The k-Sensitivity of LZ77

Gabriel Bathie Paul Huber Guillaume Lagarde Akka Zemmari

LaBRI, University of Bordeaux

LZ77 Parsing

Built greedily: each block = longest match in the already-parsed prefix + one new character.

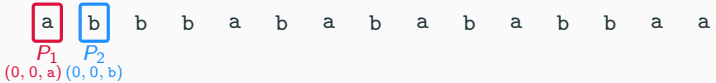
a b b b a b a b a b a b b a a

LZ77 Parsing

Built greedily: each block = longest match in the already-parsed prefix + one new character.

LZ77 Parsing

Built greedily: each block = longest match in the already-parsed prefix + one new character.

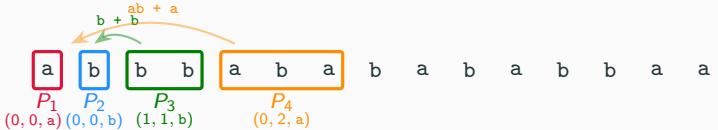


LZ77 Parsing

Built greedily: each block = longest match in the already-parsed prefix + one new character.

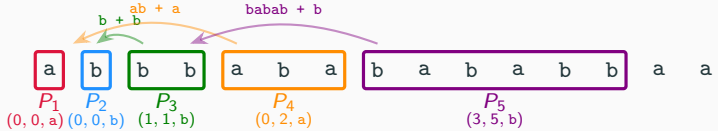
LZ77 Parsing

Built greedily: each block = longest match in the already-parsed prefix + one new character.



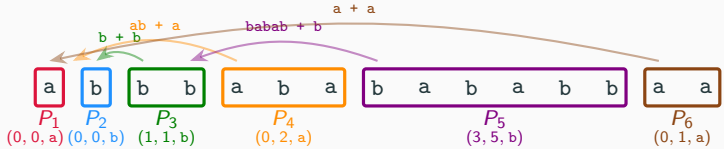
LZ77 Parsing

Built greedily: each block = longest match in the already-parsed prefix + one new character.



LZ77 Parsing

Built greedily: each block = longest match in the already-parsed prefix + one new character.



Blocks: $\mathcal{D} = \{\mathtt{a},\ \mathtt{b},\ \mathtt{bb},\ \mathtt{aba},\ \mathtt{bababb},\ \mathtt{aa}\}$ $\qquad C_{\mathrm{LZ77}}(w) = |\mathcal{D}| = 6.$

Size of the compression: $\Theta\big(C_{\mathrm{LZ77}}(w) \cdot \log|w|\big)$

Ziv-Lempel algorithms

A family of lossless data compression algorithms (LZ77, LZ78, LZW, ...) based on dictionary-based encoding.

In practice.

Used in many places:

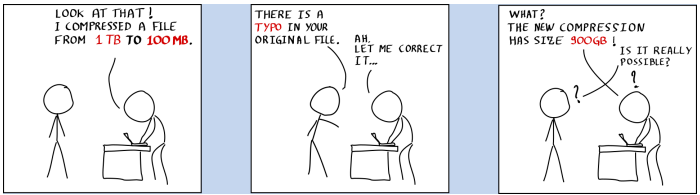
- deflate (gzip)
- PNG, GIF, TIFF images
- compress (Unix)
- DNA compression
- ...

In theory.

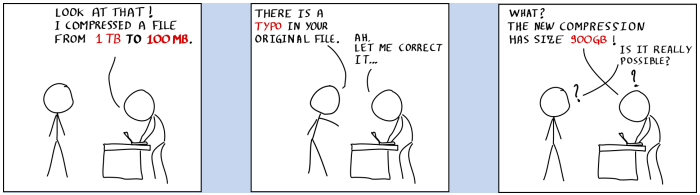
Connections with:

- Entropy of a random source
- Lyndon factorization
- Absolutely normal numbers
- Effective Hausdorff dimension
- ...

A strange scenario...



A strange scenario...



The one-bit catastrophe in LZ78

on infinite words

w — compressible ...

↓ **one bit**

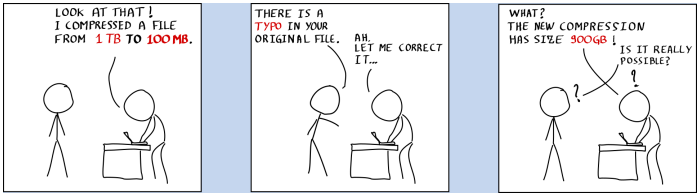
$b \cdot w$ — incompressible ...

Theorem [**L.**, Perifel '18]

$\exists\, w \in \{0,1\}^\omega$ such that

$$\rho_{\sup}(w) = 0 \qquad \rho_{\inf}(0 \cdot w) \geq \frac{1}{6075}$$

A strange scenario...



The one-bit catastrophe in LZ78

on finite words

w — $C_{LZ78}(w) = \sqrt{n}$

↓ **one bit**

$b \cdot w$ — $C_{LZ78}(b \cdot w) = n^{2/3}$

On a **1 GB** file:

$\sqrt{n} \cdot \log n \approx$ **350 KB** compressed

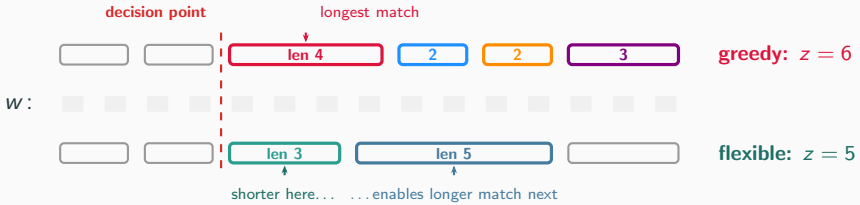
$n^{2/3} \cdot \log n \approx$ **16 MB** compressed

One bit: 350 KB $\rightarrow$ 16 MB.

Programmers knew it

Before "catastrophe" theorems, programmers started developing workarounds.

Flexible parsing:



- **Matias–Sahinalp** (SODA'99): one-step lookahead for LZ78; up to 35% gain on DNA/medical data.
- **Aronica et al.** (TCS'18), **Langiu** (PhD'12): flexible parsing minimizes the number of blocks against all LZ-like factorizations.

The LZ-miracle?

Sensitivity = an opportunity.

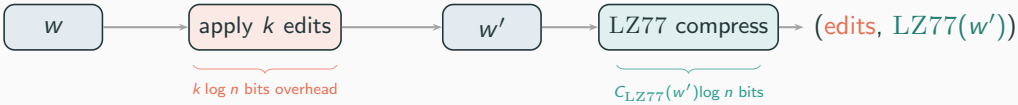
If an edit can *increase* the parse, a *clever* edit might decrease it.

The LZ-miracle?

Sensitivity = an opportunity.

If an edit can *increase* the parse, a *clever* edit might decrease it.

Idea: to compress w, first modify it, then compress.

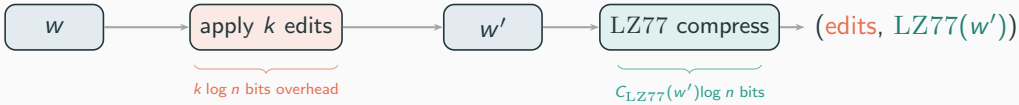


The LZ-miracle?

Sensitivity = an opportunity.

If an edit can *increase* the parse, a *clever* edit might decrease it.

Idea: to compress w, first modify it, then compress.



Goal
Find $w' \in B(w,k)$ such that $k \log n + C_{\mathrm{LZ77}}(w') \log n < C_{\mathrm{LZ77}}(w) \log n$.

High-Level goals

We focus on LZ77. Two main objectives.

1. **Understand the sensitivity** under k edits.
2. **Find efficient algorithms** to leverage this sensitivity.

The k-sensitivity of LZ77
Upper and lower bounds

Previously known: one edit

Akagi–Funakoshi–Inenaga '23
For all $w \in \Sigma^*$ and $w' \in B(w, 1)$,

$$C_{\mathrm{LZ77}}(w') \;\leq\; 2\, C_{\mathrm{LZ77}}(w).$$

Moreover, this is tight.

Previously known: one edit

Akagi–Funakoshi–Inenaga '23
For all $w \in \Sigma^*$ and $w' \in B(w, 1)$,

$$C_{\mathrm{LZ77}}(w') \;\leq\; 2\, C_{\mathrm{LZ77}}(w).$$

Moreover, this is tight.

Hence we can go from a compression C to $C/2$ with one edit.

Previously known: one edit

Akagi–Funakoshi–Inenaga '23
For all $w \in \Sigma^*$ and $w' \in B(w, 1)$,

$$C_{\mathrm{LZ77}}(w') \;\leq\; 2\, C_{\mathrm{LZ77}}(w).$$

Moreover, this is tight.

Hence we can go from a compression C to $C/2$ with one edit.

Naive extension to k edits: apply the bound k times:

$$C_{\mathrm{LZ77}}(w') \;\leq\; 2^k \cdot C_{\mathrm{LZ77}}(w).$$

Previously known: one edit

Akagi–Funakoshi–Inenaga '23
For all $w \in \Sigma^*$ and $w' \in B(w, 1)$,

$$C_{\mathrm{LZ77}}(w') \;\leq\; 2\, C_{\mathrm{LZ77}}(w).$$

Moreover, this is tight.

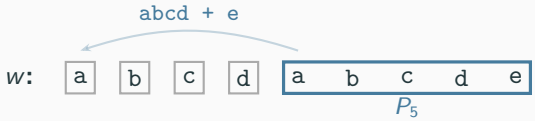
Hence we can go from a compression C to $C/2$ with one edit.

Naive extension to k edits: apply the bound k times:

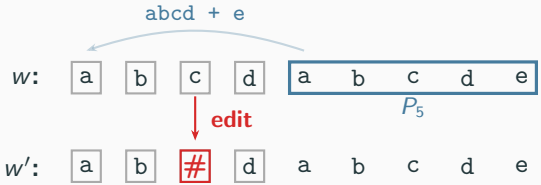
$$C_{\mathrm{LZ77}}(w') \;\leq\; 2^k \cdot C_{\mathrm{LZ77}}(w).$$

Exponential in k! Can we do better?

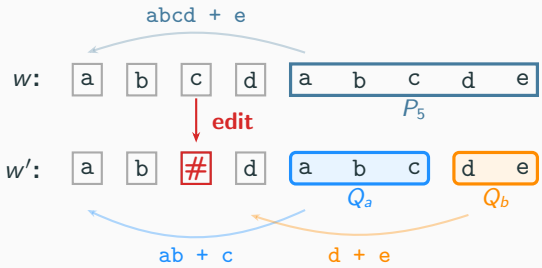
Splits



Splits

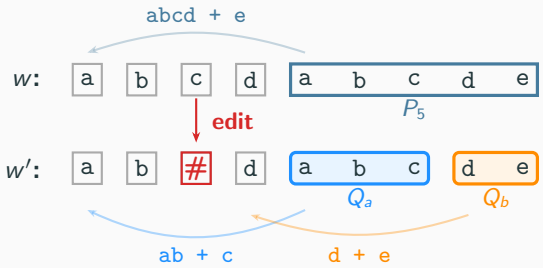


Splits



One edit in the reference of block P_j $\Rightarrow$ P_j splits into ≤ 2 pieces.

Splits



One edit in the reference of block P_j $\Rightarrow$ P_j splits into ≤ 2 pieces.

More generally, k edits in one reference block $\Rightarrow \leq k+1$ fragments (not 2^k).

Natural conjecture: k edits, each block splits into $\leq k+1$ pieces,
so $C_{\mathrm{LZ77}}(w') \leq (k+1) \cdot C_{\mathrm{LZ77}}(w)$. Is this bound tight?

First theorem: general bounds

Slightly counter-intuitive: the multiplicative factor is a constant, not linear in k.

Theorem 1
For all $k \in \mathbb{N}$, $w \in \Sigma^*$, and $w' \in B(w, k)$,

$$\boxed{C_{\mathrm{LZ77}}(w') \leq 3 \cdot C_{\mathrm{LZ77}}(w) + 4k.}$$

First theorem: general bounds

Slightly counter-intuitive: the multiplicative factor is a constant, not linear in k.

Theorem 1
For all $k \in \mathbb{N}$, $w \in \Sigma^*$, and $w' \in B(w, k)$,

$$\boxed{C_{\mathrm{LZ77}}(w') \leq 3 \cdot C_{\mathrm{LZ77}}(w) + 4k.}$$

Proposition 2
For all $k \in \mathbb{N}$ and $\varepsilon > 0$, there exist w, w' with $w' \in B(w, k)$ such that

$$C_{\mathrm{LZ77}}(w') \geq (3 \cdot C_{\mathrm{LZ77}}(w) + k)(1 - \varepsilon).$$

First theorem: general bounds

Slightly counter-intuitive: the multiplicative factor is a constant, not linear in k.

Theorem 1
For all $k \in \mathbb{N}$, $w \in \Sigma^*$, and $w' \in B(w, k)$,

$$\boxed{C_{\mathrm{LZ77}}(w') \leq 3 \cdot C_{\mathrm{LZ77}}(w) + 4k.}$$

Proposition 2
For all $k \in \mathbb{N}$ and $\varepsilon > 0$, there exist w, w' with $w' \in B(w, k)$ such that

$$C_{\mathrm{LZ77}}(w') \geq (3 \cdot C_{\mathrm{LZ77}}(w) + k)(1 - \varepsilon).$$

Hence we can go from a compression C to $\frac{C-k}{3}$ with k edits.

Second theorem: a fine-grained trichotomy

Recall: $C_{\mathrm{LZ77}}(w') \leq 3 \cdot C_{\mathrm{LZ77}}(w) + 4k$. Can we refine the constant 3?

Second theorem: a fine-grained trichotomy

Recall: $C_{\mathrm{LZ77}}(w') \leq 3 \cdot C_{\mathrm{LZ77}}(w) + 4k$. Can we refine the constant 3?

Yes, it depends on how compressible w is.

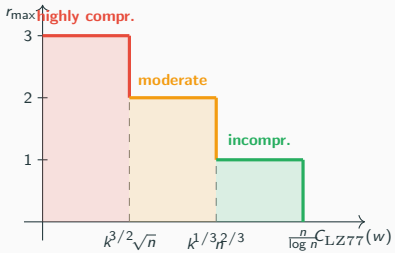
Second theorem: a fine-grained trichotomy

Recall: $C_{\mathrm{LZ77}}(w') \leq 3 \cdot C_{\mathrm{LZ77}}(w) + 4k$. Can we refine the constant 3?

Yes, it depends on how compressible w is.

Theorem 3 (Trichotomy)
Let $r = C_{\mathrm{LZ77}}(w')/C_{\mathrm{LZ77}}(w)$. *For any* $0 < \eta < 1$*:*

$$\begin{cases} C_{\mathrm{LZ77}}(w) > \dfrac{8\, k^{\frac{1}{3}} n^{\frac{2}{3}}}{\eta} & \implies r < 1 + \eta \\ C_{\mathrm{LZ77}}(w) > \dfrac{k^{\frac{3}{2}}\sqrt{n}}{\eta\sqrt{2}} & \implies r < 2 + \eta \\ \textit{otherwise} & \implies r \leq 3 \end{cases}$$



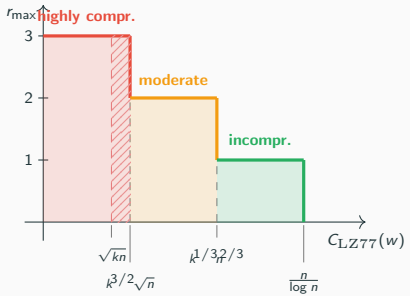
The more compressible a string, the more sensitive its $\mathrm{LZ77}$ parse can be.

Matching Lower Bounds for the Trichotomy

All three regimes are essentially tight (up to lower-order terms):

- **Incompressible** ($r \approx 1$): tight.
- **Moderate** ($r \approx 2$): tight.
- **Highly compressible** ($r \approx 3$): gap between upper bound at $k^{3/2}\sqrt{n}$ and lower bound at $\sqrt{kn}$.

Open question: close the gap in the highly compressible regime.



Algorithmic Results

The k-Modifications Problem

Input: word $w \in \{0,1\}^n$, budget $k \in \mathbb{N}$, target ratio $\rho \in [0,1]$.
Output: A word $w' \in B(w,k)$ such that $C_{\mathrm{LZ77}}(w')/C_{\mathrm{LZ77}}(w) \leq \rho$, if such a word exists; otherwise, report that no such word exists.

This is the problem we would like to solve.

The k-Modifications Problem

Input: word $w \in \{0,1\}^n$, budget $k \in \mathbb{N}$, proportion $p \in [0,1]$.
Output: A word $w' \in B(w,k)$ such that at least a fraction p of the blocks in $\mathrm{LZ77}(w')$ each contain at least 3 block-starts of $\mathrm{LZ77}(w)$, if such a word exists; otherwise, report that no such word exists.

This is a proxy for finding a w' that minimizes $C_{\mathrm{LZ77}}(w')/C_{\mathrm{LZ77}}(w)$.

Naive Approach: Brute force all $\binom{n}{k}$ edit positions $\Rightarrow$ $\mathcal{O}(n^{k+1})$.

The k-Modifications Problem

Input: word $w \in \{0,1\}^n$, budget $k \in \mathbb{N}$, proportion $p \in [0,1]$.
Output: A word $w' \in B(w,k)$ such that at least a fraction p of the blocks in $\mathrm{LZ77}(w')$ each contain at least 3 block-starts of $\mathrm{LZ77}(w)$, if such a word exists; otherwise, report that no such word exists.

This is a proxy for finding a w' that minimizes $C_{\mathrm{LZ77}}(w')/C_{\mathrm{LZ77}}(w)$.

Naive Approach: Brute force all $\binom{n}{k}$ edit positions $\Rightarrow$ $\mathcal{O}(n^{k+1})$.

Theorem 4
k-Modifications *is solvable exactly for* $k = 2$ *in time* $\mathcal{O}(n^2/p)$.

The k-Modifications Problem

Input: word $w \in \{0,1\}^n$, budget $k \in \mathbb{N}$, proportion $p \in [0,1]$.
Output: A word $w' \in B(w,k)$ such that at least a fraction p of the blocks in $\mathrm{LZ77}(w')$ each contain at least 3 block-starts of $\mathrm{LZ77}(w)$, if such a word exists; otherwise, report that no such word exists.

This is a proxy for finding a w' that minimizes $C_{\mathrm{LZ77}}(w')/C_{\mathrm{LZ77}}(w)$.

Naive Approach: Brute force all $\binom{n}{k}$ edit positions $\Rightarrow$ $\mathcal{O}(n^{k+1})$.

Theorem 4
k-Modifications is solvable exactly for $k = 2$ in time $\mathcal{O}(n^2/p)$.

Theorem 5
For fixed k, a deterministic ε-approximation runs in time:

$$\mathcal{O}\Big(n^{\lceil 2k/3\rceil+1} \cdot f(k,\varepsilon)\Big), \quad \text{where} \quad f(k,\varepsilon) = \mathcal{O}\left(\left(\frac{k^3}{\varepsilon}\right)^{2k/3}\right).$$

That's all, folks! Thank you

- **Close the gap in the third regime:**
  - Upper bound: $C_{\mathrm{LZ77}}(w) = \Theta(k^{3/2}\sqrt{n})$ vs Lower bound: $C_{\mathrm{LZ77}}(w) = \Theta(\sqrt{kn})$.
- **Complexity of k-Modifications:**
  - Our algorithms are unlikely to be optimal. How to improve?
  - NP-hard?
  - W[1]-hard parameterized by k?
- **What about other compression schemes?**
  - In particular, **LZ78** is known to have catastrophes, can we leverage them for better algorithms?

Thank you!