Testing Periodicity[⋆]

Oded Lachish[1] and Ilan Newman[2]

[1] Haifa University, Haifa, Israel 31905,
emailloded@cs.haifa.ac.il
[2] Haifa University, Haifa, Israel 31905,
emaililan@cs.haifa.ac.il

Abstract. A string $\alpha \in \Sigma^n$ is called *p-periodic*, if for every $i, j \in \{1, \ldots, n\}$, such that $i \equiv j \bmod p$, $\alpha_i = \alpha_j$, where α_i is the i-th place of α. A string $\alpha \in \Sigma^n$ is said to be $period(\leq g)$, if there exists $p \in \{1, \ldots, g\}$ such that α is *p-periodic*.

An ϵ property tester for $period(\leq g)$ is a randomized algorithm, that for an input α distinguishes between the case that α is in $period(\leq g)$ and the case that one needs to change at least ϵ-fraction of the letters of α, so that it will become $period(\leq g)$. The complexity of the tester is the number of letter-queries it makes to the input. We study here the complexity of ϵ testers for $period(\leq g)$ when g varies in the range $1, \ldots, \frac{n}{2}$. We show that there exists a surprising exponential phase transition in the query complexity around $g = \log n$. That is, for every $\delta > 0$ and for each g, such that $g \geq (\log n)^{1+\delta}$, the number of queries required and sufficient for testing $period(\leq g)$ is polynomial in g. On the other hand, for each $g \leq \frac{logn}{4}$, the number of queries required and sufficient for testing $period(\leq g)$ is only poly-logarithmic in g.

We also prove an exact asymptotic bound for testing general periodicity. Namely, that 1-sided error, non adaptive ϵ-testing of periodicity ($period(\leq \frac{n}{2})$) is $\Theta(\sqrt{n \log n})$ queries.

1 Introduction

Periodicity in strings plays an important role in several branches of CS and engineering applications. It is being used as a measure of 'self similarity' in many application regarding string algorithms (e.g. pattern matching), computational biology, data analysis and planning (e.g. analysis of stock prices, communication patterns etc.), signal and image processing and others. On the other hand, sources of 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 (sub-linear time) algorithms that test whether a long sequence is periodic or approximately periodic, and in particular, that test if it has a very short

[⋆] this research was supported by THE ISRAEL SCIENCE FOUNDATION (grant number 55/03)

period. This calls for algorithms in the framework of Combinatorial Property Testing [1]. In this framework, introduced initially by Rubinfeld and Sudan [2] and formalized by Goldreich et al. [1], one uses a randomized algorithm that queries the input at very few locations and based on this, decides whether it has a given property or it is 'far' from having the property. Indeed related questions to periodicity have already been investigated [3–5], although the focus here is somewhat different.

In [6], the authors constructs an algorithm that approximates in a certain sense the DFT (Discrete Fourier Transform) of a finite sequence in sub linear time. This is quite related but not equivalent to testing how close is a sequence 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. They mainly relate the different definitions of periodicity. They also show that there is a tolerant tester for periodicity. That is, they show a simple 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.

There are other works on sequences sketching [4] etc. but none of those seems to address directly periodicity testing.

The property of being periodic is formalized here in a very general form: Let Σ be a finite alphabet. A string $\alpha \in \Sigma^n$ is said to be *p-periodic* for an integer $p \in [n]$, if for every $i, j \in [n]$, $i \equiv j (\text{mod} p)$, $\alpha_i = \alpha_j$, where α_i is the i-th character of α. We say that α is in $period(\leq g)$, where $g \in [\frac{n}{2}]$, if it is *p-periodic* for some $p \leq g$. We say that a string is periodic if it is in $period(\leq \frac{n}{2})$.

We study here the complexity of ϵ testing the property $period(\leq g)$ when g varies in the range $1, \ldots, \frac{n}{2}$. An ϵ-tester for $period(\leq g)$ is a randomized algorithm, that for an input $\alpha \in \Sigma^n$ distinguishes between the case that α is in $period(\leq g)$ and the case that one needs to change at least an ϵ-fraction of the letters of α, so that it will be in $period(\leq g)$. The complexity of the tester is the number of letter-queries it makes to the input.

We show that there exists a surprising exponential phase transition in the complexity of testing $period(\leq g)$ around $g = \log n$. That is, for every $\delta > 0$ and for each g, such that $g \geq (\log n)^{1+\delta}$, the number of queries required and sufficient for testing $period(\leq g)$ is polynomial in g. On the other hand, for each $g \leq \frac{logn}{4}$, the number of queries required and sufficient for testing $period(\leq g)$ is only poly-logarithmic in g. We also settle the exact complexity of non-adaptive 1-sided error test for general periodicity (that is, $period(\leq n/2)$). We show that the exact complexity in this case is $\theta(\sqrt{n \log n})$. The upper bound that we prove is an improvement over the result of [3] and uses a construction of a small random set $A \subseteq [n]$ of size $\sqrt{n \log n}$ for which the multi-set $A - A = \{a - b | a, b \in A\}$ contains at least $\log n$ copies of each member of $[n/2]$. This can be trivially done with a set A of size $\sqrt{n} \log n$. We improve on the previous bound by giving up the full independence of the samples, but still retain the property that the probability of the $\log n$ copies of each number in $A - A$ behave as if being not too far from independent. A similar problem has occurred in various other situation, (e.g. [7, 8]) and thus could be interesting in its own.

The rest of the paper is organized as follows. In section 2 we introduce the necessary notations and some very basic observations. Section 3 contains an ϵ-test for *period*$(\leq g)$ that uses $\theta(\sqrt{g \log g})$ queries. In section 4 we construct an ϵ-test for *period*$(\leq g)$, $g \leq \frac{\log n}{4}$ that uses only $\tilde{O}((\log g)^6)$ queries. Sections 5 and 6 contain the corresponding lower bounds, thus showing the claimed phase transition. Finally, the special case of $g = \theta(n)$ is treated in Section 7 which contains a proof that any 1-sided error non adaptive tester for periodicity (*period*$(\leq \frac{n}{2})$) requires $\Omega(\sqrt{n \log n})$ queries. This implies that the tester in section 3 is asymptotically optimal when $g = n/2$.

2 Preliminaries

In the following Σ is a fixed size alphabet. For a string $\alpha \in \Sigma^n$ and an integer $i \in [n]$, ($[n] = \{1, \ldots, n\}$) we denote by α_i the i-th symbol of α, that is $\alpha = \alpha_1 \ldots \alpha_n$. Given a set $S \subseteq [n]$ such that $S = \{i_1, i_2, \ldots, i_m\}$ and $i_1 < i_2 < \ldots < i_m$ we define $\alpha_S = \alpha_{i_1}\alpha_{i_2} \ldots \alpha_{i_m}$. In the following Σ will be fixed. Also, unless otherwise stated, all strings are in Σ^n.

For two strings α, β we denote by $dist(\alpha, \beta)$ the Hamming distance between α and β. Namely, $dist(\alpha, \beta) = |\{i|\ \alpha_i \neq \beta_i\}|$. 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 $\mathcal{P}$.

Definition 1. *For a string α and a subset $S \subseteq [n]$ we say that α_S is homogeneous if for all $i, j \in S$, $\alpha_i = \alpha_j$.*

Property Testing

The type of algorithms that we consider here are 'property-testers' [1, 9, 10]. An ϵ-test is a randomized algorithms that accesses the input string via a 'location-oracle' which it can query: A query is done by specifying one place in the string to which the answer is the value of the string in the queried location. The complexity of the algorithm is the amount of queries it makes in the worst case. Such an algorithm is said to be an ϵ-test for a property $\mathcal{P} \subseteq \Sigma^n$ if it distinguishes with success probability at least 2/3 between the case that the input string belongs to $\mathcal{P}$ and the case that it is *ϵ-far* from $\mathcal{P}$.

Periodicity

Definition 2. *A string $\alpha \in \Sigma^n$ has period p (denoted p-periodic), if $\alpha_i = \alpha_j$ for every $i, j \in [n]$, such that $i \equiv j \bmod p$.*

Note that a string is homogeneous if and only if it is *1-periodic.*

Definition 3. *A string has the property period$(\leq g)$ if it is p-periodic for some $p \leq g$. We say that it is periodic if it has the property period$(\leq \frac{n}{2})$.*

Definition 4. *A witness that a string α is not p-periodic, denoted as p-witness, is an unordered pair $\{i, j\} \subseteq [n]$, such that $i \equiv j (mod\ p)$ and $\alpha_i \neq \alpha_j$.*

According to the definition of periodicity a string has a period $p \leq \frac{n}{2}$ if and only if there does not exist a *p-witness*.

In the same manner a witness for not having *period*$(\leq g)$ is defined as follows:

Definition 5. *A witness that a string is not in period*$(\leq g)$ *is a set of integers* $Q \subseteq [n]$ *such that for every* $p \leq g$ *there are two integers* $i, j \in Q$ *that form a p-witness.*

Fact 1. *A string has period*$(\leq g)$*, if and only if there does not exist a witness that the string is not in period*$(\leq g)$*.*

Fact 2. *If a string* $\alpha \in \Sigma^n$ *does not have a period in* $(\frac{g}{2}, g]$*, where* $g \leq \frac{n}{2}$*, then it has no period of at most* $\frac{g}{2}$*. If a string* $\alpha \in \Sigma^n$ *is* ϵ*-far from having a period in* $(\frac{g}{2}, g]$*, where* $g \leq \frac{n}{2}$*, then it is* ϵ*-far from having the property 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 must exist such $q \in (\frac{g}{2}, g]$. □

Definition 6. *For* $\alpha \in \Sigma^n$*,* $p \in [n]$ *and* $0 \leq i \leq p-1$*, let* $Z(p,i) = \{j |\ j \equiv i(mod\ p)\}$*. We call* $\alpha_{Z(p,i)}$ *the i-th p-section of* α*.*

The following obvious fact relates the distance of a string to *p-periodic* and the homogeneity of its *p-sections*.

Fact 3. *For each* α*,* $dist(\alpha, p-periodic) = \Sigma_{i=0}^{p-1} dist(\alpha_{Z(p,i)}, homogeneous)$*.*

In further sections we use the following basic ϵ-test for *p-period*.

Algorithm p-test

Input: a string $\alpha \in \Sigma^n$, the string length n, a period $p \in [\frac{n}{2}]$ and a distance parameter $0 < \epsilon < 1$;

1. Select $\frac{1}{\epsilon}$ random unordered pairs (with repetitions) $\{i, j\} \subset [n]$ such that $i \equiv j(\text{mod}p)$.
2. Reject if one of the selected pairs is a *p-witness* (namely a witness for being *non-p-periodic*). Otherwise accept.

Proposition 1. *Algorithm p-test is a 1-sided error, non-adaptive* ϵ*-test for p-periodic. Its query complexity is* $\frac{2}{\epsilon}$*.*

Proof. The query complexity is obvious. The test is 1-sided error since it rejects only if it finds a *p-witness*. To estimate its error probability let α be such that $dist(\alpha, p-periodic) \geq \epsilon n$.

Assume that p divides n (the general case is essentially the same). With this assumption $|Z(p,i)| = n/p = m$. For every $0 \leq i \leq p-1$ set $dist(\alpha_{Z(p,i)}, homogeneous) = d_i$. For fixed i and every $\sigma \in \Sigma$ let n_σ be the number of occurrences of σ in $Z(p,i)$. Assume also that we have renumbered the letters in Σ so that $n_1 \geq n_2, \ldots \geq n_k$. Then $d_i = n - n_1$ as it is easy to see that the closest homogeneous string is obtained by changing all letters different from σ_1 to σ_1 (exactly $m - n_1$ letters). Hence the number of *p-witnesses* in $Z(p,i)$, W_i, is $W_i = \frac{1}{2}(\Sigma n_j)^2 - \Sigma n_j^2$. It is easy to see that $W_i \geq m \cdot d_i/2$. Now, according to Fact 3 we get that, $dist(\alpha, p - periodic) =$ $\Sigma_{i=0}^{p-1} dist(\alpha_{Z(p,i)}, homogeneous) = \Sigma_{i=0}^{p-1} d_i \leq \frac{2}{m} \Sigma W_i \leq \frac{2}{m} W$ where W is the total number of *p-witnesses*.

It follows that $W \geq \frac{m}{2} dist(\alpha, p - periodic) \geq \frac{\epsilon n^2}{2p}$. Note however that the total number of unordered pairs i,j such that $j \equiv i(mod\ p)$ is $\frac{n^2}{2p}$. Thus we conclude that the probability that a random such pair is a *p-witness* is at least ϵ and the result follows. □

We will need the following proposition in a number of lower bound proofs. Let $m < n/8$, $S \subset [n]$, $|S| = m$ and a be an assignment $a : S \longrightarrow \{0,1\}$. We define the following distribution U/S on strings of length n. We choose every letter $\alpha_i = a(i)$ if $i \in S$. For all places $i \notin S$ we choose α_i to be '1' with probability 1/2 and '0' with probability 1/2, independently between different i's. Note that U/S is just the uniform distribution on all binary strings conditioned on the event that the projection on S is a. We have,

Proposition 2. *Let m, S, a be as above and let $\mathcal{G}(g)$ be the event that a string selected according to U/S is $\frac{1}{16}$-far from having period$(\leq g)$. Then* $\text{Prob}_{U/S}(\mathcal{G}(g)) \geq 1 - \frac{1}{n}$.

Proof: Let U be the uniform distribution over Σ^n. Note that if a string is $\frac{1}{16}$-*far* from having *period*$(\leq \frac{n}{2})$ then for every $g \leq \frac{n}{2}$ the string $\frac{1}{16}$-*far* from having *period*$(\leq g)$. Hence it is sufficient to prove the proposition only in the case that $g = \frac{n}{2}$. According to Fact 2, it is enough to prove that the probability that a string selected according to U/S is $\frac{1}{16}$-*far* from having a period in $\left(\frac{n}{4}, \frac{n}{2}\right]$, is at least $1 - \frac{1}{n}$. We prove that for each $p \in \left(\frac{n}{4}, \frac{n}{2}\right]$ the probability that a string selected according to U, is $\frac{1}{16}$-*close* to being *p-periodic*, is at most $\frac{4}{n^2}$, and therefore by the union bound we are done.

Indeed let $p \in \left(\frac{n}{4}, \frac{n}{2}\right]$, and let α be a string selected according to U/S. Since $p \in \left(\frac{n}{4}, \frac{n}{2}\right]$ there are at least $\frac{n}{4}$ *p-sections*. Because $m \leq \frac{n}{8}$ and the size of each *p-section* is at least 2, at least $\frac{3n}{16}$ of these *p-sections* contain at least 1 location not in S. We call such a location a *free location*, and define B to be the set of all integers x such that the x *p-section* contains a free location.

Then for every $i \in B$, $\text{Prob}_{U/S}(dist(\alpha_{Z(p,i)}, homogeneous) \geq 1) \geq \frac{1}{2}$ (the *free location* is different than some other location in the *p-sections*). Thus, since the *p-sections* are mutually disjoint and $|B| \geq \frac{3n}{16}$, Chernoff bound implies that, $\text{Prob}_{U/S}\left(\sum_{i \in B} dist(\alpha_{Z(p,i)}, homogeneous) \leq \frac{n}{16}\right) \leq e^{-\frac{n}{24}} \leq \frac{4}{n^2}$. □

3 Upper Bound for Testing *period*$(\leq g)$

In this section we construct a 1-sided error, non adaptive ϵ-test for *period*$(\leq g)$ that uses $O(\sqrt{g \log g}/\epsilon^2)$ queries. The construction consists of two stages. We first show a 1-sided error, non adaptive algorithm, PT, for testing periodicity ($period(\leq \frac{n}{2})$), that uses $O(\frac{\sqrt{n \log n}}{\epsilon^2})$ queries. Then we show how to use algorithm PT in order to construct the claimed algorithm for testing $period(\leq g)$.

The intuition behind the algorithm PT is simple. We first explain it by hinting to a test for $period(\leq \frac{n}{2})$ that uses $O(\frac{\sqrt{n} \log n}{\epsilon^2})$ queries. By Fact 2 α is ϵ-*far* from $period(\leq n)$ if and only if α is ϵ-*far* from being p-*periodic* for every $p \in (\frac{n}{4}, \frac{n}{2}]$. We show that if α is ϵ-*far* from $period(\leq \frac{n}{2})$, we can find a p-*witness* for each $p \in (\frac{n}{4}, \frac{n}{2}]$ with probability at least $\frac{1}{n}$. This is sufficient since by the union bound we will be done.

Suppose that we can construct a random set $Q \subseteq [n]$ of size $O(\sqrt{n} \log n)$, such that Q contains at least $\log n$ potential p-*witnesses* for each $p \in (\frac{n}{4}, \frac{n}{2}]$. Then obviously we achieve the above goal. Indeed such a set Q can be constructed e.g. by choosing Q uniformly between all sets of the required cardinality.

To reduce the size of Q, we construct such a random set Q of size $\sqrt{n \log n}$ with the above properties while giving up the independence between the $\log n$ pairs for each p. In turn we will have to show that the probability that none will be a p-*witness* is still low enough.

We use the following definition in the ϵ-test for $g = \frac{n}{2}$ (periodicity).

Definition 7. *Let* $\ell = \sqrt{n \log n}$ *and let* $J = \{I_i\}_{i=1}^{\ell}$ *be the set of pairwise disjoint intervals* $I_i = \left((i-1)\sqrt{\frac{n}{\log n}}, i\sqrt{\frac{n}{\log n}}\right]$.

We now present the ϵ-test for $period(\leq \frac{n}{2})$.

Algorithm PT

Input: a string $\alpha \in \Sigma^n$, the string length n and a distance parameter $0 < \epsilon < 1$;

Let ℓ, J be as in Definition 7.

1. Select a set of integers $T = \{t_1, \ldots, t_\ell\}$ by choosing one random point from each $I \in J$.
2. Repeat the following independently for $m = \frac{2^{10} \cdot \log n}{\epsilon^2}$ times: uniformly select an interval $I \in J$. Let J^* be the set of intervals that where selected and let $H = \cup_{I \in J^*} I$.
3. Reject if for every $p \in (\frac{n}{4}, \frac{n}{2}]$ the set $H \cup T$ contains a p-*witness*. Otherwise accept.

Theorem 1. *Algorithm PT is a 1-sided error, non-adaptive* ϵ*-test for periodicity. Its query complexity is* $O(\sqrt{n \log n}/\epsilon^2)$.

The proof is in omitted.

Theorem 2. *For any* $g \leq \frac{n}{2}$ *there is a 1-sided error, non-adaptive* ϵ*-test for* $period(\leq g)$*. Its query complexity is* $O(\sqrt{g \log g}/\epsilon^2)$.

Proof (idea): Let $\alpha \in \Sigma^n$. We think of α as being composed of $\frac{n}{2g}$ pieces of length $n' = 2g$ each. We now run the ϵ-test for periodicity for strings of length n' and for each query $q \in [n']$ we query q in one the pieces that is chosen randomly and independently for each query. We avoid further details here. □

4 Upper Bound for Testing *period*$(\leq g)$, where $g \leq \frac{\log n}{4}$

Let $g \leq \frac{\log n}{4}$, we describe an algorithm for testing whether a string has *period*$(\leq g)$ that has query complexity *poly*$(\log g)$. The technicalities are somewhat involved, however, the intuition is simple and motivated by the following reasoning: Our goal is to find a witness for not being *r-periodic* for every $r \in [g]$. One thing that we could do easily is to check whether α is *q-periodic* where q is the product of all numbers in $[g]$ (this number would be a actually too big, but for understanding the following intuition this should suffice). Now if α is not *q-periodic* then it is certainly not in *period*$(\leq g)$ and we are done. Yet as far as we know α may be far from having *period*$(\leq g)$ but is *q-periodic*. Our goal is to show that if α is far from *p-periodic*, where $p \leq g$, then there exists $q' >> g$, such that, α is far from *q'-periodic* and p divides q'. Indeed it follows form Lemma 3 that if α is far form p then there exists such a q' as above. This together with the following concept enables us to construct an *ϵ-test*.

Definition 8. *For integers n,g $\in [n]$, a gcd-cover of $[g]$ is a set $E \subseteq [n]$ such that for every $\ell \leq g$, there exists a subset $I \subseteq E$, that satisfies $\ell = \gcd(I)$.*

The importance of a *gcd-cover of* $[g]$ is the following: We prove that if $\ell = gcd(I)$ for an integer ℓ and a subset I, then, the assumption that α is far from being *ℓ-periodic* implies that it is also sufficiently far from being *t-periodic* for some $t \in I$. Using this, let E be a *gcd-cover of* $[g]$. Then we are going to query for each $t \in E$ enough pairs $i, j \in [n]$ such that $i \equiv j (\text{mod } t)$. We say that such pairs i, j cover t. Now for each $\ell \leq g$ there is a set $I_\ell \subseteq E$ such that $\ell = gcd(I_\ell)$. The pairs of queries that cover I_ℓ cover ℓ. Thus if α is 'far' from being *ℓ-periodic* then, as $gcd(I_\ell) = \ell$, there exists a period $t,\ t \in I_\ell$, such that the string is far from *t-periodic*. We thus expect that this t will distinguish between strings that are *ℓ-periodic* and those that are *$k\ell$-periodic* but far from being *ℓ-periodic*.

As the complexity of the test will depend crucially on the size of the *gcd-cover*, we need to guarantee the existence of a small one. This is done in the following Lemma, which also brings in an additional technical requirement.

Lemma 1. *For every integers n and $g \in \left[\frac{\log n}{4}\right]$, there exists a gcd-cover of $[g]$, $E \subseteq [\sqrt{n}]$, of size $O((\log g)^3)$.*

The proof of Lemma 1 is omitted.

A *gcd-cover of* $[g]$ as in the Lemma is called an *efficient gcd-cover of* $[g]$.

We next describe our algorithm.

Algorithm SPT

Input: a string $\alpha \in \Sigma^n$, a distance parameter $0 < \epsilon < 1$ and a threshold period $g \leq \frac{\log n}{4}$;

1. Set E to be an *efficient gcd-cover of* $[g]$.
2. Let $M = \frac{|E| \cdot \log(8|E|)}{2\epsilon}$. For each $\ell \in E$ select M pairs of integers uniformly and with repetitions from the set $\{\{x, y\} \mid x \equiv y \bmod \ell \text{ and } x, y \in [n]\}$. Let Q be the resulting (multi)set.

3. Reject if for each $\ell \in [g]$ the set Q contains a witness that α is not *ℓ-periodic*. Otherwise accept.

Theorem 3. *Algorithm SPT is a 1-sided error, non-adaptive ϵ-test for period($\leq g$), where $g \leq \frac{\log n}{4}$. Its query complexity is $\tilde{O}\left(\frac{(\log g)^6}{\epsilon}\right)$.*

Proof: Since for every member of E we used M queries and $|E| = O((\log g)^3)$ the estimate on the query complexity follows. The fact that the algorithm never rejects a *ℓ-periodic* string, for $\ell \leq g$ is immediate, as the algorithm rejects only when it finds a witness that the input string is not in *period*($\leq g$).

In order to compute the success probability of algorithm *SPT* we first need the following definition.

Definition 9. *A string $\alpha \in \Sigma^n$ is called ϵ-bad if for every $\ell \leq g$ and every $S \subseteq E$, where $\ell = gcd(S)$, there exists $s \in S$ for which α is $\frac{\epsilon}{2|E|}$-far from being s-periodic.*

Note: if α is *ϵ-bad* and s, ℓ as in the definition, then a witness for it showing that it is not *s-periodic* is also a witness that it is not *ℓ-periodic*.

The proof of the Theorem now follows from Lemma 2 and Lemma 3. □

Lemma 2. *Algorithm SPT rejects every α that is ϵ-bad with probability at least $\frac{3}{4}$.*

Proof. Let $\alpha \in \Sigma^n$ be *ϵ-bad*. Let $B \subseteq E$ be the set that contains every $b \in E$ such that α is $\frac{\epsilon}{2|E|}$*-far* from being *b-periodic*. By Definition 9, for every $\ell \leq g$ there exists a $b \in B$ such that $\ell \mid b$, and hence if we find a *b-witness* for each $b \in B$ - this is also a witness that α is not in *period*($\leq g$). Thus, it is enough to prove that for each $b \in B$, the probability that there is no *b-witness* in Q is at most $\frac{1}{8|E|}$, as then by the union bound we are done.

Let $\delta = \frac{\epsilon}{2|E|}$ and let b be a member of B, namely α is *δ-far* from being *b-periodic*. By Corollary 1, for a random $x, y \in [n]$, such that $x \equiv y \bmod b$ we have that $Prob(\alpha_x \neq \alpha_y) \geq \delta$. Since M random pairs are being queried for each $\ell \in E$ and in particular for b, the probability that a *b-witness* is not found is at most $(1-\delta)^M \leq \frac{1}{(8|E|)}$. □

Lemma 3. *If a string $\alpha \in \Sigma^n$ is ϵ-far from having short period($\leq g$), where $g \leq (\log n)/4$ then it is a ϵ-bad string.*

The proof of Lemma 3 is omitted.

5 Lower Bound for Testing *period*($\leq g$)

In this section we prove that every $g \leq n/2$ every adaptive, 2-sided error, $\frac{1}{32}$-test for *period*($\leq g$) uses $\Omega(\sqrt{g/(\log g \cdot \log n)})$ queries. We prove this for $\Sigma = \{0, 1\}$. This implies the same bound for every alphabet that contains at least two symbols.

Theorem 4. *Any adaptive 2-sided, error $\frac{1}{32}$-test for period($\leq g$), uses $\Omega(\sqrt{g/(\log g \cdot \log n)})$ queries.*

Proof. Fix $g \leq n/2$. We prove the theorem by using Yao's principle. That is, we construct a distribution $\mathcal{D}$ over legitimate instances (strings that are either in *period*($\leq g$), or strings that are $\frac{1}{32}$-*far* from *period*($\leq g$)) and prove that any adaptive deterministic tester, that uses $m = o(\sqrt{g/(\log g \cdot \log n)})$ queries, gives an incorrect answer with probability greater than $\frac{1}{3}$.

In order to define $\mathcal{D}$ we use auxiliary distributions $\mathcal{D}_P, \mathcal{D}_N$ and the following notation. Let $Primes = \{p \mid p \text{ is a prime such that } p \leq g\}$. According to the *Prime Number Theorem* [11–13], $|Primes| = \theta(\frac{g}{\log g})$.

We now define distributions $\mathcal{D}_P, \mathcal{D}_N$.

- $\mathcal{D}_N$ is simply the uniform distribution over Σ^n.
- An instance α of length n is selected according to distribution $\mathcal{D}_P$ as follows. Uniformly select $p \in Primes$, then uniformly select $\omega \in \Sigma^p$ and finally set $\alpha = (\omega)^{\frac{n}{p}}$.

We next define distribution $\mathcal{D}$. Let $\mathcal{G}$ be the event that α is $\frac{1}{16}$-*far* from having *period*($\leq g$). Let $\mathcal{D}_{N/\mathcal{G}}$ be the distribution D_N given that event $\mathcal{G}$ is true. A string α is selected according to distribution $\mathcal{D}$ by choosing one of the distributions $\mathcal{D}_P$,$\mathcal{D}_{N/\mathcal{G}}$ with equal probability and then selecting α according to the distribution chosen. Namely $\mathcal{D} = \frac{1}{2}\mathcal{D}_P + \frac{1}{2}\mathcal{D}_{N/\mathcal{G}}$.

We don't work directly with $\mathcal{D}$ but rather with a simpler distribution $\mathcal{D}'$ which approximates $\mathcal{D}$ well enough and is defined as follows $\mathcal{D}' = \frac{1}{2}\mathcal{D}_P + \frac{1}{2}\mathcal{D}_N$. Let $\mathcal{B}$ be the event that the tester gives an incorrect answer. We prove that $\text{Prob}_{\mathcal{D}'}(\mathcal{B}) \geq \frac{2}{5}$. This is indeed sufficient as $\text{Prob}_{\mathcal{D}}(\mathcal{B}) \geq \text{Prob}_{\mathcal{D}'}(\mathcal{B}) - \text{Prob}_{\mathcal{D}_N}(\overline{\mathcal{G}})$. Using Proposition 2, $\text{Prob}_{\mathcal{D}}(\mathcal{B}) \geq \frac{1-o(1)}{2} - \frac{1}{g} > \frac{1}{3}$.

We assume with out loss of generality that for any string of length n the tester uses the same number of queries. Hence we can view the tester as a full binary decision tree of depth m which is labeled as follows. Each node of the tree represents a query location, for each internal node one of the outgoing edges is labeled by 1 and the other by 0, where $0, 1$ represent the answers to the query, and each leaf is labeled either by "accept" or by "reject" according to the decision of the algorithm.

For each leaf l in the tree we associate a pair Q_l, f_l, where Q_l is the set of queries on the path from the root to the leaf l and $f_l : Q_l \longrightarrow \{0, 1\}$ is a mapping between each query and its answer, that is the labellings on the edges of the path from the root to the leaf l. Let L be the set of all leaves, L_0 be the set of all leaves that are labeled by "reject" and let L_1 be the set of all leaves that are labeled by "accept".

Let $h_\ell : \Sigma^n \longrightarrow \{0, 1\}$ be a function that is 1 only on strings $\alpha \in \Sigma^n$ such that for every $q \in Q_\ell$ we have $\alpha_q = f_\ell(q)$. That is $h_\ell(\alpha) = 1$ if and only if α is consistent with Q_ℓ, f_ℓ. Let $far : \Sigma^n \longrightarrow \{0, 1\}$ be a function that is 1 only on strings $\alpha \in \Sigma^n$ such that are $\frac{1}{16}$-*far* from *period*($\leq g$)). Thus the $\text{Prob}_{\mathcal{D}'}(\mathcal{B})$ is at least

$$\frac{1}{2}\Sigma_{\ell\in L_0} prob_{\mathcal{D}_P}[h_\ell(\alpha)=1]+\frac{1}{2}\Sigma_{\ell\in L_1} prob_{\mathcal{D}_N}[far(\alpha)=1\bigwedge h_\ell(\alpha)=1] \tag{1}$$

We will prove that for each $\ell\in L_0$, $prob_{\mathcal{D}_P}[h_\ell(\alpha)=1]\geq\frac{1-o(1)}{2^m}$, and for each $\ell\in L_1$, $prob_{\mathcal{D}_N}[far(\alpha)=1\bigwedge h_\ell(\alpha)=1]\geq\frac{1-o(1)}{\cdot 2^m}$. This implies that $\mathrm{Prob}_{D'}(\mathcal{B})\geq\frac{1}{2}\Sigma_{\ell\in L}\frac{1-o(1)}{\cdot 2^m}\geq\frac{1-o(1)}{2}$.

Indeed, recall that a string is selected according to D_P by first selecting $z\in Primes$ and then selecting a *z-periodic* string. For $Q\subset[n]$, $|Q|=m$ let $\mathcal{A}(Q)$ be the event that for α selected according to $\mathcal{D}_P$, there exists no $j,k\in Q$ such that $j\equiv k \bmod z$. For any fixed $i<j\in Q$ there are at most $\log n$ prime divisors of $j-i$. Hence for each $\ell\in L_0$, $prob_{\mathcal{D}_P}[\mathcal{A}(Q_\ell)]\geq 1-\frac{m^2\cdot\log n}{|Primes|}\geq 1-o(1)$. Observe that for any fixed ℓ if $\mathcal{A}(Q_\ell)$ occurs then according to the definition of $\mathcal{D}_P$, for each $q\in Q_l$, α_q is selected uniformly and independently of any other q'. Hence, $prob_{\mathcal{D}_P}[h_\ell(\alpha)=1]\geq prob_{\mathcal{D}_P}[h_\ell(\alpha)=1\wedge\mathcal{A}(Q_\ell)]\geq prob_{\mathcal{D}_P}[h_\ell(\alpha)=1\mid\mathcal{A}(Q_\ell)]\cdot Prob_{\mathcal{D}_P}[\mathcal{A}(Q_\ell)]\geq\frac{1}{2^m}(1-o(1))$.

Observe that for each $\ell\in L_1$,

$$\begin{array}{c}\mathrm{Prob}_{\mathcal{D}_N}[far(\alpha)=1\bigwedge h_\ell(\alpha)=1]\geq\\ \mathrm{Prob}_{\mathcal{D}_N}[far(\alpha)=1\mid h_\ell(\alpha)=1]\cdot\mathrm{Prob}_{\mathcal{D}_N}[h_\ell(\alpha)=1]\end{array}. \tag{2}$$

By the definition of $\mathcal{D}_N$, $\mathrm{Prob}_{\mathcal{D}_N}[h_\ell(\alpha)=1]=\frac{1}{2^m}$. Also, using the definition just before Proposition 2 we see that $\mathrm{Prob}_{\mathcal{D}_N}[far(\alpha)=1\mid h_\ell(\alpha)=1]=prob_{U(Q_\ell)}[far(\alpha)=1]\geq 1-\frac{1}{n}$ (where the last inequality is by Proposition 2). □

6 Lower Bound for Testing *period*$(\leq g)$, where $g\leq\frac{\log n}{4}$

In this section we prove that every 2-sided error, $\frac{1}{16}$-test for *period*$(\leq g)$, where $g\leq\frac{\log n}{4}$, uses $\Omega((\log g)^{\frac{1}{4}})$ queries. We prove this for $\Sigma=\{0,1\}$. This implies the same bound for every alphabet that contains at least two symbols. This also shows that the test presented in Section 4 cannot be dramatically improved.

Theorem 5. *Any 2-sided error $\frac{1}{16}$-test for period$(\leq g)$, where $g\leq\frac{\log n}{4}$, requires $\Omega((\log g)^{\frac{1}{4}})$ queries.*

Proof. The proof is very similar to the proof of Theorem 4. We just describe here the two probabilities $\mathcal{D}_\mathcal{P}$ and $\mathcal{D}_\mathcal{N}$ that are concentrated on positive inputs (those that have *period*$(\leq g)$) and negative inputs (those that are $\frac{1}{16}$-*far* from being *period*$(\leq g)$) respectively.

Let $S=\{p_1,\ldots,p_k\}$ where $k=1+\sqrt{\log g}$ and each p_i is a prime such that $2^{\sqrt{\log g}-0.9}\leq p_i\leq 2^{\sqrt{\log g}}$. According to the *Prime Number Theorem* [11–13] there exists at least $\frac{1}{8\sqrt{\log g}}\cdot 2^{\sqrt{\log g}}>2+\sqrt{\log g}$ such primes, hence the set S is well defined. Let $t=\prod_{p\in S}p$ and let $Z=\{\frac{t}{p}|\ p\in S\}$. Note that $t>g$ while $z\leq g$ for every $z\in Z$.

We now define distributions $\mathcal{D}_\mathcal{P}$, $\mathcal{D}_\mathcal{N}$ (we assume w.l.o.g that t divides n).

- We use an auxiliary distribution $\mathcal{U}$ in order to define $\mathcal{D}_{\mathcal{N}}$. To select an instance according to $\mathcal{U}$, select a random string $\omega \in \Sigma^t$ and then set $\alpha = (\omega)^{\frac{n}{t}}$. Let G be the event that α is $\frac{1}{16}$-*far* from being *period*$(\leq g)$. Then $\mathcal{D}_{\mathcal{N}}$ is defined as the distribution $\mathcal{U}|\mathcal{G}$, namely $\mathcal{U}$ given $\mathcal{G}$.
- An instance α of length n is selected according to distribution $\mathcal{D}_{\mathcal{P}}$ as follows. Uniformly a select $z \in Z$, select uniformly a string $\omega \in \Sigma^z$ and then set $\alpha = (\omega)^{\frac{n}{z}}$.

We omit further details. □

7 Lower Bound for Testing Periodicity

Let $\Sigma = \{0, 1\}$, we prove the following lower bound on the number of queries that is needed for testing *period*$(\leq \frac{n}{2})$ over Σ. This clearly implies the same lower bound for any alphabet that contains at least two letter. It also shows that the test presented in Section 3 is asymptotically optimal.

Theorem 6. *Any non-adaptive 1-sided error $\frac{1}{16}$-test for periodicity requires $\Omega(\sqrt{n \log n})$ queries.*

Proof: A 1-sided error test rejects only when the input string is not periodic. Therefore according to corollary 1 such a test rejects only if the set of queries it uses contains a witness that the string in not periodic. To prove the Theorem we use Yao's principle (the easy direction). Namely, we construct a distribution on 1/16-*far* inputs and show that any *deterministic* non-adaptive test that queries at most $\frac{\sqrt{n \log n}}{100}$ queries finds a witness for non-periodicity with probability at most 1/3.

Let U be the uniform distribution over Σ^n, and let $\mathcal{G}$ be the event that a string selected according to U is $\frac{1}{16}$-*far* from being periodic. Let D be the distribution $U/\mathcal{G}$ (U given $\mathcal{G}$). Namely, D is uniform over strings in Σ^n that are 1/16-*far* from being periodic. Let $\mathcal{B}$ be the event that the set of queries Q contains a witness. Proposition 2 $Prob_U(\mathcal{G}) \geq 1 - \frac{1}{n}$. Thus, it is enough to prove that $prob_U(\mathcal{B}) < \frac{1}{4}$, since $prob_D(\mathcal{B}) \leq \text{Prob}_U(\mathcal{B}) + \text{Prob}_U(\overline{\mathcal{G}}) \leq \frac{1}{4} + \frac{1}{n}$. The proof follows from Lemma 4 below. □

Lemma 4. *Let $Q \subseteq [n]$ be a set of $\frac{\sqrt{n \log n}}{100}$ queries. Then for α chosen according to U, the probability that Q contains a witness that α is not periodic is at most $\frac{1}{4}$.*

The proof is omitted.

References

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

2. R. Rubinfeld and M. Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal of Computing*, 25:252–271, 1996.
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*, pages 16–28, 2004.
4. 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*, pages 363–372. Morgan Kaufmann, 2000.
5. R. Krauthgamer O. Sasson. Property testing of data dimensionality. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 18–27. Society for Industrial and Applied Mathematics, 2003.
6. 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*, pages 152–161, 2002.
7. Alex Samorodnitsky and Luca Trevisan. A PCP characterization of NP with optimal amortized query complexity. In *Proc. of the 32 ACM STOC*, pages 191–199, 2000.
8. Johan Hästad and Avi Wigderson. Simple analysis of graph tests for linearity and pcp. *Random Struct. Algorithms*, 22(2):139–160, 2003.
9. E. Fischer. The art of uninformed decisions: A primer to property testing. *The computational complexity column of The Bulletin of the European Association for Theoretical Computer Science*, 75:97–126, 2001.
10. D. Ron. Property testing (a tutorial). In *Handbook of Randomized computing*, pages 597–649. Kluwer Press, 2001.
11. J. Hadamard. Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques. *Bull. Soc. Math. France*, 24:199–220, 1896.
12. V. Poussin. Recherces analytiques sur la théorie des nombres premiers. *Ann. Soc. Sci. Bruxelles*, 1897.
13. D.J. Newman. Simple analytic proof of the prime number theorem. *Amer. Math. Monthly*, 87:693–696, 1980.