Testing Periodicity*

Oded Lachish† Ilan Newman‡

Abstract

We study the string-property of being periodic and having periodicity smaller than a given bound. Let Σ be a fixed alphabet and let p, n be integers such that $p \leq \frac{n}{2}$. A length-n string over Σ, $\alpha = (\alpha_1, \ldots, \alpha_n)$, has the property $Period(p)$ if for every $i, j \in \{1, \ldots, n\}$, $\alpha_i = \alpha_j$ whenever $i \equiv j \pmod{p}$. For an integer function $g = g(n) \leq \frac{n}{2}$, the property $Period(\leq g)$ is the property of all strings that are in $Period(p)$ for some $p \leq g$. The property $Period(\leq \frac{n}{2})$ is also called *Periodicity*.

An ϵ-test for a property P of length-n strings is a randomized algorithm that for an input α distinguishes between the case that α is in P and the case where one needs to change at least an ϵ-fraction of the letters of α to get a string in P. The query complexity of the ϵ-test is the number of letter queries it makes for the worst case input string of length n. We study the query complexity of ϵ-tests for $Period(\leq g)$ as a function of g, when g varies from 1 to $\frac{n}{2}$, while ignoring the exact dependence on the proximity parameter ϵ. We show that there exists an exponential phase transition in the query complexity around $g = \log n$. That is, for every $\delta > 0$ and $g \geq (\log n)^{1+\delta}$, every two-sided, adaptive ϵ-test for $Period(\leq g)$ has a query complexity that is polynomial in g. On the other hand, for $g \leq \frac{\log n}{6}$, there exists a one-sided error, non-adaptive ϵ-test for $Period(\leq g)$, whose query complexity is poly-logarithmic in g.

We also prove that the asymptotic query complexity of one-sided error non-adaptive ϵ-tests for *Periodicity* is $\Theta(\sqrt{n \log n})$, ignoring the dependence on ϵ.

1 Introduction

Periodicity of strings plays an important role in several branches of Computer Science, engineering and social sciences. It is used as a measure of 'self similarity' in many string algorithms, computational biology, data analysis and planning (e.g., analysis of stock prices, communication patterns etc.), signal and image processing and others. Very large streams of data are now common inputs for strategy-planning or trend detection algorithms. Typically, such streams of data are either too large to store entirely in the computer memory, or so large that even linear processing time is not feasible. Thus it would be of interest to develop very fast (sublinear time) algorithms that test whether a long sequence is periodic or approximately periodic, and in particular, that test if it has a very short period. This calls for algorithms in the framework of streaming (e.g., [8]) or combinatorial property testing [6]. In property testing, introduced by Rubinfeld and Sudan [13] and formalized by Goldreich et al. [6], algorithms are randomized, they query the input at very few locations and based on this information, decide whether the input has a given property or it is 'far' from having the property. Indeed questions related to periodicity have already been investigated [3, 8, 9], although the focus here is somewhat different.

*An initial report of these results was presented at Random05.

†Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom.

‡Department of Computer Science, University of Haifa, Haifa 31905, Israel. Email: ilan@cs.haifa.ac.il, Research supported in part by an Israel Science Foundation grant number 1011/06.

In [5], the authors construct an algorithm that approximates the Discrete Fourier Transform of a finite sequence in sublinear time. This is related but not equivalent to testing how close a sequence is to being periodic. In [3], the authors study some alternative parametric definitions of periodicity that intend to 'capture the distance' of a sequence to being periodic. Their main motivation is the comparison of different measures of 'self-distance' and approximate periodicity. Most relevant to us is that the paper asserts the existence of a tolerant ϵ-test for periodicity. That is, it describes an algorithm that given, $0 \leq \epsilon_1 < \epsilon_2 \leq 1$, decides whether a sequence is ϵ_1-*close* to periodic or ϵ_2-*far* from being periodic, using $O(\sqrt{n} \cdot poly(\log n))$ queries.

Other related work are on sequences sketching in the streaming model. E.g., in [8] the authors construct an efficient sketching of a large string by a short vector, so that approximate 'trend' can be estimated (a trend is a small string that best resemble each of a given collection of substrings). Such sketches are also useful for approximating distance to periodicity, but in ℓ_2 norm rather than the hamming norm that is used below. As this requires the entire string to be read, this is less relevant for the property testing model.

Let Σ be a fixed alphabet and let p, n be integers such that $p \leq \frac{n}{2}$. The property $Period(p)$ of length-n strings over Σ contains a string $\alpha = (\alpha_1, \ldots, \alpha_n)$ if for every $i, j \in \{1, \ldots, n\}$, $\alpha_i = \alpha_j$ whenever $i \equiv j \pmod{p}$. For an integer $g \leq \frac{n}{2}$ the property $Period(\leq g)$ is the union of $Period(p)$ over all $p \leq g$. The property $Period(\leq \frac{n}{2})$ is also called *Periodicity*.

Let P be a property of length-n strings. An ϵ-test for P is a randomized algorithm that, for an input $\alpha \in \Sigma^n$, distinguishes between the case that α is in P and the case that one needs to change at least an ϵ-fraction of the positions of α to get a string in P. The query complexity of the ϵ-test is the number of letter-queries it makes for the worst case input string.

We study the query complexity of ϵ-tests for $Period(\leq g)$ when g varies from 1 to $\frac{n}{2}$. Our main focus is the dependence of the query complexity on the input length n, and the period bound g. We state the dependence of our tests on the distance parameter ϵ, but we don't try to optimize it in this respect. We note, though, that in all our tests, the dependence on $1/\epsilon$ is polynomial .

1.1 Our Results

We show that there exists an exponential phase transition in the query complexity around $g = \log n$. That is, for every $\delta > 0$ and for $g \geq (\log n)^{1+\delta}$, every two-sided, adaptive ϵ-test for $Period(\leq g)$ has a query complexity that is polynomial in g. On the other hand, for $g \leq \frac{\log n}{6}$ there exists a one-sided error, non-adaptive, ϵ-test for $Period(\leq g)$, whose query complexity is poly-logarithmic in g. We also prove that the asymptotic query complexity of one-sided error, non-adaptive, ϵ-tests for *Periodicity* is $\Theta(\sqrt{n \log n})$, ignoring the dependence on ϵ.

The results are summarized in the following theorems.

Theorem 1.1. *For every large enough n and $\epsilon \geq (\frac{256}{9} \cdot \frac{\log g}{g})^{\frac{1}{3}}$, there exists a one-sided error, non-adaptive ϵ-test for $Period(\leq g)$, whose query complexity is $O\left(\sqrt{\frac{g \log g}{\epsilon}}\right)$.*

The following theorem implies that Theorem 1.1 is the best possible up to a logarithmic factor, for large enough g.

Theorem 1.2. *For every large enough n and $g \leq \frac{n}{2}$, any adaptive, two-sided error $\frac{1}{32}$-test for $Period(\leq g)$ has query complexity $\Omega\left(\sqrt{\frac{g}{(\log g) \cdot \log n}}\right)$.*

The lower bound of Theorem 1.2 becomes irrelevant once g is approximately $\log n$ or less. The following theorem states that if $g \leq \frac{\log n}{6}$, then there is a much more efficient test for $Period(\leq g)$, compared to that implied in Theorem 1.1.

Theorem 1.3. *For every large enough n, $g \leq \frac{\log n}{6}$ and $\epsilon > \frac{7e\log^3 g}{\sqrt{n}}$ there exists a one-sided error, non-adaptive, ϵ-test for $Period(\leq g)$, that has query complexity $O(\frac{(\log g)^6 \log\log g}{\epsilon})$.*

The following theorem implies that $poly(\log g)$ is the best query complexity we can get for $Period(\leq g)$, even if g is significantly smaller than $\log n$.

Theorem 1.4. *For every large enough n and $g \leq \frac{\log n}{4}$, any two-sided error $\frac{1}{32}$-test for $Period(\leq g)$ has query complexity $\Omega(\sqrt{\frac{\log g}{\log\log g}})$.*

Finally, the following theorem, together with Theorem 1.1 for $g \geq \frac{n}{6}$, imply that there is an exact asymptotic bound of $\Theta(\sqrt{n\log n})$ for the one-sided error, non adaptive, query complexity of periodicity.

Theorem 1.5. *For every large enough n any non-adaptive, one-sided error, $\frac{1}{16}$-test for periodicity has query complexity $\Omega(\sqrt{n\log n})$.*

Uniformity and time complexity of the tests: The test suggested by Theorem 1.1, is uniform, it can get n and ϵ as inputs and decide on the position to query in $poly(n, 1/\epsilon)$ time. The test suggested by Theorem 1.3 is based on a certain subset of $[n]$ whose existence is proved using a probabilistic argument (shown in Claim 3.10). Thus, taken simply as suggested, the algorithm can be viewed either as non-uniform, namely, asserts the existence of a test while not constructing it deterministically. Alternatively, it can be viewed as a randomized uniform test that first chooses the needed object randomly (which its existence is asserted with high probability), verifies that it is as needed (so to remain 1-sided error), which can be done in $poly(n, 1/\epsilon)$ time.

The rest of the paper is organized as follows. In Section 2 we introduce the required preliminaries, and in Section 3 we prove the upper Bound stated in Theorem 1.3. Section 4 contains the proof of Theorem 1.1. In Section 5 we prove Theorems 1.2, 1.4 and 1.5.

2 Preliminaries

For two positive integers $i \leq j$ we denote $[i, j] = \{i, \ldots, j\}$, and $[n] = [1, n]$. To simplify notations, we denote $[\alpha] = [\lceil\alpha\rceil]$ (namely, we use the $[\ ]$ notation for non integers too). In the following Σ is a fixed size alphabet that contains $0, 1$. A length-n string α is a sequence of n letters from Σ, $\alpha = (\alpha_1 \ldots \alpha_n)$. Σ^n denotes the set of all length-n strings over Σ. For $i \in [n]$ we refer to α_i as the ith letter of α. Given a subset $S \subseteq [n]$ such that $S = \{i_1, i_2, \ldots, i_m\}$ and $i_1 < i_2 < \ldots < i_m$ we denote the m-length string obtained by restricting α to the positions in S as $\alpha_S = (\alpha_{i_1}\alpha_{i_2} \ldots \alpha_{i_m})$. Unless otherwise stated, a string has length n.

For two length-n strings α, β, we denote $dist_n(\alpha, \beta) = |\{i|\ \alpha_i \neq \beta_i\}|$. We omit the subscript in $dist_n$ when the length of the strings is clear from the context. For a property $\mathcal{P} \subseteq \Sigma^n$ and a string $\alpha \in \Sigma^n$, $dist(\alpha,\ \mathcal{P}) = min\{dist(\alpha, \beta)|\ \beta \in \mathcal{P}\}$ denotes the distance from α to $\mathcal{P}$. We say that α is *ϵ-far* from $\mathcal{P}$ if $dist(\alpha, \mathcal{P}) \geq \epsilon n$, otherwise we say that α is *ϵ-close* to $\mathcal{P}$. For $\sigma \in \Sigma$ we denote by $\sigma^n = (\sigma, \sigma, \ldots, \sigma)$ the length-n string in which the letter in every position is σ.

We denote by $\ln n$ the natural logarithm of n and by $\log n$ the base-2 logarithm of n. For a number g let $Primes(g)$ be the set of all primes that are smaller than or equal to g. Let $\Pi(g) = |Primes(g)|$. The *Prime Number Theorem* [7, 12, 11] states that $\lim_{g\to\infty} \frac{\Pi(g)}{g/\ln g} = 1$.

Definition 1. *A string $\alpha \in \Sigma^n$ is called homogeneous if $\alpha_i = \alpha_j$ for every $1 \leq i < j \leq n$. Namely, $\alpha = \sigma^n$ for some $\sigma \in \Sigma$.*

The string property of being homogeneous is denoted by *Homogeneous*.

Property Testing

An ϵ-test is a randomized algorithm that accesses the input string via a 'location-oracle' which it can query: On an length-n string α, a query is done by specifying an index $i \in [n]$ to which the answer is α_i. The complexity of the algorithm is the number of queries it makes for the worst case input. Such an algorithm is said to be a "one-sided-error ϵ-test" for a property $\mathcal{P} \subseteq \Sigma^n$ if it satisfies the following:

- It accepts every string in $\mathcal{P}$ with probability 1.
- It rejects every string that is ϵ-*far* from $\mathcal{P}$ with probability at least $\frac{2}{3}$.

If the acceptance probability of strings in $\mathcal{P}$ is only guaranteed to be at least $\frac{2}{3}$ instead of 1, then the test is called a *two-sided error* ϵ-test. If an ϵ-test determines all queries it makes prior to asking them, then it is said to be a *non adaptive* test, otherwise it is an adaptive test. For further material on property testing see [6, 14, 4].

Periodicity

Definition 2. [Period(p)]
A string $\alpha \in \Sigma^n$ is in $Period(p)$, if $\alpha_i = \alpha_j$ for every $i, j \in [n]$ such that $i \equiv j \pmod p$.

Note that a string is homogeneous if and only if it is in $Period(1)$.

Definition 3. [Period($\leq$ g)]
A string α is in property $Period(\leq g)$ if there exists $p \leq g$ such that $\alpha \in Period(p)$. The property of length-n strings $Period(\leq \frac{n}{2})$ is denoted by Periodicity.

Definition 4. *A p-witness for a string $\alpha \in \Sigma^n$ is a pair $(i < j)$, $i, j \in [n]$ such that $i \equiv j \pmod p$ and $\alpha_i \neq \alpha_j$.*

The following fact is a direct result of the definition of $Period(\leq g)$.

Fact 2.1. *A string α has a period p if and only if there is no p-witness for α.*

Fact 2.2. *Let α be a string in Σ^n and $r = \lfloor \frac{g}{2} \rfloor + 1$. If $\alpha \notin Period(p)$ for every $p \in [r, g]$, then $\alpha \notin Period(\leq g)$. If α is ϵ-far from $Period(p)$ for every $p \in [r, g]$, then it is ϵ-far from $Period(\leq g)$.*

Proof. Observe that if a string $\alpha \in \Sigma^n$ has period $p \leq \frac{g}{2}$ then it also has period q for every q that is a multiple of p. Note also that there always exist such $q \in [r, g]$. □

Definition 5. *For $p < n$ and $0 \leq i \leq p-1$, let $Z(p, i) = \{j \in [n] | \ j \equiv i \pmod p\}$. $Z(p, i)$ is also called the ith p-section of n. For an n-length string α, the string $\alpha_{Z(p,i)}$ is called the ith p-section of α.*

The following fact relates $dist(\alpha, Period(p))$ and the distance of its p-sections from $Homogeneous$.

Fact 2.3. *For each $\alpha \in \Sigma^n$, $dist(\alpha, Period(p)) = \Sigma_{i=0}^{p-1} dist(\alpha_{Z(p,i)}, Homogeneous)$.* □

Using the above we can relate $dist(\alpha, Period(p))$ to the number of *p-witnesses* for α.

Claim 2.4. *Let $\alpha \in \Sigma^n$ and $p \leq \frac{n}{2}$, then there are at least $\frac{n-p}{2p} \cdot dist(\alpha, Period(p))$ distinct p-witnesses for α.*

Proof. By the definition of a *p-witness*, every *p-witness* for α is a subset of some p-section $Z(p,i)$. Let $\mathcal{W}$ be the set of all *p-witnesses* for α, and let $\mathcal{W}_i$ be the set of *p-witnesses* for α that are subsets of $Z(p,i)$. As the p-sections are pairwise disjoint, we conclude that so are the $\mathcal{W}_i$'s. Thus, $|\mathcal{W}| = \sum_{i=0}^{p-1} |\mathcal{W}_i|$.

Fix i and for every $\sigma \in \Sigma$, let m_σ be the number of occurrences of σ in $\alpha_{Z(p,i)}$. Let $m' = \max_{\sigma\in\Sigma} m_\sigma$. By the definition of a *p-witness* we get,

$$|\mathcal{W}_i| = \frac{1}{2}\cdot\sum_{\sigma\in\Sigma} m_\sigma\cdot(|Z(p,i)| - m_\sigma) \geq \frac{1}{2}\cdot\sum_{\sigma\in\Sigma} m_\sigma\cdot(|Z(p,i)| - m') = \frac{|Z(p,i)| - m'}{2}\cdot|Z(p,i)| \quad (1)$$

Set $d_i = dist(\alpha_{Z(p,i)}, homogeneous)$. Observe that the homogeneous string with minimum distance to $\alpha_{Z(p,i)}$ is $\sigma^{|Z(p,i)|}$, where σ is the most frequent letter, that is, for which $m' = m_\sigma$. Thus $d_i = |Z(p,i)| - m'$ and hence $|\mathcal{W}_i| \geq \frac{1}{2}\cdot d_i \cdot |Z(p,i)|$ for every $i \in \{0,1,\ldots,p-1\}$. Hence,

$$d_i \leq \frac{2\cdot|\mathcal{W}_i|}{|Z(p,i)|} \leq \frac{2\cdot p\cdot|\mathcal{W}_i|}{n-p}, \quad (2)$$

where the last inequality is due to the fact that $|Z(p,i)| \geq \frac{n-p}{p}$. Fact 2.3 and Equation (2) imply that

$$dist(\alpha, Period(p)) = \Sigma_{i=0}^{p-1} d_i \leq \frac{2\cdot p}{n-p}\cdot\Sigma_{i=0}^{p-1}|\mathcal{W}_i| \leq \frac{2\cdot p\cdot|\mathcal{W}|}{n-p}.$$

It follows that $|\mathcal{W}| \geq \frac{n-p}{2\cdot p}\cdot dist(\alpha, Period(p))$. □

The following is an immediate corollary of Claim 2.4.

Corollary 2.5. *Let $p \leq \frac{n}{2}$ and α be ϵ-far from $Period(p)$. Then, a random pair $(i<j)$ for which $i \equiv j \pmod p$ is a p-witness with probability at least $\frac{\epsilon}{3}$.*

Proof. There are at most $p\cdot\binom{\lceil\frac{n}{p}\rceil}{2} \leq \frac{n(n+p)}{2p} = s$ pairs of the form $i \equiv j \pmod p$. The probability that a random pair is a p-witness is at least the bound on the number of p-witnesses asserted in Claim 2.4 divided by s. Hence, it is at least $\epsilon\cdot\frac{n-p}{n+p} \geq \frac{\epsilon}{3}$, where the last inequality is by the assumption that $p \leq \frac{n}{2}$. □

3 An ϵ-test for $Period(\leq g)$ for very small g

In this Section we prove Theorem 1.3. We first note that there is a one-sided error ϵ-test for $Period(\leq g)$, of complexity $O(g\log g/\epsilon)$, for any g. This is by testing for every $p \leq g$, that the input is ϵ-far from $Period(p)$, amplified so that the error probability is less than $\frac{1}{3g}$. According to Corollary 2.5, for every fixed p, this can be done using $O(\log g/\epsilon)$ queries, implying a total as claimed. By the union bound, the error probability of the whole test is bounded by $g\cdot\frac{1}{3g} \leq 1/3$. In particular, this implies that for every constant g there is a one sided error ϵ-test for $Period(\leq g)$ making $O(1/\epsilon)$ queries. Hence, in the following, we assume that n and g are large enough so that the function $\Pi(g)$ (see Section 2 for the definition) satisfies: $\Pi(g) < \frac{1.5g}{\log g}$ (while $g \leq \frac{\log n}{6}$ as required by the theorem).

We first analyze the following naive idea that captures some of the intuition of the proof, while not being part of it. Let $\alpha \in \Sigma^n$. By definition, if t is divisible by p, then a *t-witness* for α is also a *p-witness* for α. Let $t = g!$, and for the sake of simplicity assume that $g\ln g \leq \ln(\frac{n}{2})$, so that $t \leq \frac{n}{2}$. Hence, a *t-witness* for α is a witness that $\alpha \notin Period(\leq g)$. One might be tempted to think that an ϵ-test for $Period(t)$ is also an ϵ-test for $Period(\leq g)$. However, this is not the case since α may be in $Period(t)$ and far from $Period(\leq g)$. While this naive idea does not work the reason it does not work suggests the following notion that we call an (ϵ,λ)-cover of g.

Definition 6. **[(ϵ, λ)-*cover*]**
For $\epsilon, \lambda \leq 1$, a set $I \subseteq [n]$ is called an (ϵ, λ)-cover of g if it satisfies the following. For every $\alpha \in \Sigma^n$ and $p \leq g$ such that α is ϵ-far from $Period(p)$, there exists $t \in I$ such that p divides t and α is $\epsilon\lambda$-far from $Period(t)$.

In view of the observation above it should be clear why small (ϵ, λ)-covers are of interest: If α is ϵ-far from $Period(\leq g)$, α is ϵ-far from $Period(p)$ for every $p \leq g$. By the definition of (ϵ, λ)-cover, for every $p \leq g$ there is some t in the cover for which α is $\epsilon\lambda$-far from $Period(t)$. Thus, in order to test if α has $period(\leq g)$ it is enough to test α for $Period(t)$ for every t in the cover. If the size of the cover is much smaller than g, and $\epsilon\lambda$ is not too small then this would imply an efficient test.

Note that $[g]$ is always an $(\epsilon, 1)$-cover for g. Our goal is to find a smaller cover when g is small, while keeping λ large enough.

We first describe the generic algorithm for testing $Period(\leq g)$ using an (ϵ, λ)-cover of g.

Algorithm 1.
Input: $\epsilon > 0, \alpha \in \Sigma^n$. **Assumption:** *A set E that is an (ϵ, λ)-cover of g is given.*

1. *For each $t \in E$ select uniformly at random and with repetitions $q_t = \frac{6 \cdot (\ln |E| + 1)}{\epsilon \cdot \lambda}$ members from the set $W_t = \{(x < y) |\ x, y \in [n], x \equiv y \pmod{t}\}$. Let Q be the union of all the sets selected.*
2. *Reject if for each $p \in [g]$ the set Q contains a p-witness and otherwise accept.*

Claim 3.1. Algorithm 1 *is a one-sided error, non-adaptive ϵ-test for $Period(\leq g)$. Its query complexity is $O(\frac{|E| \cdot \log |E|}{\epsilon \cdot \lambda})$.*

Proof. The algorithm is clearly non-adaptive and its query complexity is as stated. The algorithm rejects only if for every $p \leq g$ it finds a p-witness. Hence the algorithm has a one-sided error.

Let $\alpha \in \Sigma^n$ be *ϵ-far* from $Period(\leq g)$. By the definition of a (ϵ, λ)-cover, for every $p \leq g$ there exists $t \in E$ such that p divides t and α is *$\epsilon\lambda$-far* from $Period(t)$. Let $E' \subseteq E$ be the set of all such t's. By Corollary 2.5, for any fixed $t \in E'$ the probability that a random member of W_t is not a t-witness is less than $1 - \frac{\epsilon\lambda}{3}$. Thus the probability that Q does not contain a t-witness is at most $(1 - \frac{\epsilon\lambda}{3})^{q_t} \leq \frac{1}{3|E|^2}$. By the union bound, the probability that there exists $t \in E$ for which Q contains no t-witnesses is at most $\frac{|E'|}{3|E|^2} \leq \frac{1}{3}$. As E is an (ϵ, λ)-cover of g, this is also the probability that there exists a $p \leq g$ for which Q contains no p-witnesses. □

In order to prove Theorem 1.3 it is enough to show how to construct a small and good enough (ϵ, λ)-cover. Lemma 3.2 states that such a cover exists for $\epsilon > \frac{7e \log^3 n}{\sqrt{n}}$ and $g \leq \frac{\log n}{6}$. This, in turn, ends the proof of Theorem 1.3. □

Lemma 3.2. *Let $g \leq \frac{\log n}{6}$ and $\epsilon > \frac{7e \log^3 g}{\sqrt{n}}$, then there exists a set E of size $O((\log g)^3)$ that is an $(\epsilon, \frac{1}{3 \cdot |E|})$-cover of g.*

To prove the lemma we need some additional machinery. For a set $I \subseteq [n]$, $\gcd(I)$ denotes the greatest common divisor of the numbers in I. We make use of the following notion.

Definition 7. **[*GCD-cover*]**
A GCD-cover of g is a set $E \subseteq [n]$ such that: for every $p \leq g$, there exists a subset $I \subseteq E$, for which $p = \gcd(I)$.

The next Lemma 3.3 and Corollary 3.4 imply that a suitable GCD-cover, E, of g is $(\epsilon, \frac{1}{3\cdot|E|})$-*cover* of g as needed in Lemma 3.2. Lemma 3.5, shows that such a suitable GCD-cover can be constructed. This will end the proof of Lemma 3.2. We next present these Lemmas and the formal proof of Lemma 3.2. The proofs of Lemmas 3.3 and 3.5 appear later in the next two sections.

Lemma 3.3. *Let $I \subseteq [n^{\frac{1}{4}}]$, $|I| \leq \frac{\epsilon\sqrt{n}}{3}$ be such that $p = \gcd(I)$, and let $\alpha \in \Sigma^n$ be ϵ-far from $Period(p)$. Then there exists $t \in I$ such that α is $\frac{\epsilon}{3\cdot|I|}$-far from $Period(t)$.*

Lemma 3.3 directly implies the following corollary.

Corollary 3.4. *Let $E \subseteq [n^{\frac{1}{4}}]$ be a GCD-cover of g with $|E| \leq \frac{\epsilon\sqrt{n}}{3}$. Then E is an $(\epsilon, \frac{1}{3\cdot|E|})$-cover of g.* □

Lemma 3.5. *Let $g \leq \frac{\log n}{6}$, then there exists a subset $E \subseteq [n^{\frac{1}{4}}]$ of size $1 + 2e\log^3 g$ that is a GCD-cover of g.*

Proof of Lemma 3.2: Lemma 3.5 asserts the existence of a set E of size $1 + 2e\log^3 g$ that is a GCD-cover of g. For $\epsilon > \frac{7e\log^3 g}{\sqrt{n}}$, $|E| \leq \frac{\epsilon\sqrt{n}}{3}$, thus, Corollary 3.4 may be applied with E to imply the lemma. □

3.1 Proof of Lemma 3.3

We first need the following claims.

Claim 3.6. *Let $s = p \cdot q$, then $dist(\alpha, Period(s)) = \Sigma_{i=0}^{p-1} dist(\alpha_{Z(p,i)}, Period(q))$.*

Proof. Fix $i,\ 0 \leq i \leq p-1$. Fact 2.3 asserts that

$$dist(\alpha, Period(s)) = \Sigma_{j=0}^{s-1} dist(\alpha_{Z(s,j)}, Homogeneous) =$$

$$\sum_{i=0}^{p-1} \sum_{j \leq s-1,\ j \equiv i \bmod p} dist(\alpha_{Z(s,j)}, Homogeneous) \tag{3}$$

Let $\beta^{(i)} = \alpha_{Z(p,i)}$ be the ith p-section of α. Observe that the q-sections of $\beta^{(i)}$ are exactly $\alpha_{Z(s,j)}$, $j \leq s-1, j \equiv i \pmod p$. Hence by Fact 2.3, for every fixed i, $0 \leq i \leq p-1$, $\sum_{j \leq s-1,\ j \equiv i \pmod p} dist(\alpha_{Z(s,j)}, Homogeneous) = dist(\alpha_{Z(p,i)}, Period(q))$. Plugging this into Equation (3) implies the claim. □

Claim 3.7. *Let $\alpha \in \Sigma^n$, $gcd(r,s) = 1$ and $r \cdot s$ divides n. If $\alpha \in Period(r)$, then $dist(\alpha, Homogeneous) = dist(\alpha, Period(s))$.*

Proof. As $r \cdot s$ divides n and $gcd(r,s) = 1$, $|Z(r,i) \cap Z(s,j)| = \frac{n}{r\cdot s}$ for every $i \in \{0,\ldots,r-1\}$ and $j \in \{0,\ldots,s-1\}$. Let $\alpha \in Period(r)$, Fact 2.3 implies that $\alpha_{Z(r,i)}$ is Homogeneous for every $i \in \{0,\ldots,r-1\}$. Let $\sigma^{(i)}$ be the unique symbol that appears in $\alpha_{Z(r,i)}$. Thus for every fixed j, and every $i = 0,\ldots r-1$, $\sigma^{(i)}$ appears $\frac{n}{r\cdot s}$ times in $\alpha_{Z(s,j)}$. We conclude that for every $\sigma \in \Sigma$, the number of times σ appears in $\alpha_{Z(s,j)}$ is fixed and independent of j. In particular, let σ^* be the letter that appears the largest number of times in $\alpha_{Z(s,0)}$, and let this number be m^*. Then σ^* appears m^* times in $\alpha_{Z(s,j)}$ for every j. Thus $dist(\alpha_{Z(s,j)}, Homogeneous) = \frac{n}{s} - m^*$ for every $j = 0,1,\ldots,s-1$, and $\beta = (\sigma^*)^{\frac{n}{s}}$ is the closest homogeneous string to each of $\alpha_{Z(s,j)}$.

This implies that $dist(\alpha, Homogeneous) = s \cdot (\frac{n}{s} - m^*) = \sum_{j=0}^{s} dist(\alpha_{Z(s,j)}, Homogeneous) = dist(\alpha, Period(s))$ where the last equality is by Fact 2.3. □

The following is a corollary of Claim 3.7, and could be viewed as a 'clean' version of Lemma 3.3. It implies that if $dist(\alpha, Period(p))$ is large, and $p = gcd(r, s)$ then if $dist(\alpha, Period(r)) = 0$ then $dist(\alpha, Period(s))$ is large.

Claim 3.8. *Let $\alpha \in \Sigma^n$, $p = gcd(r, s)$ and $r \cdot s$ divides n. If $\alpha \in Period(r)$, then*

$$dist(\alpha, Period(p)) = dist(\alpha, Period(s)).$$

Proof. We first claim that for every $i \in \{0, \ldots, p-1\}$,

$$dist(\alpha_{Z(p,i)}, Homogeneous) = dist(\alpha_{Z(p,i)}, Period(\frac{s}{p})). \tag{4}$$

Indeed, fix $i \in \{0, \ldots, p-1\}$. Since $\alpha \in Period(r)$, then $\alpha_{Z(p,i)} \in Period(\frac{r}{p})$. As $gcd(r, s) = p$, $gcd(\frac{r}{p}, \frac{s}{p}) = 1$ and the fact that $r \cdot s$ divides n implies that $\frac{r}{p} \cdot \frac{s}{p}$ divides $\frac{n}{p} = |\alpha_{Z(p,i)}|$. Thus, Claim 3.7 implies Equation (4).

Claim 3.6 in addition to Equation (4) and Fact 2.3 imply the claim. □

We can now present the proof of Lemma 3.3.

Proof. Let $I \subseteq [n^{\frac{1}{4}}], |I| \leq \frac{\epsilon\sqrt{n}}{3}$ be such that $p = \gcd(I)$. Let $\alpha \in \Sigma^n$ be *ϵ-far* from $Period(p)$. We show by induction on $|I|$ that there exists $t \in I$ such that α is $\frac{\epsilon}{3 \cdot |I|}$*-far* from $Period(t)$. This is trivial for $|I| = 1$. We assume therefore, that $|I| \geq 2$ and that the lemma holds for any I' of size smaller than $|I|$.

Let $r \in I$. If α is $\frac{\epsilon}{3 \cdot |I|}$*-far* from $Period(r)$ then we are done. Assume then that α is $\frac{\epsilon}{3 \cdot |I|}$-close to $Period(r)$. Set $s = gcd(I \setminus \{r\})$. By definition, we have $p = gcd(\{r, s\})$. Let m be the largest multiple of $r \cdot s$ that is smaller than n. The assumption that $I \subseteq [n^{\frac{1}{4}}]$ implies that $n - m \leq \sqrt{n}$. Let $\beta \in Period(r) \cap \Sigma^m$ be of minimum distance to $\alpha_{[m]}$. Using the triangle inequality we get

$$dist_m(\beta, Period(s)) \leq dist_n(\alpha, Period(s)) + dist_n(\alpha, Period(r)). \tag{5}$$

Claim 3.8 asserts that

$$dist(\beta, Period(p)) = dist(\beta, Period(s)).$$

Plugging this into Equation (5) implies

$$dist(\alpha, Period(s)) \geq dist(\beta, Period(p)) - dist(\alpha, Period(r)). \tag{6}$$

Again by the triangle inequality

$$dist_n(\alpha, Period(p)) \leq dist_m(\alpha_{[m]}, Period(p)) + n - m \leq \tag{7}$$

$dist_m(\beta, Period(p)) + (n-m) + dist_n(\alpha, Period(r)) \leq dist_m(\beta, Period(p)) + dist_n(\alpha, Period(r)) + \sqrt{n}$.

Plugging equation (7) into Equation (6) results in the following

$$dist(\alpha, Period(s)) \geq dist(\alpha, Period(p)) - 2 \cdot dist(\alpha, Period(r)) - \sqrt{n}. \tag{8}$$

Recall that $dist(\alpha, Period(p)) \geq \epsilon n$, $dist(\alpha, Period(r)) \leq \frac{\epsilon n}{3|I|}$ and that $|I| \leq \frac{\epsilon\sqrt{n}}{3}$. Plugging this into Equation (8) implies that

$$dist(\alpha, Period(s)) \geq \frac{\epsilon n \cdot (|I| - 1)}{|I|}$$

Finally, by the induction hypothesis on $I \setminus \{r\}$, there exists $t \in I \setminus \{r\}$ such that α is $\frac{\epsilon \cdot (|I|-1)}{|I|} \cdot \frac{1}{3 \cdot (|I|-1)}$*-far* from $Period(t)$. That is, α is $\frac{\epsilon}{3 \cdot |I|}$*-far* from $Period(t)$ as required. □

3.2 Proof of Lemma 3.5

Definition 8. [Prime-cover of g]
We say that a collection of sets of primes, $\mathcal{R}$ is a Prime-cover *of g if for every subset of primes $S' \subseteq Primes(g)$ with $\prod_{q\in S'} q \leq g$ the following is satisfied: For every $p \in Primes(g)$ there is an $R \in \mathcal{R}$ such that $p \in R$ and for every $q \in S' \setminus \{p\}$, $q \notin R$.*

Claim 3.9. *Let $\mathcal{R}$ be a Prime-cover of g. Then, there exists a set E that is a GCD-cover of g that satisfies the following.*

- *$z \leq 2^{1.5g}$ for every $z \in E$.*
- *$|E| \leq 1 + |\mathcal{R}| \cdot \log g$.*

Proof. For every prime $p \in Primes(g)$ let $\kappa(p)$ be the maximum integer such that $p^{\kappa(p)} \leq g$. Let $\mathcal{R}$ be a *Prime-cover* of g. For each $R \in \mathcal{R}$ we define the following set of at most $\log g + 1$ numbers.

$$y_R(i) = \prod_{r\in R} r^{min\{i,\kappa(r)\}} \cdot \prod_{q\in Primes(g)\setminus R} q^{\kappa(q)}, \quad i = 0, 1, \ldots, \lfloor \log g \rfloor.$$

We set $E = \{y_R(i) |\ i = 0, 1, \ldots \lfloor \log g \rfloor,\ R \in \mathcal{R}\}$. To see that indeed E is a *GCD-cover* let $t \in [g]$ and let $t = \prod_{p\in Primes(g)} p^{a(p)}$ be the prime power decomposition of t. Let $P(t)$ contain all the prime divisors of t, namely these primes, $p \in Primes(g)$ for which $a(p) \geq 1$. Note that since $t \leq g$, $\prod_{p\in P(t)} p \leq t \leq g$ and also $a(p) \leq \log g$ for every $p \in Primes(g)$.

By assumption $\mathcal{R}$ is a *Prime-cover* of g and thus checking the definition with respect to $S' = P(t)$, we conclude that there are subsets $R_p \in \mathcal{R}$ for each $p \in Primes(g)$ such that $p \in R_p$ while for every $q \in P(t) \setminus \{p\},\ q \notin R_p$. It is obvious that $t = gcd(\{y_{R_p}(a(p))|\ p \in Primes(g)\})$. This is so as t clearly divides each $y_{R_p}(a(p))$. On the other hand, for each $p \in Primes(g)$, $p^{a(p)+1}$ does not divide $y_{R_p}(a(p))$. Thus for any prime $p \in Primes(g)$, the largest power of p that divides all the numbers above is $p^{a(p)}$.

It remains to be shown that the bounds of the numbers in E and the size of E are as claimed. Indeed, by definition $|E| \leq 1+|\mathcal{R}|\cdot\log g$ (note that i takes possibly $(1+\log g)$ values but for each $p > 2,\ \kappa(g) < \log g$ and thus $|E|$ is indeed as claimed). Note also that for every $z \in E$, $z \leq g^{|Primes(g)|}$, as each number in every product expression is at most g. By our assumptions $|Primes(g)| = \Pi(g) \leq \frac{1.5g}{\log g}$, thus $z \leq g^{\frac{1.5g}{\log g}} = 2^{1.5g}$ for every $z \in E$. □

The following claim asserts that there is a small enough *Prime-cover* of g.

Claim 3.10. *There exists a Prime-cover of g whose size is at most $2e(\log g)^2$.*

Proof. We show a probabilistic construction of a family of $2e(\log g)^2$ sets, and prove that with strictly positive probability it is a *Prime-cover* of g. This implies the existence of such a set.

Let $\mathcal{R}$ be a set of $2e(\log g)^2$ random subsets of primes, each selected independently at random, as follows. A set R is formed by selecting each prime $p \leq g$, to be in R with probability $\frac{1}{\log g}$.

Fix $S' \subseteq Primes(g)$ with $\prod_{q\in S'} q \leq g$ and a prime $p \leq g$. Note that $|S'| \leq \log g$. A set R is *good* for (S', p) if $p \in R$ while $q \notin R$ for every $q \in S' \setminus \{p\}$. The probability that a set R drawn as above is *good* for (S', p) is at least $\frac{1}{\log g}(1 - \frac{1}{\log g})^{|S'|} \geq \frac{1}{\log g}(1 - \frac{1}{\log g})^{\log g} \geq \frac{0.9}{e \log g}$ (for large enough g). As the sets are drawn independently, the probability that none of the $2e \log^2 g$ sets is good for (S', p) is at most $(1 - \frac{0.9}{e \log g})^{3e(\log g)^2} \leq \frac{1}{g^3}$. However, there are at most g^2 pairs (S', p), as above, since for any such S', $\prod_{q\in S'} q \in [g]$ and distinct S' results in distinct products. Hence, by the union bound, with probability at least $1 - \frac{1}{g}$ for every such (S', p) there is a set in $\mathcal{R}$ that is *good* for (S', p). □

We can now conclude the proof of Lemma 3.5.

Proof. Claim 3.10 asserts that there is a *Prime-cover* of g of size $2e(\log g)^2$. In turn, Claim 3.9 asserts that there is a *GCD-cover* of g of size $1 + 2e(\log g)^3$ and in which every member is at most $2^{1.5g}$. For $g \leq \frac{\log n}{6}$ the bound on each member in the cover is $2^{1.5g} \leq n^{\frac{1}{4}}$. □

A final note on how to efficiently construct a small prime cover: The proof of Claim 3.10 uses a randomized argument to show the existence of a small prime cover. We note that verifying that a collection of $O(\log^2 g)$ sets is a prime-cover of g (for $g \leq \frac{\log n}{6}$) can be done in $poly(n)$ time. This is by checking the conditions for all $O(g^2)$ pairs (S', p) as in the proof of the claim. Thus Claim 3.10 suggests a 1-sided error randomized algorithm for constructing efficiently a prime cover. Making this 1-sided error algorithm part of the ϵ-test does not change its query complexity but makes the whole test a uniform, 1-sided error test, that runs in polynomial time.

4 An ϵ-test for $Period(\leq g)$ for large g

Algorithm 1, described in Section 3, is an efficient test only when $g \leq \frac{\log n}{6}$, implying that there is a polylog size *GCD-cover*. Here we prove Theorem 1.1 by constructing a test for larger g's. We therefore assume in the following that $g > \frac{\log n}{6}$ (although the correctness of the algorithm will not depend on this).

Conceptually, the test here will be much simpler. As our goal is to find a p-witness for every $p \in [g]$ we select a random subset $Q \subseteq [n]$ by picking each member of $[n]$ independently with some predetermined probability. If the set contains a p-witness for every $p \leq g$ the test rejects, otherwise it accepts.

We next formally describe the algorithm and prove its correctness. Let $\epsilon \geq (\frac{256}{9} \cdot \frac{\log g}{g})^{\frac{1}{3}}$.

Algorithm 2.
Input: $\epsilon > 0, \alpha \in \Sigma^n$.
Let $\nu = 4\sqrt{\frac{g \cdot \log g}{\epsilon}}$.

1. *Select a random subset $Q \subseteq [n]$ by selecting each member of $[n]$ independently with probability $\frac{\nu}{n}$.*
2. *If $|Q| > 2 \cdot \nu$ then accept and terminate.*
3. *Query α on each index in Q.*
4. *Reject if for every $\frac{g}{2} \leq p \leq g$ the set Q contains a pair that is a p-witness. Otherwise accept.*

Obviously the algorithm is non-adaptive, it has a one sided error and query complexity $2 \cdot \nu = O(\sqrt{\frac{g \cdot \log g}{\epsilon}})$. The following claim asserts that the error probability is at most $\frac{1}{3}$ and thus completes the proof of Theorem 1.1.

Claim 4.1. *Let $\frac{\log n}{6} < g \leq \frac{n}{2}$. If α is ϵ-far from $Period(\leq g)$ then Algorithm 2 rejects with probability at least $\frac{2}{3}$.*

Proof. Let $\alpha \in \Sigma^n$ be *ϵ-far* from $Period(\leq g)$ and let Q be the set selected by the algorithm at Step 1.

Let $\mathcal{G}$ be the event that Q contains a p-witness for every $\frac{g}{2} \leq p \leq g$. Let $\mathcal{B}$ be the event that $|Q| > 2 \cdot \nu$. By definition Algorithm 2 rejects α if and only if $\mathcal{G}$ occurs and $\mathcal{B}$ does not. Thus it is enough to show that $\Pr(\mathcal{B}) \leq \frac{1}{9}$ and $\Pr(\mathcal{G}) \geq \frac{8}{9}$.

Let X_i be the indicator function for the event $i \in Q$ and set $X = \sum_{i=1}^{n} X_i$. By definition, $\Pr(\mathcal{B}) = \Pr(X > 2 \cdot \nu)$. Since the events X_i are independent, Chernoff inequality [2] implies that $\Pr(X > 2 \cdot \nu) \leq (\frac{e}{4})^{\nu} \leq \frac{1}{9}$.

Let $\overline{\mathcal{G}}_p$ be the event that the set Q does not contain a pair of two elements that form a p-witness for α. By the union bound we get that $\Pr(\mathcal{G}) \geq 1 - \sum_{p \in [\frac{g}{2}, g]} \Pr(\overline{\mathcal{G}}_p)$. Thus, to conclude the statement of the claim it is sufficient to show that $\Pr(\overline{\mathcal{G}}_p) \leq \frac{1}{g^2}$ for every p such that $\frac{g}{2} \leq p \leq g$.

Fix p such that $\frac{g}{2} \leq p \leq g$. Let $\mathcal{T} = \{T_i = (x_i < y_i) |\ i = 1, \ldots, t\}$ be the set of all the p-witnesses for α. Claim 2.4 asserts that $|\mathcal{T}| = t \geq \frac{\epsilon n(n-p)}{2p}$. For each $i \in [t]$ let A_i be the indicator function for the event $x_i, y_i \in Q$ and let $\overline{A}_i$ be its complement. Let μ be the expectation of $\sum_{i=1}^{t} A_i$. For distinct i, j let $T_i \sim T_j$ if $\{x_i, y_i\} \cap \{x_j, y_j\} \neq \emptyset$, namely, that T_i, T_j share an element. Let $\Delta = \sum_{T_i \sim T_j} \Pr(A_i \wedge A_j)$. Then by Janson inequality (see, e.g., [2] Chapter 8),

$$\Pr(\overline{\mathcal{G}}_p) = \Pr(\wedge_{i=1}^{t}\ \overline{A}_i) \leq e^{-\mu + \frac{\Delta}{2}}$$

Obviously, $\Pr(A_i) = \frac{\nu^2}{n^2}$ for every $i \in [t]$. Hence, using that $p \leq g \leq \frac{n}{2}$, we conclude that

$$\mu = \frac{\nu^2 \cdot t}{n^2} \geq \frac{\epsilon n \cdot (n-p)}{2p} \cdot \frac{\nu^2}{n^2} \geq \frac{\epsilon \nu^2}{4p}$$

Observe that for $i \neq j$, $T_i \sim T_j$ if and only if $|\{x_i, y_i, x_j, y_j| = 3$. Consequently, we get that for $T_i \sim T_j$, $\Pr(A_i \wedge A_j) = (\frac{\nu}{n})^3$. Note that if $T_i \sim T_j$ then x_i, y_i, x_j, y_j are all elements of the same p-section of n. Any fixed subsection $Z(p,k)$ contains $\binom{|Z(p,k)|}{3} \leq \frac{n(n+p)(n-p)}{6p^3}$ such pairs. Thus there are, all together, at most $\frac{n(n+p)(n-p)}{6p^2}$ such pairs. Hence

$$\Delta \leq \frac{(n+p)(n-p)\nu^3}{6p^2n^2} \leq \frac{\nu^3}{6p^2}.$$

Our choice of ν and ϵ implies that $\mu \geq \Delta$ and hence $\Pr(\overline{\mathcal{G}}_p) \leq e^{-\frac{\mu}{2}} \leq e^{-2\log g} \leq \frac{1}{g^2}$ (where the inequality comes from substituting g for p as $p \leq g \leq n/2$). □

5 Lower Bounds

We prove here the lower bounds stated in Theorem 1.2, 1.4 and 1.5. All the lower bounds are shown for the case that $\Sigma = \{0, 1\}$, which implies the same lower bound for any alphabet that contains at least two letters.

Let $\mathcal{U}_n$ denote the uniform distribution over $\{0,1\}^n$. The following claim will be used repeatedly.

Claim 5.1. *Let $n \geq 2^9$ and $g \leq \frac{n}{2}$. Let α be a string drawn from $\mathcal{U}_n$. Then α is $\frac{1}{16}$-far from $Period(\leq g)$ with probability at least $\frac{8}{9}$.*

Proof. We shall show that for every integer $p \leq g$ the string α is $\frac{1}{16}$-*close* to $Period(p)$ with probability at most $\frac{2}{9 \cdot n}$. This is sufficient since by the union bound and Fact 2.2, the probability that α is $\frac{1}{16}$-*close* to $Period(\leq g)$ is at most $\frac{1}{9}$.

Fix $p \leq \frac{n}{2}$. Then the number of strings that are $\frac{1}{16}$-*close* to $Period(p)$ is at most $2^p \cdot \binom{n}{n/16}$, as there are 2^p ways to choose the periodic string. Using $\binom{n}{\beta n} \leq 2^{H_2(\beta)n}$, where $H_2(x) = -x \log x - (1-x)\log(1-x)$ is the entropy function, we conclude that the number of strings that are $\frac{1}{16}$-*close* to $Period(p)$ is at most $2^{n/2} \cdot 2^{H_2(1/16)n} = o(2^n/n)$. □

5.1 A Lower Bound on the One-Sided Error Non-Adaptive Query Complexity of *Periodicity*

This subsection is devoted to the proof of Theorem 1.5. Let $m = \sqrt{\frac{n \log n}{50}}$. Yao's principle [16], adapted for one-sided error, asserts that to prove a lower bound of m on the query complexity of a one-sided error, non-adaptive test for $Periodicity$, it is sufficient to show the following: For every large enough n, and $\epsilon = \frac{1}{16}$ there exists a distribution $\mathcal{D}$ that is concentrated on strings that are ϵ-far from $Periodicity$ such that for an α chosen according to $\mathcal{D}$, any fixed set $Q \subseteq [n]$ with $|Q| < m$ (Q is thought as the query set of the non adaptive deterministic test) contains no p-witness for some p, $\lfloor \frac{n}{4} \rfloor + 1 \leq p \leq \frac{n}{2}$, with probability more than $\frac{1}{3}$.

Let $\mathcal{U} = \mathcal{U}_n$ be the uniform distribution on $\{0,1\}^n$, and Far be the event that the string chosen according to $\mathcal{U}$ is ϵ-far from $Periodicity$. Let $\mathcal{D} = (\mathcal{U} |\ Far)$, that is, the uniform distribution conditioned on Far. By definition, $\mathcal{D}$ is concentrated on ϵ-far inputs. We note that $\mathcal{U}$ well approximates $\mathcal{D}$ in the following sense. For every event A, $\Pr_{\mathcal{D}}(A) \geq \Pr_{\mathcal{U}}(A \cap Far) \geq \Pr_{\mathcal{U}}(A) - (1 - \Pr_{\mathcal{U}}(Far)) \geq \Pr_{\mathcal{U}}(A) - \frac{1}{9}$, where the last inequality is by Claim 5.1. In the following, unless specifically stated, all probabilities will be with respect to $\mathcal{U}$.

For $Q \subseteq [n]$, $|Q| < m$, let $\mathcal{W}$ be the set of all unordered pairs $\{i,j\} \subseteq Q$. For every $p \in [\lfloor \frac{n}{4} \rfloor + 1, \frac{n}{2}]$ let $\mathcal{W}_p \subseteq \mathcal{W}$ be the set of all $\{i,j\} \in \mathcal{W}$ such that $i \equiv j \pmod{p}$ and let Q_p be the union of all the members of $\mathcal{W}_p$. Observe that if there exists a Q_p such that α_i has the same value for every $i \in Q_p$ then no pair of elements of Q is a p-witness for α. Thus, we only need to show that for α drawn according to $\mathcal{D}$ this happened with probability more than $\frac{1}{3}$. However, by the note above, it is enough to show that this is true with probability more than $\frac{4}{9}$ for an α drawn according to $\mathcal{U}$. Indeed this is shown in the following.

For every $p \leq g$ let A_p be the event that α_i has the same value for every $i \in Q_p$. We next show that there exists a subset $J \subseteq [\lfloor \frac{n}{4} \rfloor + 1, \ldots, \lfloor \frac{n}{2} \rfloor]$ such that: (1) $|J| \geq 4 \cdot n^{\frac{1}{3}}$, (2) $|Q_p| \leq \frac{\log n}{3}$ for every $p \in J$, and (3): the sets $Q_p,\ p \in J$ are pairwise disjoint. Assuming the existence of such J, for every $p \in J$, $\Pr(A_p) \geq 2^{-|Q_p|} \geq n^{-\frac{1}{3}}$, where the last inequality is by (2). Since the events $A_p,\ p \in J$ are independent by property (3) above, we get,

$$\Pr_{\alpha \in \mathcal{D}}(\cup_{p \in J} A_p) \geq 1 - \left(1 - n^{-\frac{1}{3}}\right)^{4 \cdot n^{\frac{1}{3}}} \geq \frac{8}{9}.$$

Thus, it only remains to be shown that such a set J exists. Observe that for every $i,j \in \mathcal{W}$ there are at most two distinct integers $p,q \in [\lfloor \frac{n}{4} \rfloor + 1, \frac{n}{2}]$ such that $i \equiv j \pmod{p}$ and $i \equiv j \pmod{q}$. Since there are at most $\binom{m}{2} \leq \frac{n \log n}{100}$ unordered pairs in $\mathcal{W}$, then, by averaging, there exists a set $\mathcal{I} \subseteq [\lfloor \frac{n}{4} \rfloor + 1, \frac{n}{2}]$ of size at least $\frac{n}{8}$ such that for every $p \in \mathcal{I}$ the set $\mathcal{W}_p$ has size at most $\frac{\log n}{6}$. Note that if $|\mathcal{W}_p| \leq \frac{\log n}{6}$ then $|Q_p| \leq \frac{\log n}{3}$. Let $J \subseteq \mathcal{I}$ be the set that is constructed as follows. We pick an arbitrary $p \in \mathcal{I}$ and add it to J, then we remove p from $\mathcal{I}$ and we also remove from $\mathcal{I}$ any q such that $Q_p \cap Q_q \neq \emptyset$. We repeat this until $\mathcal{I}$ becomes empty. Obviously, for every $p,q \in J$, $Q_p \cap Q_q = \emptyset$ as required by (3) above. Thus, we only need to show that J is large enough.

Observe that for every p that is inserted into J, every $i \in Q_p$ may result in a deletion of at most $2m$ q's from $\mathcal{I}$ since for each $j \in Q$, $\{i,j\}$ could belong to at most two W_q's. Consequently, since $|Q_p| \leq \frac{\log n}{3}$ for every $p \in J$, it follows that for every p that is inserted to J at most $\frac{\log n}{3} \cdot 2m = o(n^{\frac{2}{3}})$ potential members are removed from $\mathcal{I}$. Since at the beginning $|\mathcal{I}| \geq \frac{n}{8}$ we get that the size of J is as required. $\square$

5.2 Lower Bounds on the Query Complexity of adaptive two sided error tests for $Period(\leq g)$

In this subsection we prove Theorem 1.2 and 1.4. Both lower bounds deal with adaptive two-sided error ϵ-tests. Yao's principle [16] in this case states the following: To prove that any *2-sided error* ϵ-test for a property $\mathcal{P}$ requires more than m queries, it is enough to show that there exists a distribution $\mathcal{D}$ over inputs in $\mathcal{P}$, and inputs that are ϵ-far from $\mathcal{P}$, such that any *deterministic* ϵ-test for $\mathcal{P}$ that uses m queries, fails on an input drawn from $\mathcal{D}$ with probability greater than $\frac{1}{3}$.

For both proofs we use a distribution $\mathcal{D}$ that is constructed out of two separate distributions $\mathcal{D}_P$ and $\mathcal{D}_N$. The distribution $\mathcal{D}_P$ is over strings in $Period(\leq g)$ and the distribution $\mathcal{D}_N$ is over strings that are ϵ-far from $Period(\leq g)$. A string is drawn from $\mathcal{D}$ by first selecting uniformly at random one of the distribution $\mathcal{D}_P$,$\mathcal{D}_N$ and then returning a string drawn from the selected distribution. Thus formally $\mathcal{D} = \frac{1}{2}\mathcal{D}_P + \frac{1}{2}\mathcal{D}_N$.

We may assume without loss of generality, that any test making m queries in the worst case, is making m queries in every run (otherwise just add dummy queries). Thus such a test is a deterministic decision tree that all its leaves are at depth m.

Let $\alpha \in \{0,1\}^n$ be an input to the ϵ-test, let $M \subseteq [n]$ be the set of queries used by the ϵ-test on α, and let $\eta \in \{0,1\}^{|M|}$ be the set of answers, that is $\alpha_M = \eta$. The tuple (M, η) is denoted as the *interaction* of the tree with α. Thus, if the interaction of the decision tree with α is (M, η) then for every $\beta \in \Sigma^n$ such that $\beta_M = \eta$, the interaction with β will also be (M, η), and in particular, β will arrive to the same leaf of the decision tree and will be classified (either reject or accept) the same as α. Let $\mathcal{A}$ be the set of all accepting interactions (these that end in an accepting leaf) and $\mathcal{R}$ the set of all rejecting interactions (these that end in a rejecting leaf).

The proof of both lower bounds uses the following claim.

Claim 5.2. *Let* $\mathrm{A}lg$ *be a deterministic adaptive* ϵ*-test for* $Period(\leq g)$*, whose query complexity is* m*. Let* $\mathcal{D}$ *be a distribution as described above. Assume that for every interaction* (M, η) *of* Alg*,* $\Pr_{\alpha \in \mathcal{D}_N}[\alpha_M = \eta] > \frac{2}{3 \cdot 2^{|M|}}$ *and* $\Pr_{\alpha \in \mathcal{D}_P}[\alpha_M = \eta] > \frac{2}{3 \cdot 2^{|M|}}$*. Then,* $\mathrm{A}lg$ *errs with probability greater than* $\frac{1}{3}$ *on a random* α *that is chosen according to* $\mathcal{D}$*.*

Proof. The test Alg errs on a string α if it was drawn from $\mathcal{D}_P$ and there exists $(M, \eta) \in \mathcal{R}$ such that $\alpha_M = \eta$, or if it was drawn from $\mathcal{D}_N$ and there exists $(M, \eta) \in \mathcal{A}$ such that $\alpha_M = \eta$. Hence,

$$\Pr_{\alpha \in \mathcal{D}}(\ Alg \text{ errs }) \geq \frac{1}{2} \cdot \sum_{(M,\eta) \in \mathcal{A}} \Pr_{\alpha \in \mathcal{D}_N}[\alpha_M = \eta] + \frac{1}{2} \cdot \sum_{(M,\eta) \in \mathcal{R}} \Pr_{\alpha \in \mathcal{D}_P}[\alpha_M = \eta].$$

Using the lower bound assumed in the claim and the fact that $|\mathcal{A}| + |\mathcal{R}| = 2^{|M|}$ we get,

$$\Pr_{\alpha \in \mathcal{D}}(\ Alg \text{ errs }) > \frac{1}{2} \cdot |\mathcal{A}| \cdot \frac{2}{3 \cdot 2^{|M|}} + \frac{1}{2} \cdot |\mathcal{R}| \cdot \frac{2}{3 \cdot 2^{|M|}} = \frac{1}{3}.$$

□

For this section let Far be the event that contains all strings that are $\frac{1}{32}$-far from $Period(\leq g)$ and let $\mathcal{U} = \mathcal{U}_n$ be the uniform distribution over $\{0,1\}^n$. In the proofs of Theorems 1.2 and Theorem 1.4, the same distribution $\mathcal{D}_N$ is used (but with different n). It is defined as $\mathcal{D}_N = (\mathcal{U} |\ Far)$, that is, $\mathcal{D}_N$ is uniform distribution over all strings $\alpha \in \{0,1\}^n$ conditioned on the event Far. Claim 5.3 below asserts that the distribution $\mathcal{D}_N$ meets the assumptions in Claim 5.2 regarding $\mathcal{D}_N$.

Claim 5.3. *Let $g \leq \frac{n}{2}$ and let (M, η) be an interaction such that $|M| \leq \frac{n}{32}$. Then,*

$$\Pr_{\alpha \in \mathcal{D}_N}[\alpha_M = \eta] > \frac{2}{3 \cdot 2^{|M|}}$$

Proof. For any event A let $\Pr[A] = \Pr_{\mathcal{U}}[A]$. By Bayes rule,

$$\Pr_{\alpha \in \mathcal{D}_N}[\alpha_M = \eta] = \Pr[(\alpha_M = \eta) \mid Far] = \Pr[Far \mid (\alpha_M = \eta)] \cdot \frac{\Pr[\alpha_M = \eta]}{\Pr[Far]}. \tag{9}$$

Since $\Pr[\alpha_M = \eta] \geq \frac{1}{2^{|M|}}$ and $\Pr[Far] \leq 1$ we get that $\Pr_{\alpha \in \mathcal{D}_N}[(\alpha_M = \eta)] \geq \frac{1}{2^{|M|}} \cdot \Pr[Far \mid (\alpha_M = \eta)]$. Thus to complete the proof we only need to show that $\Pr[Far \mid (\alpha_M = \eta)] > \frac{2}{3}$.

One way of selecting a string uniformly from the set of all strings $\alpha \in \{0,1\}^n$ such that $\alpha_M = \eta$, is to first uniformly choose $\alpha \in \{0,1\}^n$ according to $\mathcal{U}$ and then to change the letters in α whose indices are in M so that $\alpha_M = \eta$. By Claim 5.1, with probability greater than $\frac{2}{3}$, a string drawn according to $\mathcal{U}$ is $\frac{1}{16}$-*far* from $Period(\leq g)$. By the triangle inequality if we change at most $\frac{n}{32}$ letters of a string that is $\frac{1}{16}$-*far* from $Period(\leq g)$ we get a string that is $(\frac{1}{16} - \frac{1}{32})$-*far* from $Period(\leq g)$. Thus, $\Pr[Far \mid (\alpha_M = \eta)] > \frac{2}{3}$. □

5.2.1 Proof of Theorem 1.2

Fix $g \leq \frac{n}{2}, \epsilon = \frac{1}{32}$ and let Alg be a deterministic adaptive, two sided error, ϵ-test for $Period(\leq g)$. Assume that Alg has query complexity $m = \frac{1}{4} \cdot \sqrt{\frac{g}{\log g \cdot \log n}}$. The distribution $\mathcal{D}_N$ was already defined above. We next describe the distributions $\mathcal{D}_P$.

Recall that $Primes(g)$ is the set of all primes that are smaller or equal to g and $\Pi(g) = |Primes(g)|$. A string is chosen according to $\mathcal{D}_P$ by uniformly selecting $p \in Primes(g)$, then uniformly and independently choosing $\omega \in \{0,1\}^p$ and finally setting α to be the concatenation of ω to itself enough times until a total length of n is obtained (possibly concatenating a prefix of ω at the end if p does not divide n).

Fix (M, η) to be any interaction of Alg. Let $Bad(M) = \{q \in Primes(g) | \ \exists i, j \in M, i \equiv j \pmod{q}\}$. If $p \notin Bad(M))$ then the α_i's are independent for every $i \in M$. Hence, in this case $\alpha_M = \eta$ with probability $\frac{1}{2^{|M|}}$. Consequently, $\Pr_{\alpha \in \mathcal{D}_P}[\alpha_M = \eta]$ is at least the probability that the prime is selected from $Primes(g) \setminus Bad(M)$ times $\frac{1}{2^{|M|}}$. Thus, we only need to show that $\frac{|Primes(g) \setminus Bad(M)|}{|Primes(g)|} > \frac{2}{3}$.

For every $q \in [n]$ there are at most $\log n$ different primes that divide it. Thus the size of $Bad(M)$ is at most $\log n \cdot \binom{m}{2} < \frac{g}{32 \cdot \log g}$. As noted in Section 2, $\Pi(g) \longrightarrow \frac{g}{\ln g}$ and hence $\Pi(g) \geq \frac{g}{\log g}$ for large enough g. This implies that $|Primes(g) \setminus Bad(M)| > \frac{2 \cdot \Pi(g)}{3}$. □

5.2.2 Proof of Theorem 1.4

Fix $g \leq \frac{\log n}{4}$. Since the theorem is asymptotic, we may assume that g is large enough so that $e^{0.9 \log g} \geq 2g$ and so that $r = \Pi(2 \log g) - \Pi(\log g) \geq 0.9 \frac{\log g}{\ln \log g}$. The first assumption is obviously correct for large enough g, while the second is by substituting $\Pi(2 \log g)$ with $(1 - \delta)$ times its limit formula, as given in Section 2, and $\Pi(\log g)$ by $1 + \delta$ times the limit formula, for small enough δ.

Let Alg be a potential deterministic adaptive, two-sided error, $(1/32)$-test for $Period(\leq g)$, that has query complexity $m \leq \sqrt{\frac{2}{3} \cdot \frac{1 + \log g}{1 + \log \log g}}$. Let $p_1, p_2, \ldots$ be the sequence of prime numbers in increasing order. The assumptions on r, g above imply that

$$\Pi_{p_i \in [\log g, 2 \log g]} p_i \geq (\log g)^r \geq e^{r \ln \log g} \geq e^{0.9 \log g} \geq 2g. \tag{10}$$

Let z be the largest integer such that $p_z \leq 2\log g$ and let y be the largest integer such that $\kappa = \prod_{i\in[y,z]} p_i > 2g$. Observe that such a $y \geq 2$ exists by Equation (10).

Note also that $\kappa \leq (2\log g)^{z-y+1}$ which implies that $z - y + 1 \geq \frac{1+\log g}{1+\log\log g}$. In addition, by the definition of κ and y it follows that $\kappa \leq 2g \cdot 2\log g$ which by our assumption on g implies that $\kappa \leq n/2$.

Let $S = \{\frac{\kappa}{p_j},\ j \in [y,z]\}$, thus $|S| \geq z - y + 1$. Note that by our assumption on m we get

$$\binom{m}{2} < m^2/2 \leq |S|/3 \tag{11}$$

We next describe the distributions $\mathcal{D}_P$,$\mathcal{D}_N$.

Both distributions $\mathcal{D}_P$ and $\mathcal{D}_N$, are over strings in $Period(\kappa)$. Each will be constructed by choosing $w \in \{0,1\}^\kappa$ according to the corresponding distributions $\mathcal{D}_P^*$ and $\mathcal{D}_N^*$ that are defined below, and then concatenating w to itself enough times to form a string of length n. Note that if $w \in Period(\leq g)$, then the resulting length-n string is in $Period(\leq g)$ w.r.t. $\{0,1\}^n$. Also, if w is ϵ-far from $Period(p)$, for any $p \leq g$, or from $Period(\leq g)$, then the corresponding length-n string is ϵ-far from $Period(p)$, or $Period(\leq g)$, respectively. Consequently, we may assume that Alg uses only queries in $[\kappa]$. This enables us to simplify notations by treating the input string as if it had length κ.

For this section let Far be the event that $w \in \{0,1\}^\kappa$ is $\frac{1}{32}$-far from $Period(\leq g)$. The distributions $\mathcal{D}_N^*$, $\mathcal{D}_P^*$ are defined as follows: $\mathcal{D}_N^* = (\mathcal{U}_\kappa \mid Far)$, where $\mathcal{U}_k$ is the uniform distribution on length-κ strings. The distribution $\mathcal{D}_P^*$ is defined by: first a random $s \in S$ is chosen, then a random $w \in \{0,1\}^s$ is chosen. Finally, w is concatenated to itself enough times so to form a word of length κ. Thus, as explained before, we may restrict ourselves to the words of length κ generated according to $\mathcal{D}_N^*$ and $\mathcal{D}_P^*$.

Since $\kappa > 2g$ we may apply Claim 5.3 which asserts that for any interaction (M,η) of Alg, $\Pr_{\alpha\in\mathcal{D}_N^*}[(\alpha_M = \eta)] > \frac{2}{3\cdot 2^{|M|}}$.

Fix (M,η) to be any possible interaction of Alg. Let $Bad(M)$ be the set of all t for which there exists $i,j \in M$ such that $i = j \pmod t$. If $s \notin Bad(M)$ then the α_i's are independent for every $i \in M$. Hence, in this case $\alpha_M = \eta$ with probability $\frac{1}{2^{|M|}}$. Consequently, $\Pr_{\alpha\in\mathcal{D}_P^*}[(\alpha_M = \eta)]$ is at least the probability of selecting $s \in S \setminus Bad(M))$ times $\frac{1}{2^{|M|}}$.

Observe that each member of $Bad(M)$ is divisible by at most one member of S, since the lowest common multiplier of any two elements in S is κ. Thus $|Bad(M)| \leq \binom{m}{2}$ which by Equation (11) implies that $|Bad(M) \leq \frac{1}{3}\cdot|S|$. Thus $\Pr[s \in Bad(M)] < \frac{1}{3}$ which implies that $\Pr_{\alpha\in\mathcal{D}_P^*}[(\alpha_M = \eta)] > \frac{2}{3}\cdot\frac{1}{2^{|M|}}$. Claim 5.2 ends the proof of the Theorem. □

References

[1] N. Alon, O. Goldreich, J. Håstad and R. Peralta, Simple Construction of Almost k-wise Independent Random Variables. Random Struct. & Algorithms, 3(3) 289-304, 1992, and Addendum at same journal 4(1) 119–120, 1993.

[2] N. Alon and J. H. Spencer, **The Probabilistic Method**, Second Edition, Wiley, New York, 2000.

[3] F. Ergun, S. Muthukrishnan, and C. Sahinalp. Sub-linear methods for detecting periodic trends in data streams. In *LATIN 2004, Proc. of the 6th Latin American Symposium on Theoretical Informatics*, 16–28, 2004.

[4] E. Fischer, The art of uninformed decisions: A primer to property testing, *Current Trends in Theoretical Computer Science: The Challenge of the New Century, G. Paun, G. Rozenberg and A. Salomaa (editors), World Scientific Publishing* (2004), Vol. I 229-264.

[5] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss, Near-optimal sparse Fourier representations via sampling. In *STOC 2002, Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, 152–161, 2002.

[6] S. Goldwasser O. Goldreich and D. Ron, Property testing and its connection to learning and approximation. *Journal of the ACM*, 45:653–750, 1998.

[7] J. Hadamard, Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques. *Bull. Soc. Math. France*, 24:199–220, 1896.

[8] P. Indyk, N. Koudas, and S. Muthukrishnan, Identifying representative trends in massive time series data sets using sketches. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, 363–372. Morgan Kaufmann, 2000.

[9] R. Krauthgamer, O. Sasson, Property testing of data dimensionality. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, 18–27, 2003.

[10] J. Naor and M. Naor, Small-bias probability spaces: efficient construction and applications, *SIAM J. Comput.* 22(4): 838-856, 1993.

[11] D.J. Newman, Simple analytic proof of the prime number theorem. *Amer. Math. Monthly*, 87:693–696, 1980.

[12] V. Poussin, Recherces analytiques sur la théorie des nombres premiers. *Ann. Soc. Sci. Bruxelles*, 1897.

[13] R. Rubinfeld and M. Sudan, Robust characterization of polynomials with applications to program testing. *SIAM Journal of Computing*, 25:252–271, 1996.

[14] D. Ron, Property testing (a tutorial). In *Handbook of Randomized computing*, 597–649. Kluwer Press, 2001.

[15] A. Samorodnitsky and L. Trevisan, A PCP characterization of NP with optimal amortized query complexity. In *Proc. of the 32 ACM STOC*, 191–199, 2000.

[16] A. C. Yao, Probabilistic computations: Towards a unified measure of complexity. In *FOCS '17: Proceedings of the 17th Annual Symposium on Foundations of Computer Science*, 222–227, 1977.