

Theorem (Man01)

For any string s of length n and $k \geq 0$,

$$|\mathrm{BW0}_{\mathrm{RL}}(s)| \leq (5 + 3\mu)H_k^*(s)n + (1 + o(1)) \log n + g_k$$

for some g_k depending only on the alphabet size and k.

Kaplan, Landau and Verbin, 2006

Theorem (KLV06)

For any string s of length n, $k \geq 0$ and $\lambda > 1$,

$$|\mathrm{BW0}(s)| \leq \lambda H_k(s) n + \left(\log \zeta(\lambda) + \mu\right) n + (1 + o(1)) \log n + \lambda g_k$$

for some g_k depending only on the alphabet size and k.

Kaplan, Landau and Verbin, 2006

Theorem (KLV06)

For any string s of length n and $k \geq 0$,

$$|\mathrm{BW0}(s)| \leq 4.45 H_k(s) n + (0.08 + \mu) n + (1 + o(1)) \log n + g_k$$

for some g_k depending only on the alphabet size and k.

Kaplan, Landau and Verbin, 2006

Theorem (KLV06)

For any string s of length n, $k \geq 0$ and $\lambda > 1$,

$$|\text{Order0}(\text{DC}(\text{BWT}(s)))| \leq \lambda H_k(s)n + \left(\log(\zeta(\lambda) - 1) + \mu\right) n + (1 + o(1)) \log n + \lambda g_k$$

for some g_k depending only on the alphabet size and k.

Kaplan, Landau and Verbin, 2006

Theorem (KLV06)

For any string s of length n and $k \geq 0$,

$$|\text{Order0}(\text{DC}(\text{BWT}(s)))| \leq 1.73 H_k(s) n + \mu n + (1 + o(1)) \log n + g_k$$

for some g_k depending only on the alphabet size and k.

Outline

H, the empirical entropy

Definition (see Man01 for details)

The kth-order empirical entropy of a string s of length n is

$$H_k(s) = \begin{cases} (1/n) \sum_i n_i \log(n/n_i) & \text{if } k = 0, \\ (1/n) \sum_{|w|=k} H_0(w_s)|w_s| & \text{if } k \geq 1. \end{cases}$$

H, the empirical entropy

Example

$$H_0(\text{abracadabra}) = \frac{1}{11}\left(5\log\frac{11}{5} + 2\log\frac{11}{2} + \log 11 + \log 11 + 2\log\frac{11}{2}\right)$$

H, the empirical entropy

Example

$$H_0(\text{abracadabra}) \approx 2.04\,,$$

$$H_1(\text{abracadabra}) = \frac{1}{11}\left(4H_0(\text{bcdb}) + 2H_0(\text{rr}) + H_0(\text{a}) + H_0(\text{a}) + 2H_0(\text{aa})\right)$$

H, the empirical entropy

Example

$$
\begin{aligned}
H_0(\text{abracadabra}) &\approx 2.04\,, \\
H_1(\text{abracadabra}) &\approx 0.55\,, \\
H_2(\text{abracadabra}) &= 0\,.
\end{aligned}
$$

H, the empirical entropy

Example

$$\begin{aligned} H_0(\text{abracadabra}) &\approx 2.04\,, \\ H_1(\text{abracadabra}) &\approx 0.55\,, \\ H_2(\text{abracadabra}) &= 0\,. \end{aligned}$$

Example

$$H_0(\mathrm{a}^n) = 0\,.$$

H^*, the modified empirical entropy

Definition (see Man01 for details)

The 0th-order modified empirical entropy of a string s of length n is

$$H_0^*(s) = \begin{cases} 0 & \text{if } n = 0, \\ (\lfloor \log n \rfloor + 1)/n & \text{if } n \geq 1 \text{ and } H_0(s) = 0, \\ H_0(s) & \text{otherwise.} \end{cases}$$

Low-entropy strings

Usually, μn is a minor concern and $O(\log n)$ is negligible — but not always:

Low-entropy strings

Usually, μn is a minor concern and $O(\log n)$ is negligible — but not always:

- μn can be exponentially larger than $H_k^*(s)n$,

Low-entropy strings

Usually, μn is a minor concern and $O(\log n)$ is negligible — but not always:

- μn can be exponentially larger than $H_k^*(s)n$,
- the constant hidden in the $O(\log n)$ term is often about h^{k+1}

Low-entropy strings

Usually, μn is a minor concern and $O(\log n)$ is negligible — but not always:

- μn can be exponentially larger than $H_k^*(s)n$,
- the constant hidden in the $O(\log n)$ term is often about h^{k+1} (e.g., for 4th-order compression of ASCII, about 2^{40}).

Low-entropy strings

Usually, μn is a minor concern and $O(\log n)$ is negligible — but not always:

- μn can be exponentially larger than $H_k^*(s)n$,
- the constant hidden in the $O(\log n)$ term is often about h^{k+1} (e.g., for 4th-order compression of ASCII, about 2^{40}).

We want a *small* ratio.

Outline

Distance Coding's output

Lemma (following KLV06)

For any string s of length n and $\lambda > 1$,

$$\begin{aligned} & H_0(\mathrm{DC}(s))|\mathrm{DC}(s)| \\ & \quad \leq \quad (\lambda + 0.01)\left(H_0(s)n + \log n\right) + \log(\zeta(\lambda) - 1)|\mathrm{DC}(s)| + g \end{aligned}$$

for some g depending only on the alphabet size.

Outline

The maximal runs in a string

Lemma (MN05)

The number of maximal runs in a string s of length n is at most $H_0(s)n + 1$.

Corollaries for DC

Corollary

For any string s of length n and $\lambda > 1$,

$$\begin{aligned} & H_0(\mathrm{DC}(s))|\mathrm{DC}(s)| \\ \leq \quad & \left(\lambda + \log(\zeta(\lambda) - 1) + 0.01\right)\left(H_0(s)n + \log n\right) + g \end{aligned}$$

for some g depending only on the alphabet size.

Corollaries for DC

Corollary

For any string s of length n,

$$H_0(\mathrm{DC}(s))|\mathrm{DC}(s)| \leq 1.74\left(H_0(s)n + \log n\right) + g$$

for some g depending only on the alphabet size.

Corollaries for DC

Corollary

For any string s of length n and $k \geq 0$,

$$|\text{Order0}(\text{DC}(\text{BWT}(s)))| \leq (1.74 + \mu)H_k(s)n + O(\log n)$$

with the hidden coefficient depending only on the alphabet size and k.

Outline

Local optimality

Definition (Man01)

A compression algorithm A is *locally λ-optimal* if, for any string s and partition $s_1, \ldots, s_t$ of s,

$$|A(s)| \leq \lambda \sum_{i=1}^{t} H_0^*(s_i)|s_i| + gt$$

for some g depending only on the alphabet size.

Local optimality

Lemma (Man01)

If A is locally λ-optimal then, for any string s of length n and $k \geq 0$,

$$|A(\mathrm{BWT}(s))| \leq \lambda H_k^*(s)n + (1 + o(1)) \log n + g_k$$

for some g_k depending only on the alphabet size and k.

Outline

DC + Pfx is *not* locally optimal

Example

Consider $s = s_1 s_2 s_3$ with $s_1 = s_3 = a_1 \dots a_{h-1}$ and $s_2 = a_h^n$:

$$\sum_{i=1}^{3} H_0^*(s_i)|s_i| = \log n + O(h \log h)$$

but $|\mathrm{Pfx}(\mathrm{DC}(s))| > (h-1)\log n$.

Avoiding long jumps

To use escapes:

1. whenever DC would write 1, we write $\langle 1, 1 \rangle$;
2. whenever DC would write a distance, we compare the cost of writing the distance to the cost of writing an escape and a re-entry, and do whichever is cheaper;
3. an escape is $\langle 1, 1 \rangle$;
4. a re-entry is $\langle 1, 2, \ell, a_i \rangle$.

Computing DC_{ESC}

abdbcrraaaa

$\Downarrow$

first character: $\langle 1, 1 \rangle$

other characters: 2, 3, 5, 6

Computing DC_{ESC}

~~abdber~~raaaa

$\Downarrow$

first character: $\langle 1, 1 \rangle$

other characters: 2, 3, 5, 6

distances: $\langle 1, 1 \rangle$

Computing DC_{ESC}

abdbcrraaaa

$\Downarrow$

first character: $\langle 1, 1 \rangle$

other characters: 2, 3, 5, 6

distances: $\langle 1, 1 \rangle, 2, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle$

Computing DC_{ESC}

abdbcrraaaa

$\Downarrow$

first character: $\langle 1, 1 \rangle$

other characters: 2, 3, 5, 6

distances: $\langle 1, 1 \rangle, 2, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 2, 2, 1 \rangle$

Computing DC_{ESC}

abdbcrraaaa

$\Downarrow$

first character: $\langle 1, 1 \rangle$

other characters: 2, 3, 5, 6

distances: $\langle 1, 1 \rangle, 2, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 2, 2, 1 \rangle$

last run: 4

Computing DC_{ESC}

abdbcrraaaa

$\Downarrow$

first character: $\langle 1, 1 \rangle$

other characters: 2, 3, 5, 6

distances: $\langle 1, 1 \rangle, 2, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 1 \rangle, \langle 1, 2, 2, 1 \rangle$

last run: 4

$\Downarrow$

1, 1, 2, 3, 5, 6, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 4

Outline

A stepping-stone

Consider the algorithm $\mathrm{DC}^*_{\mathrm{ESC}}$ that takes a partition $s_1, \ldots, s_t$ of s as an argument — whereas $\mathrm{DC}_{\mathrm{ESC}}$ takes *only* s — and uses escapes and re-entries only to replace distances that cross between partition elements.

$$\begin{aligned} & |\mathrm{Order0}(\mathrm{DC}_{\mathrm{ESC}}(s))| \\ \leq \quad & |\mathrm{Order0}(\mathrm{DC}^*_{\mathrm{ESC}}(s_1, \ldots, s_t))| \end{aligned}$$

A stepping-stone

Consider the algorithm $\mathrm{DC}^*_{\mathrm{ESC}}$ that takes a partition $s_1, \ldots, s_t$ of s as an argument — whereas $\mathrm{DC}_{\mathrm{ESC}}$ takes *only* s — and uses escapes and re-entries only to replace distances that cross between partition elements.

$$
\begin{aligned}
&|\mathrm{Order0}(\mathrm{DC}_{\mathrm{ESC}}(s))| \\
&\quad\leq \quad |\mathrm{Order0}(\mathrm{DC}^*_{\mathrm{ESC}}(s_1, \ldots, s_t))| \\
&\quad\leq \quad \sum_{i=1}^{t} |\mathrm{Order0}(\mathrm{DC}(s_i))| + gt \\
&\quad\leq \quad \sum_{i=1}^{t} \Big(\big(\lambda + 0.01\big) \big(H_0(s_i)|s_i| + \log n\big) + \\
&\qquad\qquad \big(\log(\zeta(\lambda) - 1) + \mu\big) |\mathrm{DC}(s_i)| \Big) + gt\,.
\end{aligned}
$$

Another lemma

Lemma

For any string s of length n and positive constants a, b and ϵ,

$$a(H_0(s)n+\log n)+b|\mathrm{DC}(s)|+O(1) \leq \max(2a, a+b+\epsilon)H_0^*(s)n+O(1)\,.$$

$\mathrm{DC}_{\mathrm{ESC}}$ + Order0 is locally optimal

Lemma

For any string s and partition $s_1, \ldots, s_t$ of s,

$$|\mathrm{Order0}(\mathrm{DC}_{\mathrm{ESC}}(s))| \leq (2.7 + \mu) \sum_{i=1}^{t} H_0^*(s_i)|s_i| + gt$$

for some g depending only on the alphabet size.

Outline

An entropy-only bound

Theorem

For any string s of length n and $k \geq 0$,

$$|\mathsf{Order0}(\mathsf{DC}_{\mathsf{ESC}}(\mathsf{BWT}(s)))| \leq (2.7+\mu)H_k^*(s)n+(1+o(1))\log n+g_k$$

for some g_k depending only on the alphabet size and k.