On the Number of Non-equivalent Parameterized Squares in a String

presented at SPIRE 2024

Rikuya Hamai Kazushi Taketsugu
Yuto Nakashima Shunsuke Inenaga Hideo Bannai

2024/12/17

Number of distinct squares

Square = string represented by xx

Q. How many distinct squares can a string of length n contain?

# of distinct squares = 5
# of occ. of squares = 9

Considering the distinctness is more interesting than the occurrences because the maximum number of occ. of squares is $\Theta(n^2)$ for a string $\mathtt{a}^n$.

Distinct square conjecture [Fraenkel and Simpson, 1998]
The maximum number of distinct squares is less than n.

$< 2n$	[Fraenkel and Simpson, 1998]	$< 11n/6$	[Deza et al., 2015]
$< 2n - \Theta(\log n)$	[Ilie, 2007]	$< 1.5n$	[Thierry, 2020]
$< 95n/48$	[Lam, 2013]	$< n - \sigma + 1$	[Brlek & Li, 2022]

Squares in other matching models

Square = concatenation of two equal strings

Q. What happens if “equal” is changed from exact matching to another matching model (equivalence)?

Exact matching

cb**abac**ababcbd
  ||
  abac

Parameterized matching

cb**abac**a**babc**bd
  ||| |||
  abac abac

bac
↕↕↕
abc

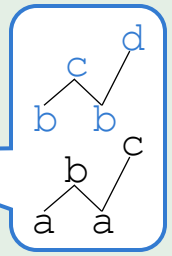
Abelian matching

cbabacab**abcb**d
  ||| |||
  cabb cabb

#(a) = 1
#(b) = 2
#(c) = 1

Order-preserving matching

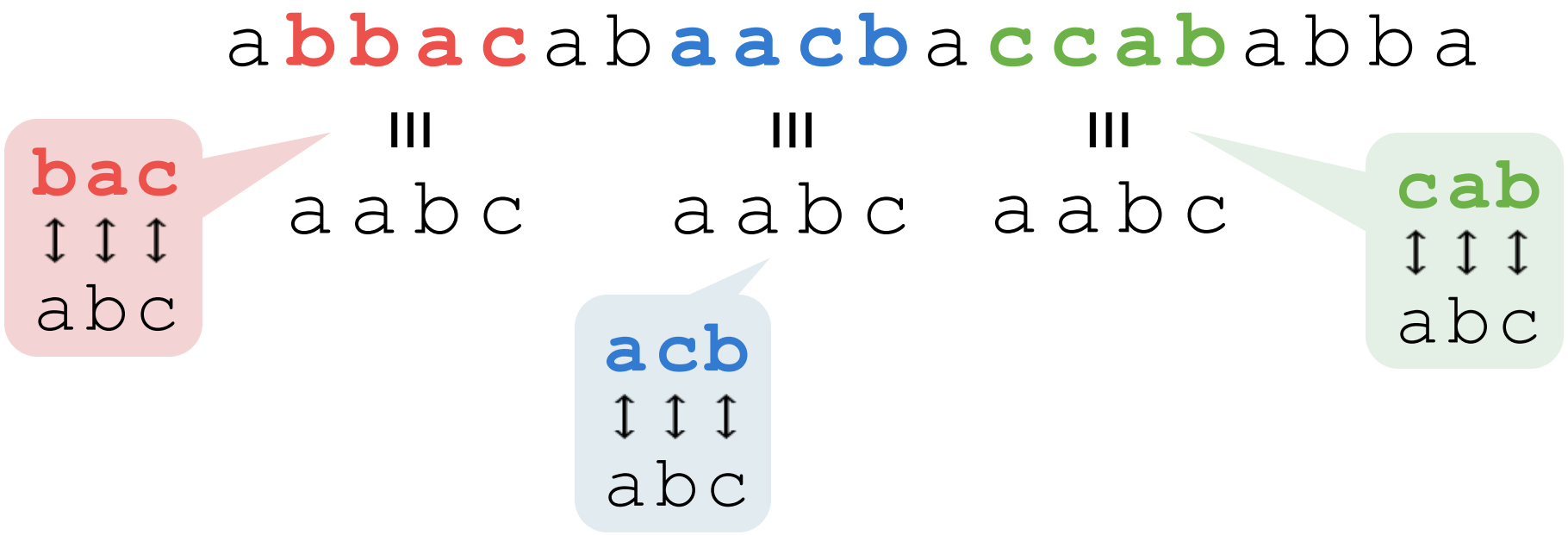
cb**abac**aba**bcbd**
  ||| |||
  abac abac



Parameterized equivalence [Baker, 1996]

Definition. Two strings x and y of length k are said to be **parameterized equivalent** if there is a bijection f on Σ such that $f(x[i]) = y[i]$ for all $1 \leq i \leq k$. We write $x \equiv y$ if two strings x and y are parameterized equivalent.

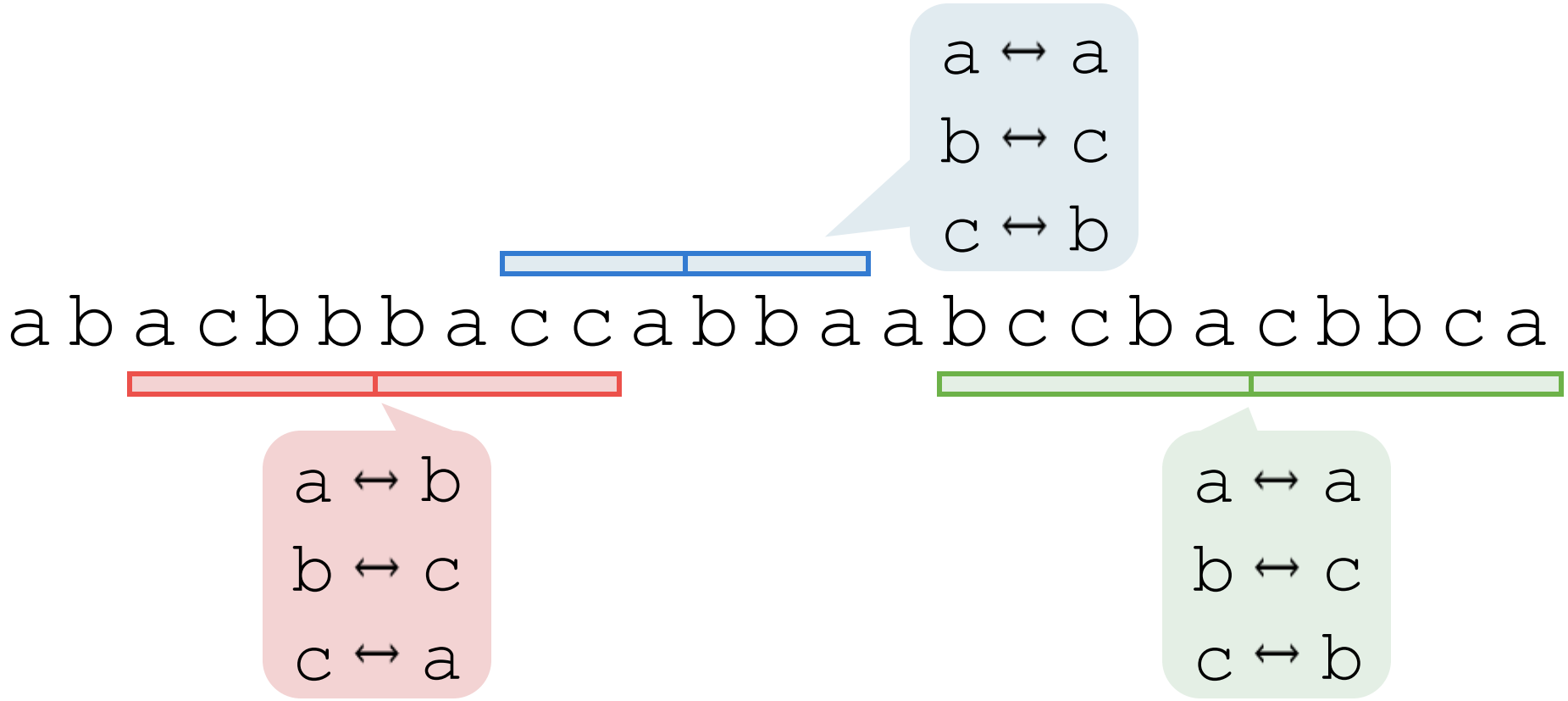
E.g.



Parameterized square (P-square)

Definition. A string w is called a parameterized square when $w = xy$ for strings x, y and x and y are parameterized equivalent.

E.g.



- We sometimes use “P” to denote the term “parameterized” (e.g., P-square).

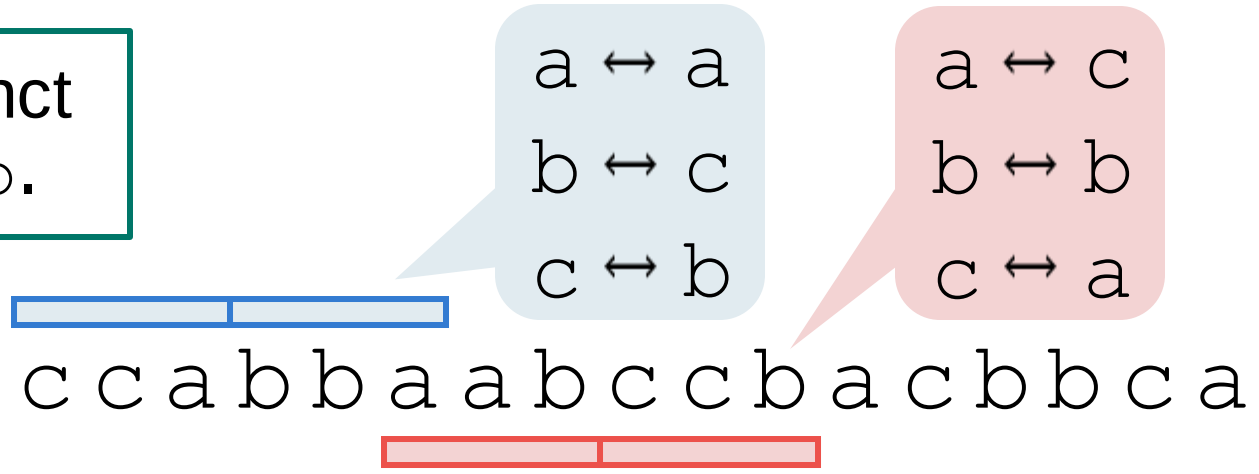
Distinct squares in other models [Kociumaka et al., 2016]

There are two types of the distinctness.

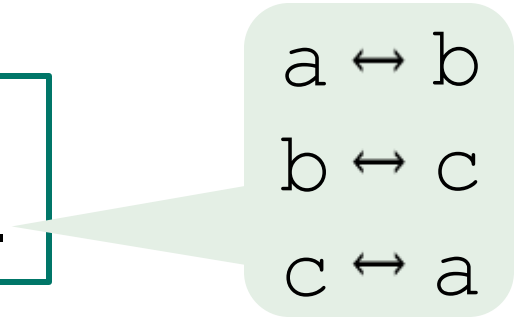
A) # of p-squares that are distinct as strings.

B) # of p-squares that are non-equivalent in the equivalence model.

In case A, they are distinct since $ccabba \neq aabccb$.



In case B, they are not distinct since $ccabba \neq aabccb$ but $ccabba \equiv aabccb$.



Distinct squares in other models [Kociumaka et al., 2016]

There are two types of the distinctness.

A) # of p-squares that are distinct as strings.

B) # of p-squares that are non-equivalent in the equivalence model.

	distinct as strings		non-equivalent	
	upper bound	lower bound	upper bound	lower bound
Parameterized	$\leq 2(\sigma!)^2 n$ [Kociumaka et al., 2016]	–	$\leq 2\sigma! n$ [Kociumaka et al., 2016]	–
Order-preserving	$O(\sigma n)$ [Gawrychowski et al., 2023]	$\Omega(\sigma n)$ [Gawrychowski et al., 2023]	$O(\sigma n)$ [Gawrychowski et al., 2023]	–
Abelian	$O(\sigma n)$ [Kociumaka et al., 2016]	$\Omega(\sigma n)$ [Kociumaka et al., 2016]	$\leq (n-1)^{11/6}$ [Kociumaka et al., 2016]	$\Omega(n^{1.5})$ [Kociumaka et al., 2016]

n : string length, σ : alphabet size

Distinct squares in other models [Kociumaka et al., 2016]

There are two types of the distinctness.

A) # of p-squares that are distinct as strings.

B) # of p-squares that are non-equivalent in the equivalence model.

	distinct as strings		non-equivalent	
	upper bound	lower bound	upper bound	lower bound
Parameterized	$\leq 2(\sigma!)^2 n$ [Kociumaka et al., 2016]	–	$\leq 2\sigma! n$ [Kociumaka et al., 2016]	–
Order-preserving	$O(\sigma n)$ [Gawrychowski et al., 2023]	$\Omega(\sigma n)$ [Gawrychowski et al., 2023]	$O(\sigma n)$ [Gawrychowski et al., 2023]	–
Abelian	$O(\sigma n)$ [Kociumaka et al., 2016]	$\Omega(\sigma n)$ [Kociumaka et al., 2016]	$\leq (n-1)^{11/6}$ [Kociumaka et al., 2016]	$\Omega(n^{1.5})$ [Kociumaka et al., 2016]

Theorem 1. Any string of length n that contains σ distinct characters can contain at most σn non-equivalent parameterized squares.

Upper bound (main theorem)

Theorem 2. For any string s that contains σ distinct characters, there can be at most σ prefixes of s that are P-squares and have no other parameterized occurrences in s.

at most σ parameterized squares for each of n positions

Theorem 1. Any string of length n that contains σ distinct characters can contain at most σn non-equivalent parameterized squares.

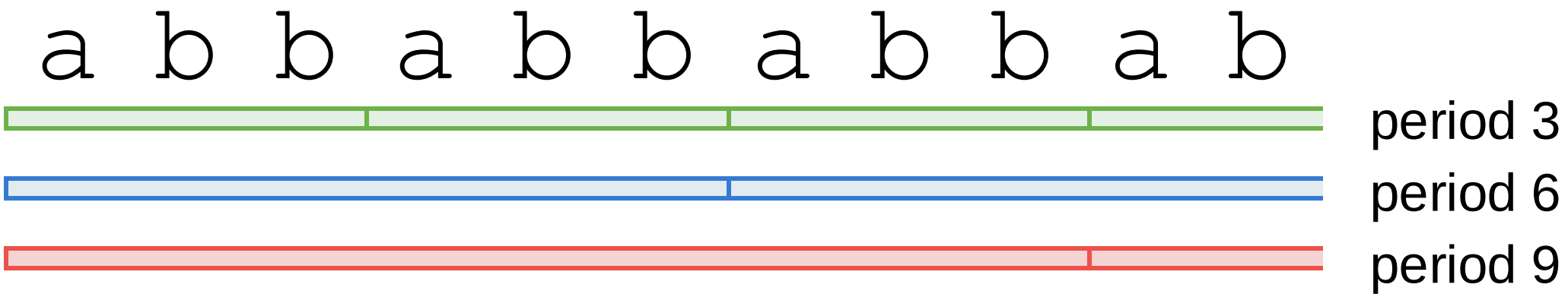
- We consider the **periodicity** in the parameterized equivalent model.

Standard period

Definition. An integer p is said to be a period of string w

1. if $w[i] = w[i+p]$ for all i satisfying $1 \leq i \leq |w| - p$,
   or equivalently,
2. if $w[1..|w|-p] = w[p+1..|w|]$.

E.g.

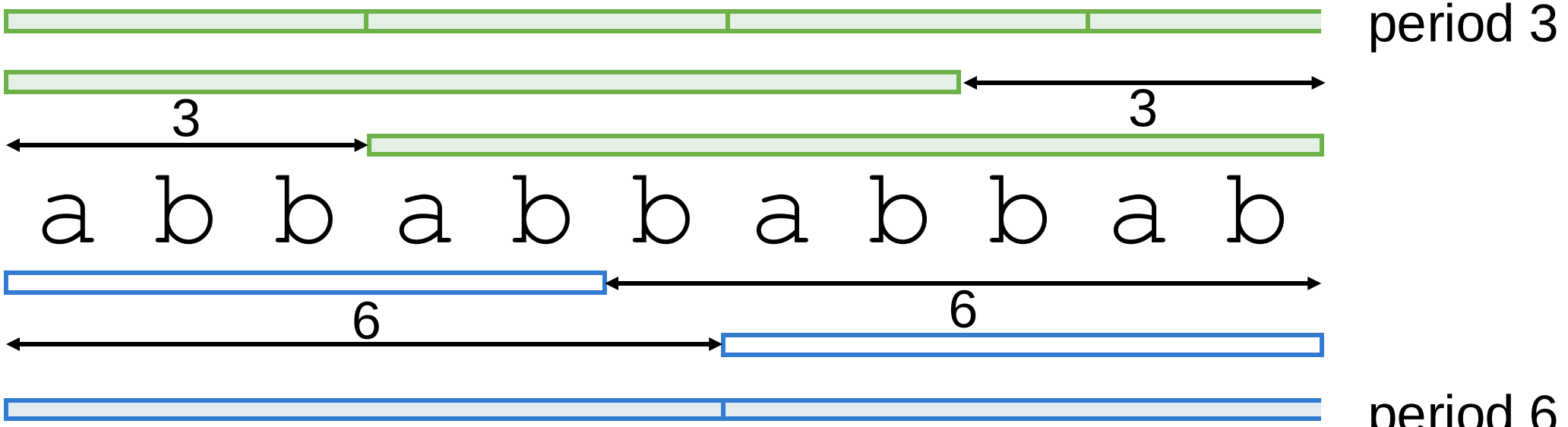


Standard period

Definition. An integer p is said to be a period of string w

1. if $w[i] = w[i+p]$ for all i satisfying $1 \le i \le |w| - p$,
   or equivalently,
2. if $w[1..|w|-p] = w[p+1..|w|]$.

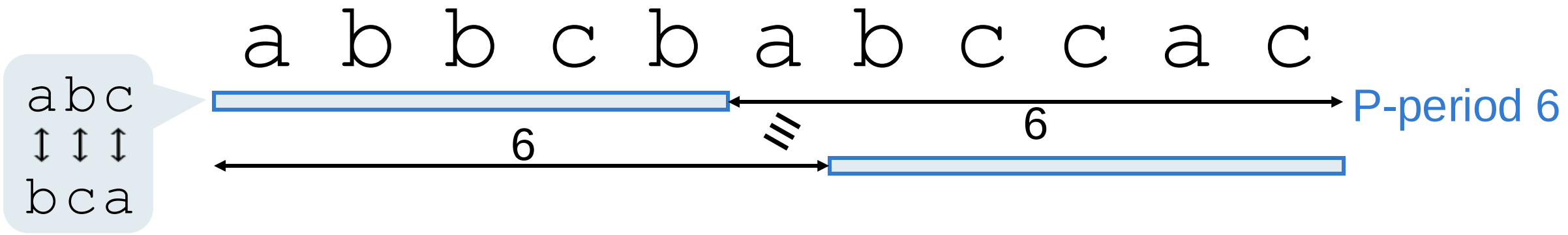
E.g.



Parameterized period (P-period)

Definition. An integer p is said to be a P-period of string w
1. if $f(w[i]) = w[i+p]$ for all i satisfying $1 \le i \le |w| - p$ with a bijection f,
   or equivalently,
2. if $w[1..|w|-p] \equiv w[p+1..|w|]$.

E.g.



- We sometimes use $p \|_f w$
  to denote the fact "p is a P-period of w with a bijection f"
  (we just write $p \| w$ when no confusions occur).

Properties on P-period

To prove our main result, we use the following lemmas about P-periods.

Periodicity lemma

Lemma 1. Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \geq 2$. If $p + q + \min(p, q)\cdot(|\Sigma_s| - 2) \leq |s|$, then $\gcd(p, q)$ is a P-period of s.

Period connection lemma

Lemma 2. Let x, y, z be strings s.t. p is a P-period of xy with bijection f and a P-period of yz with bijection g. If $|y| \geq p\cdot(|\Sigma_{xyz}| - 1) + 1$, then p is a P-period of xyz with bijection f $(= g)$.

Period extension lemma

Lemma 3. Let s be a string that has a P-period q, and s' be a substring of s that has a P-period p. If $p\cdot(|\Sigma_s| - 2) + q + 1 \leq |s'|$ and $q = kp$ for some integer k, then p is a P-period of s.

Properties on P-period

Proofs are omitted in this talk.
I will give an intuition of these lemmas.

To prove our main result, we use the follo

Periodicity lemma

Lemma 1. Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \geq 2$. If $p + q + \min(p, q)\cdot(|\Sigma_s| - 2) \leq |s|$, then $\gcd(p, q)$ is a P-period of s.

Period connection lemma

Lemma 2. Let x, y, z be strings s.t. p is a P-period of xy with bijection f and a P-period of yz with bijection g. If $|y| \geq p\cdot(|\Sigma_{xyz}| - 1) + 1$, then p is a P-period of xyz with bijection f $(= g)$.

Period extension lemma

Lemma 3. Let s be a string that has a P-period q, and s' be a substring of s that has a P-period p. If $p\cdot(|\Sigma_s| - 2) + q + 1 \leq |s'|$ and $q = kp$ for some integer k, then p is a P-period of s.

Parameterized periodicity lemma

Periodicity lemma for standard period

Lemma. [Fine & Wilf, 1965] Let s be a string that has periods p and q.
If $p + q - \gcd(p, q) \leq |s|$, then $\gcd(p, q)$ is a period of s.

(Previous) Periodicity lemmas for P-period

Lemma. [Apostolico & Giancarlo, 2007]
Let s be a string that has a P-period of p with bijection f and a P-period q with bijection g. If $p + q \leq |s|$ and $f \circ g = g \circ f$, then $\gcd(p, q)$ is a P-period of s.

Lemma. [Ideguchi et al., 2023]

without commutativity

Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \geq 2$.
If $p + q + \min(p, q) \cdot (|\Sigma_s| - 1) \leq |s|$, then $\gcd(p, q)$ is a P-period of s.

Parameterized periodicity lemma

Lemma. [Ideguchi et al., 2023]
Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \geq 2$.
If $p + q + \min(p, q)\cdot(|\Sigma_s| - 1) \leq |s|$, then $\gcd(p, q)$ is a P-period of s.

- We give an improved version of the above lemma.

Lemma 1. Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \geq 2$.
If $p + q + \min(p, q)\cdot(|\Sigma_s| - 2) \leq |s|$, then $\gcd(p, q)$ is a P-period of s.

- Our result gives a tight bound when $\sigma = 2$ [Apostolico & Giancarlo, 2007].
  - When $\sigma = 2$, f and g always commute.

Tighter versions

- There exist tighter versions for each alphabet size.

Alphabet size	Lower bound of $\lvert s \rvert$	
$\sigma = 2$	$\geq p + q$ tight!	[Apostolico & Giancarlo, 2007]
$\sigma = 3$	$\geq p + q + 1$ tight!	[Our work (not published)]
$\sigma = 4$	$\geq p + q + \lfloor \min(p, q) / 2 \rfloor$ tight!	[Our work (not published)]
$\sigma = 5$	$\geq p + q + \lfloor \min(p, q) / 2 \rfloor + 1$ tight!	[Our work (not published)]
$\sigma \geq 6$	$\geq p + q + \min(p, q) \cdot (\sigma - 3)$	[Our work (not published)]

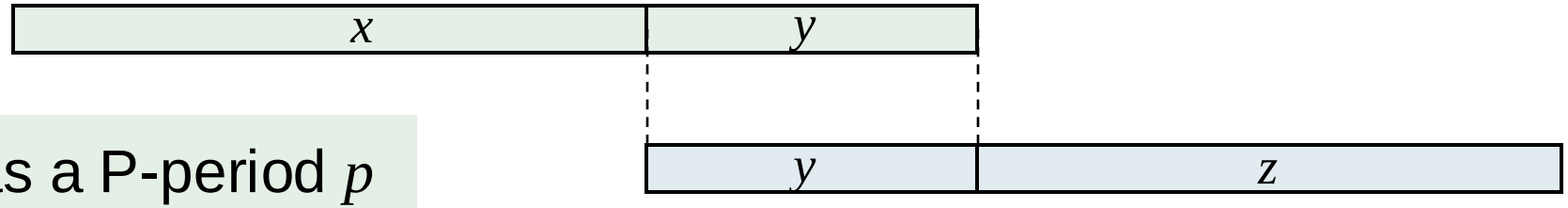
Lemma 1. Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \geq 2$. If $p + q + \min(p, q) \cdot (|\Sigma_s| - 2) \leq |s|$, then $\gcd(p, q)$ is a P-period of s.

In this work, we only use this general version.

Parameterized period connection lemma

Lemma 2. Let x, y, z be strings s.t. p is a P-period of xy with bijection f and a P-period of yz with bijection g. If $|y| \geq p \cdot (|\Sigma_{xyz}| - 1) + 1$, then p is a P-period of xyz with bijection f $(= g)$.

Almost all the characters occur in y.



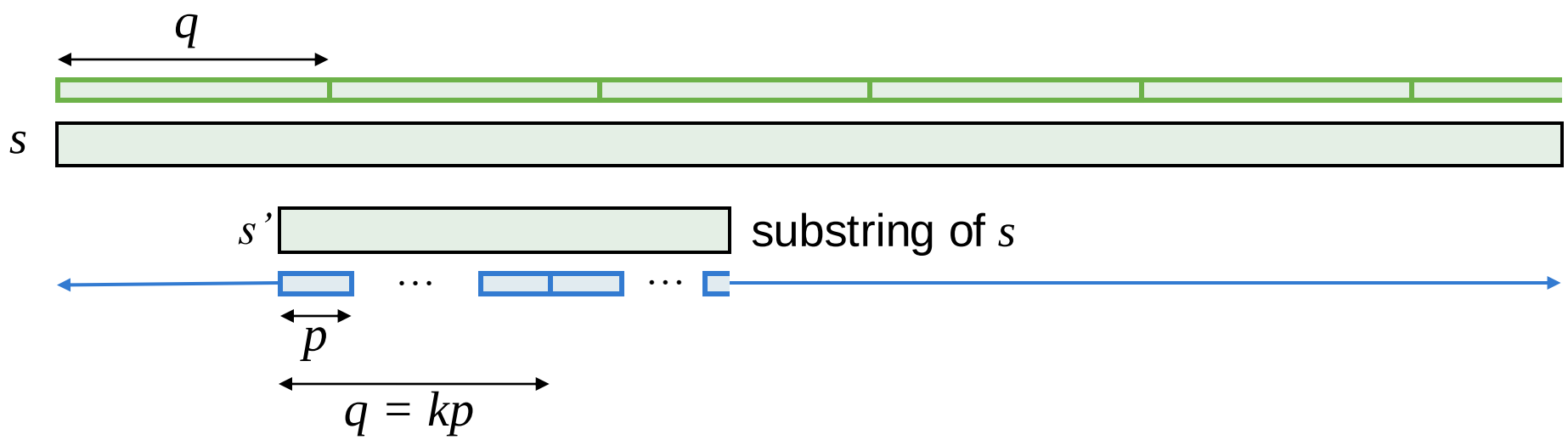
xy has a P-period p with bijection f.

yz has a P-period p with bijection g.

- Intuitively, if the overlapping part y is sufficiently long, the whole string has the same period p and they use the same bijection.

Parameterized period extension lemma

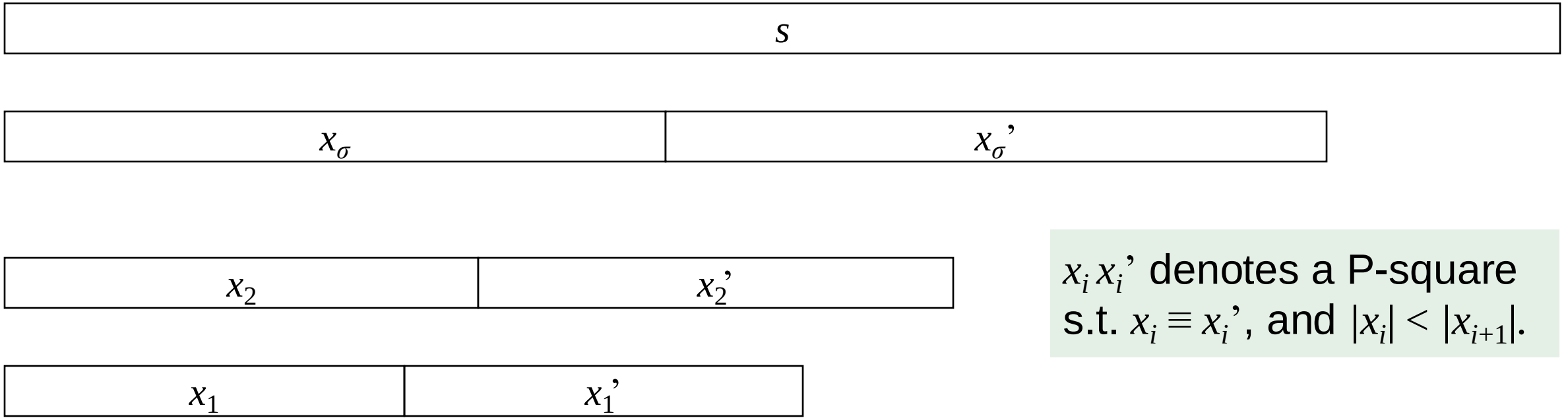
Lemma 3. Let s be a string that has a P-period q, and s' be a substring of s that has a P-period p. If $p \cdot (|\Sigma_s| - 2) + q + 1 \leq |s'|$ and $q = kp$ for some integer k, then p is a P-period of s.



- Intuitively, a P-period of a substring which is sufficiently long can extend to the whole string, if the whole string has a P-period that is a multiple of the substring's P-period.

Proof of σ prefixes theorem

Theorem 2. For any string s that contains σ distinct characters, there can be at most σ prefixes of s that are P-squares and have no other parameterized occurrences in s.

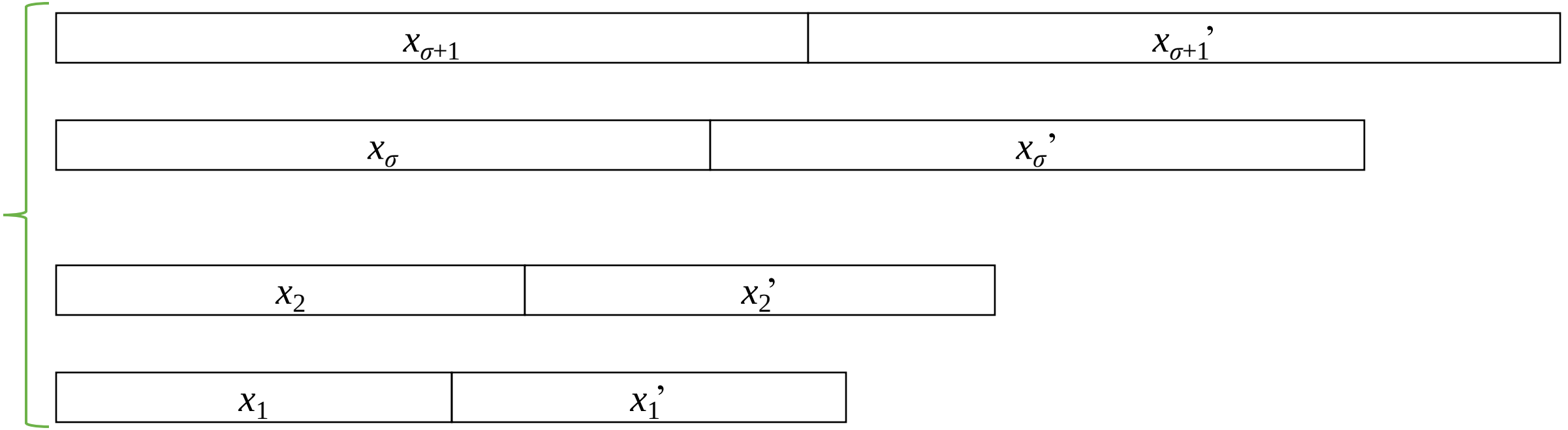


$x_i\, x_i$' denotes a P-square s.t. $x_i \equiv x_i$', and $|x_i| < |x_{i+1}|$.

Proof of σ prefixes theorem

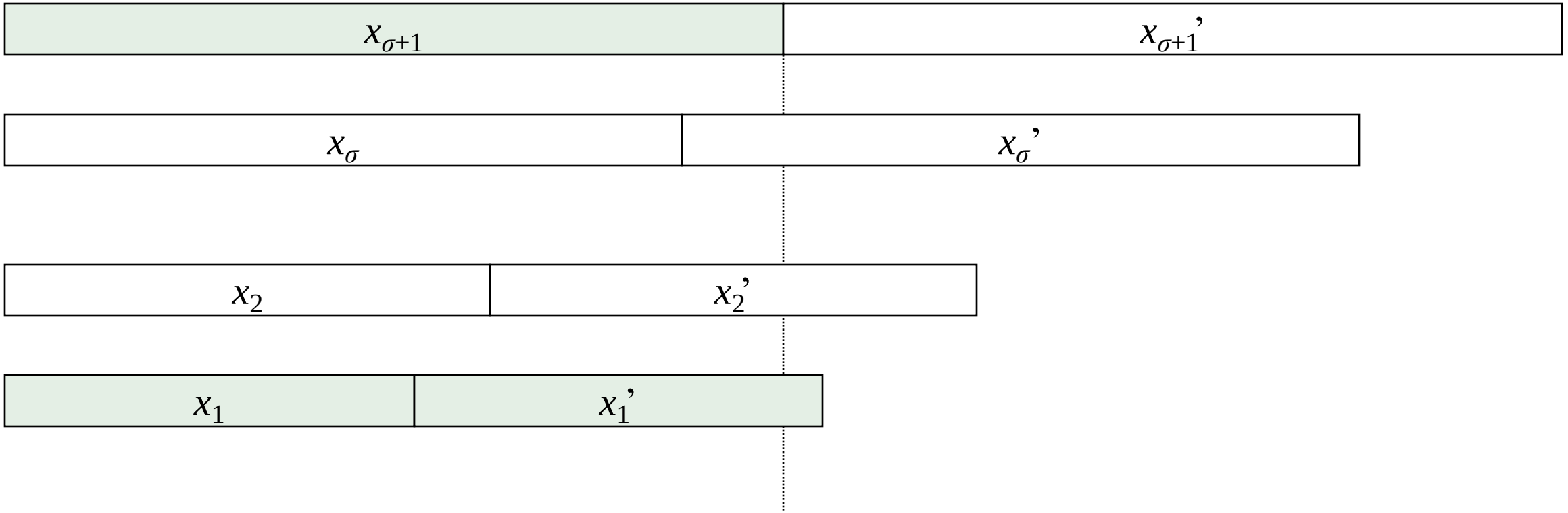
- Assume on the contrary that there exists a string s that contains σ distinct characters s.t. the number of parameterized squares of the prefixes of s that have no parameterized occurrences other than the prefixes is $\sigma + 1$.

$\sigma + 1$ prefix squares



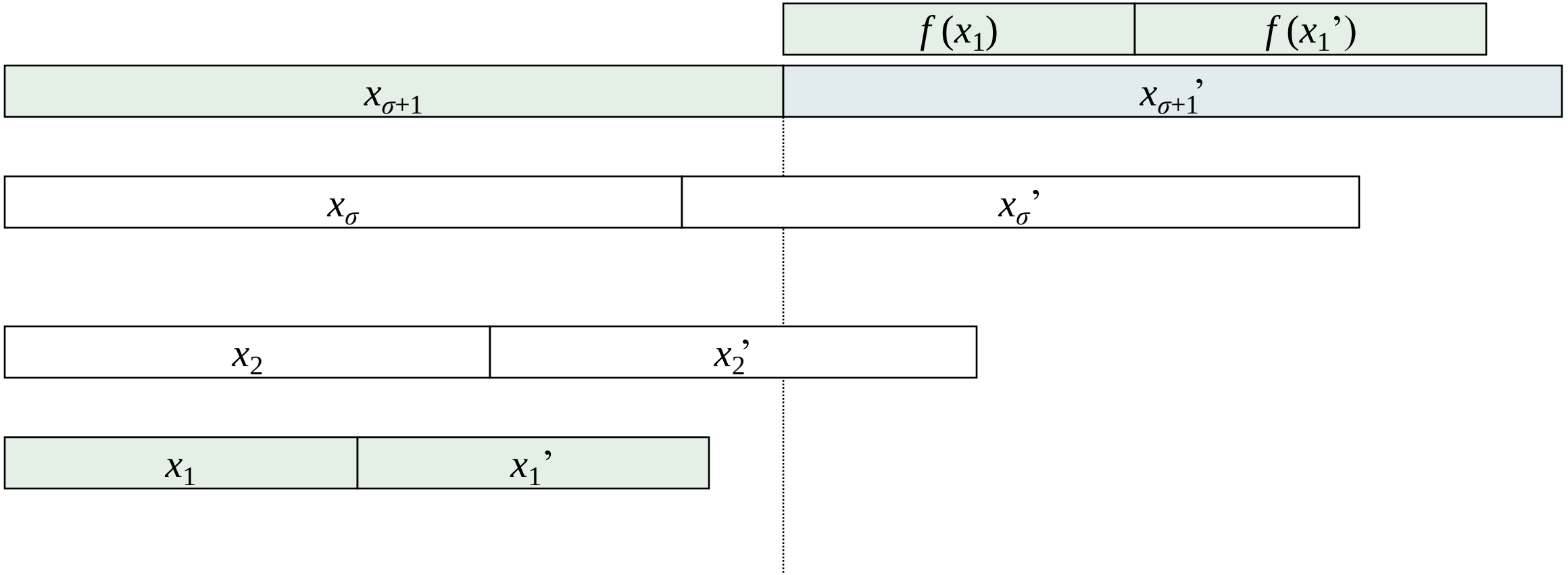
Proof of σ prefixes theorem

- Since $x_1 x_1$' has no parameterized equivalent occurrences other than the prefix, then $|x_1 x_1'| > |x_{\sigma+1}|$.
  (otherwise, $x_{\sigma+1}$' has a prefix that is parameterized equivalent to $x_1 x_1$'.)



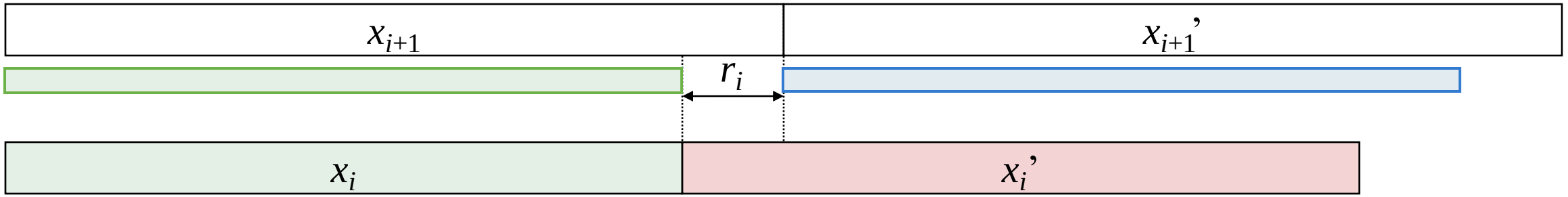
Proof of σ prefixes theorem

- Since x_1x_1' has no parameterized equivalent occurrences other than the prefix, then $|x_1x_1\text{'}| > |x_{\sigma+1}|$.
  (otherwise, $x_{\sigma+1}$' has a prefix that is parameterized equivalent to x_1x_1'.)



Proof of σ prefixes theorem

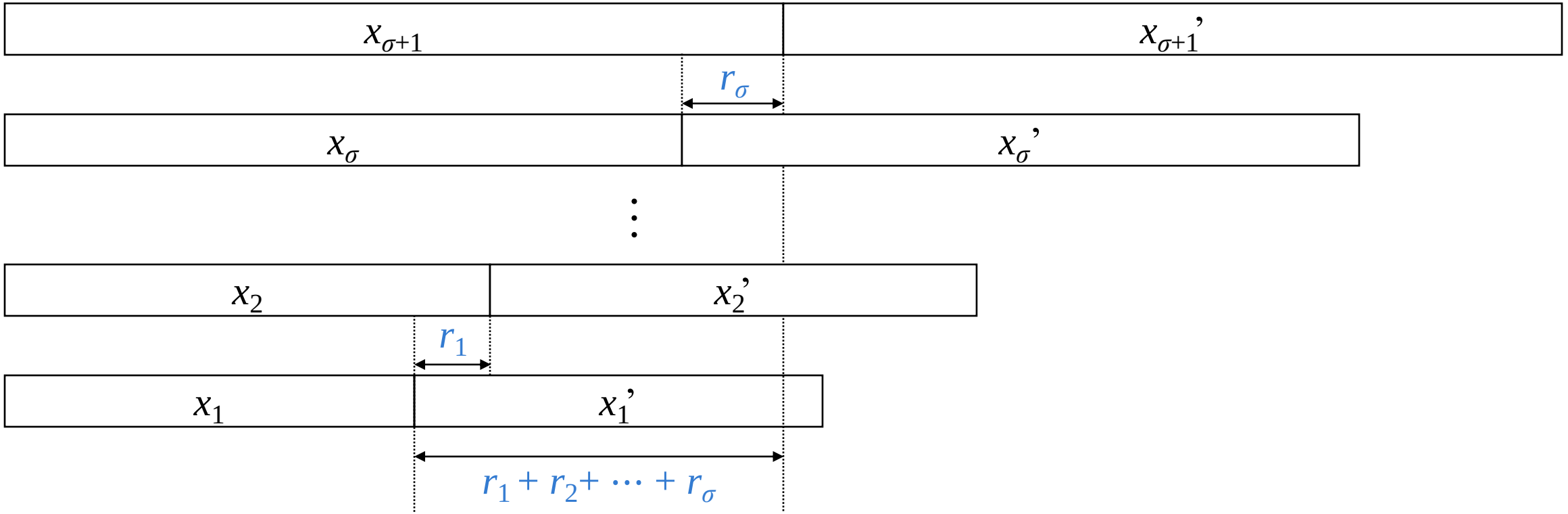
- Let $r_i = |x_{i+1}| - |x_i|$.
- $r_i \parallel x_i$' (r_i is a period of x_i') because $x_i \equiv x_i' \equiv x_{i+1}[1..|x_i|] \equiv x_{i+1}'[1..|x_i|]$.



- Hence, $r_i \parallel x_i$ also holds for any i.
- This implies that $r_i \parallel x_1$ holds for any i since $r_i < |x_1|$ and x_1 is a prefix of x_i.

Proof of σ prefixes theorem

- Hence, $r_i \parallel x_i$ also holds for any i.
- This implies that $r_i \parallel x_1$ holds for any i since $r_i < |x_1|$ and x_1 is a prefix of x_i.



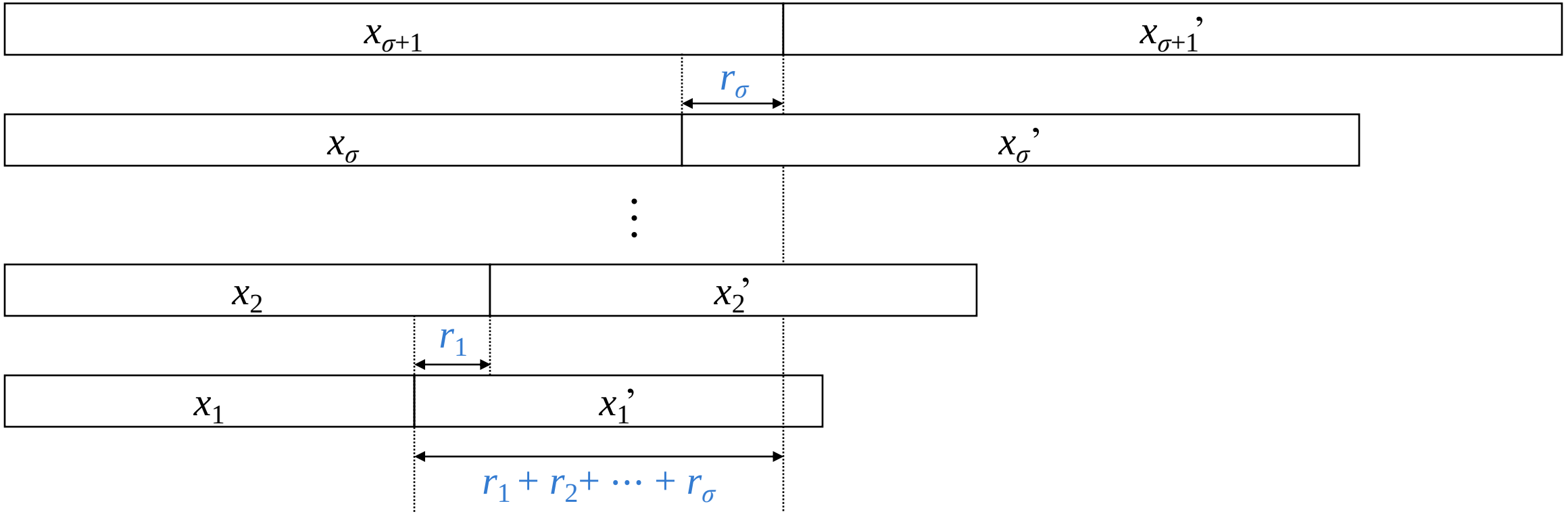
Proof of σ prefixes theorem

Claim 1. The smallest P-period $p(x_1)$ is a period of x_σ.

- $r_i \parallel x_1$
- $p(x_1) \le r_i$

- Since the condition of the periodicity, we have

$$p(x_1)\cdot(\sigma-1) + r_\sigma \le r_1 + \cdots + r_\sigma \le |x_1'| = |x_1| \quad \Leftrightarrow \quad p(x_1)\cdot(\sigma-2) + p(x_1) + r_\sigma \le |x_1|.$$



Proof of σ prefixes theorem

Claim 1. The smallest P-period $p(x_1)$ is a period of x_σ.

- $r_i \parallel x_1$
- $p(x_1) \le r_i$

- Since the condition of the periodicity, we have

$$p(x_1)\cdot(\sigma-1) + r_\sigma \le r_1 + \cdots + r_\sigma \le |x_1'| = |x_1| \quad \Leftrightarrow \quad \boxed{p(x_1)\cdot(\sigma-2) + p(x_1) + r_\sigma \le |x_1|.}$$

Lemma 1. Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \ge 2$. If $p + q + \min(p, q)\cdot(|\Sigma_s| - 2) \le |s|$, then $\gcd(p, q)$ is a P-period of s.

- By combining with Lemma 1, we have $\gcd(p(x_1), r_\sigma) \parallel x_1$.
- Since $p(x_1)$ is the smallest P-period of x_1, $r_\sigma = c \cdot p(x_1)$ for some integer $c \ge 1$.

Proof of σ prefixes theorem

Claim 1. The smallest P-period $p(x_1)$ is a period of x_σ.

- $r_i \parallel x_1$
- $p(x_1) \leq r_i$

- Since the condition of the periodicity, we have

$$p(x_1)\cdot(\sigma - 1) + r_\sigma \leq r_1 + \cdots + r_\sigma \leq |x_1'| = |x_1| \quad \Leftrightarrow \quad p(x_1)\cdot(\sigma - 2) + p(x_1) + r_\sigma \leq |x_1|.$$

- This inequation also implies that $p(x_1)\cdot(\sigma - 2) + r_\sigma + 1 \leq |x_1|$.

Lemma 3. Let s be a string that has a P-period q, and s' be a substring of s that has a P-period p. If $p\cdot(|\Sigma_s| - 2) + q + 1 \leq |s'|$ and $q = kp$ for some integer k, then p is a P-period of s.

- By combining with Lemma 3 and $r_\sigma = c \cdot p(x_1)$, we have $p(x_1) \parallel x_\sigma$ (Claim 1 holds).

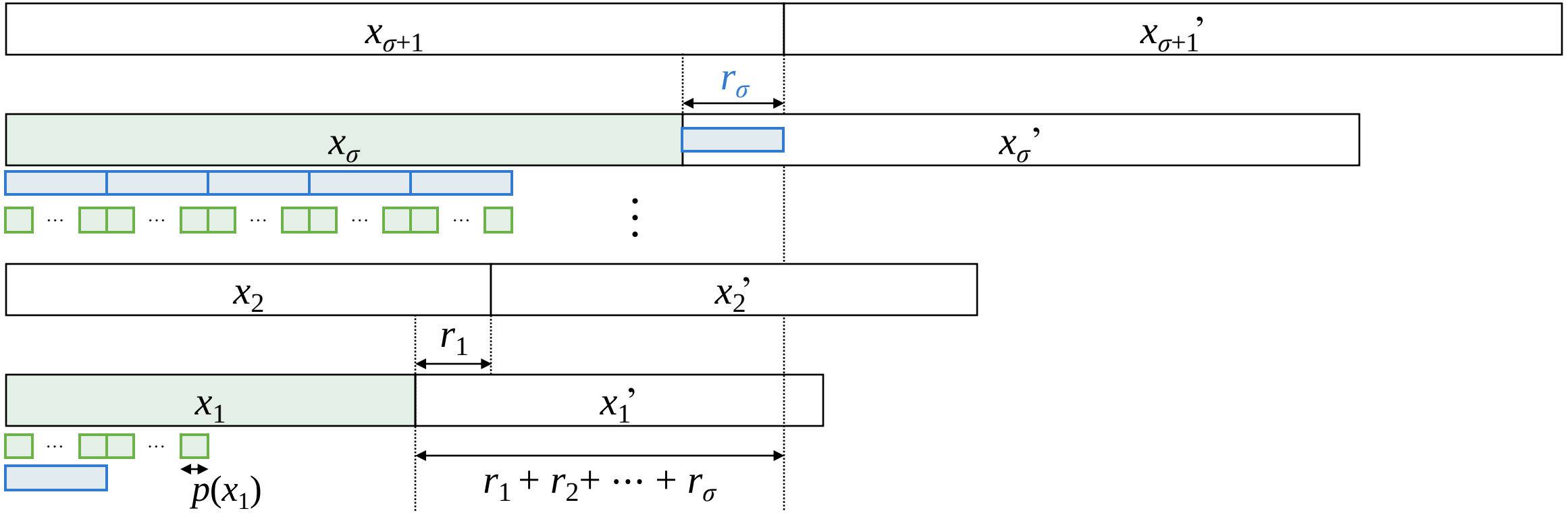
Proof of σ prefixes theorem

Claim 1. The smallest P-period $p(x_1)$ is a period of x_σ.

- $r_i \parallel x_1$
- $p(x_1) \le r_i$

- Since the condition of the periodicity, we have

$$p(x_1)\cdot(\sigma-1) + r_\sigma \le r_1 + \cdots + r_\sigma \le |x_1'| = |x_1| \quad \Leftrightarrow \quad p(x_1)\cdot(\sigma-2) + p(x_1) + r_\sigma \le |x_1|.$$



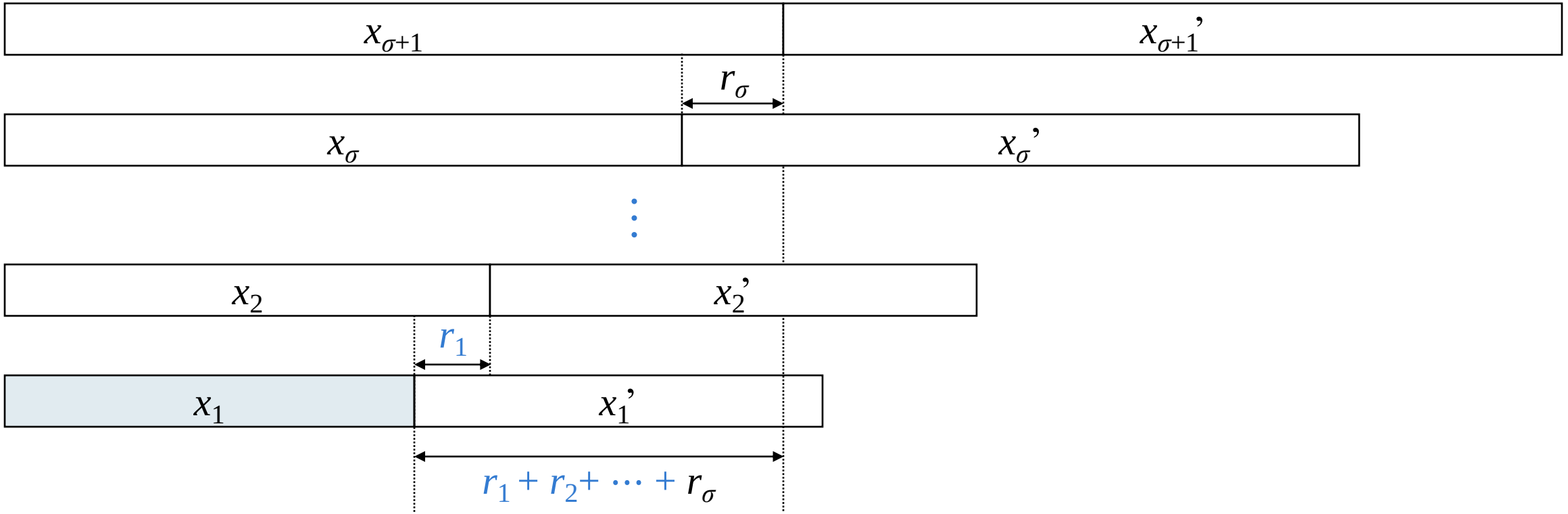
Proof of σ prefixes theorem

Claim 2. The smallest P-period $p(x_1)$ is a period of $x_{\sigma+1}$.

- $r_i \parallel x_1$
- $p(x_1) \le r_i$

- Since the condition of the periodicity and $|x_1| \parallel x_\sigma$, we have

$$p(x_1)\cdot(\sigma-1) + |x_1| \le r_1 + \cdots + r_{\sigma-1} + |x_1| = |x_\sigma| \iff \boxed{p(x_1)\cdot(\sigma-2) + p(x_1) + |x_1| \le |x_\sigma|.}$$



Proof of σ prefixes theorem

Claim 2. The smallest P-period $p(x_1)$ is a period of $x_{\sigma+1}$.

- $r_i \parallel x_1$
- $p(x_1) \leq r_i$

- Since the condition of the periodicity and $|x_1| \parallel x_\sigma$, we have

$$p(x_1)\cdot(\sigma-1) + |x_1| \leq r_1 + \cdots + r_{\sigma-1} + |x_1| = |x_\sigma| \Leftrightarrow \boxed{p(x_1)\cdot(\sigma-2) + p(x_1) + |x_1| \leq |x_\sigma|.}$$

Lemma 1. Let s be a string that has P-periods p and q and satisfies $|\Sigma_s| \geq 2$. If $p + q + \min(p, q)\cdot(|\Sigma_s| - 2) \leq |s|$, then $\gcd(p, q)$ is a P-period of s.

- By combining with Lemma 1, we have $\gcd(p(x_1), |x_1|) \parallel x_\sigma$.
- Since $p(x_1)$ is the smallest P-period of x_1, $|x_1| = d \cdot p(x_1)$ for some integer $d \geq 1$.

Proof of σ prefixes theorem

Claim 2. The smallest P-period $p(x_1)$ is a period of $x_{\sigma+1}$.

- $r_i \parallel x_1$
- $p(x_1) \leq r_i$

- Since the condition of the periodicity and $|x_1| \parallel x_\sigma$, we have

$$p(x_1)\cdot(\sigma-1) + |x_1| \leq r_1 + \cdots + r_{\sigma-1} + |x_1| = |x_\sigma| \;\Leftrightarrow\; p(x_1)\cdot(\sigma-2) + p(x_1) + |x_1| \leq |x_\sigma|.$$

- This inequation also implies that $p(x_1)\cdot(\sigma-2) + |x_1| + 1 \leq |x_\sigma|$.

Lemma 3. Let s be a string that has a P-period q, and s' be a substring of s that has a P-period p. If $p\cdot(|\Sigma_s| - 2) + q + 1 \leq |s'|$ and $q = kp$ for some integer k, then p is a P-period of s.

- By combining with Lemma 3 and $|x_1| = d \cdot p(x_1)$, we have $p(x_1) \parallel x_{\sigma+1}$ (Claim 2 holds).

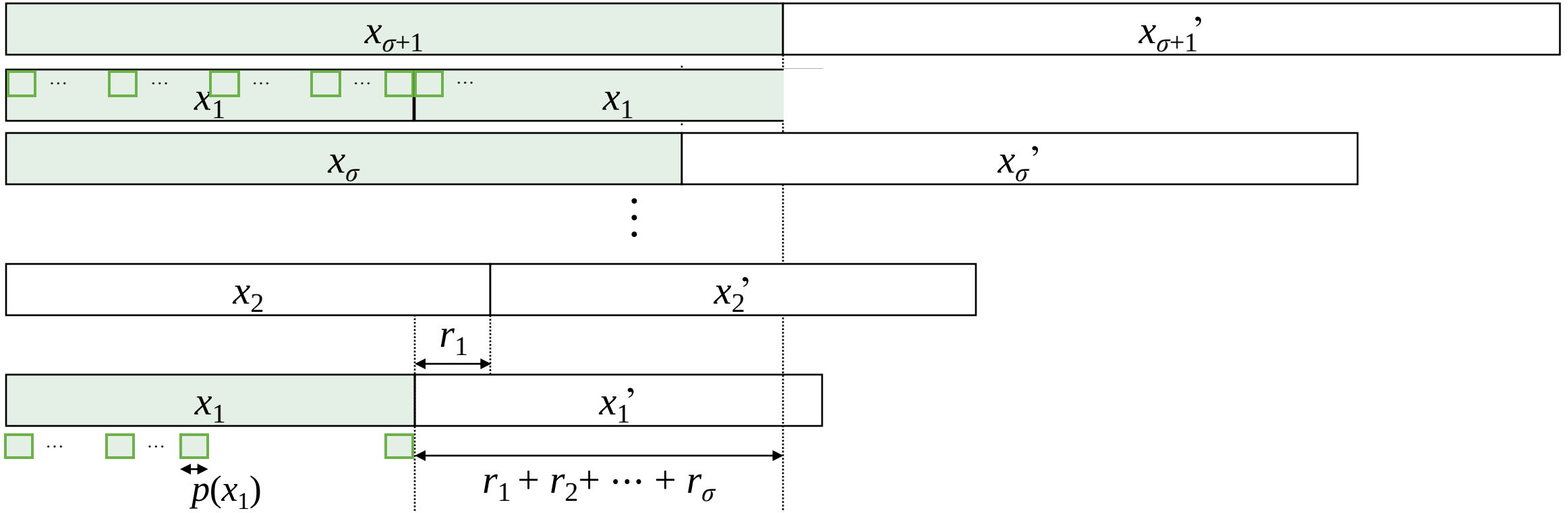
Proof of σ prefixes theorem

Claim 2. The smallest P-period $p(x_1)$ is a period of $x_{\sigma+1}$.

- $r_i \parallel x_1$
- $p(x_1) \le r_i$

- Since the condition of the periodicity and $|x_1| \parallel x_\sigma$, we have

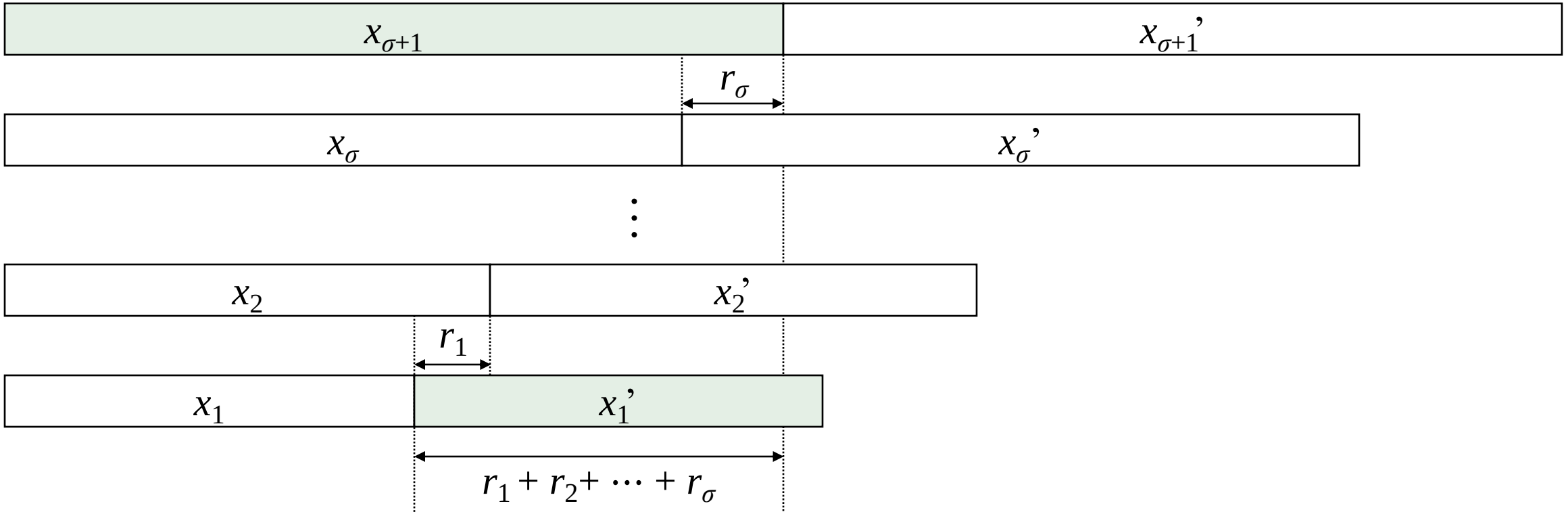
$$p(x_1)\cdot(\sigma-1) + |x_1| \le r_1 + \cdots + r_{\sigma-1} + |x_1| = |x_\sigma| \;\Leftrightarrow\; p(x_1)\cdot(\sigma-2) + p(x_1) + |x_1| \le |x_\sigma|.$$



Proof of σ prefixes theorem

We can use Lemma 2 for $x_{\sigma+1}$ and x_1' because

- $x_{\sigma+1}$ and x_1' have a period $p(x_1)$ (by Claim 2 & definition).
- $x_{\sigma+1}$ and x_1' overlap each other, and the length is $r_1 + \cdots + r_\sigma$.



Proof of σ prefixes theorem

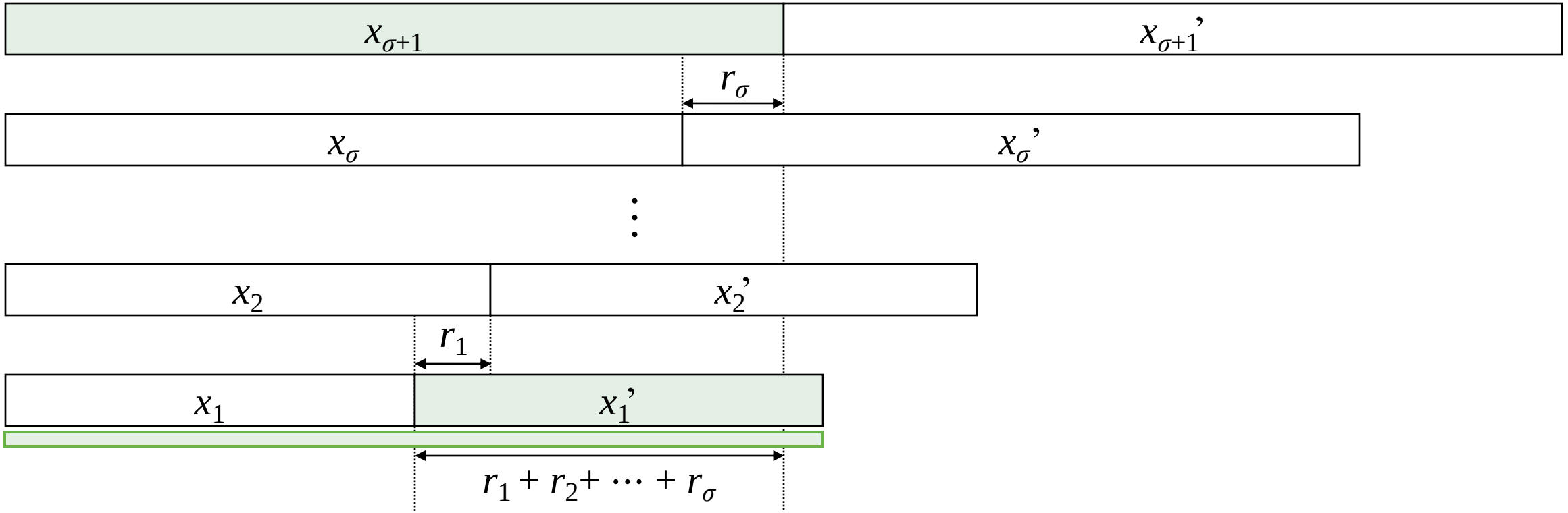
We can use Lemma 2 for $x_{\sigma+1}$ and x_1' because

- $x_{\sigma+1}$ and x_1' have a period $p(x_1)$ (by Claim 2 & definition),
- $x_{\sigma+1}$ and x_1' overlap each other, and
- the overlap length is $r_1 + \cdots + r_\sigma \geq p(x_1)\cdot(\sigma - 1) + 1$.

Lemma 2. Let x, y, z be strings s.t. p is a P-period of xy with bijection f and a P-period of yz with bijection g. If $|y| \geq p\cdot(|\Sigma_{xyz}| - 1) + 1$, then p is a P−period of xyz with bijection f $(= g)$.

Proof of σ prefixes theorem

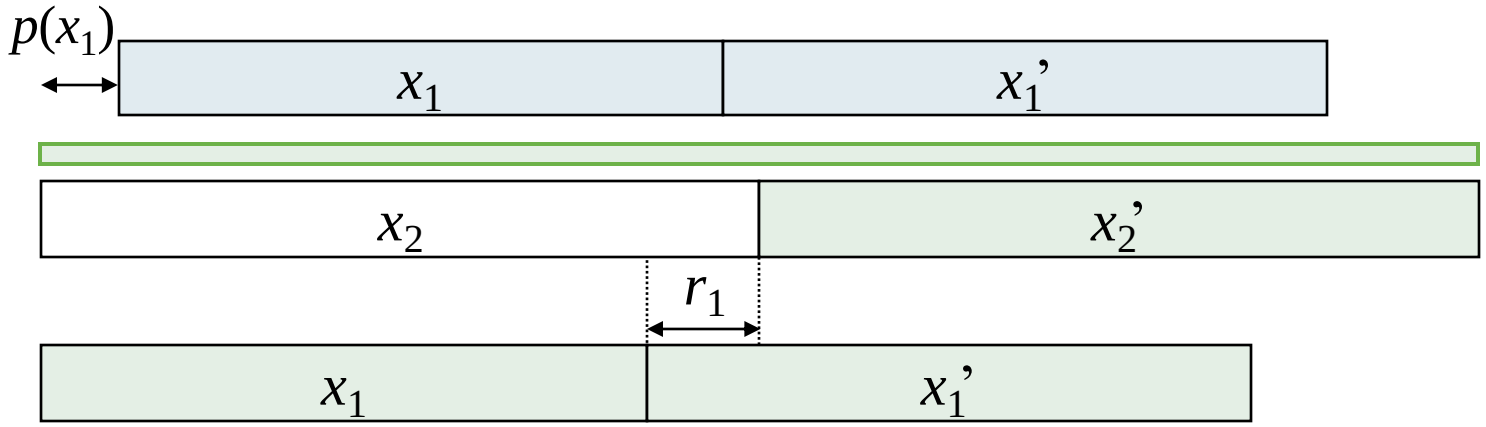
By applying Lemma 2 for $x_{\sigma+1}$ and x_1', we have $p(x_1) \parallel x_1x_1$'.



Proof of σ prefixes theorem

By a similar argument for x_1x_1' and x_2', we have $p(x_1) \parallel x_2x_2$'.

Then, x_1x_1' has another parameterized equivalent occurrence in x_2x_2'.



Theorem 2. For any string s that contains σ distinct characters, there can be at most σ prefixes of s that are P-squares and have no other parameterized occurrences in s.

Open questions

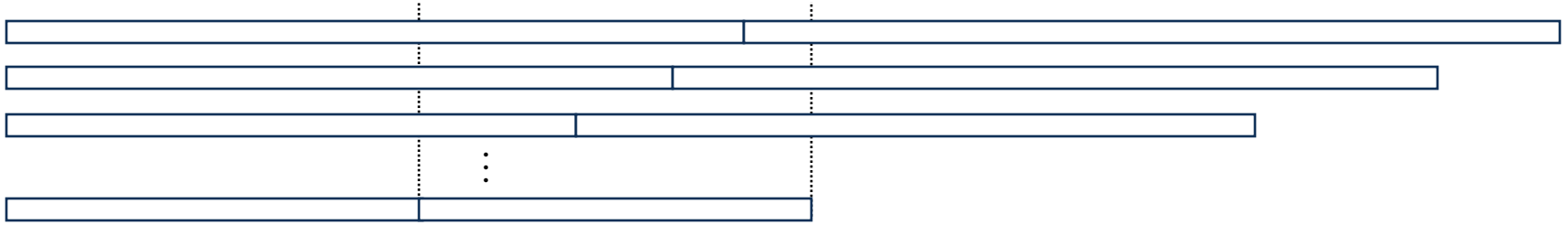
Theorem 1. Any string of length n that contains σ distinct characters can contain at most σn non-equivalent parameterized squares.

- Is this upper bound is tight?
  - How can we improve this bound?
    (There is a lower bound for the σ prefixes property.)
  - How can we construct a family of strings that gives a good lower bound?
- Can we improve bounds in another distinctness (distinct as words)?

Example of σ prefixes

Theorem. For any σ, there exists a string with alphabet size σ such that there are σ prefixes that are parameterized squares and have no other parameterized occurrence.

$(c_1 \cdots c_{\sigma-2}\, c_{\sigma-1})\,(c_1 \cdots c_{\sigma-2}\, c_{\sigma-1})\, c_\sigma\, (c_2 \cdots c_{\sigma-2}\, c_{\sigma-1}\, c_\sigma)\,(c_2 \cdots c_{\sigma-2}\, c_{\sigma-1}\, c_\sigma)\, c_1$



$\sigma = 5$

1 2 3 4 1 2 3 4 5 2 3 4 5 2 3 4 5 1