Upper Bounds on Distinct Maximal (Sub-)Repetitions in Compressed Strings

Julian Pape-Lange

18.08.2021

TECHNISCHE UNIVERSITÄT
CHEMNITZ

A *repetition* is a substring which is at least 2 times as long as its minimum period.

A *repetition* is a substring which is at least 2 times as long as its minimum period.

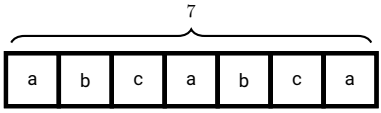


Figure: A repetition.

A *repetition* is a substring which is at least 2 times as long as its minimum period.

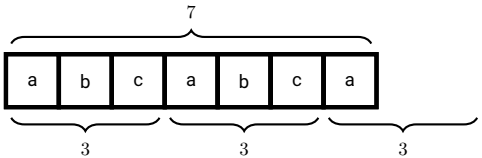


Figure: A repetition.

A repetition is *maximal* if its minimum period cannot be extended.

A repetition is *maximal* if its minimum period cannot be extended.

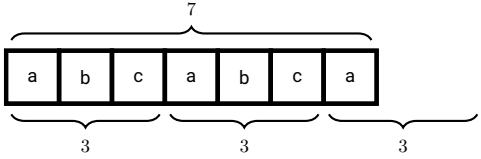


Figure: A maximal repetition.

A repetition is *maximal* if its minimum period cannot be extended.

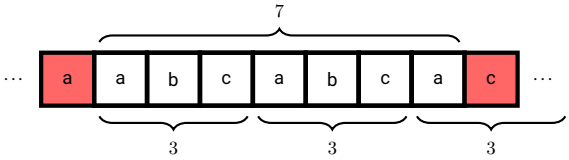


Figure: A maximal repetition.

A repetition is *maximal* if its minimum period cannot be extended.

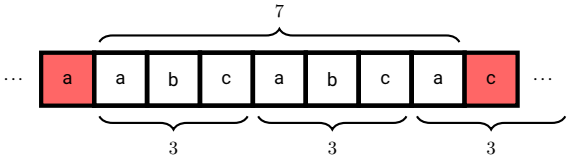


Figure: A maximal repetition.

In a string S, there are at most $|S|$ maximal repetitions.

Each LZ77-factor is

Each LZ77-factor is

- a single character or

Each LZ77-factor is

- a single character or
- a substring with an earlier occurrence.

Each LZ77-factor is

- a single character or
- a substring with an earlier occurrence.

Figure: The LZ77-decomposition of $banana$.

Each LZ77-factor is

- a single character or
- a substring with an earlier occurrence.

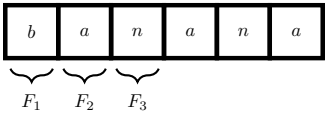


Figure: The LZ77-decomposition of $banana$.

Each LZ77-factor is

- a single character or
- a substring with an earlier occurrence.

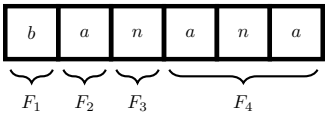


Figure: The LZ77-decomposition of $banana$.

Each LZ77-factor is

- a single character or
- a substring with an earlier occurrence.

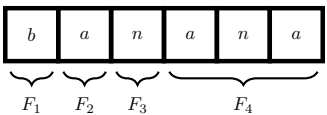


Figure: The LZ77-decomposition of $banana$.

Hope: If a string compressible, it contains few *distinct* maximal repetitions.

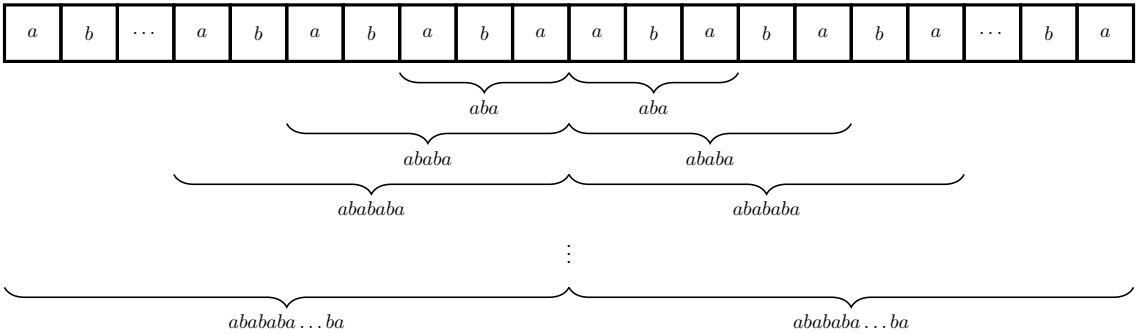


Figure: The string $((ab)^k a)^2$ with its k distinct maximal repetitions with exponent 2.

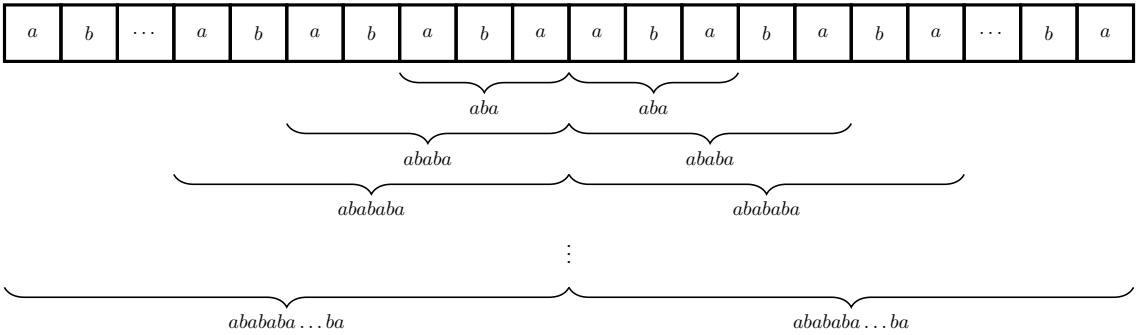


Figure: The string $((ab)^k a)^2$ with its k distinct maximal repetitions with exponent 2.

$$((ab)^k a)^2 = a \cdot b \cdot (ab)^{k-\frac{1}{2}} \cdot (ab)^{k+\frac{1}{2}}$$

A *δ-repetition* is a substring which is at least $2 + \delta$ times as long as its minimum period.

A *δ-repetition* is a substring which is at least $2 + \delta$ times as long as its minimum period.

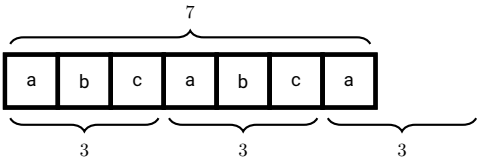


Figure: A 0.3-repetition.

A *δ-repetition* is a substring which is at least $2 + \delta$ times as long as its minimum period.

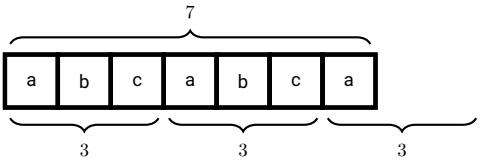


Figure: A 0.3-repetition.

$$\frac{7}{3} \approx 2 + 0.33$$

A δ-repetition is *maximal* if its minimum period cannot be extended.

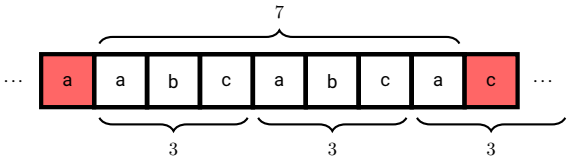


Figure: A maximal 0.3-repetition.

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Components of the proof:

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Components of the proof:

- For any index t, there are at most $\left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$ maximal δ-repetitions whose extensions contain $t-1$ and t.

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Components of the proof:

- For any index t, there are at most $\left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$ maximal δ-repetitions whose extensions contain $t-1$ and t.
- Each maximal δ-repetitions has an occurrence whose extensions crosses LZ77-factors.

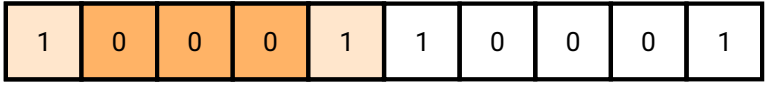


Figure: The string “1000110001”.

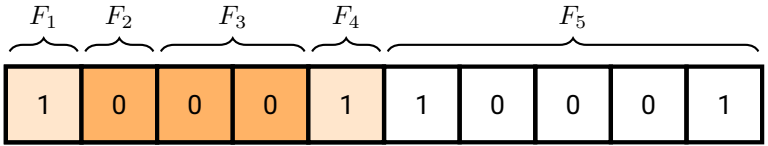


Figure: The string “1000110001”.

Lemma

Let $S[s_1..e_1]$ and $S[s_2..e_2]$ be two maximal δ-repetitions with minimum periods p_1 and p_2.
There is no number L such that $p_1, p_2 \in \left[L, (1+\frac{\delta}{4})L\right)$ and such that the intersection of $S[s_1..e_1]$ and $S[s_2..e_2]$ contains at least $\left(2+\frac{\delta}{2}\right)L$ characters.

Lemma

Let $S[s_1..e_1]$ and $S[s_2..e_2]$ be two maximal δ-repetitions with minimum periods p_1 and p_2.
There is no number L such that $p_1, p_2 \in \left[L, (1+\frac{\delta}{4})L\right)$ and such that the intersection of $S[s_1..e_1]$ and $S[s_2..e_2]$ contains at least $\left(2+\frac{\delta}{2}\right)L$ characters.

Proof.

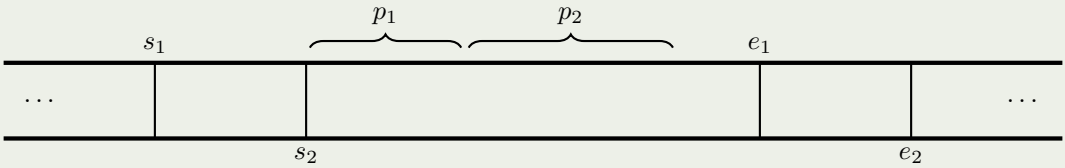
Assume, there is such a number L.

Lemma

Let $S[s_1..e_1]$ and $S[s_2..e_2]$ be two maximal δ-repetitions with minimum periods p_1 and p_2. There is no number L such that $p_1, p_2 \in \left[L, (1+\frac{\delta}{4})L\right)$ and such that the intersection of $S[s_1..e_1]$ and $S[s_2..e_2]$ contains at least $\left(2+\frac{\delta}{2}\right)L$ characters.

Proof.

Assume, there is such a number L.

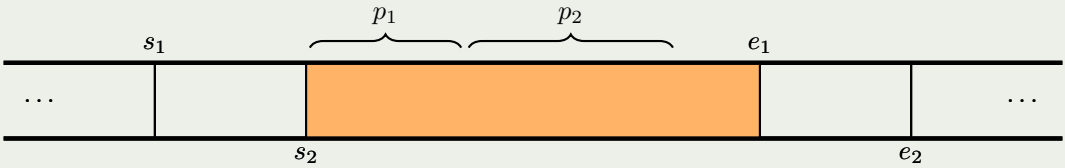


Lemma

Let $S[s_1..e_1]$ and $S[s_2..e_2]$ be two maximal δ-repetitions with minimum periods p_1 and p_2. There is no number L such that $p_1, p_2 \in \left[L, (1+\frac{\delta}{4})L\right)$ and such that the intersection of $S[s_1..e_1]$ and $S[s_2..e_2]$ contains at least $\left(2+\frac{\delta}{2}\right)L$ characters.

Proof.

Assume, there is such a number L.

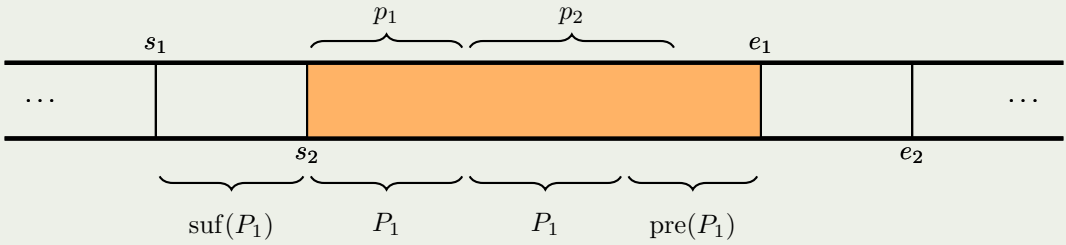


Lemma

Let $S[s_1..e_1]$ and $S[s_2..e_2]$ be two maximal δ-repetitions with minimum periods p_1 and p_2. There is no number L such that $p_1, p_2 \in \left[L, (1+\frac{\delta}{4})L\right)$ and such that the intersection of $S[s_1..e_1]$ and $S[s_2..e_2]$ contains at least $\left(2+\frac{\delta}{2}\right)L$ characters.

Proof.

Assume, there is such a number L.

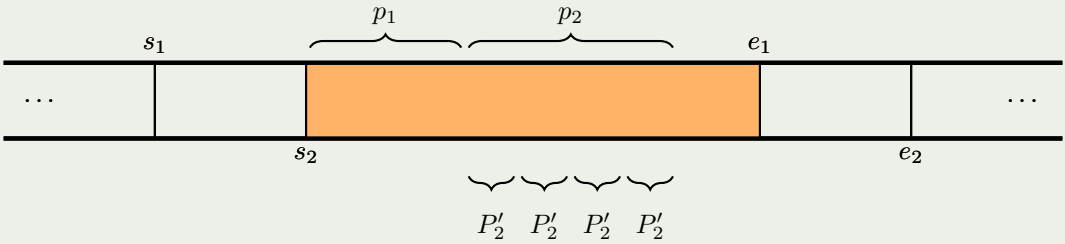


Lemma

Let $S[s_1..e_1]$ and $S[s_2..e_2]$ be two maximal δ-repetitions with minimum periods p_1 and p_2. There is no number L such that $p_1, p_2 \in \left[L, (1+\frac{\delta}{4})L\right)$ and such that the intersection of $S[s_1..e_1]$ and $S[s_2..e_2]$ contains at least $\left(2+\frac{\delta}{2}\right)L$ characters.

Proof.

Assume, there is such a number L.



Theorem

The number of maximal δ-repetitions whose extensions contain $t-1$ and t is bounded by

$$\left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor .$$

Theorem

The number of maximal δ-repetitions whose extensions contain $t-1$ and t is bounded by

$$\left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor.$$

Assume, there are more such maximal δ-repetitions.

Lemma

Let $S[s_1..e_1]$ and $S[s_2..e_2]$ be two maximal δ-repetitions with minimum periods p_1 and p_2.
There is no number L such that $p_1, p_2 \in \left[L, (1+\frac{\delta}{4})L\right)$ and such that the intersection of $S[s_1..e_1]$ and $S[s_2..e_2]$ contains at least $\left(2+\frac{\delta}{2}\right)) L$ characters.

Theorem

The number of maximal δ-repetitions whose extensions contain $t-1$ and t is bounded by

$$\left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor.$$

Assume, there are more such maximal δ-repetitions.
Then, there is a number L such that more than $\left\lfloor 3 + \frac{6}{\delta} \right\rfloor$ of these maximal δ-repetitions have minimum periods in $\left[L, (1+\frac{\delta}{4})L\right)$.

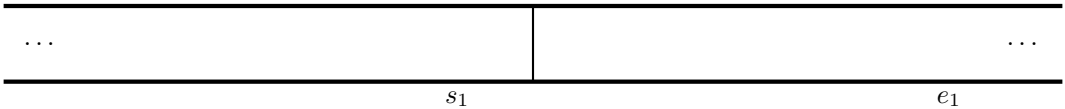


Figure: A maximal 0.5-repetitions.

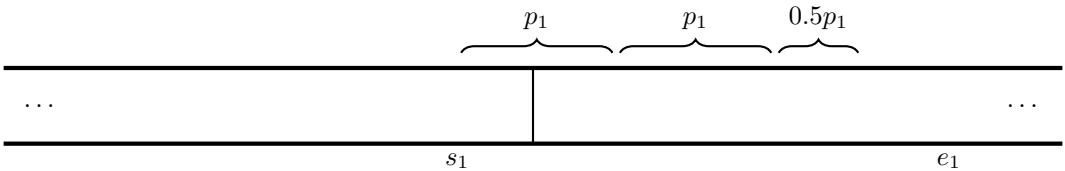


Figure: A maximal 0.5-repetitions.

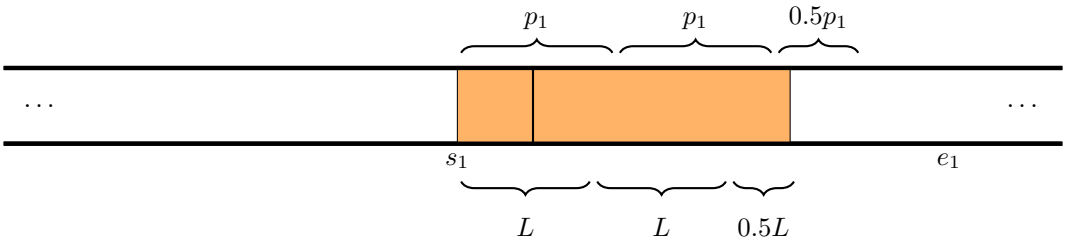


Figure: A maximal 0.5-repetitions.

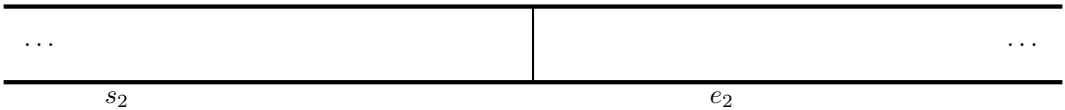


Figure: A maximal 0.5-repetitions.

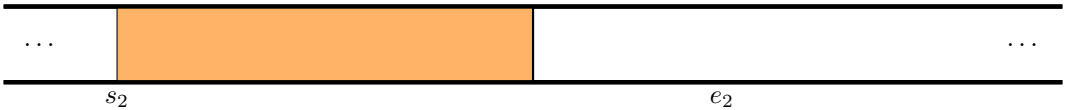


Figure: A maximal 0.5-repetitions.

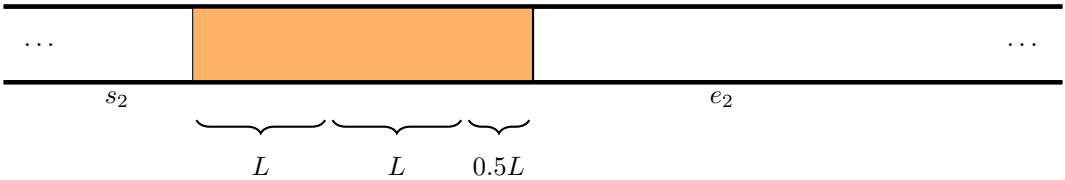


Figure: A maximal 0.5-repetitions.

Theorem

The number of maximal δ-repetitions whose extensions contain $t-1$ and t is bounded by

$$\left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor .$$

Assume, there are more such maximal δ-repetitions.
Then, there is a number L such that more than $\left\lfloor 3 + \frac{6}{\delta} \right\rfloor$ of these maximal δ-repetitions have minimum periods in $\left[L, (1+\frac{\delta}{4})L\right)$.
Then, there are at least two of these maximal δ-repetitions which also have an intersection with at least $\left(2 + \frac{\delta}{2})\right) L$ characters.

Theorem

The number of maximal δ-repetitions whose extensions contain $t-1$ and t is bounded by

$$\left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor .$$

Assume, there are more such maximal δ-repetitions.
Then, there is a number L such that more than $\lfloor 3 + \frac{6}{\delta} \rfloor$ of these maximal δ-repetitions have minimum periods in $\left[L, (1+\frac{\delta}{4})L\right)$.
Then, there are at least two of these maximal δ-repetitions which also have an intersection with at least $\left(2 + \frac{\delta}{2})\right) L$ characters.
This is not possible.

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Corollary

The number of distinct maximal δ-repetitions is

- *for large δ in* $\mathcal{O}\left(\frac{z \log |S|}{\log \delta}\right)$

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Corollary

The number of distinct maximal δ-repetitions is

- *for large δ in* $\mathcal{O}\left(\frac{z \log |S|}{\log \delta}\right)$ *and*
- *for small δ in* $\mathcal{O}\left(z \frac{6}{\delta} \frac{\log |S|}{\log(1+\frac{\delta}{4})}\right)$

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Corollary

The number of distinct maximal δ-repetitions is

- *for large δ in* $\mathcal{O}\left(\frac{z \log |S|}{\log \delta}\right)$ *and*
- *for small δ in* $\mathcal{O}\left(z \frac{6}{\delta} \frac{\log |S|}{\log(1+\frac{\delta}{4})}\right) \subseteq \mathcal{O}\left(\frac{z \log |S|}{\delta^2}\right)$

Theorem

The number of distinct maximal δ-repetitions is bounded by $z \left\lfloor 3 + \frac{6}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{4}}(|S|) \right\rfloor$.

Corollary

The number of distinct maximal δ-repetitions is

- *for large δ in* $\mathcal{O}\left(\frac{z \log |S|}{\log \delta}\right)$ *and*
- *for small δ in* $\mathcal{O}\left(z \frac{6}{\delta} \frac{\log |S|}{\log(1+\frac{\delta}{4})}\right) \subseteq \mathcal{O}\left(\frac{z \log |S|}{\delta^2}\right)$ *and*
- *for fixed δ in* $\mathcal{O}(z \log |S|)$.

Let f, g be functions such that $\mathcal{O}(f(z)g(\log|S|))$ is an upper bound for the number of maximal repetitions. Let further $g(\log x)$ be bounded by $c(\log x)^i$ for some constants c and i.

Let f, g be functions such that $\mathcal{O}(f(z)g(\log|S|))$ is an upper bound for the number of maximal repetitions. Let further $g(\log x)$ be bounded by $c(\log x)^i$ for some constants c and i.
Then, $f(z) \in \Omega(z)$.

Let f, g be functions such that $\mathcal{O}(f(z)g(\log|S|))$ is an upper bound for the number of maximal repetitions. Let further $g(\log x)$ be bounded by $c(\log x)^i$ for some constants c and i.
Then, $f(z) \in \Omega(z)$.
Also, $f(x)g(\log x) \in \Omega(x)$.

Let f, g be functions such that $\mathcal{O}(f(z)g(\log|S|))$ is an upper bound for the number of maximal repetitions. Let further $g(\log x)$ be bounded by $c(\log x)^i$ for some constants c and i.
Then, $f(z) \in \Omega(z)$.
Also, $f(x)g(\log x) \in \Omega(x)$.
Therefore, this upper bound is asymptotically optimal for fixed δ.

A *δ-subrepetition* is a substring which is at least $1 + \delta$ times as long as its minimum period.

A δ-*subrepetition* is a substring which is at least $1 + \delta$ times as long as its minimum period.
The number of maximal δ-subrepetitions is in $\mathcal{O}\left(\frac{|S|}{n}\right)$

A δ-*subrepetition* is a substring which is at least $1 + \delta$ times as long as its minimum period.
The number of maximal δ-subrepetitions is in $\mathcal{O}\left(\frac{|S|}{n}\right)$

Theorem

The number of distinct maximal δ-subrepetitions is bounded by $z\left\lfloor 3 + \frac{4}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{2q}}(|S|) \right\rfloor$.

A δ-*subrepetition* is a substring which is at least $1 + \delta$ times as long as its minimum period.
The number of maximal δ-subrepetitions is in $\mathcal{O}\left(\frac{|S|}{n}\right)$

Theorem

The number of distinct maximal δ-subrepetitions is bounded by $z\left\lfloor 3 + \frac{4}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{2q}}(|S|) \right\rfloor$.

Corollary

The number of distinct maximal δ-subrepetitions is

- *for small δ in* $\mathcal{O}\left(\frac{zq \log |S|}{\delta^2}\right)$

A δ-*subrepetition* is a substring which is at least $1 + \delta$ times as long as its minimum period.
The number of maximal δ-subrepetitions is in $\mathcal{O}\left(\frac{|S|}{n}\right)$

Theorem

The number of distinct maximal δ-subrepetitions is bounded by $z\left\lfloor 3 + \frac{4}{\delta} \right\rfloor \cdot \left\lfloor \log_{1+\frac{\delta}{2q}}(|S|) \right\rfloor$.

Corollary

The number of distinct maximal δ-subrepetitions is

- *for small δ in* $\mathcal{O}\left(\frac{zq \log |S|}{\delta^2}\right)$ *and*
- *for fixed δ and q in* $\mathcal{O}(z \log |S|)$.

- Improve the upper bounds by the factor δ.

- Improve the upper bounds by the factor δ.
- Give an efficient algorithm for finding maximal δ-subrepetitions in compressed strings.