Periodicity Property Testing on Strings with Wildcards

Carl Barton
Birkbeck, University of London, UK

Panagiotis Charalampopoulos
King's College London, UK

Taha El Ghazi
DIENS, École normale supérieure de Paris, PSL Research University, France

Jonas Ellert
CWI, Amsterdam, The Netherlands

Oded Lachish
Birkbeck, University of London, UK

Tatiana Starikovskaya
DIENS, École normale supérieure de Paris, PSL Research University, France

Abstract

In this work, we study periodicity in strings with wildcards. A string T with at most k wildcards is called strongly (p,k)-periodic if the wildcards in T can be replaced with alphabet symbols to obtain a string with period p, and weakly (p,k)-periodic if $T[i]$ matches $T[i+p]$ for all i. Intuitively, both generalize to $(\leq g,k)$-periodicity, which is the property of being (p,k)-periodic for some $p \in [1\,..\,g]$.

An ε-tester for a property $\mathcal{P}$ is an algorithm that distinguishes between strings that satisfy $\mathcal{P}$ and strings where one needs to change at least an ε-fraction of the symbols to obtain a string that satisfies $\mathcal{P}$. We study one-sided error testers, where strings satisfying $\mathcal{P}$ must *always* be accepted, while strings that are ε-far must be rejected with probability at least $2/3$. The complexity of a tester is the worst-case number of symbols of an input of length n it must read to make the decision.

We design the following testers for $p, g \leq n/2$:

1. An ε-tester for strong (p,k)-periodicity with complexity $\tilde{\mathcal{O}}_\varepsilon(1)$ [1].
2. An ε-tester for strong $(\leq g,k)$-periodicity with complexity $\tilde{\mathcal{O}}_\varepsilon(\sqrt{g})$.
3. An ε-tester for weak (p,k)-periodicity with complexity $\tilde{\mathcal{O}}_\varepsilon(\min(k, \frac{n}{k+p}))$.
4. An ε-tester for weak $(\leq g,k)$-periodicity with complexity $\tilde{\mathcal{O}}_\varepsilon(\min(k+\sqrt{gk}, \frac{n}{\sqrt{k}}))$.

Additionally, we show a lower bound on the complexity of ε-testers for weak $(\leq g,k)$-periodicity, implying that our tester for weak $(\leq g,k)$-periodicity is optimal up to a multiplicative $(\varepsilon^{-1}\log(gk))^{\mathcal{O}(1)}$ factor for a wide range of g and k. Finally, our tester for strong $(\leq g,k)$-periodicity generalizes the one of [Lachish and Newman; Algorithmica 2011] for strings without wildcards, matching (up to polylogarithmic factors) the unconditional lower bound of $\tilde{\Omega}(\sqrt{g})$ in said work for constant ε.

2012 ACM Subject Classification Theory of computation → Probabilistic computation; Mathematics of computing → Combinatorics on words

Keywords and phrases periodicity, property testing, wildcards

Digital Object Identifier 10.4230/LIPIcs.CPM.2026.28

Funding Parts of the presented ideas were conceived during a research visit funded by a Royal Society International Exchanges Award. Taha El Ghazi, Jonas Ellert, and Tatiana Starikovskaya were partially funded by grant ANR20-CE48-0001 from the French National Research Agency. Jonas Ellert was partially supported by an ERCIM Alain Bensoussan fellowship.

[1] Hereafter, $\tilde{\mathcal{O}}_\varepsilon$ and $\tilde{\Omega}_\varepsilon$ means that we hide factors polylogarithmic in n, k, g and polynomial in ε.

1 Introduction

In this work, we study property testers for periodicity in the presence of wildcards. An *ε-tester* is an algorithm that, given an input string, must decide whether it satisfies a given property, while querying as few symbols of the input as possible. Often, it is impossible to make an exact decision without querying the entire input; instead, one wants to distinguish between strings that have the property and those that are *far* from having the property, where far means that one has to change an ε-fraction of the symbols of the input to obtain a string satisfying the property (for a given ε). The tester does not access the input directly and instead uses oracle queries of the form: what is the symbol at a given position? The number of queries a tester makes is referred to as its *complexity*. We study one-sided error testers, which *always* accept strings satisfying the property, while rejecting strings that are ε-far from having the property with probability at least $2/3$.

A string $T[0\,..\,n)$ without wildcards has a *period* p if $T[i] = T[i+p]$ for all $i \in [0\,..\,n-p)$. Property testing for periodicity in such strings has been previously studied in [9, 10] and [15]. Ergün et al. [9, 10] studied the problem using three generalizations of periodicity that lead to different interpretations of approximate periodicity. They show that the resulting definitions of approximate periodicity are constant approximations of each other and that it is possible to distinguish between the case when T has an ε-approximate period $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$ and T is ε-far from having a period $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$ using $\mathcal{O}(\sqrt{n} \cdot \operatorname{polylog} n)$ queries, even for constant ε. Lachish and Newman [15] showed a property tester that decides whether a string has a period $p \leq n/2$ and has complexity $\mathcal{O}(1/\varepsilon)$. They further showed that, to decide whether there is a period $p \in [1\,..\,g]$ for $g \leq n/2$, it suffices to query $\mathcal{O}(\varepsilon^{-1}\sqrt{g \log g})$ symbols.

In this work, we focus on strings with wildcards. A *wildcard*, denoted by $\diamond$, matches all symbols from the alphabet, intuitively representing a symbol whose exact value is irrelevant or unavailable. We consider two natural extensions of the notion of periodicity to strings containing wildcards: *strong periodicity* and *weak periodicity*.

A string $T[0\,..\,n)$ with at most k wildcards is said to have a *strong period* (p, k) if there exists a wildcard-free string $T'[0\,..\,n)$ that matches T and has period p. Equivalently, T has a strong period (p, k) if its wildcards can be replaced by alphabet symbols so that the resulting string T' satisfies $T'[i] = T'[i+p]$ for all $i \in [0\,..\,n-p)$.

For *weak periodicity*, we relax this requirement: we say that T has a *weak period* (p, k) if $T[i] \cong T[i+p]$ for all $i \in [0\,..\,n-p)$, where $a \cong b$ if and only if $a = b$ or at least one of a and b is a wildcard. For example, $\texttt{xxx} \diamond \texttt{yy}$ is weakly $(1, 1)$-periodic (but not strongly). Both notions of periodicity with wildcards have previously been studied extensively in the context of generalizing the Fine–Wilf periodicity lemma [11]; see [2–7, 14, 18–20].

Our ε-testers. For a string $T[0\,..\,n)$, integers $p \leq n/2$, $g \leq n/2$, and k, and a value ε, we design the following testers:

1. An ε-tester for strong (p, k)-periodicity with complexity $\mathcal{O}(\varepsilon^{-2})$; for $k \leq \varepsilon n/32$, we achieve complexity $\mathcal{O}(\varepsilon^{-1})$. See Lemma 9.
2. An ε-tester for strong $(\leq g, k)$-periodicity with complexity $\mathcal{O}(\varepsilon^{-1}\sqrt{g \log g})$; for $k \leq \varepsilon n/32$, we achieve complexity $\mathcal{O}(\varepsilon^{-1/2}\sqrt{g \log g})$. See Lemma 10.
3. An ε-tester for weak (p, k)-periodicity with complexity $\tilde{\mathcal{O}}_\varepsilon(\min(k, \frac{n}{k+p}))$. See Corollary 23.
4. An ε-tester for weak $(\leq g, k)$-periodicity with complexity $\tilde{\mathcal{O}}_\varepsilon(\min(k + \sqrt{gk}, \frac{n}{\sqrt{k}}))$. See Corollary 24.

The starting point of our testers is a conceptual partition of the input string T into equivalence classes for a fixed period p. Each equivalence class is a subsequence of T consisting of the symbols of T whose positions are in the same class modulo p. If T is strongly

(p, k)-periodic, then every equivalence class must be unary. Intuitively, a witness against strong periodicity is a pair of positions in an equivalence class in which the symbols do not match. On a very high level, we show that for far instances the number of such witnesses is large, which allows detecting them with a few queries. For weakly (p, k)-periodic strings, every equivalence class contains at most $k + 1$ maximal unary factors separated by wildcards. Here, a witness consists of $k + 1$ non-matching pairs that do not overlap. Intuitively, each wildcard can "fix" one such pair (for example, consider $a \diamond b \diamond c$ which is weakly $(1, 2)$-periodic and has two such pairs), but we would need $k + 1$ wildcards to fix $k + 1$ non-overlapping non-matching pairs, which we simply do not have. To show that the number of such witnesses in far instances is large, we demonstrate that a far instance can be decomposed into $k + 1$ factors, where each factor is far from unary. We then need to find one non-matching pair in each factor, which can be done with a few queries.

Lower bounds. Lachish and Newman [15] showed a lower bound on the complexity of testers for periodicity on strings without wildcards. Namely, they showed that every 1/32-tester for $(\leq g, 0)$-periodicity must query $\Omega(\sqrt{g}/\log n)$ symbols of the input, for every g. This implies that, for constant k and ε, the tester in Lemma 10 is optimal up to polylogarithmic factors.

Finally, in Theorem 31, we show a lower bound on the complexity of testers for weak $(\leq g, k)$-periodicity. In a simplified form, it states that any 1/64-tester for weak $(\leq g, k)$-periodicity has complexity $\tilde{\Omega}(\min(\sqrt{gk}, \max(\sqrt{n}, \frac{n}{k})))$. In the regime when g and k are sufficiently small, for example, when $g, k \leq \sqrt{n}$, our tester for weak $(\leq g, k)$-periodicity (Corollary 24) matches this bound up to polylogarithmic factors.

Other related work. Awofeso et al. [1] studied property testing for quasiperiodicity. The quasiperiodicity studied in this work is a cover, where they presented an algorithm that can decide if a string has a cover of length q using $\mathcal{O}(\varepsilon^{-1}q^3 \log q)$ queries. Chan et al. [8] and then Jin and Kociumaka [13] studied property testers for pattern matching.

2 Preliminaries

For integers $i, j \in \mathbb{Z}$, we write $[i\,..\,j] = [i\,..\,j+1) = (i-1\,..\,j] = (i-1\,..\,j+1)$ to denote $\{k \in \mathbb{Z} : i \leq k \leq j\}$, and $[i]$ to denote $[0\,..\,i)$. Let Primes denote the set of prime numbers.

An *alphabet* Σ is a finite set whose elements are called *symbols*. We consider a special symbol $\diamond \notin \Sigma$ called a *wildcard* and denote $\Sigma \cup \{\diamond\}$ by $\Sigma_\diamond$. For two symbols $a, b \in \Sigma_\diamond$ we say that a matches b and write $a \cong b$ if either $a, b \in \Sigma$ and $a = b$, or at least one of a and b is a wildcard. Otherwise, we say that a and b do not match and write $a \ncong b$.

A length-n string T over an alphabet $\Sigma_\diamond$ is a sequence of n symbols from $\Sigma_\diamond$. We write $T[0\,..\,n)$ to denote that T is a string of length $|T| = n$. For $i, j \in [0\,..\,n)$, we write $T[i]$ to denote the i-th symbol in T, and $T[i\,..\,j] = T[i\,..\,j+1) = T(i-1\,..\,j] = T(i-1\,..\,j+1)$ to denote the sequence $T[i]T[i+1]\ldots T[j]$ if $i \leq j$, and the *empty string* ε otherwise. $T[i\,..\,j]$ is called a *factor* of T. We say that T is a k-wildcard string if it contains at most k wildcards.

Consider two strings $T_1, T_2 \in \Sigma_\diamond^n$. We say that T_1 and T_2 match and write $T_1 \cong T_2$ if, for all $i \in [n]$, $T_1[i] \cong T_2[i]$.

Strong and weak periods. We now give formal definitions of strong and weak periods of strings containing wildcards. Intuitively, a string is strongly periodic if we can replace all its wildcards with symbols from Σ to obtain a periodic string in Σ^*.

▶ **Definition 1** (Strongly periodic string). *Let $T[0\,..\,n) \in \Sigma_\diamond^*$, $k \in [0\,..\,n]$, and $p \in [n]$. We say that T is strongly (p,k)-periodic or that T has a strong period (p,k) if T contains at most k wildcards and there is a string $Q \in \Sigma^p$ such that $T \cong Q^\infty[0\,..\,n)$, where Q^∞ is an infinite string obtained by concatenating infinitely many copies of Q.*

Weak periodicity is a more relaxed notion that, intuitively, allows wildcards to act as transition points.

▶ **Definition 2** (Weakly periodic string). *Let $T[0\,..\,n) \in \Sigma_\diamond^*$, $k \in [0\,..\,n]$, and $p \in [n]$. We say that T is weakly (p,k)-periodic or that T has a weak period (p,k) if and only if T contains at most k wildcards and $T[0\,..\,n-p) \cong T[p\,..\,n)$.*

Property testers. In the property testing model, the goal is to decide whether the input has a certain property or is far from having it by querying as few of the input symbols as possible. We next formalize what it means for a string to be far from having some property.

▶ **Definition 3** (ε-far). *We say that a string T is* ε-far *from having a property $\mathcal{P}$ if one needs to change at least $\varepsilon|T|$ symbols of T to obtain a string that has property $\mathcal{P}$.*

▶ **Definition 4** (Property tester). *A one-sided* ε-property tester *for a property $\mathcal{P}$ is an algorithm that receives as an input a string T and:*

- *returns "yes" if T satisfies the property $\mathcal{P}$ with probability 1,*
- *returns "no" with probability at least $2/3$ if T is ε-far from satisfying the property,*
- *returns "yes" or "no" otherwise.*

In this work, we focus on *non-adaptive* property testers, meaning that the symbols to query are selected offline, i.e. the, position of the i-th queried symbol does not depend on the first $i-1$ queried symbols. Consistent with the literature, we define the *complexity* of a tester as the number of symbols of the input string that they query in the worst case.

Structure of strings with wildcards. Before describing periodicity property testers, we first introduce additional notation used throughout the paper. For analyzing whether a string $T[0\,..\,n)$ is (p,k)-periodic for some p and k, we define the following strings $T_0, \ldots, T_{p-1}$: for $i \in [n]$, we have $T_{i \bmod p}[\lfloor i/p \rfloor] = T[i]$, i.e., T_j is obtained by restricting T to positions equivalent to $j \pmod p$. We refer to strings $T_0, \ldots, T_{p-1}$ as *equivalence classes*.

We will analyze (factors of) equivalence classes via symbol frequencies. For any string $S \in \Sigma_\diamond^*$ and any symbol $s \in \Sigma_\diamond$, we define $\#_S(s)$ to be the number of occurrences of s in S. If $s \in \Sigma$, then we further define $\#_S(\bar{s}) = |S| - \#_S(s) - \#_S(\diamond)$, i.e., the number of occurrences of symbols that are neither s nor $\diamond$. The *majority symbol* of S is $s \in \Sigma$ that maximizes $\#_S(s)$; if multiple symbols satisfy this definition, then we choose the one that occurs first in S. For this uniquely defined majority symbol s, we write $\#_S^+ = \#_S(s)$ and $\#_S^- = \#_S(\bar{s})$. All other symbols are called *minority symbols*.

▶ **Proposition 5.** *For every string $S \in \Sigma_\diamond^*$, there is a bipartition A, B of Σ such that $\sum_{s\in A} \#_S(s) \geq \#_S^-/2$ and $\sum_{s\in B} \#_S(s) \geq \max\{\#_S^-/2, (|S| - \#_S(\diamond))/2\}$.*

Proof. Let s^* be the majority symbol of S. We start with empty sets A and B and process the symbols in $\Sigma \setminus \{s^*\}$ in an arbitrary order. We add each symbol we encounter to the set that currently contributes fewer positions to S. Clearly, at any point in time, the terms $\sum_{s\in A} \#_S(s)$ and $\sum_{s\in B} \#_S(s)$ differ by at most $\#_S^+$. We assume, without loss of generality, that in the end we have $b := \sum_{s\in B} \#_S(s) \leq a := \sum_{s\in A} \#_S(s)$, meaning that $a \geq \#_S^-/2$.

Finally, we insert s^* to B. Then, we have $\sum_{s\in B} \#_S(s) = b + \#_S^+ \geq a \geq \#_S^-/2$, which concludes the proof since $\sum_{s\in B} \#_S(s) \geq a$ and $a + \sum_{s\in B} \#_S(s) = |S| - \#_S(\diamond)$ imply that $\sum_{s\in B} \#_S(s) \geq \max\{\#_S^-/2, (|S| - \#_S(\diamond))/2\}$. ◀

3 Strong Periodicity

We start by introducing the notion of witnesses of strong non-periodicity.

▶ **Definition 6** (Witnesses). *Let $T \in \Sigma_\diamond^n$, and $p \geq 1$. We say that $(i, j) \in [n]^2$ is a p-*witness *if $j \equiv i \pmod p$ and $T[i] \not\approx T[j]$. We denote the set of all p-witnesses of T by $W_p(T)$.*

▶ **Observation 7.** *A string $T \in \Sigma_\diamond^n$ is strongly (p, k)-periodic if and only if it is a k-wildcard string and $W_p(T) = \emptyset$.*

In strings that are far from periodic the number of p-witnesses is large:

▶ **Lemma 8.** *Let $k \in [n]$ and $T \in \Sigma_\diamond^n$ be a k-wildcard string that is ε-far from being strongly (p, k)-periodic. We have $|W_p(T)| \geq \frac{\varepsilon^2 n^2}{2p}$. Further, if $k \leq \varepsilon n/32$, then $|W_p(T)| \geq \frac{\varepsilon n^2}{16p}$.*

Proof.

Case I (unbounded k). There are at least εn positions $i \in [0\,..\,n)$ such that $T[i] \in \Sigma$ and $T[i]$ is a minority symbol of $T_{i \bmod p}$. Otherwise, we could simply replace each minority symbol to obtain a strongly (p, k)-periodic string, contradicting the fact that T is ε-far. Let us fix such a position i, and set $s := T[i]$ and $j := i \bmod p$. Our fixed symbol can be paired with each symbol from $\Sigma \setminus \{s\}$ in T_j to give a witness pair, and the number of such symbols is $\#_{T_j}^- + \#_{T_j}^+ - \#_{T_j}(s) \geq \#_{T_j}^-$. In particular, by summing over all minority symbols from Σ, we conclude that each equivalence class contributes at least $(\#_{T_j}^-)^2/2$ unordered pairs, resulting in $(\#_{T_j}^-)^2/2$ p-witnesses (where we divide by 2 to avoid double-counting). Then, by the Cauchy–Schwarz inequality, we have $\sum_{j=0}^{p-1}(\#_{T_j}^-)^2/2 \geq \left(\sum_{j=0}^{p-1} \#_{T_j}^-\right)^2/(2p) \geq \varepsilon^2 n^2/(2p)$.

Case II ($k \leq \varepsilon n/32$). The bound on k implies that there are $\leq \varepsilon p/8$ classes with $\geq n/(4p)$ wildcards. The combined length of these classes is $\leq \varepsilon p/8 \cdot \lceil n/p \rceil \leq \varepsilon n/4$. We claim that there are $\geq 3\varepsilon n/4$ positions $i \in [0\,..\,n)$ such that $T[i] \in \Sigma$ is not the majority symbol of $T_{i \bmod p}$ and $\#_{T_{i \bmod p}}(\diamond) < n/(4p)$. Otherwise, we could simply overwrite the at most $\varepsilon p/8$ classes that have $\geq n/(4p)$ wildcards with unary strings, using $\varepsilon n/4$ substitutions, and then replace the minority symbols in the remaining classes with the majority symbol using another $< 3\varepsilon n/4$ substitutions. This contradicts that T is ε-far from being strongly (p, k)-periodic. Among the at least $3\varepsilon n/4$ positions specified above, fix a position i, and let $s := T[i]$ and $j := i \bmod p$. The symbol s can be paired with every symbol from $\Sigma \setminus \{s\}$ in T_j to give a p-witness. The number of such symbols is $|T_j| - \#_{T_j}(\diamond) - \#_{T_j}(s) \geq n/p - n/(4p) - n/(2p) \geq n/(4p)$. Thus, in total, we have $\geq 3\varepsilon n/4 \cdot (n/(4p))/2 \geq \frac{\varepsilon n^2}{16p}$ p-witnesses (we have again divided by 2 to avoid double-counting). ◀

3.1 Strong Periodicity Testers

We now show testers for strong periodicity.

▶ **Lemma 9.** *For all $\varepsilon \in (0, 1)$ and $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$, Algorithm 1 is a one-sided ε-property tester for strong (p, k)-periodicity with complexity $\mathcal{O}(\varepsilon^{-2})$. The complexity is $\mathcal{O}(\varepsilon^{-1})$ if $k \leq \varepsilon n/32$.*

Algorithm 1 A strong (p,k)-periodicity tester for a fixed candidate period $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$, a k-wildcard string $T \in \Sigma_\diamond^n$

1: $r \leftarrow \begin{cases} \lceil 32/\varepsilon \rceil, & \text{if } k \leq \varepsilon n/32, \\ \lceil 4/\varepsilon^2 \rceil, & \text{otherwise} \end{cases}$
2: sample r pairs $(x,y) \in [n]^2$, $x \equiv y \pmod p$, independently and uniformly at random
3: **for** each sampled pair (x,y) **do**
4: **if** $T[x] \not\approx T[y]$ **then**
5: **return** "no"
6: **return** "yes"

Proof. The algorithm never returns "no" if it receives a strongly (p,k)-periodic string by Observation 7. Next, consider a k-wildcard string $T \in \Sigma_\diamond^n$ that is ε-far from being strongly (p,k)-periodic. Since the number of pairs $(x,y) \in [n]^2$ such that $x \equiv y \pmod p$ is at most $\frac{n^2}{p}$, the probability that a randomly selected pair belongs to $W_p(T)$ is at least $\frac{|W_p(T)|}{n^2/p}$. Hence, the probability that our algorithm returns "yes" is at most $\big(1 - |W_p(T)|/(n^2/p)\big)^r \leq \exp\big(-r \cdot |W_p(T)|/(n^2/p)\big) \leq e^{-2} \leq 1/3$, where the penultimate step is due to Lemma 8. ◀

We now generalize the algorithm to test for strong $(\leq g,k)$-periodicity. A key observation from the algorithm above is that roughly $p/(\varepsilon^2 n)$ independent pairs of positions suffice to discover a p-witness. By sampling around $\sqrt{g/(\varepsilon^2 n)}$ *positions*, we can still form roughly $g/(\varepsilon^2 n) \geq p/(\varepsilon^2 n)$ pairs. By slightly increasing the sampling rate, we can account for the fact that the pairs are not entirely independent, and also boost the success probability enough to succeed for all $p \in [1\,..\,g]$ with high enough probability. This results in the lemma below, whose proof is provided in Appendix B.

Algorithm 2 A strong $(\leq g,k)-$periodicity tester for $g \in [2\,..\,\lfloor \frac{n}{2} \rfloor]$.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $g \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$, a k-wildcard string $T \in \Sigma_\diamond^n$

1: $r \leftarrow \begin{cases} 1024\sqrt{16 \cdot g \ln g/(\varepsilon n^2)}, & \text{if } k \leq \varepsilon n/32, \\ 1024\sqrt{2 \cdot g \ln g/(\varepsilon^2 n^2)}, & \text{otherwise.} \end{cases}$
2: sample each position $i \in [n]$ independently with probability r
3: $I \leftarrow$ the set of the sampled positions
4: **if** $|I| \geq 2rn$ **then return** "yes"
5: **for** $p \in (\lfloor g/2 \rfloor\,..\,g]$ **do**
6: **if** I does not contain a p-witness **then return** "yes"
7: **return** "no"

▶ **Lemma 10.** *For $g \in [2\,..\,\lfloor \frac{n}{2} \rfloor]$ and $\varepsilon \geq (\frac{\ln g}{g})^{1/6}$, Algorithm 2 is a one-sided ε-property tester for strong $(\leq g,k)$-periodicity with complexity $\mathcal{O}(\varepsilon^{-1}\sqrt{g \log g})$.*

If $k \leq \varepsilon n/32$, for $g \in [2\,..\,\lfloor \frac{n}{2} \rfloor]$ and $\varepsilon \geq (\frac{\ln g}{g})^{1/3}$, Algorithm 2 is a one-sided ε-property tester for strong $(\leq g,k)$-periodicity with complexity $\mathcal{O}(\varepsilon^{-1/2}\sqrt{g \log g})$.

▶ Remark 11. For constant g, Lemma 10 requires that ε is at least constant. However, in this case, we can obtain a tester for small g and arbitrary ε by simply invoking Lemma 9 for each $p \in [1\,..\,g]$ (a few times to ensure that the error probability does not blow up).

4 Weak Periodicity

We first introduce the notion of witnesses for weak periodicity. For Lemma 14 below, we essentially apply Proposition 5 to each class of T that does not contain too many wildcards.

▶ **Definition 12.** *Let $T \in \Sigma_\diamond^*$ be a k-wildcard string and $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$.*

- *A* p-witness *is a pair (a, b) such that $a \equiv b \pmod p$ and $T[a] \not\cong T[b]$.*
- *An* explicit p-witness *is a p-witness (a, b) such that $b = a + p$.*
- *An* implicit p-witness *is a set of $k + 1$ pairs $\{(a_0, b_0), \ldots, (a_k, b_k)\}$ such that*
  - *for each $i \in [0\,..\,k]$, either $a_i = b_i$ and $T[a_i] = T[b_i] = \diamond$, or (a_i, b_i) is a p-witness, and*
  - *for each distinct $i, j \in [0\,..\,k]$, if $a_i \equiv a_j \pmod p$ then $[a_i\,..\,b_i) \cap [a_j\,..\,b_j) = \emptyset$ (disjointness condition).*

▶ **Lemma 13.** *Weakly (p, k)-periodic strings admit neither explicit nor implicit p-witnesses.*

Proof. Assume that some string T is weakly (p, k)-periodic. In this case, it does not contain explicit p-witnesses because $T[i] \cong T[i + p]$ for all i. Furthermore, assume towards a contradiction that there exists an implicit p-witness $\{(a_0, b_0), \ldots, (a_k, b_k)\}$. For every (a_i, b_i) such that $a_i = b_i$ and $T[a_i] = T[b_i] = \diamond$, we obviously have $T[m_i] = \diamond$ for $m_i := a_i$. For every (a_i, b_i) which is a p-witness, there exists $m_i \in (a_i\,..\,b_i)$ with $m_i \equiv a_i \pmod p$ and $T[m_i] = \diamond$ (as otherwise the class $T_{a_i \bmod p}$ cannot be weakly 1-periodic). By the disjointness condition, all m_i are distinct, and hence T contains $k + 1 > k$ wildcards, a contradiction. ◀

▶ **Lemma 14.** *For $k \le \varepsilon n/16$, consider a k-wildcard string $T \in \Sigma_\diamond^n$ and $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$. If T is ε-far from being weakly (p, k)-periodic, then there is a set $J \subseteq [0\,..\,p)$ and, for each $j \in J$, sets $A_j, B_j \subseteq [0\,..\,|T_j|)$ satisfying:*

- *for every $j \in J$ and $(a, b) \in A_j \times B_j$, $T_j[a] \not\cong T_j[b]$;*
- *$\sum_{j \in J} |A_j| \ge \varepsilon n/4$, and, for every $j \in J$, $|B_j| \ge n/(8p)$;*
- *if each position in $[n]$ is sampled independently with probability $r \le p/(2n)$, the sample hits A_j (resp. B_j) independently with probability at least $r|A_j|/2$ (resp. $r|B_j|/2$).*

Proof. Consider the equivalence classes $T_0, \ldots, T_{p-1}$. For each j, define the set $M_j \subseteq [0\,..\,|T_j|)$ to contain the positions corresponding to the minority symbols of T_j, i.e., $|M_j| = \#^-_{T_j}$. We have $\sum_{j=0}^{p-1} |M_j| \ge \varepsilon n$. Otherwise, we could merely replace all minority symbols with the majority ones to obtain a (p, k)-periodic string, contradicting the fact that T is ε-far. Let $J \subseteq [0\,..\,p)$ be the set containing exactly the indices $j \in [0\,..\,p)$ such that T_j contains at most $n/(4k)$ wildcards. Note that there are at most $4kp/n$ classes that contain more than $n/(4k)$ wildcards and they contribute at most $4kp/n \cdot \lceil n/p \rceil \le 8k \le \varepsilon n/2$ minority elements. Furthermore, $|J| \ge p - 4kp/n \ge p/2$ (where the second step is due to $k < n/8$). We have shown that $\sum_{j \in J} |M_j| \ge \varepsilon n/2$. For each class T_j with $j \in J$, by Proposition 5, we can bipartition its positions into sets A_j and B_j of sizes $|A_j| \ge |M_j|/2$ and $|B_j| \ge (|T_j| - n/(4p))/2 \ge n/(8p)$ such that every pair in $A_j \times B_j$ yields a mismatch in T_j. For the final point, all sets are disjoint and thus hit independently. We observe that $|A_j| \le |T_j| \le n/(2p)$. Hence, $r|A_j| \le 1$. The probability of not missing A_j is at least $1 - (1 - r)^{|A_j|} \ge \frac{r|A_j|}{1 + r|A_j|} \ge r|A_j|/2$ (using $1 - x \le e^{-x} \le 1/(1 + x)$). The proof for B_j is symmetric. ◀

4.1 Testing Weak (p, k)-Periodicity Through Implicit Witnesses

We start with testers for a fixed period p.

▶ **Lemma 15.** *For $p > 2560k/\varepsilon$, Algorithm 3 is a non-adaptive one-sided error ε-tester for weak (p, k)-periodicity with complexity $\mathcal{O}(\varepsilon^{-1}k)$.*

Algorithm 3 A weak (p, k)-periodicity tester for fixed large period p.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $p \in [1 \,..\, \lfloor \frac{n}{2} \rfloor]$ with $p \geq 2560k/\varepsilon$, a k-wildcard string $T \in \Sigma_\diamond^n$

1: sample each position $i \in [n]$ independently with probability $r = 1280k/(\varepsilon n)$
2: $I \leftarrow$ the set of the sampled indices
3: **if** $|I| \geq 2rn$ **then return** "yes"
4: $I' \leftarrow \emptyset$
5: **for** $i \in I$ **do**
6: $\quad I' \leftarrow I' \cup \{i'\}$, where $i' \in [n]$, $i' \equiv i \pmod p$ is chosen uniformly at random
7: **if** $I \cup I'$ contains an implicit p-witness **then return** "no"
8: **return** "yes"

Proof. Observe that the algorithm only returns "no" if it finds an implicit p-witness. By Lemma 13, a weakly (p, k)-periodic string contains no implicit p-witness; thus, the tester has one-sided error, i.e., it answers "yes" for every weakly (p, k)-periodic string.

Now, assume that T is ε-far from being weakly (p, k)-periodic. If the algorithm answers "yes", then at least one of the following holds:

- the algorithm samples more than $2rn$ positions in the first stage;
- the set of sampled positions does not contain an implicit p-witness.

The probability of the first event is at most $\exp(-rn/3) \leq \exp(-400) < 0.01$ via a Chernoff bound (Fact 33 invoked with $\delta = 1$). We only query the input if $|I| \leq 2rn$ and thus $|I \cup I'| \leq 4rn$. Hence, the complexity is as claimed. For the second event, we show that, with constant probability, we find $k + 1$ (non-explicit) p-witnesses in $k + 1$ different equivalence classes. These pairs form an implicit p-witness.

We use the construction from Lemma 14, consisting of $J \subset [0 \,..\, p)$ and, for $j \in J$, sets A_j, B_j such that $T_j[a] \not\approx T_j[b]$ for every $(a, b) \in A_j \times B_j$. Since we sample with rate $r \leq p/(2n)$, by Lemma 14, we expect to hit each set A_j independently with probability at least $r|A_j|/2$. If we hit A_j, then we also sample another position from the same equivalence class in the second stage. This position is in B_j with probability at least $|B_j|/|T_j| \geq 1/16$ due to Lemma 14. Hence, with probability at least $r|A_j|/32$, we hit both A_j and B_j.

Finally, the expected number of $j \in J$ for which we succeed, i.e., for which we hit both A_j and B_j and thus find a witness pair, is at least $\mu \geq \sum_{j=0}^{p-1} \frac{r|A_j|}{32} \geq \frac{r\varepsilon n}{128} = 10k$, where the middle step follows from Lemma 14. Since each j succeeds independently, we can apply a Chernoff bound (Fact 33) with $\delta = 9/10$ to show that we succeed for less than $k + 1$ classes with probability at most $2\exp(-\delta^2\mu/3) \leq 2\exp(-2.7k) \leq 0.15$. Otherwise, the at least $k+1$ discovered pairs constitute an implicit witness. Finally, using the union bound for the two sources of error, we answer "yes" for an ε-far instance with probability $\leq 0.16 < 1/3$. ◀

In the approach above, we guarantee that the p-witnesses are disjoint by finding them in distinct equivalence classes. Hence, the approach becomes unsuitable for small p. For example, if $p \leq k$, then we have to find multiple witnesses in one equivalence class.

We propose a different approach for small p. In this case, let $c \in \mathbb{N}_{\geq 1}$ and assume that $p \leq ck/\varepsilon$, and $k \leq \varepsilon^2 n/(8c)$. We introduce the following factorization scheme that will be used to detect witness pairs: first, we split $T[0 \,..\, n)$ into equivalence classes $T_0, \ldots, T_{p-1}$, and then we read and factorize each T_j from left to right. Assume that we have already factorized $T_j[0 \,..\, i)$. The next factor is $T_j[i \,..\, i + \ell)$, where ℓ is the minimal value such that either $i + \ell = |T_j|$ or $\#^-_{T_j[i \,..\, i+\ell)} \geq \varepsilon^2 n/(8ck) \geq 1$. Hence, each factor is the shortest prefix of

the remaining string that has exactly $\lceil \varepsilon^2 n/(8ck) \rceil$ occurrences of minority symbols. Further, we call a factor with exactly $\lceil \varepsilon^2 n/(8ck) \rceil$ occurrences of minority symbols a *complete factor*. This factorization only serves the proof and is not actually computed.

▶ **Lemma 16.** *If T is ε-far from being weakly (p,k)-periodic, then T has at least $k+1$ complete factors over all the equivalence classes, assuming that $p \leq ck/\varepsilon$ and $k \leq \varepsilon^2 n/(8c)$ for some $c \in \mathbb{N}_{\geq 1}$.*

Proof. We first show that T has at least $p+k+1$ factors. If T has at most $p+k$ factors, we can obtain a weakly (p,k)-periodic string with εn substitutions as follows. We first substitute all minority symbols of each factor with its majority symbol. Since $\varepsilon^2 n/(8ck) \geq 1$, we have $\lceil \varepsilon^2 n/(8ck) \rceil \leq \varepsilon^2 n/(4ck)$. Hence, substituting the minority symbols requires at most

$$(p+k) \cdot \varepsilon^2 n/(4ck) \leq (ck/\varepsilon + k) \cdot \varepsilon^2 n/(4ck) \leq 2ck/\varepsilon \cdot \varepsilon^2 n/(4ck) \leq \varepsilon n/2$$

substitutions. We perform another k substitutions to replace each wildcard with the majority symbol of its surrounding factor. Afterwards, each factor is over a unary alphabet. Since there are at most $p+k$ factors, and each of the p equivalence classes contributes at least one factor, there are only k boundaries between factors. For each boundary, we perform one substitution to make one of the boundary symbols into a wildcard. As a result, the entire string is weakly (p,k)-periodic. The total number of substitutions is at most $\varepsilon n/2 + 2k \leq \varepsilon n$. Hence, T has at least $p+k+1$ factors, and thus at least $k+1$ complete factors. ◀

Algorithm 4 A weak (p,k)-periodicity tester for fixed small period p.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $p \in [1 \,..\, \lfloor \frac{n}{2} \rfloor]$ with $k \leq \varepsilon^2 n/(8c)$ and $p \leq ck/\varepsilon$, a k-wildcard string $T \in \Sigma_\diamond^n$

1: sample each position $i \in [n]$ independently with probability $r = 16ck \cdot \ln(20(k+1))/(\varepsilon^2 n)$
2: $I \leftarrow$ the set of the sampled indices
3: **if** $|I| \geq 2rn$ **then return** "yes"
4: **if** I contains an implicit p-witness **then return** "no"
5: **return** "yes"

▶ **Lemma 17.** *Let $c \in \mathbb{N}_{\geq 1}$. For $p \leq ck/\varepsilon$, and $k \leq \varepsilon^2 n/(8c)$, Algorithm 4 is a non-adaptive one-sided error ε-tester for weak (p,k)-periodicity with complexity $\mathcal{O}(ck \log k/\varepsilon^2)$.*

Proof. Since we only answer "no" if we find an implicit p-witness, we always answer "yes" for periodic instances due to Lemma 13. Assume now that T is ε-far from being weakly (p,k)-periodic. Consider a complete factor of T. By Lemma 16, there are $\geq k+1$ of those. By Proposition 5 there is a bipartition A, B of the alphabet such that the symbols of A (resp. B) correspond to a set of at least $\varepsilon^2 n/(16ck)$ positions of the factor. Crucially, if we pair up any position with a symbol from A with any position with a symbol from B, we obtain a witness pair. Furthermore, since factors in the same equivalence class do not overlap, a set of $k+1$ witness pairs over the $k+1$ different factors forms an implicit witness. As a result, if the algorithm answers "yes", then at least one of the following holds:

- the algorithm samples more than $2rn$ positions;
- the algorithm misses a witness pair in at least one of the complete factors.

The probability of the first event is at most $\exp(-rn/3) \leq \exp(-16 \ln 40/3) < 0.01$ via a Chernoff bound (Fact 33 invoked with $\delta = 1$). We only query the input if $|I| \leq 2rn$, and thus the complexity is as claimed. The probability of not sampling a position corresponding to a

symbol in A (analogously in B) in a fixed factor is at most $(1-r)^{\varepsilon^2 n/(16ck)} \leq e^{-r\varepsilon^2 n/(16ck)} = 1/(20(k+1))$. Hence, by the union bound, we fail for at least one factor with probability at most 0.1. Combining the two sources of failure by the union bound, we answer "yes" for a far instance with probability at most 0.2. ◀

4.2 Testing Weak $(\leq g, k)$-Periodicity Through Implicit Witnesses

We now present a tester for weak $(\leq g, k)$-periodicity which uses implicit witnesses.

Algorithm 5 A tester for weak $(\leq g, k)$-periodicity for small g.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $c, n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $g \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$ with $g \leq ck/\varepsilon$ and $k \leq \varepsilon^2 n/(8c)$, a k-wildcard string $T \in \Sigma_\diamond^n$

1: sample each $i \in [n]$ independently with probability $r = 16ck \cdot \ln(20(g(k+1)))/(\varepsilon^2 n)$
2: $I \leftarrow$ the set of the sampled indices
3: **if** $|I| \geq 2rn$ **then return** "yes"
4: **for** $p \in [1\,..\,g]$ **do**
5: **if** I does not contain an implicit p-witness **then return** "yes"
6: **return** "no"

▶ **Lemma 18.** *Let $c \in \mathbb{N}_{\geq 1}$. For $g \leq ck/\varepsilon$ and $k \leq \varepsilon^2 n/(8c)$, Algorithm 5 is a non-adaptive one-sided error ε-tester for weak $(\leq g, k)$-periodicity with complexity $\mathcal{O}(ck \log(gk)/\varepsilon^2)$.*

Proof. The analysis is similar to that of Algorithm 4, but with a higher sampling rate. By Lemma 13, the algorithm answers "yes" for strings that have a weak period $p \in [1\,..\,g]$.

Otherwise, i.e., for an input that is ε-far from being weakly $(\leq g, k)$-periodic, the algorithm only answers "yes" if it samples more than $2rn$ indices, or if, for a certain $p \leq g$, it does not find a witness pair in at least one complete factor. The first event occurs with probability at most $\exp(-rn/3) \leq \exp(-16 \ln 40/3) < 0.01$ via a Chernoff bound (Fact 33 invoked with $\delta = 1$). For a fixed $p \leq g$ and a fixed complete factor, the second event occurs with probability at most $1/(20(k+1) \cdot g)$ (via the same calculation as in Lemma 17). Therefore, over all g possible periods and their $2 \cdot (k+1) \cdot g$ sets of positions, the union bound implies that we fail for at least one set with probability at most 0.1. If T is indeed ε-far from being weakly $\leq g$-periodic, and thus also from being weakly (p, k)-periodic for every $p \in [1\,..\,g]$, then with probability at least 0.8 we can report a witness for every p. ◀

Algorithm 6 A tester for weak $(\leq g, k)$-periodicity for $g \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $g \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$ with $k \leq \varepsilon^2 n/(8 \cdot \lceil 4096 + (\ln g)/k \rceil)$, a k-wildcard string $T \in \Sigma_\diamond^n$

1: $g' \leftarrow \min(g, \lfloor 4096(k + \ln g)/\varepsilon \rfloor)$
2: $c \leftarrow \lceil 4096 + (\ln g)/k \rceil$
3: run Algorithm 5 with parameters c and g' and input T three times
4: **if** Algorithm 5 returned "yes" every time **then return** "yes"
5: **if** $g' = g$ **then return** "no"
6: $q \leftarrow g' + 1$
7: sample each $i \in [n]$ independently with probability $r = \sqrt{gq}/(2n)$
8: $I \leftarrow$ the set of the sampled indices
9: **if** $|I| \geq 2rn$ **then return** "yes"
10: **for** $p \in [q\,..\,g]$ **do**
11: **if** I does not contain an implicit p-witness **then return** "yes"
12: **return** "no"

▶ **Lemma 19.** *For $k \leq \varepsilon^2 n/(8 \cdot \lceil 4096 + (\ln g)/k \rceil)$, Algorithm 6 is a non-adaptive one-sided error ε-tester for weak $(\leq g, k)$-periodicity with complexity $\mathcal{O}(\log^2 g/\varepsilon^2 + k \log(gk)/\varepsilon^2 + \sqrt{g \cdot (k + \log g)}/\varepsilon)$.*

Proof. We only answer "no" if both of the following events occur: Algorithm 5 returns "no" at least once (which by Lemma 18 implies that T has no period $\leq g'$), and we find an implicit p-witness for every period $p \in (g' \,..\, g]$. Hence, we never return "no" if the input string is weakly $(\leq g, k)$-periodic. It remains to handle strings that are ε-far from all periods in $[1 \,..\, g']$. We have to show that we accept them with probability at most $2/3$.

First, by Lemma 18, if T is ε-far from being weakly $(\leq g', k)$-periodic, then the probability of all three calls to Algorithm 5 returning "yes" is $1/3^3 < 0.05$. The number of queries is $\mathcal{O}(ck \ln(g'k)/\varepsilon^2) = \mathcal{O}(ck \ln(gk)/\varepsilon^2)$, which is bounded by $\mathcal{O}((\ln^2 g)/\varepsilon^2)$ if $k = \mathcal{O}(\ln g)$; otherwise, $c = \mathcal{O}(1)$ and hence the number of queries is $\mathcal{O}(k \ln(gk)/\varepsilon^2)$.

We next bound the probability of sampling at least $2rn$ indices. This happens with probability at most $\exp(-rn/3) \leq \exp(-q/6) < \exp(-4096/6) < 0.01$ via a Chernoff bound (Fact 33 with $\delta = 1$). We only query the input if the total number of samples is at most $2rn$. Hence, the complexity is as claimed.

If we enter the second stage, then let $\mathcal{G}_p$ be the event that the algorithm does not find an implicit p-witness. By the union bound, we answer "yes" due to one of these events with probability at most $\sum_{p \in [q \,..\, g]} \Pr[\mathcal{G}_p]$. Hence, we can conclude the proof by showing that $\Pr[\mathcal{G}_p] \leq 1/(9g)$, which implies that we answer "yes" with probability at most $1/9$. Then, via the union bound, the error probabilities from all sources sum to $0.05 + 0.01 + 1/9 < 1/3$.

From now on, fix any $p \in [q \,..\, g]$. We perform the analysis with a lower sampling rate $r' := \sqrt{pq}/(2n) \leq r$, which can only lower our success probability. We use the construction from Lemma 14, consisting of $J \subset [0 \,..\, p)$ and, for $j \in J$, sets A_j, B_j such that $T_j[a] \ncong T_j[b]$ for every $(a, b) \in A_j \times B_j$. Due to $r' \leq p/(2n)$, by Lemma 14, we expect to hit each set A_j (resp. B_j) independently with probability $r'|A_j|/2$ (resp. $r'|B_j|/2$). For any $j \in J$, we hit both A_j and B_j (and thus discover a witness pair) with probability at least $r'^2|A_j| \cdot |B_j|/4 \geq r'^2 n|A_j|/(32p)$. The expected number of different j for which we are successful is $\mu \geq \sum_{j=0}^{p-1} \frac{r'^2 n|A_j|}{32p} \geq \frac{r'^2 \varepsilon n^2}{128p} = \frac{q\varepsilon}{512} \geq 8k + 8 \ln g$, where the second step is due to Lemma 14. Finally, applying the Chernoff bound from Fact 33 with $\delta = 7/8$, we are successful for at most $k + \ln g \leq (1 - \delta)\mu$ classes with probability at most $\exp(-\delta^2 \mu/2) \leq \exp(-3k - 3 \ln g) < 1/(9g)$. ◀

4.3 Testing Weak Periodicity Through Explicit Witnesses

▶ **Lemma 20.** *Consider a k-wildcard string $T \in \Sigma_\diamond^*$ that is ε-far from being (p, k)-periodic, where $k \leq \varepsilon n/4$. There are at least $(p + k) \cdot \varepsilon/8$ explicit p-witnesses.*

Proof. At least $\varepsilon p/2$ of equivalence classes of T are not weakly $(1, k)$-periodic. Otherwise we could simply overwrite these classes with unary strings using less than $\varepsilon p/2 \cdot \lceil n/p \rceil \leq \varepsilon n$ substitution. This would result in a weakly (p, k)-periodic string, which contradicts the fact that T is ε-far. Consequently, $\varepsilon p/2$ classes have at least one explicit p-witness each.

We now prove that the number of explicit witnesses exceeds $\varepsilon k/2$. To do so, we further greedily factorize each equivalence class from left to right into maximal factors of the form $S \in \{s, \diamond\}^*$ for any $s \in \Sigma$. We then obtain a string T' without wildcards by replacing each wildcard in T with the symbol of its surrounding factor. This requires $k \leq \varepsilon n/4$ substitutions and results in a string T' that is $3\varepsilon/4$-far from being weakly (p, k)-periodic. Furthermore, if T' contains $k + x$ explicit witnesses, then T contains at least x explicit witnesses.

Let t be the total number of maximal unary factors over all equivalence classes in T'. Since T' is $3\varepsilon/4$-far from being weakly (p,k)-periodic, we have $t > k+1$, or else it could be made into a weakly (p,k)-periodic string by adding at most k wildcards at factor boundaries. We say that a maximal unary factor is short if it is among the shortest $t-k>0$ maximal unary factors, and long otherwise. We claim that the short factors must contribute at least $\varepsilon n/2$ symbols overall. Otherwise, we can again show that $3\varepsilon n/4$ substitutions are sufficient to make T' periodic: We replace each symbol of a short factor with the symbol populating the closest long factor in its equivalence class (or with a if such run is undefined). This effectively merges each short factor with a long one. Every remaining run is either an extension of one of the original k long factor, or an equivalence class that consists of a single factor. The total number of long factors in all classes that contain at least two long factors is at most k. Hence, we can fix their at most k boundaries with wildcards. The total number of substitutions is less than $\varepsilon n/2 + k \leq 3\varepsilon n/4$.

Since short runs contribute $\geq \varepsilon n/2$ symbols, there is a short run of length $\ell \geq \varepsilon n/(2t-2k)$. Consequently, all long factors have length at least ℓ. The total length of long runs is therefore at least $k \cdot \ell \geq k \cdot \frac{\varepsilon n}{2t-2k}$. On the other hand, it is upper-bounded by n, which implies $k \cdot \varepsilon/2 + k \leq t$. Hence, the number of explicit p-witnesses in T' is at least $k + \varepsilon k/2$, which implies that T contains at least $\varepsilon k/2$ explicit p-witnesses, concluding the proof. ◀

Algorithm 7 A tester for weak (p,k)-periodicity.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $p \in [1\,..\,\lfloor \frac{n}{2} \rfloor]$ with $k \leq \varepsilon n/4$,
a k-wildcard string $T \in \Sigma_\diamond^n$

1: **for** $\ell = 1$ to $\lceil \frac{2n}{\varepsilon(p+k)} \rceil$ **do**
2: draw $i \in [0\,..\,n-p)$ uniformly at random
3: **if** $T[i] \not\approx T[i+p]$ **then return** "no"
4: **return** "yes"

The following two lemmas follow from Lemma 20 in a straightforward manner.

▶ **Lemma 21.** *Algorithm 7 is a non-adaptive ε-tester for weak (p,k)-periodicity with complexity $\mathcal{O}(\varepsilon^{-1}n/(k+p))$.*

Proof. Since weakly (p,k)-periodic strings do not contain explicit p-witnesses (see Lemma 13), they are accepted by Algorithm 7. Consider now a string that is ε-far from being weakly (p,k)-periodic. By Lemma 20, the probability that a given pair sampled by the algorithm is an explicit witness is at least $\frac{\varepsilon(p+k)}{n}$. Hence, the probability that the algorithm does not find an explicit witness is at most $(1-\frac{\varepsilon(p+k)}{n})^{\frac{2n}{\varepsilon(p+k)}} \leq e^{-2} \leq 1/3$. ◀

▶ **Lemma 22.** *Algorithm 8 is a non-adaptive ε-tester for weak $(\leq g,k)$-periodicity with complexity $\mathcal{O}\left(n\sqrt{\frac{\log g}{\varepsilon k}}\right)$.*

Proof. First, consider a weakly (p,k)-periodic string for some $1 \leq p \leq g$. It does not contain an explicit p-witness, and hence is accepted by Algorithm 8. Now, consider a string T that is ε-far from being weakly (p,k)-periodic for all $1 \leq p \leq g$. Fix one p, $1 \leq p \leq g$. By Lemma 20, T has at least $\varepsilon(k+p)/8$ explicit p-witnesses. Furthermore, since there are at most 2 explicit p-witnesses sharing the same index, there are at least $\varepsilon(k+p)/16$ explicit witnesses that are mutually disjoint. Let X be the number of these explicit p-witnesses sampled by the algorithm. X is a sum of independent indicator random variables, and therefore we have

$$\Pr(X=0) \leq (1-r^2)^{\varepsilon(k+p)/16} \leq e^{-r^2\varepsilon(k+p)} = e^{-(k+p)\ln(100g)/k} \leq \frac{1}{100g}.$$

Algorithm 8 A tester for weak $(\leq g, k)$-periodicity.

Input: values $\varepsilon \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_{\geq 0}$, $k \in [n]$, $g \in [1 \,..\, \lfloor \frac{n}{2} \rfloor]$ with $k \leq \varepsilon n/4$, a k-wildcard string $T \in \Sigma_\diamond^n$

1: sample each $i \in [n]$ independently with probability $r = 4\sqrt{\frac{\ln(100g)}{\varepsilon k}}$
2: $I \leftarrow$ the set of the sampled indices
3: **if** $|I| \geq 10rn$ **then return** "yes"
4: **for** $p \in [1 \,..\, g]$ **do**
5: **if** I does not contain an explicit p-witness **then return** "yes"
6: **return** "no"

Hence, by the union bound, the probability that the algorithm misses an explicit witness for at least one p such that T is ε-far is at most 0.01. Furthermore, by Fact 33 invoked with $\delta = 9$, the probability that the algorithm samples more than $10rn$ positions is at most 0.01. By the union bound, the probability that the algorithm answers "no" is at least $0.98 \geq 2/3$. ◀

4.4 Testers for Weak Periodicity: Wrap-up

▶ **Corollary 23.** *For $k \leq \varepsilon n/4$, there is a non-adaptive ε-tester for weak (p, k)-periodicity with complexity $\mathcal{O}(\min(\varepsilon^{-2} k \log k, \varepsilon^{-1} n/(k + p)))$.*

Proof. We first show that there is a tester with complexity $\mathcal{O}(\varepsilon^{-2} k \log k)$. If $k > \varepsilon^2 n/320$ then the claimed number of queries exceeds n, and we simply query the whole string. Otherwise, at least one of $p > 2560k/\varepsilon$ or $p \leq 40k/\varepsilon$ holds, and hence we can use one of Lemma 15 and Lemma 17 with $c = 40$. The final claim follows by combining with Lemma 21. ◀

▶ **Corollary 24.** *For $k \leq \varepsilon n/4$, there is a non-adaptive ε-tester for weak $(\leq g, k)$-periodicity with complexity $\mathcal{O}(\min(\log^2 g/\varepsilon^2 + k \log(gk)/\varepsilon^2 + \sqrt{g \cdot (k + \log g)/\varepsilon}, n\sqrt{\frac{\log g}{\varepsilon k}}))$.*

Proof. We first show that there is a tester with complexity $\mathcal{O}(\log^2 g/\varepsilon^2 + k \log(gk)/\varepsilon^2 + \sqrt{g \cdot (k + \log g)/\varepsilon})$. If $k > \varepsilon^2 n/(8 \cdot \lceil 4096 + (\ln g)/k \rceil)$, then $k \log(gk)/\varepsilon^2 = \Omega(n)$, and we can simply query the whole string in the claimed complexity. Otherwise, we use Lemma 19. The final claim follows by combining this result with Lemma 22. ◀

5 Lower Bounds for Weak Periodicity Testing

The proof of the following proposition is deferred to Appendix B.

▶ **Proposition 25.** *Let $M \subseteq [0 \,..\, n)$, $m = |M|$, and $g \in [2 \,..\, \lfloor \frac{n}{2} \rfloor]$. For $p \in [2 \,..\, g]$, define the set $M_p := \{(i, j) \in M^2 \mid i < j \text{ and } i \equiv j \pmod p\}$. If we choose $p \in (\lfloor g/2 \rfloor \,..\, g] \cap \mathit{Primes}$ uniformly at random, then for all $x > 0$, $\Pr(|M_p| > x) \leq (m^2 \log_2 n \ln g)/(gx)$.*

▶ **Proposition 26.** *Let $T[0 \,..\, n)$ be a string with at most k wildcards, $M \subseteq [0 \,..\, n)$ a subset of its positions, and $f : M \to \Sigma$ satisfy $T[i] = f(i)$ for $i \in M$. If M contains neither an implicit nor an explicit p-witness, then the symbols $T[i]$, $i \notin M$, can be chosen so that T is weakly (p, k)-periodic.*

Proof. Consider an equivalence class T_j of T. If T_j contains only undefined symbols, we fill it with an arbitrary symbol of Σ and exclude it from further consideration. Otherwise, we divide T_j from left to right into maximal factors such that, for each such factor, all its

symbols are either equal to a fixed symbol of Σ or undecided. For two neighboring factors $F_1 = T_j[x\,..\,y)$ and $F_2 = T_j[y\,..\,z)$, we cannot have both the symbols adjacent to the boundary defined (i.e., we cannot have both $(y-1)p+(j \bmod p), yp+(j \bmod p) \in M$). Otherwise, the maximality property implies that $F_1[|F_1|-1]$ and $F_2[1]$ must be different, and hence we have an explicit p-witness in M. Let k_j be the number of factors in T_j. We have $\sum_j (k_j - 1) \leq k$, or otherwise we can form an implicit p-witness in M by taking pairs of positions of defined symbols in neighboring factors. Now, we explain how to complete T. We place a wildcard in one of the boundary positions of each pair of neighboring factors, which requires at most k wildcards. Then, we fill each factor with the single symbol from Σ occurring in it. The resulting string is weakly (p,k)-periodic and contains at most k wildcards. ◀

Next, we define and analyze an auxiliary distribution $\mathcal{F}$ on strings in Σ^n. We start by drawing a potential period $p \in [\lfloor g/2 \rfloor\,..\,g] \cap \mathsf{Primes}$ uniformly at random. The classes T_j, for $j \in [0\,..\,\lceil 8\varepsilon p \rceil)$, are chosen to be unary: for all i, $T_j[i] = j+2$. The remaining classes are constructed as follows. For each $j \in (\lceil 8\varepsilon p \rceil\,..\,p)$, we consider a partition $T_j = T_j^0 T_j^1 T_j^2 T_j^4$, where each T_j^i is of length $\lfloor |T_j|/4 \rfloor$ or $\lceil |T_j|/4 \rceil$, T_j^0 contains only zeros, and T_j^1 contains only ones. T_j^2 is constructed using a two-state Markov chain as follows: $T_j[2] = 1$ and, for every position $i \in [1\,..\,|T_j^2|)$, we assign $T_j^2[i] = 1 - T_j^2[i-1]$ with probability $r = 100(p+k)/n$, and $T_j^2[i] = T_j^2[i-1]$ otherwise. T_j^3 is defined in the same way as T_j^2, but starting with a 0.

▶ **Lemma 27.** *A random string T of length n drawn according to $\mathcal{F}$ satisfies each of the following for integers k and g with $k < g/4 < g \leq n/(64 \ln n)$ and a constant $\varepsilon < 1/64$:*

- *T is ε-far from a weakly (p',k)-periodic string for all $p' \neq p$ with $p' \leq g$ with probability 1 and ε-far from a weakly (p,k)-periodic string with probability at least 0.8;*
- *T has less than $2000(p+k)$ explicit p-witnesses with probability at least 0.99.*

Proof. The lemma follows directly from Claims 28, 29, and 30 shown below.

▷ Claim 28. T is ε-far from a weakly (p',k)-periodic string for any $p' \neq p$ with $p' \leq g$.

Proof. For every $j \in [0\,..\,\lceil 8\varepsilon p \rceil)$ we have $T_j[i] = T[i \cdot p + j] \neq T[i \cdot p + j - p'] = T_{j-p' \bmod p}[i]$, because $j - p' \bmod p \neq j$, meaning that $T[i \cdot p + j - p']$ is in a different class modulo p (either in a different class among the first $\lceil 8\varepsilon p \rceil$ classes or in a binary class). This implies that in order to obtain a weak (p',k)-periodic string from T one needs to change at least half of the symbols of the classes T_j, $j \in [0\,..\,\lceil 8\varepsilon p \rceil)$, or $8\varepsilon p \cdot (\lfloor n/p \rfloor/2 - 1) \geq \varepsilon n$ symbols in total. Hence T is ε-far from every period in $([1\,..\,2p] \setminus \{p\}) \supseteq [1\,..\,g] \setminus \{p\}$. ◁

▷ Claim 29. With probability at least 0.99, T has less than $2000(p+k)$ explicit p-witnesses.

Proof. Let N_j be the number of pairs $(a, a+1)$ such that $a \geq |T_j^0 T_j^1|$ and $T_j[a] \neq T_j[a+1]$. Let $N = \sum_j (2 + N_j) = 2p + \sum_j N_j$, which is an upper bound on the total number of explicit p-witnesses in T. We have $N_j \sim \mathsf{Bin}(\lceil |T_j|/2 \rceil, r) + X_j$, where $X_j \in [-5\,..\,5]$ takes care of the ceilings and floors, and the transition between T_j^2 and T_j^3. Hence, we have $\sum_j N_j \sim \mathsf{Bin}(\lceil (n-p)/2 \rceil, r) + Xp$, where $X \in [-5\,..\,5]$. Then, by Fact 33, $\mathsf{Bin}(\lceil (n-p)/2 \rceil, r) \leq 5rn \leq 1000(p+k)$ with probability 0.99. As we have $2p + 1000(p+k) + 5p < 2000(p+k)$, the claim follows. ◁

▷ Claim 30. With probability at least 0.8, T is ε-far from a weakly (p,k)-periodic string.

Proof. To make T weakly (p,k)-periodic, we must ensure that each equivalence class T_j can be partitioned into k_j factors of identical symbols separated by $k_j - 1$ wildcards, where $\sum_j (k_j - 1) \leq k$. We show that we cannot achieve this by changing at most εn symbols

with good probability. Since $k < g/4$ and $\lfloor g/2 \rfloor \leq p$, there are at least $p/4$ classes T_j such that $k_j = 0$. Among those, there are at least $p/8$ classes where we can change at most $\varepsilon 8n/p$ symbols. We consider one such class T_j with $j > \lceil 8\varepsilon p \rceil$, which is possible since $\varepsilon < 1/64$. It contains at least $|T_j^0|$ zeros and at least $|T_j^1|$ ones. Hence, we must change at least $\lfloor n/4p \rfloor > \varepsilon 8n/p$ symbols to obtain a unary string. ◁

◀

▶ **Theorem 31.** *Every non-adaptive one-sided ε-tester for weak $(\leq g, k)$-periodicity has complexity $\Omega(\min(\sqrt{gk/(\log n \log g)}, \max(\sqrt{n/\log g}, n/k)))$, assuming $\varepsilon < 1/64$ and $k < g/4 < g \leq n/(64 \ln n)$.*

Proof. We use Yao's principle adapted for one-sided algorithms: to prove that a one-sided tester requires at least m queries, we show a distribution $\mathcal{D}$ on k-wildcard strings for which every *deterministic* tester that performs m queries either answers "no" on a weakly $(\leq g, k)$-periodic string with non-zero probability, or answers "yes" on an ε-far string with probability greater than $1/3$ (when the string is drawn via $\mathcal{D}$). The distribution $\mathcal{D}$ is defined as follows: with probability $3/4$, a string is drawn according to $\mathcal{F}$, and with probability $1/4$ we draw a weakly (p, k)-periodic string uniformly at random.

By Proposition 26, if the set of queried positions does not form either explicit or implicit p'-witnesses for some $p' \leq g$, the input can be completed as a weakly (p', k)-periodic string. Thus, assuming that the tester does not return "no" on weakly (p', k)-periodic strings with positive probability, the tester must return "yes" if the set of queried symbols contains neither explicit nor implicit p'-witnesses for some $p' \in [1\,..\,g]$.

Henceforth, we assume that the input string drawn according to $\mathcal{D}$ is ε-far from a weakly $(\leq g, k)$-periodic string, which by Lemma 27 happens with probability at least 0.6. It suffices to upper bound the probability that a deterministic algorithm hits either an implicit or an explicit p'-witness for all $p' \in [1\,..\,g]$ on such string. By Lemma 27, this probability is upper-bounded by the probability that an algorithm hits an implicit or an explicit p-witness.

Fix an arbitrary set $M \subseteq [0\,..\,n)$ of size m. If $m < 0.01 \cdot \sqrt{gk/(\log_2 n \ln g)}$, then by Lemma 27 the probability that $|M_p| \geq k$ and hence the probability that the algorithm hits an implicit p-witness is at most $(0.01)^2$ (regardless of the actual string).

Now note that the space of explicit p-witnesses can be defined as $[n]$, since for each explicit p-witness (a, b) we have $b = a + p$. Consequently, if $m < n/(20000(p + k))$, then by Lemma 27, the expected number of explicit p-witnesses hit by a deterministic algorithm is less than $1/10$, which implies that a deterministic algorithm misses an explicit p-witness with probability at least 0.9 by Markov's inequality (Fact 32).

Finally, let $M_p = \{(i, j) \in M^2 \mid i < j \text{ and } i \equiv j \pmod p\}$. By Proposition 25, we have $\Pr(|M_p| > 1000m^2 \ln g/g) < 0.001$. If $m \leq \sqrt{gn/(4 \cdot 10^{14}(p + k) \ln g)}$, then $\Pr(|M_p| > n/(2 \cdot 10^4(p + k))) < 0.001$. Furthermore, by Lemma 27, the expected number of explicit p-witnesses hit by M is less than $|M_p| \cdot 2000(p+k)/n < (n/(2 \cdot 10^4(p+k))) \cdot 2000(p+k)/n < 1/10$ conditional on $|M_p| < n/(2 \cdot 10^4(p + k)))$. Hence, the probability that that a deterministic algorithm misses an explicit p-witness conditional on $|M_p| < n/(2 \cdot 10^4(p + k)))$ is at least 0.9 by Markov's inequality (Fact 32). By the union bound, the probability that an algorithm misses all explicit witnesses in this case is at least 0.8.

Thus, if any of these conditions of m is satisfied, the algorithm fails with probability at least $0.6 \cdot 0.8 > 1/3$, a contradiction. The statement follows. ◀

References

1 Christine Awofeso, Ben Bals, Oded Lachish, and Solon P. Pissis. Testing quasiperiodicity. In *Proceedings of the International Symposium on String Processing and Information Retrieval (SPIRE 2026)*, pages 1–9, 2026. doi:10.1007/978-3-032-05228-5_1.

2 Jean Berstel and Luc Boasson. Partial words and a theorem of Fine and Wilf. *Theoretical Computer Science*, 218(1):135–141, 1999. doi:10.1016/S0304-3975(98)00255-2.

3 F. Blanchet-Sadri. Periodicity on partial words. *Computers & Mathematics with Applications*, 47(1):71–82, 2004. doi:10.1016/S0898-1221(04)90006-5.

4 F. Blanchet-Sadri, Deepak Bal, and Gautam Sisodia. Graph connectivity, partial words, and a theorem of Fine and Wilf. *Information and Computation*, 206(5):676–693, 2008. Special Issue: The 17th International Conference on Concurrency Theory (CONCUR 2006). doi:10.1016/j.ic.2007.11.007.

5 F. Blanchet-Sadri, Travis Mandel, and Gautam Sisodia. Periods in partial words: An algorithm. *Journal of Discrete Algorithms*, 16:113–128, 2012. doi:10.1016/j.jda.2012.04.001.

6 Francine Blanchet-Sadri, Taktin Oey, and Timothy D. Rankin. Fine and Wilf's theorem for partial words with arbitrarily many weak periods. *Int. J. Found. Comput. Sci.*, 21(5):705–722, 2010. doi:10.1142/S0129054110007519.

7 Francine Blanchet-Sadri and Nathan D. Wetzler. Partial words and the critical factorization theorem revisited. *Theor. Comput. Sci.*, 385(1-3):179–192, 2007. doi:10.1016/J.TCS.2007.06.012.

8 Timothy M. Chan, Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, and Ely Porat. Approximating text-to-pattern Hamming distances. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC 2020)*, pages 643–656. ACM, 2020. doi:10.1145/3357713.3384266.

9 Funda Ergun, S. Muthukrishnan, and Cenk Sahinalp. Periodicity testing with sublinear samples and space. *ACM Trans. Algorithms*, 6(2), April 2010. doi:10.1145/1721837.1721859.

10 Funda Ergün, S. Muthukrishnan, and Süleyman Cenk Sahinalp. Sublinear methods for detecting periodic trends in data streams. In *LATIN 2004*, volume 2976 of *LNCS*, pages 16–28, 2004. doi:10.1007/978-3-540-24698-5_6.

11 Nathan J. Fine and Herbert S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16(1):109–114, 1965. doi:10.2307/2034009.

12 Svante Janson, Tomasz Łuczak, and Andrzej Ruciński. *Random Graphs*, chapter Exponentially Small Probabilities, pages 25–51. Wiley-Interscience, 2000. doi:10.1002/9781118032718.ch2.

13 Ce Jin and Tomasz Kociumaka. Near-optimal property testers for pattern matching. In *66th IEEE Annual Symposium on Foundations of Computer Science (FOCS 2025)*, pages 1639–1647. IEEE Computer Society, 2025. doi:10.1109/FOCS63196.2025.00086.

14 Tomasz Kociumaka, Jakub Radoszewski, Wojciech Rytter, and Tomasz Waleń. A periodicity lemma for partial words. *Inf. Comput.*, 283(C), February 2022. doi:10.1016/j.ic.2020.104677.

15 Oded Lachish and Ilan Newman. Testing periodicity. *Algorithmica*, 60(2):401–420, 2011. doi:10.1007/s00453-009-9351-y.

16 Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2nd edition, 2017.

17 J. Barkley Rosser and Lowell Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois Journal of Mathematics*, 6(1), March 1962. doi:10.1215/ijm/1255631807.

18 A. M. Shur and Yu. V. Gamzova. Partial words and the interaction property of periods. *Izvestiya: Mathematics*, 68(2):405–428, 2004. doi:10.1070/IM2004v068n02ABEH000480.

19 Arseny M. Shur and Yulia V. Konovalova. On the periods of partial words. In *Mathematical Foundations of Computer Science 2001*, pages 657–665, 2001. doi:10.1007/3-540-44683-4_57.

20 W.F. Smyth and Shu Wang. A new approach to the periodicity lemma on strings with holes. *Theoretical Computer Science*, 410(43):4295–4302, 2009. doi:10.1016/j.tcs.2009.07.010.

A Probability Inequalities Used in This Work

▶ **Fact 32** (Markov's inequality). *Let X be a nonnegative random variable and $t > 0$.*

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

▶ **Fact 33** (Multiplicative Chernoff bounds, see e.g. [16]). *Let $X_1, \ldots, X_n$ be independent random variables taking values in $\{0, 1\}$ and let $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}[X]$. Then, for any $0 < \delta \leq 1$,*

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\delta^2\mu}{3}\right), \quad \text{and} \quad \Pr[X \leq (1-\delta)\mu] \leq \exp\left(-\frac{\delta^2\mu}{2}\right),$$

and for any $\delta > 0$,

$$\Pr[X \geq (1+\delta)\mu] \leq \exp\left(-\frac{\delta^2\mu}{2+\delta}\right).$$

▶ **Fact 34** (The extended Janson inequality [12, Theorem 2.14 and Remark 2.15]). *Let Ω be a finite universal set and let R be a random subset of Ω given by $\Pr[r \in R] = p_r$, these events are mutually independent over $r \in \Omega$.*

Let $\{A_i\}_{i \in I}$ be subsets of Ω, where I a finite index set, and B_i be the event $A_i \subseteq R$. Let X_i be the indicator random variable for B_i and $X = \sum_i X_i$ be the number of $A_i \subseteq R$.

For $i, j \in I$ we write $i \sim j$ if $i \neq j$ and $A_i \cup A_j \neq \emptyset$. Define $\Delta = \frac{1}{2}\sum_{i \sim j} \Pr[B_i \wedge B_j]$ and $\mu = \mathbb{E}[X]$, then $\Pr[\wedge_{i \in I} \overline{B_i}] \leq e^{-\mu^2/(\mu+2\Delta)}$.[2]

B Omitted Proofs

▶ **Lemma 10.** *For $g \in [2 \,..\, \lfloor \frac{n}{2} \rfloor]$ and $\varepsilon \geq (\frac{\ln g}{g})^{1/6}$, Algorithm 2 is a one-sided ε-property tester for strong $(\leq g, k)$-periodicity with complexity $\mathcal{O}(\varepsilon^{-1}\sqrt{g \log g})$.*

If $k \leq \varepsilon n/32$, for $g \in [2 \,..\, \lfloor \frac{n}{2} \rfloor]$ and $\varepsilon \geq (\frac{\ln g}{g})^{1/3}$, Algorithm 2 is a one-sided ε-property tester for strong $(\leq g, k)$-periodicity with complexity $\mathcal{O}(\varepsilon^{-1/2}\sqrt{g \log g})$.

Proof. Assume that the given string is strongly (p, k)-periodic with $p \in [1 \,..\, g]$. Let $h = \lfloor g/p \rfloor$, then $ph \in (\lfloor g/2 \rfloor \,..\, g]$. By the definition of witnesses, every ph-witness is also a p-witness. By Observation 7, the tester does not find a ph-witness and returns "yes". We have shown that the tester accepts all strings that have a strong (p, k)-period for some $p \in [1 \,..\, g]$.

Now, assume T is ε-far from being strongly $(\leq g, k)$-periodic. If the algorithm answers "yes", then at least one of the following holds:

- the algorithm samples $|I| \geq 2rn$ positions (henceforth called event $\mathcal{B}$);
- I does not contain a p-witness for some $p \in (\lfloor g/2 \rfloor \,..\, g]$ (henceforth called event $\mathcal{G}$).

We will show that $\Pr[\mathcal{G}] < 1/9$ and $\Pr[\mathcal{B}] < 1/9$. Then, by the union bound, the probability that at least one of the events occurs (and hence the probability to answer "yes") is at most $2/9 < 1/3$. The probability of $\mathcal{B}$ is at most $\exp(-rn/3) \leq \exp(-1024/3) < 0.01$ via a Chernoff bound (Fact 33 invoked with $\delta = 1$). We only query the input if $|I| \leq 2rn$ and hence the complexity is as claimed.

[2] The original claim in [12] states $2\bar{\Delta}$ instead of $\mu + 2\Delta$; for $t = \mu$, this can be strengthened to $\bar{\Delta}$ by observing that $\varphi(-t/\mu) = \varphi(-1) = 1$.

For estimating the probability of $\mathcal{G}$, let $\mathcal{G}_p$ be the event that the algorithm does not find a p-witness. Then, by the union bound, $\Pr[\mathcal{G}] \leq \sum_{p\in(\lfloor g/2\rfloor\,..\,g]} \Pr[\mathcal{G}_p]$. Hence, we can conclude the proof by showing that $\Pr[\mathcal{G}_p] \leq 1/(9g)$, which implies $\Pr[\mathcal{G}] \leq 1/9$.

Let W_p be the set of p-witnesses. Lemma 8 implies that $|W_p| \geq 1024^2 \cdot g \ln g/(pr^2)$. (One gets this bound by observing that, in both cases for k, the product of r^2 and the corresponding bound from Lemma 8 equals $1024^2 \cdot g \ln g/p$). For a fixed p-witness, the probability of sampling both of its positions is r^2. Hence, the expected number of sampled p-witnesses is at least $\mu := r^2|W_p| \geq 1024^2 \cdot g \ln g/p \geq 1024^2 \cdot \ln g$.

By the extended Janson inequality (Fact 34), the probability of not sampling any p-witness is at most $\exp(-\mu^2/(\mu + 2\Delta))$. Here, Δ is a correlation term that accounts for the fact that pairs are not sampled independently. To explain Δ, we introduce the following notation. For two potential witnesses $w_1 := (a, b) \in [0\,..\,n)^2$ and $w_2 := (c, d) \in [0\,..\,n)^2$, we write $w_1 \sim w_2$ if and only if $|\{a, b, c, d\}| = 3$, i.e., if they share exactly one position. Note that the probability of sampling both w_1 and w_2 is r^3. Also, the three distinct positions of w_1 and w_2 are in the same equivalence class. Now we can define Δ and provide an upper bound using

$$2\Delta := \sum_{\substack{w_1, w_2 \in W_p \\ w_1 \sim w_2}} r^3 \leq \sum_{\substack{w_1, w_2 \in [0\,..\,n)^2 \\ w_1 \sim w_2}} r^3 \leq p \cdot \binom{\lceil n/p \rceil}{3} r^3 \leq \frac{r^3 n^3}{p^2} \leq \frac{4r^3 n^3}{g^2}.$$

For our range of ε, we have $r \leq 1024 \cdot \sqrt{16 \cdot g^{4/3} \ln^{2/3} g/n^2}$, which readily implies $2\Delta \leq 1024^3 \cdot 64 \cdot \ln g$. Finally, if $\mu \geq 2\Delta$ then $\mu^2/(\mu+2\Delta) \geq \mu/2 \geq \frac{1}{2} \cdot 1024^2 \cdot \ln g$. Otherwise, $\mu^2/(\mu + 2\Delta) \geq \mu^2/(4\Delta) \geq 1024^4 \ln g/(1024^3 \cdot 128) \geq 8 \ln g$. Hence, $\exp(-\mu^2/(\mu + 2\Delta)) \leq \exp(-8 \ln g) = 1/g^8 \leq 1/(9g)$ for $g \geq 2$, which concludes the proof. ◀

▶ **Proposition 35.** *For every integer $g \geq 2$, we have $|[\lfloor g/2\rfloor\,..\,g] \cap \mathsf{Primes}| > g/(10 \ln g)$.*

Proof. The number of primes in $[\lfloor g/2\rfloor\,..\,g]$ is at least $1 \geq g/(10 \ln g)$ for $2 \leq g \leq 16$. For $g > 16$, it is at least $g/\ln g - (g/2)(\ln(g/2) - 3/2) \geq g/(10 \ln g)$, see Theorem 2 and Corollary 1 of [17]. ◀

▶ **Proposition 25.** *Let $M \subseteq [0\,..\,n)$, $m = |M|$, and $g \in [2\,..\,\lfloor \frac{n}{2} \rfloor]$. For $p \in [2\,..\,g]$, define the set $M_p := \{(i, j) \in M^2 \mid i < j \text{ and } i \equiv j \pmod p\}$. If we choose $p \in (\lfloor g/2\rfloor\,..\,g] \cap \mathsf{Primes}$ uniformly at random, then for all $x > 0$, $\Pr(|M_p| > x) \leq (m^2 \log_2 n \ln g)/(gx)$.*

Proof. By Proposition 35, $[2\,..\,g]$ contains at least $g/(10 \ln g)$ primes. For every $(i, j) \in M_p$, we know that p divides $d := j - i$. By the prime factorization theorem, d has at most $\log_2 d \leq \log_2 n$ distinct prime factors, and thus each pair $(i, j) \in M^2$ participates in the set M_p for at most $\log_2 n$ different choices of p. Hence, $\sum_{p\in[1\,..\,g]\cap\mathsf{Primes}} |M_p| \leq m^2 \cdot \log_2 n$. Therefore, $\mathbb{E}[|M_p|] \leq 10 \cdot (m^2 \log_2 n \ln g)/g$. The claim follows by Markov's inequality (Fact 32). ◀