AppliedMath

Article

Continued Fractions and Probability Estimations in Shor's Algorithm: A Detailed and Self-Contained Treatise

Johanna Barzen and Frank Leymann *

Institute of Architecture of Application Systems, University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany; johanna.barzen@iaas.uni-stuttgart.de

* Correspondence: frank.leymann@iaas.uni-stuttgart.de

Abstract: Shor's algorithm for prime factorization is a hybrid algorithm consisting of a quantum part and a classical part. The main focus of the classical part is a continued fraction analysis. The presentation of this is often short, pointing to text books on number theory. In this contribution, we present the relevant results and proofs from the theory of continued fractions in detail (even in more detail than in text books), filling the gap to allow a complete comprehension of Shor's algorithm. Similarly, we provide a detailed computation of the estimation of the probability that convergents will provide the period required for determining a prime factor.

Keywords: quantum algorithms; quantum computing; continued fractions; hybrid quantum algorithms

check for updates

Citation: Barzen, J.; Leymann, F. Continued Fractions and Probability Estimations in Shor's Algorithm: A Detailed and Self-Contained Treatise. *AppliedMath* **2022**, *2*, 393–432. https://doi.org/10.3390/appliedmath2030023

Academic Editor: Darin J. Ulness

Received: 21 May 2022
Accepted: 12 July 2022
Published: 18 July 2022

1. Introduction

Shor's algorithm [1] for prime factorization is generally considered as a major milestone and a breakthrough in quantum computing: it solves a practically very relevant problem (which is, e.g., an underpinning of cryptography) with an exponential speedup compared to classical methods.

The algorithm is based on the fact that determining a divisor and finally a prime factor of a natural number $n \in \mathbb{N}$ can be reduced to finding the period p of the modular exponentiation function $f(x) = a^x mod n$ for an a with $0 < a < n$ (see Section 3.2.1).

The overall algorithm is hybrid, consisting of classical computations and a quantum computation. The classical computations are computing greatest common divisors with the Euclidian algorithm, and perform a continuous fraction analysis. A detailed discussion of the latter is one of the two foci of this contribution (see Section 2).

The quantum part mainly consists of: (i) creating an entangled state based on an oracle computing the modular exponentiation function f above, (ii) performing a quantum Fourier transform (QFT) on this state, and (iii) measuring it. The oracle produces the following state:

$$|a\rangle|b\rangle = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} |x\rangle|f(x)\rangle \tag{1}$$

After applying the quantum Fourier transform and a measurement, the first part (i.e., the $|a\rangle$-part) of the quantum register is in state

$$\frac{1}{\sqrt{NA}} \sum_{j=0}^{A-1} \omega_N^{jpy} |y\rangle \tag{2}$$

In this state, the searched period p already appears in its amplitude. The measured value y can then be used with high probability (see Section 3.4, Theorem 16) to compute the period p of the modular exponentiation function f by analyzing the convergents of a continued fraction (see Section 3.4.1) and finally, based on the period, a prime factor (see

Section 3.2.1). A detailed discussion on how this is achieved is the second focus of this contribution (see Section 3).

Structure of the Article

The article is structured as follows: in Section 2 we cover all details about continued fractions that are required to comprehend the corresponding aspect of Shor's algorithm.

Section 2.1 defines the notion of a continued fraction, gives examples of how to compute the continued fraction representation of a rational number, and demonstrates how to compute the number that a continued fraction (and thus convergents) represents.

Convergents as the fundamental tool in the theory of continued fractions are detailed in Section 2.2: after defining the term, basic theorems about convergents such as the recursion theorem, two sign theorems, monotony properties, convergent comparison, nesting of a number by its convergents, and several distance estimations are proven.

Next, the brief Section 2.3 presents infinite regular continued fractions to represent non-rational numbers. A corresponding algorithm is provided to compute such continued fractions.

Section 2.4 gives several upper bounds and lower bounds for the difference between a number and its convergents. Exploiting one of these bounds, the convergence of the convergents of an infinite regular continued fraction of a number to this number is proven. Semiconvergents are defined and corresponding monotony properties are given.

Best approximations of a real number are introduced in Section 2.5. It is proven that best approximations of the second kind are convergents and vice versa (Lagrange's theorem). Best approximations of the first kind are proven to be convergents or semiconvergents (another theorem by Lagrange). Finally, Legendre's theorem is presented, which is the main result about continued fractions required by Shor's algorithm: it allows the implication that a given fraction is a convergent of another number.

Section 3 is devoted to estimating the probability that convergents can be used to compute periods, i.e., that Legendre's theorem can be applied.

At the beginning of Section 3, Section 3.1 proves a lower bound and an upper bound for the secant lengths of the unit circle. This estimation is central for estimating the aforementioned probability.

Section 3.2 contains many different estimations of parameters that appear in the measurement result of Shor's algorithm. In Section 3.2.1, we recall the very basics of modular arithmetic, relate this to group theory, and use Lagrange's theorem from group theory to prove that the period of the modular exponentiation function in Shor's algorithm is less than the number to be factorized (Lemma 8). Intervals of consecutive multiples of the period are studied in Section 3.2.2: it is shown that multiples of N are sparsely scattered across these intervals (Note 12). This implies that measurement results are somehow centered around multiples of N/p (Corollary 9). The cardinality of arguments in the superposition that build the pre-image of a certain $f(x)$ is estimated in Section 3.2.3. Section 3.2.4 proves bounds of phases of amplitudes relevant for computing the probability of measurement results as a geometric sum.

Finally, Section 3.3 computes this probability: it is proven that a measurement result is close to a multiple of N/p with probability of approximately $4/\pi^2$ (Lemma 10).

Section 3.4 shows that this measurement result fulfills the assumption of Legendre's theorem (Theorem 15). Thus, by computing convergents, the period can be determined (Theorem 16 and Section 3.4.1).

Section 3.5 sketches how the main results contribute to the proof of Shor's algorithm. Its purpose is to avoid getting lost in the huge amount of low-level details.

A brief conclusion and discussion of related work ends this contribution with Section 4.

2. Continued Fractions

2.1. Definition of Continued Fractions and Their Computation

We define the notion of continued fractions and give an example of how to compute them.

Definition 1. An expression of the form

$$a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cfrac{b_3}{\ddots}}} \tag{3}$$

with $a_i, b_i \in \mathbb{C}$ is called an *infinite* continued fraction.

If, in this expression, it is $b_i = 1$ for all i, $a_0 \in \mathbb{Z}$, and $a_i \in \mathbb{N}$ for i≥1, the expression is called a *regular* continued fraction.

A *finite* regular continued fraction (simply called a *continued fraction*) satisfies, in addition, the condition $\exists N \in \mathbb{N} \forall k \in \mathbb{N} : a_{N+k} = 0$ (convention: "1/0 = 0").

A continued fraction is, thus, the following expression:

$$[a_0; a_1, \cdots, a_N] \overset{def}{=} a_0+\cfrac{1}{a_1+\cfrac{1}{\ddots+\cfrac{1}{a_{N-1}+\frac{1}{a_N}}}} \tag{4}$$

A continued fraction of a rational number a/b is computed as follows: the integer part $\lfloor a/b \rfloor$ becomes $a_0 \in \mathbb{Z}$, leaving the non-negative rational remainder $x_1/y_1 \in \mathbb{Q}$. The latter is now written as $1/(y_1/x_1)$, resulting in

$$a_0+\frac{1}{\left(\frac{y_1}{x_1}\right)}$$

Next, the integer part $\lfloor y_1/x_1 \rfloor$ becomes a_1, leaving a rational remainder that is treated as before. This processing stops until the rational remainder is zero. Figure 1 gives an example of the processing.

$$\left.\begin{aligned}\frac{67}{47}&=1+\frac{20}{47}=1+\cfrac{1}{\frac{47}{20}}=1+\cfrac{1}{2+\frac{7}{20}}=1+\cfrac{1}{2+\cfrac{1}{\frac{20}{7}}}\\&=1+\cfrac{1}{2+\cfrac{1}{2+\frac{6}{7}}}=1+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{\frac{7}{6}}}}=1+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{1+\frac{1}{6}}}}\end{aligned}\right\} \Rightarrow \frac{67}{47}=[1;2,2,1,6]$$

Figure 1. Example of a straightforward computation of a continued fraction.

Beside this straightforward proceeding to compute continued fractions, the well-known Euclidian algorithm can be used for this purpose too. Figure 2 gives a corresponding example; it should be self-descriptive.

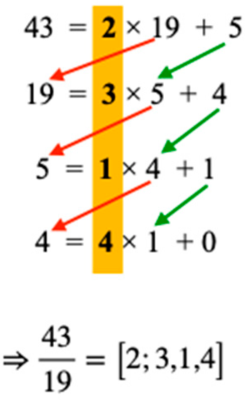


Figure 2. Using the Euclidian algorithm to compute a continued fraction.

Formally, a continued fraction can always be reduced such that its last element is greater than or equal to 2.

Note 1. Let $[a_0; a_1, \ldots, a_N]$ be a continued fraction. Then:

$$[a_0; a_1, \ldots, a_N] = \left[a_0; a_1, \ldots, a_{N-1} + \frac{1}{a_N}\right] \tag{5}$$

Especially, it can always be achieved that a continued fraction $[a_0; a_1, \ldots, a_N]$ satisfies $a_N \geq 2$.

Proof. The following simple computation proves the first claim:

$$\begin{aligned}
[a_0; a_1, \ldots, a_N] &= a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{N-1} + \frac{1}{a_N}}}} \\
&= a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{\left(a_{N-1} + \frac{1}{a_N}\right)}}} \\
&= \left[a_0; a_1, \ldots, a_{N-1} + \frac{1}{a_N}\right]
\end{aligned}$$

Furthermore, if $a_N = 1$ in $[a_0; a_1, \ldots, a_N]$ then $a_{N-1} + 1/a_N \geq 2$. This is because, by definition, $a_k \geq 1$ for $1 \leq \mathrm{k} \leq \mathrm{N}$. □

Equation (5) implies a straightforward way to compute the value represented by a continued fraction $[a_0; a_1, \ldots, a_N]$: see Figure 3.

$$\begin{aligned}
[2;3,1,4] &= [2;3,1+1/4] = [2;3,5/4] = [2;3+1/(5/4)] \\
&= [2;3+4/5] = [2;19/5] = [2+1/(19/5)] = [2+5/19] \\
&= [43/19] = \frac{43}{19}
\end{aligned}$$

Figure 3. Computing the value of a continued fraction based on Equation (5).

2.2. Convergents

Next, we define the "workhorses" of the theory of continued fractions.

Definition 2. $[a_0; a_1, \ldots, a_m]$ is called m-th convergent of the continued fraction $[a_0; a_1, \ldots, a_N]$ for $0 \leq \mathrm{m} \leq \mathrm{N}$, or m-th convergent of the infinite regular continued fraction $[a_0; a_1, \ldots]$.

Convergents can be computed recursively based on the following theorem:

Theorem 1. (*Recursion Theorem*)

Define:

- $p_0 = a_0;$
- $p_1 = a_1 a_0 + 1;$
- $p_n = a_n p_{n-1} + p_{n-2}$ for n $\geq$ 2;

and define:

- $q_0 = 1;$
- $q_1 = a_1;$
- $q_n = a_n q_{n-1} + q_{n-2}$ for n $\geq$ 2.

Then, for every convergent $[a_0; a_1, \ldots, a_n]$*, it is:*

$$[a_0; a_1, \ldots, a_n] = \frac{p_n}{q_n} \tag{6}$$

Proof (by induction). Let n = 0, 1: Then, $[\mathrm{a}_0] = \mathrm{a}_0 = \frac{\mathrm{p}_0}{\mathrm{q}_0}$ and $[\mathrm{a}_0; \mathrm{a}_1] = \mathrm{a}_0 + \frac{1}{\mathrm{a}_1} = \frac{\mathrm{a}_0\mathrm{a}_1+1}{\mathrm{a}_1} = \frac{\mathrm{p}_1}{\mathrm{q}_1}$.

Induction hypothesis: $[a_0; a_1, \ldots, a_n] = \frac{p_n}{q_n} = \frac{a_n p_{n-1}+p_{n-2}}{a_n q_{n-1}+q_{n-2}}$.

Induction step n $\rightarrow$ n + 1: According to Note 1, it is

$$[a_0; a_1, \ldots, a_n, a_{n+1}] = \left[a_0; a_1, \ldots, a_n + \frac{1}{a_{n+1}}\right]$$

and the last continued fraction has n elements, i.e., the induction hypothesis applies:

$$\begin{aligned} a_0; a_1, \ldots, a_n + \tfrac{1}{a_{n+1}}] &= \frac{\left(a_n + \frac{1}{a_{n+1}}\right) p_{n-1} + p_{n-2}}{\left(a_n + \frac{1}{a_{n+1}}\right) q_{n-1} + q_{n-2}} = \frac{\frac{a_n a_{n+1}+1}{a_{n+1}} p_{n-1} + p_{n-2}}{\frac{a_n a_{n+1}+1}{a_{n+1}} q_{n-1} + q_{n-2}} \\ &= \frac{(a_n a_{n+1}+1) p_{n-1} + a_{n+1} p_{n-2}}{(a_n a_{n+1}+1) q_{n-1} + a_{n+1} q_{n-2}} \\ &= \frac{a_{n+1}(a_n p_{n-1} + p_{n-2}) + p_{n-1}}{a_{n+1}(a_n q_{n-1} + q_{n-2}) + q_{n-1}} \overset{(A)}{=} \frac{a_{n+1} p_n + p_{n-1}}{a_{n+1} q_n + q_{n-1}} \\ &\overset{(B)}{=} \frac{p_{n+1}}{q_{n+1}} \end{aligned}$$

Here, *(A)* is valid because of the induction hypothesis, and (*B*) is the definition of p_{n+1} and q_{n+1}. □

The recursion theorem implies the often used.

Corollary 1. Numerators and denominators of convergents of a continued fraction $[a_0; a_1, \ldots, a_N]$ with $a_0 \geq 0$ are strictly monotonically increasing:

$$p_n > p_{n-1} \text{ and } q_n > q_{n-1} \text{ for all } n \in \mathbb{N}.$$

Proof (by induction). Let n = 1: By definition, $p_0 = a_0$, $p_1 = a_1 a_0 + 1$. Because $a_i \geq 1$ for i $\geq$ 1, and $a_0 \geq 0$, it is $p_1 > p_0 \geq 0$. Similarly, $q_1 > q_0 > 0$

Now, $p_n = a_n p_{n-1} + p_{n-2}$ and $q_n = a_n q_{n-1} + q_{n-2}$ for n $\geq$ 2. With $a_n \geq 1$ by definition, and $p_{n-1} > p_{n-2}$ ($\geq$1) as well as $q_{n-1} > q_{n-2}$ ($\geq$1) by induction hypothesis, the claim follows. □

The next theorem is about the sign of a combination of the numerators and denominators of consecutive convergents of a continued fraction.

Theorem 2. *(Sign Theorem)*
For $[a_0; a_1, \ldots, a_n] = \frac{p_n}{q_n}$*, the following holds:*

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1} \tag{7}$$

Proof (by induction). For n = 1, it is

$$p_1 q_0 - p_0 q_1 = (a_1 a_0 + 1) \cdot 1 - a_0 \cdot a_1 = 1 = (-1)^0$$

Induction step n → n + 1:

$$\begin{aligned} p_{n+1} q_n - p_n q_{n+1} &= (a_{n+1} p_n + p_{n-1}) q_n - p_n (a_{n+1} q_n + q_{n-1}) \\ &= a_{n+1} p_n q_n + p_{n-1} q_n - p_n a_{n+1} q_n - p_n q_{n-1} \\ &= p_{n-1} q_n - p_n q_{n-1} = -(p_n q_{n-1} - p_{n-1} q_n) \\ &\overset{(A)}{=} -(-1)^{n-1} = (-1)^n \end{aligned}$$

(A) uses the induction hypothesis. □

In case the numerators and denominators stem from the n-th convergent and the (n − 2)-nd convergent, the last n-th element of the convergent becomes part of the equation.

Theorem 3. *(Second Sign Theorem)*
For $[a_0; a_1, \ldots, a_n] = \frac{p_n}{q_n}$*, the following holds:*

$$p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n \tag{8}$$

Proof. It is $\mathrm{p_n = a_n p_{n-1} + p_{n-2}}$ and $\mathrm{q_n = a_n q_{n-1} + q_{n-2}}$.

Multiplying the first equation by q_{n-2} and the second equation by p_{n-2} results in $q_{n-2} p_n = q_{n-2} a_n p_{n-1} + q_{n-2} p_{n-2}$ and $p_{n-2} q_n = p_{n-2} a_n q_{n-1} + p_{n-2} q_{n-2}$. Next, both equations are subtracted:

$$\begin{aligned} p_n q_{n-2} - p_{n-2} q_n &= q_{n-2} a_n p_{n-1} + q_{n-2} p_{n-2} - p_{n-2} a_n q_{n-1} - p_{n-2} q_{n-2} \\ &= q_{n-2} a_n p_{n-1} - p_{n-2} a_n q_{n-1} \\ &= a_n (p_{n-1} q_{n-2} - p_{n-2} q_{n-1}) \\ &\overset{(A)}{=} (-1)^n a_n \end{aligned}$$

where (A) is implied by the sign theorem (Theorem 2) and considering $(-1)^{n-2} = (-1)^n$.
□

The sign theorem immediately yields the important.

Corollary 2. Numerator and denominator of a convergent are co-prime.

Proof. Let t be a divisor of $\mathrm{p_n}$ and $\mathrm{q_n}$, i.e., $t|p_n$ and $t|q_n$. Then, $t|(p_n q_{n-1} - p_{n-1} q_n)$, but $(p_n q_{n-1} - p_{n-1} q_n) = (-1)^{n-1}$ according to the sign theorem. Thus, t = ±1. □

Convergents can be represented as a sum of fractions with alternating sign and whose denominators consist of products of two consecutive denominators from the recursion theorem.

Theorem 4. *(Representation as a Sum)*

Each convergent can be represented as a sum:

$$[a_0; a_1, \ldots, a_n] = a_0 + \frac{1}{q_1 q_0} - \frac{1}{q_2 q_1} + \cdots + (-1)^{n-1} \frac{1}{q_n q_{n-1}} \tag{9}$$

Proof. Let $[\mathrm{a_0; a_1, \ldots, a_n}] = \frac{\mathrm{p_n}}{\mathrm{q_n}}$. Since $-\frac{p_i}{q_i} + \frac{p_i}{q_i} = 0$, we can write

$$[a_0; a_1, \ldots, a_n] = \frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} + \frac{p_{n-1}}{q_{n-1}} - \frac{p_{n-2}}{q_{n-2}} + \frac{p_{n-2}}{q_{n-2}} - \cdots + \frac{p_1}{q_1} - \frac{p_0}{q_0} + \frac{p_0}{q_0}$$

Computing the differences results in

$$\begin{aligned} [a_0; a_1, \ldots, a_n] &= \tfrac{p_n q_{n-1} - q_n p_{n-1}}{q_n q_{n-1}} + \tfrac{p_{n-1} q_{n-2} - q_{n-1} p_{n-2}}{q_{n-1} q_{n-2}} + \cdots + \tfrac{p_1 q_0 - q_1 p_0}{q_1 q_0} + \tfrac{p_0}{q_0} \\ &\overset{(A)}{=} \tfrac{(-1)^{n-1}}{q_n q_{n-1}} + \tfrac{(-1)^{n-2}}{q_{n-1} q_{n-2}} + \cdots + \tfrac{(-1)^0}{q_1 q_0} + a_0 \end{aligned}$$

where the sign theorem is applied in (A) and the last term $a_0 = p_0/q_0$ is the recursion theorem. □

The next theorem is key for many estimations in the domain of continued fractions.

Theorem 5. *(Monotony Theorem)*

Let $x_n \overset{def}{=} \frac{p_n}{q_n} = [a_0; a_1, \ldots, a_n]$ *denote the n-th convergent. Then:*

$$x_{2n} < x_{2n+2}$$

and

$$x_{2n+1} > x_{2n+3}$$

I.e., even convergents are strictly monotonically increasing, and odd convergents are strictly monotonically decreasing.

Proof. We compute the following difference, where (A) again uses $-\frac{p_i}{q_i} + \frac{p_i}{q_i} = 0$:

$$\begin{aligned} x_n - x_{n-2} &= \tfrac{p_n}{q_n} - \tfrac{p_{n-2}}{q_{n-2}} \overset{(A)}{=} \tfrac{p_n}{q_n} - \tfrac{p_{n-1}}{q_{n-1}} + \tfrac{p_{n-1}}{q_{n-1}} - \tfrac{p_{n-2}}{q_{n-2}} \\ &= \tfrac{p_n q_{n-1} - q_n p_{n-1}}{q_n q_{n-1}} + \tfrac{p_{n-1} q_{n-2} - q_{n-1} p_{n-2}}{q_{n-1} q_{n-1}} \\ &\overset{(B)}{=} \tfrac{(-1)^{n-1}}{q_n q_{n-1}} + \tfrac{(-1)^{n-2}}{q_{n-1} q_{n-2}} = \tfrac{(-1)^{n-1} q_{n-2} + (-1)^{n-2} q_n}{q_n q_{n-1} q_{n-2}} \\ &= \tfrac{(-1)^{n-2} q_n - (-1)^{n-2} q_{n-2}}{q_n q_{n-1} q_{n-2}} = \tfrac{(-1)^{n-2} (q_n - q_{n-2})}{q_n q_{n-1} q_{n-2}} \\ &= \tfrac{(-1)^n (q_n - q_{n-2})}{q_n q_{n-1} q_{n-2}} \overset{(C)}{=} \tfrac{(-1)^n a_n q_{n-1}}{q_n q_{n-1} q_{n-2}} = \tfrac{(-1)^n a_n}{q_n q_{n-2}} \end{aligned}$$

(B) is because of the sign theorem, and (C) follows from $q_n = a_n q_{n-1} + q_{n-2}$, i.e., the recursion theorem.

Now, because of $a_n, q_n, q_{n-2} > 0$, it is $\frac{a_n}{q_n q_{n-2}} > 0$. Thus, $\frac{(-1)^n a_n}{q_n q_{n-2}} > 0$ for n even and $\frac{(-1)^n a_n}{q_n q_{n-2}} < 0$ for n odd. This implies $x_n = \frac{(-1)^n a_n}{q_n q_{n-2}} + x_{n-2} > x_{n-2}$ for n even as well as $x_n = \frac{(-1)^n a_n}{q_n q_{n-2}} + x_{n-2} < x_{n-2}$ for n odd. □

While even convergents are increasing and odd convergence are decreasing, all even convergents are smaller than all odd convergents. This is the content of the next very important theorem.

Theorem 6. *(Convergents Comparison Theorem)*
For $0 \leq 2n, 2m+1 \leq N$, it is $x_{2n} < x_{2m+1}$

Proof. As before, using the sign theorem in *(A)*, we obtain

$$x_n - x_{n-1} = \frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{p_n q_{n-1} - q_n p_{n-1}}{q_n q_{n-1}} \overset{(A)}{=} \frac{(-1)^{n-1}}{q_n q_{n-1}} = \frac{(-1)^{n-1}}{\beta_n}$$

with $\beta_n := q_n q_{n-1}$. Because $q_n, q_{n-1} > 0$, it is $\beta_n > 0$, i.e., the sign of $\frac{(-1)^{n-1}}{\beta_n}$ is in fact $(-1)^{n-1}$.

Thus, $x_{2n+1} - x_{2n} = \frac{(-1)^{2n}}{\beta_{2n+1}} > 0$, and we get $x_{2n+1} = \frac{(-1)^{2n}}{\beta_{2n+1}} + x_{2n} > x_{2n}$. This shows that an even convergent x_{2n} is strictly smaller than its immediate succeeding odd convergent x_{2n+1}.

But what about an arbitrary odd convergent x_{2m+1}? For n < m, the monotony theorem (Theorem 6) yields $x_{2n} < x_{2m}$ and we showed before that $x_{2m} < x_{2m+1}$; thus, $x_{2n} < x_{2m+1}$.

For n > m, the monotony theorem yields $x_{2m+1} > x_{2n+1}$ and with $x_{2n+1} > x_{2n}$ we see $x_{2n} < x_{2m+1}$. □

The following often-used corollary computes the difference of two immediately succeeding convergents by mean of the denominators of the convergents, while the difference of the n-th convergent and the (n − 2)-nd convergent adds the n-th element of the n-th convergent as a factor.

Corollary 3.

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n-1}}{q_n q_{n-1}} \tag{10}$$

and

$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n a_n}{q_n q_{n-2}} \tag{11}$$

Proof. Equation (10) is the first equation from the proof of Theorem 6. The second equation follows because of

$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{p_n q_{n-2} - p_{n-2} q_n}{q_n q_{n-2}} \overset{(A)}{=} \frac{(-1)^n a_n}{q_n q_{n-2}}$$

where *(A)* is because of the second sign theorem (Theorem 3). □

We already saw that the even convergents are strictly monotonically increasing, that the odd convergents are strictly monotonically decreasing, and that each even convergent is less than all odd convergents. According to the next theorem, the value of a continued fraction lies between the even convergents and the odd convergents, i.e., this value is larger than all even convergents and smaller than all odd convergents. The situation is depicted in Figure 4.

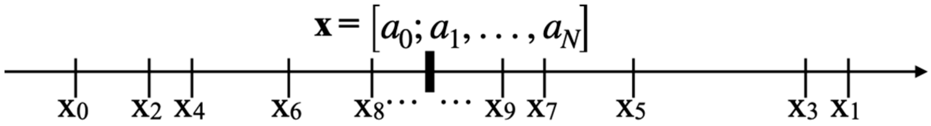


Figure 4. Nesting of the value of a continued fraction by its convergents.

Note that the notion of the value of a continued fraction is defined for finite continued fractions. In Section 4, this notion will also be defined for regular infinite continued fractions.

Theorem 7. *(Nesting Theorem)*

Let x be the value of the continued fraction $[a_0; a_1, \ldots, a_N]$ and let x_k be its convergents. Then:

$$\forall m, n < N : x_{2m} < x < x_{2n+1} \tag{12}$$

Proof. The value of x is the convergent with the highest index N, i.e., $x = x_N = [a_0; a_1, \ldots, a_N]$.

Let N = 2k be even. Since even convergents are strictly monotonically increasing, we know that $\forall 2m < N : x_{2m} < x_{2k} = x_N = x$, and according to the convergent comparison theorem (Theorem 6), we know $\forall 2n+1 : x = x_N = x_{2k} < x_{2n+1}$.

Let N = 2k + 1 be odd. Since odd convergents are strictly monotonically decreasing, we know that $\forall 2n+1 < N : x_{2n+1} > x_{2k+1} = x_N = x$, and according to the convergent comparison theorem (Theorem 6), we know $\forall 2m : x = x_N = x_{2k+1} > x_{2m}$. □

Because the value of a continued fraction is nested within its even convergents and odd convergents, the distance of this value from any of its convergents can be estimated by the distance of two consecutive convergents:

Theorem 8. *(Distance Theorem)*

Let $x = [a_0; a_1, \ldots, a_N]$ and let x_k be its convergents. Then:

$$\forall n : |x - x_n| < |x_{n-1} - x_n| \tag{13}$$

and

$$\forall n : |x - x_n| < |x_{n+1} - x_n| \tag{14}$$

Proof. Let n be even. Then, $x_n < x < x_{n-1}$, i.e., $x - x_n < x_{n-1} - x_n$. Additionally, it is $x - x_n > 0$ and $x_{n-1} - x_n > 0$. Thus, $|x - x_n| < |x_{n-1} - x_n|$ for n even.

Now, let n be odd. It is $x_{n-1} < x < x_n$, which implies $x - x_n > x_{n-1} - x_n \Leftrightarrow -(x_n - x) > -(x_n - x_{n-1}) \Leftrightarrow x_n - x < x_n - x_{n-1}$. Because of $x_n - x > 0$ and $x_n - x_{n-1} > 0$, it is $|x_n - x| < |x_n - x_{n-1}| \Leftrightarrow |x - x_n| < |x_{n-1} - x_n|$ for n odd.

Together, this proves Equation (13). Equation (14) is proven similarly. □

Figure 5 shows the corresponding geometric situation for an even n.

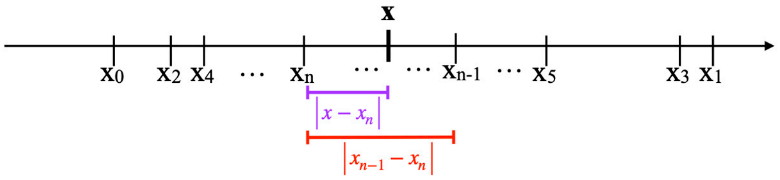


Figure 5. The distance between two succeeding convergents is greater than the distance of a convergent and the value of its continued fraction.

Similarly, the difference between any two arbitrary convergents can be estimated by the difference of the convergent with the smaller index and its immediate predecessor:

Theorem 9. *(Difference Theorem)*

Let $x = [a_0; a_1, \ldots, a_N]$ and let x_k be its convergents. Then:

$$\forall m > n : |x_m - x_n| < |x_{n-1} - x_n| \tag{15}$$

Proof. Let n be even, e.g., n = 2k.

Let m = 2t be even. By Theorem 6, even convergents are smaller than all odd convergents, i.e., $x_{2t} < x_{2k-1}$ for any $t \in \mathbb{N}$. Thus, $x_m - x_n = x_{2t} - x_{2k} << x_{2k-1} - x_{2k} = x_{n-1} - x_n$.

Let m = 2t − 1 be odd. By the monotony theorem (Theorem 5), odd convergents are strictly monotonically decreasing, i.e., $x_{2t-1} < x_{2k-1}$ for each t > k. Thus, $x_m - x_n = x_{2t-1} - x_{2k} < x_{2k-1} - x_{2k} = x_{n-1} - x_n$.

For n odd, the proof is analogous. □

The geometry of the last theorem is depicted in Figure 6.

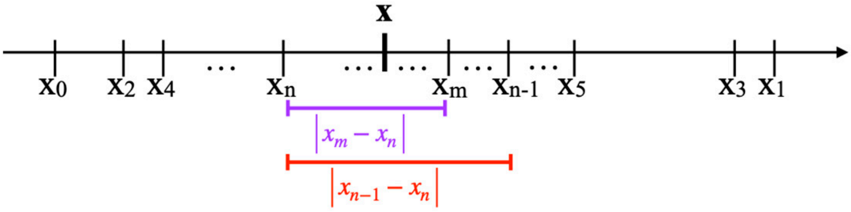


Figure 6. The distance between any two convergents is smaller than the distance between the convergent with the smaller index and its immediate predecessor.

In several calculations, the size of the denominator of a convergent must be estimated:

Lemma 1. *(Size of Denominators)*

For the denominator q_n of a convergent $\frac{p_n}{q_n} = [a_0; a_1, \dots, a_n]$, the following holds:

$$\forall n : q_n \geq n \tag{16}$$

and

$$\forall n > 3 : q_n > n \tag{17}$$

Proof. By definition, $q_0 = 1 > 0$, and $q_1 = a_1 \geq 1$ because $a_i \in \mathbb{N}$, and finally,

$$q_2 \overset{(A)}{=} a_2 q_1 + q_0 \overset{(B)}{=} a_2 q_1 + 1 \overset{(C)}{\geq} q_1 + 1 \overset{(D)}{\geq} 2$$

(A) holds because of the recursion theorem (Theorem 1), *(B)* is by definition of q_0, *(C)* is because $a_2 \in \mathbb{N}$, and *(D)* has been seen just before (i.e., $q_1 \geq 1$). This proves the lemma for $n \leq 2$.

The proof for n ≥ 3 is by induction. It is

$$q_n \overset{(A)}{=} a_n q_{n-1} + q_{n-2} \overset{(B)}{\geq} q_{n-1} + q_{n-2} \overset{(C)}{\geq} q_{n-1} + (n-2) \overset{(D)}{\geq} q_{n-1} + 1 \overset{(E)}{\geq} n$$

where *(A)* is the recursion theorem, *(B)* is because of $a_n \in \mathbb{N}$, *(C)* is by induction hypothesis applied to q_{n-2}, *(D)* is because n ≥ 3, and *(E)* is by induction hypothesis applied to q_{n-1}. This proves Equation (16).

Equation (17) is proven by induction again. Let n > 3. The argumentation is exactly as before, with the exception of *(D)*:

$$q_n \overset{(A)}{=} a_n q_{n-1} + q_{n-2} \overset{(B)}{\geq} q_{n-1} + q_{n-2} \overset{(C)}{\geq} q_{n-1} + (n-2) \overset{(D)}{>} q_{n-1} + 1 \overset{(E)}{\geq} n$$

(D) holds because n > 3, i.e., $n - 2 > 1$. □

In fact, denominators of a convergent grow much faster than the inequation $q_n > n$ may indicate:

Lemma 2. *(Geometric Growth of Denominators)*
Let q_n $(n \geq 2)$ be the denominator of the convergent $\frac{p_n}{q_n} = [a_0; a_1, \ldots, a_n]$. Then:

$$q_n \geq 2^{\frac{n-1}{2}} \tag{18}$$

Proof. It is $\mathrm{q_k} = \mathrm{a_k q_{k-1}} + \mathrm{q_{k-2}} > \mathrm{q_{k-1}} + \mathrm{q_{k-2}} \overset{(A)}{>} 2\mathrm{q_{k-2}}$, with *(A)* because, according to corollary 1, denominators are strictly monotonically increasing, i.e., $q_{k-1} > q_{k-2}$.

By induction, it is $q_{2k} \geq 2^k q_0$, and then $2^k q_0 \overset{(A)}{=} 2^k \overset{(B)}{\geq} 2^{\frac{(2k)-1}{2}}$ with *(A)* because $q_0 = 1$, and *(B)* follows from

$$2^k = 2^{\frac{2k}{2}} \geq \frac{1}{\sqrt{2}} 2^{\frac{2k}{2}} = 2^{\frac{2k}{2} - \frac{1}{2}} = 2^{\frac{2k-1}{2}}$$

Similarly, by induction, it is $q_{2k+1} \geq 2^k q_1$ and then $2^k q_1 \overset{(A)}{\geq} 2^k = 2^{\frac{(2k+1)-1}{2}}$ with (A) because of $q_1 \in \mathbb{N}$.

With $n = 2k$ and $n = 2k + 1$, respectively, Equation (18) is implied. □

2.3. Convergence of Infinite Regular Continuous Fractions

In Section 2.1, we presented an algorithm to compute the continued fraction representation of a rational number. Next, we show how to compute such a representation for a non-rational number (Algorithm 1).

Algorithm 1 Continued Fraction Representation of Non-Rational Number

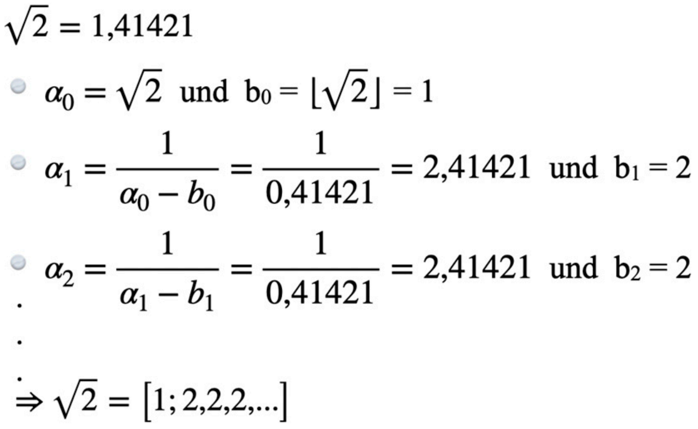
1. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Define:
   - $\alpha_0 := \alpha$ and $b_0 := \lfloor \alpha_0 \rfloor$;
   - $\alpha_i := \frac{1}{\alpha_{i-1} - b_{i-1}}$ and $b_i := \lfloor \alpha_i \rfloor$ for i ≥ 1.
2. Then, $[b_0; b_1, b_2, \ldots]$ is the continued fraction representation of α. Each α_i is called the *i-th complete quotient* of α.

The above algorithm does not terminate, i.e., the continued fraction representation of a non-rational number is infinite. This is the content of the following note:

Note 2.
In Algorithm 1, it is $\alpha_i \notin \mathbb{Z}$.

Proof (by induction).
n = 0: Then, by definition, $\alpha_0 = \alpha \notin \mathbb{Z}$.
Induction hypothesis: $\alpha_n \notin \mathbb{Z}$.
n → n + 1: Assume $\alpha_n - b_n \in \mathbb{Z} \Rightarrow (\alpha_n - b_n) = k \in \mathbb{Z} \Rightarrow \alpha_n = k + b_n \in \mathbb{Z}$, which is a contradiction to the hypothesis! Thus, $\alpha_n - b_n \notin \mathbb{Z} \Rightarrow \alpha_{n+1} := \frac{1}{\alpha_n - b_n} \notin \mathbb{Z}$. □

Figure 7 gives the computation of the continued fraction representation of $\sqrt{2}$:



$\sqrt{2} = 1{,}41421$

- $\alpha_0 = \sqrt{2}$ und $b_0 = \lfloor\sqrt{2}\rfloor = 1$
- $\alpha_1 = \dfrac{1}{\alpha_0 - b_0} = \dfrac{1}{0{,}41421} = 2{,}41421$ und $b_1 = 2$
- $\alpha_2 = \dfrac{1}{\alpha_1 - b_1} = \dfrac{1}{0{,}41421} = 2{,}41421$ und $b_2 = 2$

.
.
.

$\Rightarrow \sqrt{2} = [1;2{,}2{,}2{,}\ldots]$

Figure 7. Computing the continued fraction of $\sqrt{2}$.

2.4. Bounds Expressed by Denominators of Convergents

In the following, we give upper bounds and lower bounds of the approximations of a number by the convergents of its continued fraction representation by means of the denominators of the convergents.

First, we start with estimations of upper bounds:

Lemma 3. *(Upper Bounds)*

Let p_n/q_n be a convergent of the continued fraction representation of x. Then:

$$\left|x - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}} < \frac{1}{q_n^2} \leq \frac{1}{n^2} \tag{19}$$

Proof. With $x_n = p_n/q_n$, it is $|x - x_n| < |x_{n+1} - x_n|$ (see Theorem 8, Equation (14)). According to Corollary 3 (Equation (10)), it is

$$x_{n+1} - x_n = \frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} = \frac{(-1)^n}{q_n q_{n+1}}$$

Thus,

$$|x - x_n| < |x_{n+1} - x_n| = \left|\frac{(-1)^n}{q_n q_{n+1}}\right| = \frac{1}{q_n q_{n+1}} \overset{(A)}{<} \frac{1}{q_n^2} \overset{(B)}{\leq} \frac{1}{n^2}$$

where (A) holds because of $q_{n+1} > q_n$ (Corollary 1), and (B) is true because of $q_n \geq n$ (Lemma 1). □

An immediate consequence of this theorem is the convergence of the sequence of the convergents of a continued fraction to the value of the continued fraction. This, by the way, is the origin of the name "convergents".

Corollary 4. The series (p_n/q_n) of the convergents of the continued fraction representation of $x \in \mathbb{R} \setminus \mathbb{Q}$ converges to x:

$$lim \frac{p_n}{q_n} = x$$

Proof. The claim follows immediately from $\left|x - \frac{p_n}{q_n}\right| < \frac{1}{n^2}$. □

Often, two fractions are compared by means of their mediant ("mediant" means "somewhere in between").

Definition 3. For $\mathrm{a/b, c/d} \in \mathbb{Q}$ and b, d > 0, the term $\frac{\mathrm{a+c}}{\mathrm{b+d}}$ is called the *mediant* of the two fractions.

The following simple inequation is often used.

Note 3. *(Mediant Property)*

Let $a/b, c/d \in \mathbb{Q}$ and b, d > 0 and $\frac{a}{b} < \frac{c}{d}$.

Then:

$$\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d} \tag{20}$$

Proof. It is $\frac{a}{b} < \frac{c}{d} \Rightarrow ad < bc \Rightarrow bc - ad > 0$ and $b, d > 0 \Rightarrow b(b+d) > 0$. This implies $\frac{a+c}{b+d} - \frac{a}{b} = \frac{b(a+c)-a(b+d)}{b(b+d)} = \frac{bc-ad}{b(b+d)} > 0$ and thus $\frac{a}{b} < \frac{a+c}{b+d}$. The inequation $\frac{a+c}{b+d} < \frac{c}{d}$ follows similarly. □

Mediants of convergents that are weighted in a certain way are another important concept for computing bounds:

Definition 4. The term $x_{n,t} = \frac{tp_{n+1}+p_n}{tq_{n+1}+q_n}$ with $1 \leq t \leq a_{n+2}$ is called the (n,t)-th semiconvergent.

Semiconvergents of an even n are strictly monotonically increasing, and semiconvergents of an odd n are strictly monotonically decreasing. This is the content of the following lemma.

Lemma 4. *(Monotony of Semiconvergents)*

Let n be even. Then, $x_{n,t} < x_{n,t+1}$.

Let n be odd. Then, $x_{n,t} > x_{n,t+1}$.

Proof. A simple calculation and the use of the sign theorem (Theorem 2) results in

$$\begin{aligned} x_{n,t+1} - x_{n,t} =& \frac{(t+1)p_{n+1}+p_n}{(t+1)q_{n+1}+q_n} - \frac{tp_{n+1}+p_n}{tq_{n+1}+q_n} \\ =& \frac{(-1)^n}{((t+1)q_{n+1}+q_n)(tq_{n+1}+q_n)} \end{aligned}$$

The denominator of the last fraction is always positive. Thus, the last term is positive iff n is even (i.e., $x_{n,t+1} - x_{n,t} > 0$), and it is negative iff n is odd (i.e., $x_{n,t+1} - x_{n,t} < 0$). □

In order to simplify proofs in what follows, the following conventions are used:

$$p_{-1} \overset{def}{=} 1 \text{ and } q_{-1} \overset{def}{=} 0 \tag{21}$$

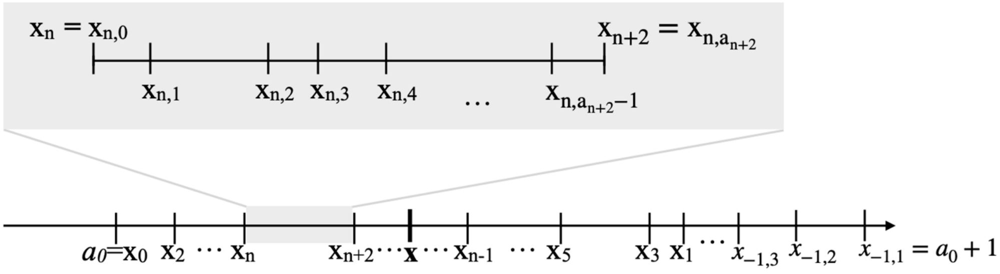
With this, $x_{-1,1} = \frac{p_0+p_{-1}}{q_0+q_{-1}} = \frac{a_0+1}{1+0} = a_0 + 1$ becomes a semiconvergent. Now, $x_1 = \frac{p_1}{q_1} \overset{(A)}{=} \frac{a_1a_0+1}{a_1} = a_0 + \frac{1}{a_1} \overset{(B)}{\leq} a_0 + 1 = x_{-1,1}$ where (A) is the recursion theorem and (B) follows because $a_1 \geq 1$; thus, $x_1 \leq x_{-1,1}$.

Furthermore, it is $x_{-1,t} = \frac{tp_0+p_{-1}}{tq_0+q_{-1}} = \frac{ta_0+1}{t \cdot 1+0} = \frac{ta_0+1}{t} = a_0 + \frac{1}{t}$ for $1 \leq t \leq a_1$.

Putting things together, it is

$$x_{-1,1} = a_0 + 1 > a_0 + \frac{1}{2} > \cdots > a_0 + \frac{1}{a_1} = x_1 \tag{22}$$

Based on this, we can refine Figure 4, which depicts the nesting and ordering of convergents by including semiconvergents: Between two succeeding convergents (e.g., x_n and x_{n+2} in Figure 8, the corresponding semiconvergents ordered according to Lemma 4 are nested (in increasing order as shown for an even n in Figure 8). Furthermore, beyond $x_1 = a_0 + \frac{1}{a_1}$, the semiconvergents $x_{-1,t}$ are added.



$x_n = x_{n,0}$
$x_{n+2} = x_{n,a_{n+2}}$
$x_{n,1}$
$x_{n,2}$ $x_{n,3}$
$x_{n,4}$
$\dots$
$x_{n,a_{n+2}-1}$
$a_0 = x_0$ x_2 $\cdots$ x_n
$x_{n+2} \cdots \mathbf{x} \cdots x_{n-1}$ $\cdots$ x_5
x_3 x_1 $\cdots$ $x_{-1,3}$ $x_{-1,2}$ $x_{-1,1} = a_0 + 1$

Figure 8. Nesting of convergents and semiconvergents (n even).

Now, we are prepared to prove a lower bound of the approximation of a number by the convergents of its continued fraction representation by means of the denominators of the convergents.

Lemma 5. *(Lower Bounds)*

Let p_n/q_n be a convergent of the continued fraction representation of x. Then:

$$\left| x - \frac{p_n}{q_n} \right| > \frac{1}{(q_n + q_{n+1})q_n} \tag{23}$$

Proof. The proof is based on the following claims:

Claim 1. n even $\Rightarrow \frac{p_n}{q_n} < \frac{p_{n+1}+p_n}{q_{n+1}+q_n} < x < \frac{p_{n+1}}{q_{n+1}}$.

Proof. $\frac{p_{n+1}+p_n}{q_{n+1}+q_n}$ is the mediant of $\frac{p_{n+1}}{q_{n+1}}$ and $\frac{p_n}{q_n}$. Thus, the mediant property (Note 3) shows that $\frac{p_n}{q_n} < \frac{p_{n+1}+p_n}{q_{n+1}+q_n} < \frac{p_{n+1}}{q_{n+1}}$. Then:

$$\frac{p_n}{q_n} < \frac{p_{n+1}+p_n}{q_{n+1}+q_n} \overset{(A)}{<} \frac{2p_{n+1}+p_n}{2q_{n+1}+q_n} < \dots < \frac{a_{n+2}p_{n+1}+p_n}{a_{n+2}q_{n+1}+q_n} \overset{(B)}{=} \frac{p_{n+2}}{q_{n+2}}$$

where (A) follows by the monotony of even semiconvergents (Lemma 4), and (B) is the recursion theorem. Because of Theorem 7 (note that n + 2 is even and n + 1 is odd), it is $\frac{p_{n+2}}{q_{n+2}} < x < \frac{p_{n+1}}{q_{n+1}}$. This proves Claim 1. $\square_{(claim1)}$

Claim 2. n odd $\Rightarrow \frac{p_{n-1}}{q_{n-1}} < x < \frac{p_{n+1}+p_n}{q_{n+1}+q_n} < \frac{p_n}{q_n}$.

Proof. As before, $\frac{p_{n+1}+p_n}{q_{n+1}+q_n}$ is the mediant of $\frac{p_{n+1}}{q_{n+1}}$ and $\frac{p_n}{q_n}$. Because n is odd, it is $\frac{p_{n+1}}{q_{n+1}} < \frac{p_n}{q_n}$ (Theorem 7). Thus, $\frac{p_{n+1}}{q_{n+1}} < \frac{p_{n+1}+p_n}{q_{n+1}+q_n} < \frac{p_n}{q_n}$ because of the mediant property (Note 3). Then:

$$\frac{p_n}{q_n} > \frac{p_{n+1}+p_n}{q_{n+1}+q_n} \overset{(A)}{>} \frac{2p_{n+1}+p_n}{2q_{n+1}+q_n} > \dots > \frac{a_{n+2}p_{n+1}+p_n}{a_{n+2}q_{n+1}+q_n} \overset{(B)}{=} \frac{p_{n+2}}{q_{n+2}}$$

where (A) follows by the monotony of odd semiconvergents (Lemma 4), and (B) is the recursion theorem. Because of Theorem 7 (note that n $-$ 1 is even and n + 2 is odd), it is $\frac{p_{n-1}}{q_{n-1}} < x < \frac{p_{n+2}}{q_{n+2}}$, and because n is odd, it is $\frac{p_{n+2}}{q_{n+2}} < \frac{p_n}{q_n}$. This proves Claim 2. $\square_{(claim2)}$

With Claim 1, for even n, it is $\frac{p_n}{q_n} < \frac{p_{n+1}+p_n}{q_{n+1}+q_n} < x \Rightarrow x - \frac{p_n}{q_n} > \frac{p_n+p_{n+1}}{q_n+q_{n+1}} - \frac{p_n}{q_n}$.

With Claim 2, for n odd, it is $x < \frac{p_{n+1}+p_n}{q_{n+1}+q_n} < \frac{p_n}{q_n} \Rightarrow \frac{p_n}{q_n} - x > \frac{p_n}{q_n} - \frac{p_n+p_{n+1}}{q_n+q_{n+1}} \Leftrightarrow -\left(x - \frac{p_n}{q_n}\right) > -\left(\frac{p_n+p_{n+1}}{q_n+q_{n+1}} - \frac{p_n}{q_n}\right)$.

Thus, for any k $\in \mathbb{N}$: $\left|x - \frac{p_k}{q_k}\right| > \left|\frac{p_{k+1}+p_k}{q_{k+1}+q_k} - \frac{p_k}{q_k}\right|$. Next, we compute

$$\begin{aligned} \tfrac{p_k+p_{k+1}}{q_k+q_{k+1}} - \tfrac{p_k}{q_k} &= \tfrac{(p_k+p_{k+1})q_k-(q_k+q_{k+1})p_k}{(q_k+q_{k+1})q_k} = \tfrac{p_{k+1}q_k-p_kq_{k+1}}{(q_k+q_{k+1})q_k} \\ &\overset{(A)}{=} \tfrac{(-1)^k}{(q_k+q_{k+1})q_k} \end{aligned}$$

where (A) is the sign theorem (Theorem 2).

This implies $\left|x - \frac{p_k}{q_k}\right| > \left|\frac{(-1)^k}{(q_k+q_{k+1})q_k}\right| = \frac{1}{(q_k+q_{k+1})q_k}$. □

Because of $q_{k+1} > q_k$ (Corollary 1), it is

$$q_k + q_{k+1} < 2q_{k+1} \Leftrightarrow \frac{1}{2q_{k+1}} < \frac{1}{q_k + q_{k+1}} \Leftrightarrow \frac{1}{2q_kq_{k+1}} < \frac{1}{(q_k + q_{k+1})q_k}$$

Using the last inequality in Lemma 5 (Lower Bounds) and using Lemma 3 (Upper Bounds), we obtain the concluding theorem of this section:

In summary, we have proved the following:

Theorem 10. *(Bounds of Approximations by Convergents)*

Let p_k/q_k be a convergent of the continued fraction representation of x. Then:

$$\frac{1}{2q_kq_{k+1}} < \frac{1}{(q_k + q_{k+1})q_k} < \left|x - \frac{p_k}{q_k}\right| < \frac{1}{q_kq_{k+1}} < \frac{1}{q_k^2} \tag{24}$$

□

2.5. Best Approximations

Our goal is to approximate a real number by a rational number as good as possible while keeping the denominator of the rational number “small”. Keeping the denominator small is important because in practice, every real number can only be given up to a certain degree of precision, and this is achieved by means of a huge denominator and corresponding numerator. i.e., approximating a real number by a rational number with a huge denominator is canonical, but finding a small denominator is a problem.

This is captured by the following:

Definition 5. A fraction $p/q \in \mathbb{Q}$ is called a *best approximation (of the first kind)* of $\alpha \in \mathbb{R}$:$\Leftrightarrow$ $\forall c/d \in \mathbb{Q} : d \leq q \Rightarrow \left|\alpha - \frac{c}{d}\right| > \left|\alpha - \frac{p}{q}\right|$ (assuming c/d $\neq$ p/q).

Often, the addition “of the first kind” is omitted. By definition, a best approximation of a real number can only be improved if the denominator of the given approximation is increased.

If p/q is a best approximation of α, then $\left|\alpha - \frac{p}{q}\right| = \frac{1}{q}|q\alpha - p|$ is small and, thus, $|q\alpha - p|$ is small. Measuring the goodness of an approximation this way results in the following:

Definition 6. A fraction $p/q \in \mathbb{Q}$ is called a *best approximation of the second kind* of $\alpha \in \mathbb{R}$:$\Leftrightarrow$ $\forall c/d \in \mathbb{Q} : d \leq q \Rightarrow |d\alpha - c| > |q\alpha - p|$ (assuming c/d $\neq$ p/q).

The question is whether every best approximation is also a best approximation of the second kind. Now, 1/3 is a best approximation of 1/5 because the only possible fractions for c/d, with d $\leq$ 3 = q, are 0, 1/2, 2/3, and 1, and these numbers satisfy $\left|\frac{1}{5} - \frac{c}{d}\right| > \left|\frac{1}{5} - \frac{1}{3}\right|$.

Next, we observe that $\left|1 \cdot \frac{1}{5} - 0\right| < \left|3 \cdot \frac{1}{5} - 1\right|$ with 1 < 3. Thus, with $d = 1$ and $q = 3$ (i.e., $d < q$) and $\alpha = 1/5$, we found a fraction $c/d = 0/1$ with $|d\alpha - c| < |q\alpha - p|$! As a consequence, although 1/3 is a best approximation of the first kind of 1/5, it is not a best approximation of the second kind.

Thus, not all best approximations of the first kind are best approximations of the second kind. But the reverse holds true:

Lemma 6. *(Every 2nd Kind Best Approximation is a 1st Kind Best Approximation)*

If $p/q \in \mathbb{Q}$ is a best approximation of the second kind of $\alpha \in \mathbb{R}$, then p/q is also a best approximation of the first kind of α.

Proof (by contradiction). Assume p/q is not a best approximation of the first kind. Then, $\left|\alpha - \frac{c}{d}\right| \leq \left|\alpha - \frac{p}{q}\right|$ for a fraction c/d with d < q. Multiplying both inequations results in $d\left|\alpha - \frac{c}{d}\right| \leq q\left|\alpha - \frac{p}{q}\right| \Leftrightarrow |d\alpha - c| \leq |q\alpha - p|$, which is a contradiction because p/q is a best approximation of the second kind. □

The next simple estimation about the distance of two fractions by means of the product of their denominators is often used.

Note 4. *(Distance of Fractions)*

Let $\frac{a}{b}, \frac{p}{q} \in \mathbb{Q}$ with $\frac{a}{b} \neq \frac{p}{q}$. Then:

$$\left|\frac{p}{q} - \frac{a}{b}\right| \geq \frac{1}{qb} \tag{25}$$

Proof. With $a, p \in \mathbb{Z}$ and $b, q \in \mathbb{N}$, it is $pb - aq \in \mathbb{Z}$. Also, $pb - aq \neq 0$ because otherwise $pb = aq \Leftrightarrow \frac{p}{q} = \frac{a}{b}$ which contradicts the premise. Thus, $|pb - aq| \in \mathbb{N}$, i.e., $|pb - aq| \geq 1$. This implies

$$\left|\frac{p}{q} - \frac{a}{b}\right| = \left|\frac{pb - aq}{qb}\right| = \frac{|pb - aq|}{|qb|} \geq \frac{1}{qb}$$

where $|qb| = qb$ because $b, q \in \mathbb{N}$. □

Next, we prove that every best approximation of the second kind is a convergent.

Theorem 11. *(2nd Kind Best Approximations are Convergents)*

Let a/b be a best approximation of the second kind of $x \in \mathbb{R}$, and let $x = [a_0; a_1, \cdots]$ be the continued fraction representation of x.

Then a/b is a convergent of x.

Proof. Being a best approximation of the second kind of x, a/b satisfies, by definition, $|dx - c| > |bx - a|$ for d ≤ b.

Claim 1. $\frac{a}{b} \geq a_0 = x_0$.

Proof (by contradiction). Assume $\frac{a}{b} < a_0 \Rightarrow -a_0 < -\frac{a}{b} \Rightarrow x - a_0 < x - \frac{a}{b}$; thus, $|x - a_0| < \left|x - \frac{a}{b}\right| \overset{(A)}{\leq} b\left|x - \frac{a}{b}\right| = |bx - a|$, where (A) holds because $b \in \mathbb{N}$, i.e., $1 \leq$ b. This implies $|1 \cdot x - a_0| \leq |bx - a|$, which contradicts $|dx - c| > |bx - a|$ for d ≤ b (with $d = 1 \leq b$ and $c = a_0$). This means that $\frac{a}{b} \geq a_0 = \frac{a_0}{1} \overset{(B)}{=} \frac{q_0}{q_0} = x_0$, (B) is because of the recursion theorem. $\square_{(claim1)}$

Thus, the geometric situation is as depicted in Figure 9, i.e., a/b is in the grey shaded area being greater than or equal to the convergent x_0. This will be refined in what follows.

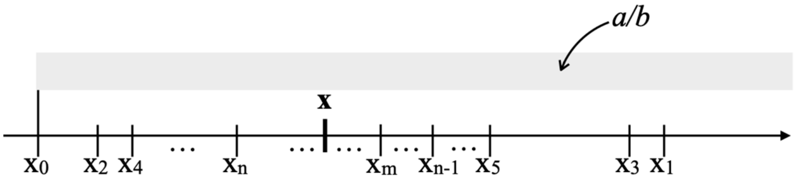


Figure 9. Any best approximation of the second kind is in the grey shaded area, i.e., greater than or equal to the convergent x_0.

Next, we proceed with a proof by contradiction assuming that a/b is not a convergent of x.

Assumption. $\frac{a}{b} \neq \frac{q_k}{q_k} = x_k$ for $k \in \mathbb{N}$.

According to Claim 1, $\frac{a}{b} \geq a_0 = x_0$. Thus, one of the following must hold:

$$\text{(i) } \frac{a}{b} \in \left]\frac{p_{k-1}}{q_{k-1}}, \frac{p_{k+1}}{q_{k+1}}\right[\text{ for k } \geq 1$$

or

$$\text{(ii) } \frac{a}{b} > \frac{p_1}{q_1} = x_1$$

This situation is shown in Figure 10.

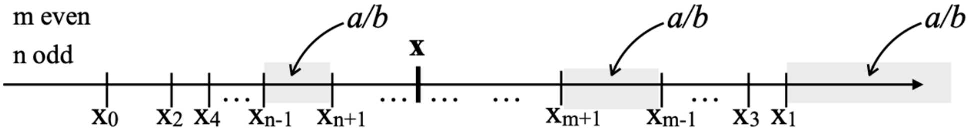


Figure 10. If a best approximation of the second kind is not a convergent, it is within the indicated grey shaded areas.

Case (1). If (i) is true, then

$$\left|\frac{a}{b} - \frac{p_{k-1}}{q_{k-1}}\right| < \left|x - \frac{p_{k-1}}{q_{k-1}}\right| \overset{(Th8)}{<} \left|\frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}}\right| \overset{(C)}{=} \frac{1}{q_k q_{k-1}}$$

where (Th8) is Theorem 8, Equation (14), and (*C*) is from Corollary 3, Equation (10). Furthermore, $\left|\frac{a}{b} - \frac{p_{k-1}}{q_{k-1}}\right| \overset{(D)}{\geq} \frac{1}{bq_{k-1}}$, with (*D*) because of Note 4 (Distance of Fractions).

Together, $\frac{1}{bq_{k-1}} \leq \left|\frac{a}{b} - \frac{p_{k-1}}{q_{k-1}}\right| < \frac{1}{q_k q_{k-1}} \Rightarrow \frac{1}{b} < \frac{1}{q_k} \Rightarrow b > q_k$ (iii).

Also, if (i) is true, then $\left|x - \frac{a}{b}\right| \geq \left|\frac{p_{k+1}}{q_{k+1}} - \frac{a}{b}\right| \overset{(E)}{\geq} \frac{1}{bq_{k+1}}$, where (*E*) is again using Note 4. This implies $b\left|x - \frac{a}{b}\right| \geq \frac{1}{q_{k+1}} \Rightarrow |bx - a| \geq \frac{1}{q_{k+1}}$ (iv).

Lemma 3 (Upper Bounds) tells us that $\left|x - \frac{p_k}{q_k}\right| < \frac{1}{q_k q_{k+1}}$ which is equivalent to $q_k\left|x - \frac{p_k}{q_k}\right| < \frac{1}{q_{k+1}} \Leftrightarrow |q_k x - p_k| < \frac{1}{q_{k+1}} \Rightarrow |q_k x - p_k| < |bx - a|$ (see (iv) just before). Since $q_k < b$ (see (iii) above), this is a contradiction to a/b being a best approximation of the second kind of x. Thus, Case (1) does not occur.

Case (2). This case is shown in Figure 11. Then, $\left|x - \frac{a}{b}\right| > \left|\frac{p_1}{q_1} - \frac{a}{b}\right| \overset{(F)}{=} \frac{1}{bq_1}$, where (F) again uses Note 4. This implies $|bx - a| > \frac{1}{q_1} \overset{(G)}{=} \frac{1}{a_1}$ (v) with (G) using the recursion theorem.

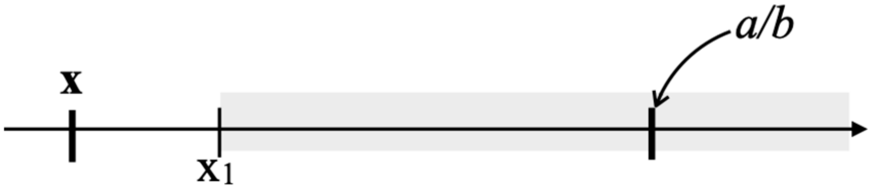


Figure 11. Pictorial representation of Case (2).

Now, $x - a_0 = \frac{1}{a_1 + \frac{1}{a_2 + \ddots}} \leq \frac{1}{a_1}$, where the last inequality holds because of $\frac{1}{a_2 + \ddots} > 0$; thus, $|x - a_0| \leq \frac{1}{a_1} \overset{(H)}{<} |bx - a|$, (H) based on (v) before. This means that $|1 \cdot x - a_0| < |bx - a|$ with $1 \leq b$, i.e., a/b is not a best approximation of the second kind of x, which is a contradiction. Thus, Case (2) does not occur either.

Consequently, the assumption is wrong and there is a $k \in \mathbb{N}$ with $\frac{a}{b} = \frac{q_k}{q_k} = x_k$, i.e., a/b is a convergent. □

So, every best approximation of the second kind is a convergent. The next theorem proves the reverse, i.e., that every convergent is a best approximation of the second kind.

Theorem 12. *(Lagrange, 1798—Convergents are 2nd Kind Best Approximations)*

Let p_n/q_n be a convergent of $x = [a_0; a_1, \cdots, a_N]$, $x \neq a_0 + \frac{1}{2}$, and $n \neq 0$. Then, for $d \leq q_n$ and $\frac{c}{d} \neq \frac{p_n}{q_n}$ it is $|dx - c| > |q_n x - p_n|$, i.e., the convergent is a best approximation of the second kind of x.

The cases $x = a_0 + \frac{1}{2}$ and $n = 0$ are excluded because the convergent $\frac{p_0}{q_0} = \frac{a_0}{1}$ is not a best approximation of the second kind of $x = a_0 + \frac{1}{2}$: it is $|1 \cdot x - (a_0 + 1)| = \left|a_0 + \frac{1}{2} - a_0 - 1\right| = \frac{1}{2}$ and $|1 \cdot x - a_0| = \left|a_0 + \frac{1}{2} - a_0\right| = \frac{1}{2}$, which implies $|1 \cdot x - (a_0 + 1)| = |1 \cdot x - a_0|$. Setting $d := 1 \leq q_0$, $c := a_0 + 1$ results in $|d \cdot x - c| = |1 \cdot x - (a_0 + 1)| = |1 \cdot x - a_0| = |q_0 \cdot x - p_0|$. If $\frac{p_0}{q_0}$ would be a best approximation of the second kind of x, then $|1 \cdot x - (a_0 + 1)| > |1 \cdot x - a_0|$ would hold.

The proof of Lagrange's theorem is very technical. First, the expression $|y_0 x - z_0|$ is analyzed to find the smallest integral numbers y_0 and z_0 such that the expression is minimized under the constraint $y_0 \in \{q_0, \ldots, q_k\}$, i.e., y_0 is a denominator of a convergent. It is shown both that z_0/y_0 is a best approximation of the second kind of x, and that $z_0 = p_k$ and $y_0 = q_k$.

Proof. Let $k \in \mathbb{Z}$ and let p_k/q_k be a convergent. First, we are looking for the smallest numbers $y_0, z_0 \in \mathbb{Z}$ with $y_0 \in \{q_0, \ldots, q_k\}$ such that $|y_0 x - z_0|$ is minimal.

Step 1. Pick an arbitrary $z \in \mathbb{Z}$, and based on this we determine $y_0 \in \{q_0, \ldots, q_k\}$.

It is $\underset{y}{min}|yx - z| = 0 \Leftrightarrow y = \frac{z}{x}$, but in general $y \notin \mathbb{Z}$. Looking for a solution $y_0 \in \{q_0, \ldots, q_k\} \subseteq \mathbb{Z}$ that minimizes $|y_0 x - z|$ results in the following potential positions of z/x with respect to the denominators $q_0, \ldots, q_k$ (see Figure 12):

- Case 1: $z/x > q_k$. Then, $y_0 = q_k$ is the solution;
- Case 2: $z/x < q_0$. Then, $y_0 = q_0$ is the solution; Let $q_i \leq z/x \leq q_{i+1}$ for 1 ≤ i ≤ k.
- Case 3: For $|q_{i+1}x - z| < |q_i x - z|$ (i.e., z/x is closer to q_{i+1} than to q_i), $y_0 = q_{i+1}$ is the solution, and for $|q_{i+1}x - z| > |q_i x - z|$ (i.e., z/x is closer to q_i than to q_{i+1}), $y_0 = q_i$ is the solution;

- Case 4: For $|q_{i+1}x - z| = |q_i x - z|$ (i.e., z/x is exactly in the middle between q_i and q_{i+1}), $y_0 = q_i$ is the solution because $q_i < q_{i+1}$, and we are looking for the smallest y_0, especially $y_0 \geq q_0 = 1$. $\square_{(step1)}$

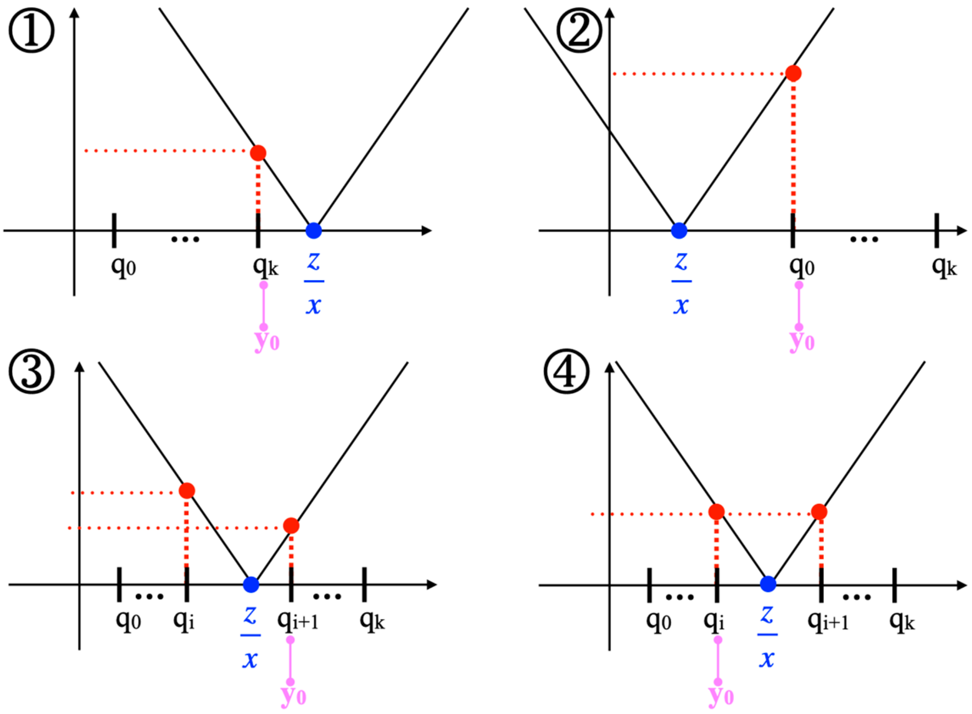


Figure 12. The potential positions of z/x with respect to the denominators $q_0, \ldots, q_k$.

Step 2. Based on the y_0 found, we determine z_0 next. It is $\underset{z}{min}|y_0 x - z| = 0 \Leftrightarrow z = y_0 x$, but in general, $z \notin \mathbb{Z}$. In solving the minimization problem within $\mathbb{Z}$ (i.e., $z_0 := \underset{z \in \mathbb{Z}}{argmin}|y_0 x - z|$), the following cases can be distinguished (see Figure 13):

- Case 0: It may happen that $y_0 x \in \mathbb{Z}$. Then, choose $z_0 = y_0 x$;
- Case 1: $y_0 x$ is between two integral numbers s and t, i.e., $s < y_0 x < t$. For $|y_0 x - s| > |y_0 x - t|$ (i.e., $y_0 x$ is closer to t than to s), $z_0 = t$ is the solution; and for $|y_0 x - s| < |y_0 x - t|$ (i.e., $y_0 x$ is closer to s than to t), $z_0 = s$ is the solution;
- Case 2: For $|y_0 x - s| = |y_0 x - t|$ (i.e., $y_0 x$ is exactly in the middle between t and s), $z_0 = s$ is the solution because $s < t$, and we are looking for the smallest z_0. $\square_{(step2)}$

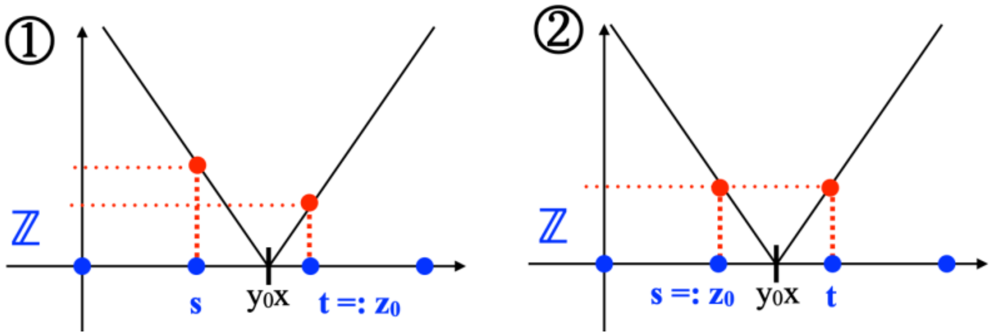


Figure 13. The potential positions of $y_0 x$.

Claim 1. z_0 is uniquely determined.

Proof (by contradiction). Assume there exists a $\widetilde{z}_0 \in \mathbb{Z}$ with $\widetilde{z}_0 \neq z_0$ and $\left|x - \frac{z_0}{y_0}\right| = \left|x - \frac{\widetilde{z}_0}{y_0}\right|$. This can only happen iff one term is positive and the other is negative, i.e., for example, if $x - \frac{z_0}{y_0} > 0$ and $x - \frac{\widetilde{z}_0}{y_0} < 0$, and then $x - \frac{z_0}{y_0} = \frac{\widetilde{z}_0}{y_0} - x$, i.e., $x = \frac{z_0 + \widetilde{z}_0}{2y_0}$.

As an intermediate step we prove:

Claim 2. $z_0 + \widetilde{z}_0$ and $2y_0$ are co-prime, i.e., $gcd(z_0 + \widetilde{z}_0, 2y_0) = 1$

Proof (by contradiction). Let $\widetilde{z}_0 + z_0 = Lp$ and $2y_0 = Lq$ with $L > 1$. Then, $x = \frac{z_0+\widetilde{z}_0}{2y_0} = \frac{Lp}{Lq} \Rightarrow x = \frac{p}{q}$ and thus

$$\text{(i) } |qx - p| = \left|q\frac{p}{q} - p\right| = 0$$

Assume $L > 2$. Then, with $2y_0 = Lq$ and $L/2 > 1$, it follows:

$$\text{(ii) } y_0 = \frac{L}{2}q > q$$

Now, y_0 has been determined in Step 1 to satisfy $y_0 = \underset{y}{argmin}\ |yx - z|$ for a given z, especially for $z = p$, i.e., $y_0 = \underset{y}{argmin}\ |yx - p|$. Because $0 = \underset{y}{min}\ |yx - p|$ and $|qx - p| = 0$, it must be $q = y_0$. This is a contradiction because $q < y_0$ according to (ii) before. Thus, $1 < L \leq 2$, i.e., $L = 2$.

With $L = 2$ and $2y_0 = Lq$, we get $y_0 = q$, which implies. By definition of z_0, $|qx - p| = |y_0x - p| > |y_0x - z_0|$. However, $|qx - p| = 0$ (see (i) above); thus, $0 > |y_0x - z_0|$, which is a contraction. $\square_{(claim2)}$

We continue the proof of Claim 1: It is $\frac{z_0+\widetilde{z}_0}{2y_0} = x$ and also $x = \frac{p_N}{q_N}$, i.e., $\frac{z_0+\widetilde{z}_0}{2y_0} = \frac{p_N}{q_N}$. Because $gcd(z_0 + \widetilde{z}_0, 2y_0) = 1$ according to Claim 2, it follows that $p_N = z_0 + \widetilde{z}_0$ and $q_N = 2y_0$.

Now, let $N \geq 2$. Then, it is $2y_0 = q_N \overset{(A)}{=} a_N q_{N-1} + q_{N-2}$ ((A) uses the recursion theorem (Theorem 1)), and with Note 1, it is $a_N \geq 2$. Thus, $2y_0 \geq 2q_{N-1} + q_{N-2}$ $\Rightarrow y_0 \geq q_{N-1} + \frac{q_{N-2}}{2} \Rightarrow q_{N-1} \leq y_0 - \frac{q_{N-2}}{2} \overset{(B)}{<} y_0$ ((B) is because $q_{N-2} > 0$). Now:

$$|q_{N-1}x - p_{N-1}| = \left|q_{N-1}\frac{p_N}{q_N} - p_{N-1}\right| = \frac{1}{q_N}|q_{N-1}p_N - p_{N-1}q_N| \overset{(C)}{=} \frac{1}{q_N} = \frac{1}{2y_0} \overset{(D)}{\leq} \frac{1}{2}$$

where (C) holds because of the sign theorem and (D) because $y_0 \geq 1$ (see the end of the proof of Step 1).

Furthermore,

$$\begin{aligned} |y_0x - z_0| = \left|y_0\tfrac{z_0+\widetilde{z}_0}{2y_0} - z_0\right| &= \left|\tfrac{z_0+\widetilde{z}_0}{2} - z_0\right| = \tfrac{1}{2}|z_0 + \widetilde{z}_0 - 2z_0| \\ &= \tfrac{1}{2}|\widetilde{z}_0 - z_0| \overset{(E)}{\geq} \tfrac{1}{2}(iii) \end{aligned}$$

where (E) is true because $\widetilde{z}_0 \neq z_0$ and, thus, $|\widetilde{z}_0 - z_0| \geq 1$ for integral numbers $\widetilde{z}_0$ and z_0. Together, we obtained $|y_0x - z_0| \geq \frac{1}{2} \geq |q_{N-1}x - p_{N-1}|$, which is a contradiction to the choice of y_0 and z_0! This proves Claim 1 for $N \geq 2$.

Now, let $N = 1$ and choose $a_1 = 2$ (based on Note 1, the highest element of a continued fraction is always greater than or equal 2, thus $a_1 \geq 2$). Then

$$x = [a_0; a_1] = \frac{p_1}{q_1} \overset{(F)}{=} \frac{a_1 a_0 + 1}{a_1} = \frac{2a_0 + 1}{2} = a_0 + \frac{1}{2}$$

((F) is the recursion theorem) which has been excluded from the theorem.

Thus, let $N = 1$ and $a_1 > 2$. Then

$$|1 \cdot x - a_0| \overset{(G)}{=} |q_0 x - p_0| = \left| q_0 \frac{p_1}{q_1} - p_0 \right| = \frac{1}{q_1} |q_0 p_1 - q_1 p_0| \overset{(H)}{=} \frac{1}{q_1} \overset{(G)}{=} \frac{1}{a_1} < \frac{1}{2}$$

where (G) applies the recursion theorem and (H) the sign theorem. Because of (iii), it is $|y_0 x - z_0| \geq \frac{1}{2}$, i.e., together, $|q_0 x - p_0| < |y_0 x - z_0|$ which contradicts the definition of y_0 and z_0! This proves Claim 1 for $N = 1$. $\square_{(claim1)}$

Next, we observe

Claim 3. $\frac{z_0}{y_0}$ is a best approximation of the second kind of x.

Otherwise: $|bx - a| \leq |y_0 x - z_0|$ for an $\frac{a}{b} \neq \frac{z_0}{y_0}$ with $b \leq y_0$, which contradicts the definition of y_0 and z_0! $\square_{(claim3)}$

According to Theorem 11, $\frac{z_0}{y_0}$ is a convergent of x, i.e., $\frac{z_0}{y_0} = \frac{p_s}{q_s}$ for an $s \leq k$. If $s = k$, the proof is done. Thus, we assume $s < k$.

Claim 4. For $s < k$, it is $\frac{1}{q_s + q_{s+1}} \geq \frac{1}{q_k + q_{k-1}}$.

Proof. $s < k \Rightarrow s \leq k - 1 \Rightarrow q_s \leq q_{k-1}$ (Corollary 1: denominators are monotonically increasing). Similarly, $s < k \Rightarrow s + 1 \leq k \Rightarrow q_{s+1} \leq q_k$. Together, this implies $q_k + q_{k-1} \geq q_s + q_{s+1}$. $\square_{(claim4)}$

Next, we get

$$|q_s x - p_s| = q_s \left| x - \frac{p_s}{q_s} \right| \overset{(I)}{>} q_s \frac{1}{(q_s + q_{s+1}) q_s} = \frac{1}{q_s + q_{s+1}} \overset{(J)}{\geq} \frac{1}{q_k + q_{k-1}}$$

where (I) is Lemma 5 (Lower Bounds) and (J) is Claim 4.

Furthermore, $|q_k x - p_k| = q_k \left| x - \frac{p_k}{q_k} \right| \overset{(K)}{<} q_k \frac{1}{q_k q_{k+1}} = \frac{1}{q_{k+1}}$, where (K) holds because of Lemma 3 (Upper Bounds).

With $\frac{z_0}{y_0} = \frac{p_s}{q_s}$ and the definition of $y_0 (= q_s)$ and $z_0 (= p_s)$ (i.e., the minimizing property), it is $|q_s x - p_s| = |y_0 x - z_0| \leq |q_k x - p_k| \Rightarrow \frac{1}{q_k + q_{k-1}} \leq \frac{1}{q_{k+1}}$, which implies $q_{k+1} < q_k + q_{k-1}$. This is a contradiction; because of the recursion theorem, it is $q_{k+1} = a_{k+1} q_k + q_{k-1} \overset{(L)}{\geq} q_k + q_{k-1}$, where (L) holds with $a_k \geq 1$. Thus, $s = k$ which proves the overall theorem. $\square$

Putting the last two theorems together yields:

Corollary 5. a/b is a best approximation of the second kind of x $\Leftrightarrow$ x is a convergent of x. $\square$

According to Theorem 12, every convergent is a best approximation of the second kind, and each best approximation of the second kind is also a best approximation of the first kind (Lemma 6). We keep this observation as:

Note 5. Every convergent is a best approximation of the first kind. $\square$

But are best approximations of the first kind also always convergents? Not quite: the next theorem proves that a best approximation of the first kind is a convergent or a semiconvergent.

Theorem 13. *(Lagrange, 1798—1st Kind Best Approximations are Convergents or Semiconvergents)* *Let a/b be a best approximation of the first kind of $x = [a_0; a_1, \cdots, a_N]$. Then a/b* is *a convergent or a semiconvergent of x.*

Proof. By definition, it is $\left|x - \frac{c}{d}\right| > \left|x - \frac{a}{b}\right|$ for $\frac{c}{c} \neq \frac{a}{b}$ and $d \leq b$.

Claim 1. $a/b > a_0$.

Otherwise: $\frac{a}{b} \leq a_0 = \frac{a_0}{1}$; thus, $x - a_0 \leq x - \frac{a}{b}$. Now, $x - a_0 = \cfrac{1}{a_1 + \ddots} > 0$; thus, $0 < x - a_0 \leq x - \frac{a}{b} \Rightarrow \left|x - \frac{a_0}{1}\right| \leq \left|x - \frac{a}{b}\right|$. Because $1 \leq b$, we obtained a contradiction since a/b is a best approximation of the first kind. $\square_{(claim 1)}$

Claim 2. $a/b < a_0 + 1$.

Otherwise: $\frac{a}{b} \geq a_0 + 1$ and based on the geometric situation depicted in Figure 8, it follows that $\left|x - \frac{a_0+1}{1}\right| \leq \left|x - \frac{a}{b}\right|$ with $1 \leq b$, which contradicts a/b being a best approximation of the first kind. $\square_{(claim 2)}$

Consequently, a/b lies between $x_0 = a_0$ and $x_{-1,1} = a_0 + 1$ (see Equation (22)), i.e.,

$$x_0 = a_0 < \frac{a}{b} < a_0 + 1 = x_{-1,1} \tag{26}$$

and is, thus, covered by the set of intervals defined by the convergents and semiconvergents of x (see Figure 8).

Assumption. a/b is neither a convergent nor a semiconvergent.

This results in the following cases:

- Case 1: a/b lies between two semiconvergents $x_{k-1,r}$ and $x_{k-1,r+1}$;
- Case 2: a/b lies between two convergents x_k and x_{k+2};
- Case 3: a/b lies between a convergent and a semiconvergent.

We will show that all three cases lead to a contradiction, i.e., the assumption must be false; thus, the theorem is proven.

Case 1. a/b lies between $x_{k-1,r} = \frac{rp_k+p_{k-1}}{rq_k+q_{k-1}}$ and $x_{k-1,r+1} = \frac{(r+1)p_k+p_{k-1}}{(r+1)q_k+q_{k-1}}$.

Then,

$$\left|\frac{a}{b} - \frac{rp_k + p_{k-1}}{rq_k + q_{k-1}}\right| < | \frac{(r+1)p_k + p_{k-1}}{(r+1)q_k + q_{k-1}} - \frac{rp_k + p_{k-1}}{rq_k + q_{k-1}} | \overset{(A)}{=} \frac{1}{((r+1)q_k + q_{k-1})(rq_k + q_{k-1})}$$

where (A) results from the same computation performed in the proof of Lemma 4.

Furthermore, it is

$$\text{(i)} \left|\frac{a}{b} - \frac{rp_k + p_{k-1}}{rq_k + q_{k-1}}\right| = \frac{|a(rq_k + q_{k-1}) - b(rp_k + p_{k-1})|}{b(rq_k + q_{k-1})} \overset{(B)}{\geq} \frac{1}{b(rq_k + q_{k-1})}$$

where (B) is seen to be valid as follows: $a(rq_k + q_{k-1}) - b(rp_k + p_{k-1}) \in \mathbb{Z}$ and, thus, $|a(rq_k + q_{k-1}) - b(rp_k + p_{k-1})| \in \mathbb{N}_0$; if it would be zero, the first modulus in (i) would be zero, i.e., $a/b = x_{k-1,r}$ which contradicts the assumption of the claim, which in turn implies $|a(rq_k + q_{k-1}) - b(rp_k + p_{k-1})| \geq 1$.

Together,

$$\frac{1}{b(rq_k + q_{k-1})} < \frac{1}{((r+1)q_k + q_{k-1})(rq_k + q_{k-1})} \Rightarrow \frac{1}{b} < \frac{1}{(r+1)q_k + q_{k-1}},$$

thus,

$$\text{(ii) } b > (r+1)q_k + q_{k-1}$$

Because of the monotony of the sequence of semiconvergents $(x_{s,t})_t$ (Lemma 4), it is for an odd k (i.e., $k-1$ even) $x_{k-1,r} < x_{k-1,r+1}$ (see the geometric situation in Figure 14);

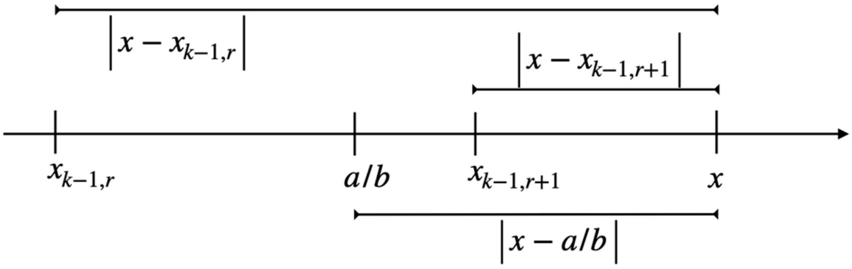


Figure 14. Distances within an interval of semiconvergents (k odd).

thus,

$$\left|x-\frac{a}{b}\right| > \left|x-\frac{(r+1)p_k+p_{k-1}}{(r+1)q_k+q_{k-1}}\right|$$

But with (ii), it is $(r+1)q_k+q_{k-1} < b$; thus, a/b is not a best approximation of the first kind to x, which is a contradiction. k even leads to a contradiction too, i.e., Case (1) is not possible $\square_{(case1)}$

Case 2. a/b lies between x_k and x_{k+2}.

Then, $\left|\frac{a}{b}-\frac{p_k}{q_k}\right| < \left|\frac{p_k}{q_k}-\frac{p_{k+2}}{q_{k+2}}\right| \overset{(C)}{=} \frac{a_{k+2}}{q_kq_{k+2}} < \frac{1}{q_kq_{k+2}}$ where (C) is Equation (11) from Corollary 3, and with Note 4, it is $\left|\frac{a}{b}-\frac{p_k}{q_k}\right| \geq \frac{1}{bq_k}$.

Together, $\frac{1}{bq_k} < \frac{1}{q_kq_{k+2}} \Rightarrow \frac{1}{b} < \frac{1}{q_{k+2}} \Rightarrow b > q_{k+2}$. Because of the geometric situation shown in Figure 15, it is $\left|x-\frac{a}{b}\right| > \left|x-\frac{p_{k+2}}{q_{k+2}}\right|$, which is a contradiction to a/b being a best approximation of the first kind to x and $b > q_{k+2}$. $\square_{(case2)}$

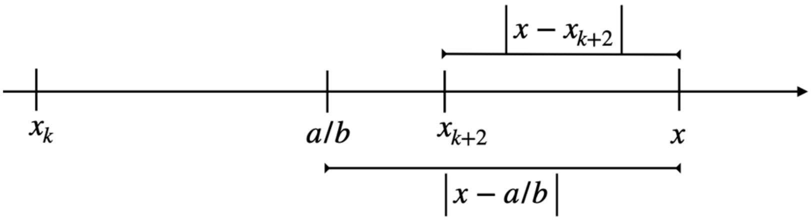


Figure 15. Distances within an interval of convergents (k even).

Case 3. a/b lies between a convergent and a semiconvergent.

This implies that a/b lies between x_k and $x_{k,1}$ (see Figure 8), otherwise a/b would lie between two semiconvergents, which has already been covered in Case 1.

Thus, $\left|\frac{a}{b}-\frac{p_k}{q_k}\right| < \left|x_k - x_{k,1}\right|$, but

$$\begin{aligned}\left|x_k - x_{k,1}\right| &= \left|\frac{p_k}{q_k}-\frac{p_{k+1}+p_k}{q_{k+1}+q_k}\right| = \left|\frac{p_k(q_{k+1}+q_k)-q_k(p_{k+1}+p_k)}{q_k(q_{k+1}+q_k)}\right| \\ &= \left|\frac{p_kq_{k+1}-q_kp_{k+1}}{q_k(q_{k+1}+q_k)}\right| \overset{(D)}{=} \frac{1}{q_k(q_{k+1}+q_k)}\end{aligned}$$

where (D) is the sign theorem. I.e., it is $\left|\frac{a}{b}-\frac{p_k}{q_k}\right| < \frac{1}{q_k(q_{k+1}+q_k)}$. As before, with Note 4, it is $\left|\frac{a}{b}-\frac{p_k}{q_k}\right| \geq \frac{1}{bq_k} \Rightarrow \frac{1}{bq_k} < \frac{1}{q_k(q_{k+1}+q_k)} \Rightarrow b > q_{k+1}+q_k$.

The geometric situation from Figure 16 reveals $\left|x-\frac{a}{b}\right| > \left|x-\frac{p_k+p_{k-1}}{q_k+q_{k-1}}\right|$, which is a contradiction to a/b being a best approximation of the first kind to x and $b > q_{k+1} + q_k$. $\square_{(case3)}$ $\square$

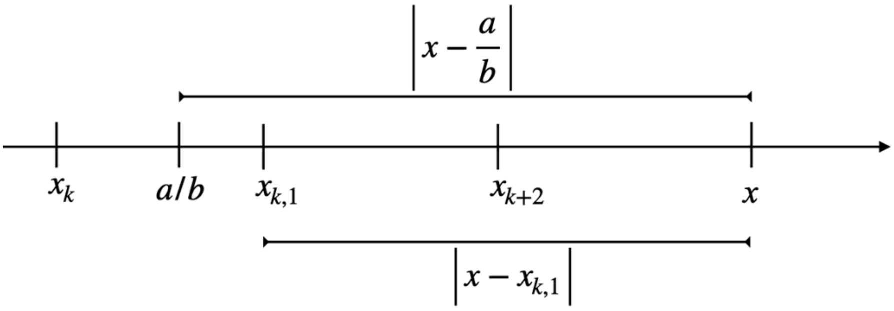


Figure 16. Situation in which a/b is between a convergent and its first semiconvergent (k even).

Finally, we give a simple criterion that allows us to prove that a given fraction is a convergent of another real number. This theorem is a cornerstone of computing a prime factor with Shor's algorithm.

Theorem 14. *(Legendre, 1798—Convergent Criterion)*
Let $\left|x-\frac{a}{b}\right| < \frac{1}{2b^2} \Rightarrow a/b$ *is a convergent of* x.

Proof. We show that a/b is a best approximation of the second kind of x. Theorem 11 then proves the claim.

Let $|dx-c| \le |bx-a|$ for $\frac{a}{b} \ne \frac{c}{d}$ and $d > 0$. We need to prove $d > b$.

Now, $|bx-a| = b\left|x-\frac{a}{b}\right| < b\frac{1}{2b^2} = \frac{1}{2b}$. This implies $|dx-c| < \frac{1}{2b} \Leftrightarrow d\left|x-\frac{c}{d}\right| < \frac{1}{2b} \Leftrightarrow \left|x-\frac{c}{d}\right| < \frac{1}{2db}$. Thus,

$\left|\frac{c}{d}-\frac{a}{b}\right| = \left|\frac{c}{d}-x+x-\frac{a}{b}\right| \le \left|\frac{c}{d}-x\right| + \left|x-\frac{a}{b}\right| < \frac{1}{2db} + \frac{1}{2b^2} = \frac{b+d}{2db^2}$

With Note 4 (Distance of Fractions), it is also $\left|\frac{c}{d}-\frac{a}{b}\right| \ge \frac{1}{db}$. Together, it is

$$\frac{1}{db} < \frac{b+d}{2db^2} \Leftrightarrow 1 < \frac{b+d}{2b} \Leftrightarrow 2b < b+d \Leftrightarrow d > b. \square$$

3. Probability of the Occurrence of Convergents

3.1. Estimating Secant Lengths

In this part, we use the main arguments of [2].

In order to estimate the probability of the occurrence of a certain state after having performed the quantum Fourier transform, we need the following estimation of a lower bound and an upper bound of the length of a secant of the unit circle:

Lemma 7. *(Secant Length Estimation)*
If $\varphi \in [-\pi, \pi]$ *then* $\frac{2|\varphi|}{\pi} \le \left|1-e^{i\varphi}\right| \le |\varphi|$.

Proof. The upper bound follows from elementary geometry, namely that the length of a secant is less than or equal to the length of the corresponding arc of a circle (see Figure 17).

The length of the arc determined by the angle φ on a circle of radius r is $r\varphi$, i.e., if the circle is a unit circle, the length of the arc (green in the Figure) is φ.

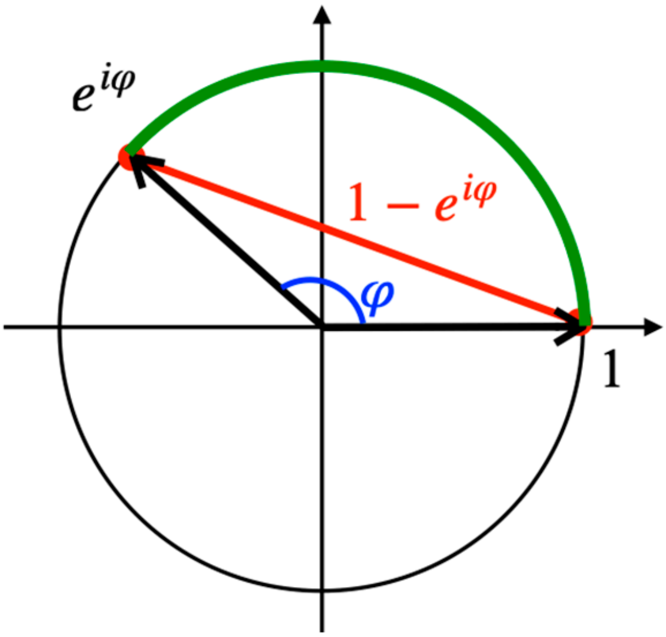


Figure 17. The length of a secant is smaller than the arc of the corresponding unit circle.

A secant of the unit circle (red in the Figure 17) can be defined by the two complex numbers on the unit circle (black in the Figure 17) that are the endpoints of the secant. Thus, the length of this secant is the difference of these complex numbers. One of these points can always be 1 because a corresponding rotation is length-preserving; the other point is then $e^{i\varphi}$, where φ is the angle of the arc cut by the secant. The length of this secant is then $\left|1-e^{i\varphi}\right|$.

This proves the inequality $\left|1-e^{i\varphi}\right| \leq |\varphi|$. $\square_{(upperbound)}$

Next, we compute

$$\begin{aligned} \left|1-e^{i\varphi}\right| &\overset{(A)}{=} |1-cos\varphi - isin\varphi| \overset{(B)}{=} \sqrt{(1-cos\varphi)^2+sin^2\varphi} \\ &= \sqrt{1-2cos\varphi+cos^2\varphi+sin^2\varphi} = \sqrt{2-2cos\varphi} \\ &= \sqrt{2}\sqrt{1-cos\varphi} \overset{(C)}{=} \sqrt{2}\sqrt{2sin^2\tfrac{\varphi}{2}} \\ &\overset{(D)}{=} 2sin\tfrac{\varphi}{2} \end{aligned}$$

where (A) uses Euler's formula, (B) is the definition of the modulus of a complex number with $Re = 1 - cos\ \varphi$ and $Im = -sin\ \varphi$, (C) is the double-angle formula, and (D) assumes that $sin\frac{\varphi}{2} \geq 0$.

To estimate a lower bound for $sin\frac{\varphi}{2}$, we analyze the function $f(x) = sin\ x - \frac{2x}{\pi}$. From elementary calculus, it is known that a function ψ is concave on $D \subseteq \mathbb{R}$ if and only if its second derivative is not positive on D, i.e., $\psi'' \leq 0$ on D.

(Reminder: ψ is *concave* on D:⇔ $\forall x, y \in D \forall t \in [0,1] : \psi(tx+(1-t)y) \geq t\psi(x) + (1-t)\psi(y)$,

i.e., for any two points on the graph of ψ, the secant between these points is below the graph, Figure 18).

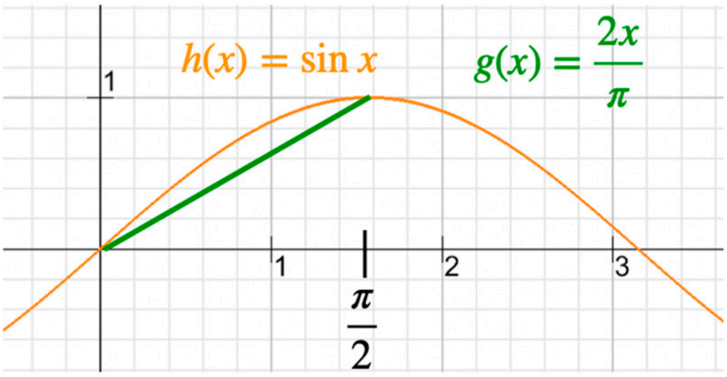


Figure 18. The graphs of sin x and 2x/π.

With $d^2 sin\ x/dx^2 = -sin\ x \leq 0$ especially for $x \in [0, \pi/2]$, i.e., $sin\ x$ is concave on $x \in [0, \pi/2]$. Thus, the secant between $sin\ 0 = 0$ and $sin \frac{\pi}{2} = 1$ is below the graph of $sin\ x$ (orange in Figure 18). However, this secant is given by $g(x) = \frac{2}{\pi}x$ (green in Figure 18). Thus, it is $\frac{2}{\pi}x \leq sinx$ for $x \in [0, \pi/2]$, i.e., with $\varphi := 2x$ we get $\frac{\varphi}{\pi} \leq sin\frac{\varphi}{2}$ for $\varphi \in [0, \pi]$, and this implies $2\frac{\varphi}{\pi} \leq 2sin\frac{\varphi}{2}$ for $\varphi \in [0, \pi]$.

Now, $|1 - e^{i\varphi}| = 2sin\frac{\varphi}{2}$ (see the computation before) implies $|1 - e^{i\varphi}| \geq 2\frac{\varphi}{\pi}$ for $\varphi \in [0, \pi]$.

Furthermore, $sin\ x$ is convex on $x \in [-\pi/2, 0]$; thus, an argument analogous to the above shows that $|1 - e^{i\varphi}| \geq 2\frac{|\varphi|}{\pi}$ for $\varphi \in [-\pi, \pi]$. $\square_{(lowerbound)}$ □

3.2. Estimating Amplitude Parameters

As stated in the introduction, the quantum part of Shor's algorithm produces in its final step the following quantum state via a measurement:

$$\frac{1}{\sqrt{NA}} \sum_{j=0}^{A-1} \omega_N^{jpy} |y\rangle \tag{27}$$

Thus, according to the Born rule, the probability $P(y)$ of this particular state $|y\rangle$ is the square of the modulus of the amplitude of $|y\rangle$, i.e.,

$$P(y) = \left| \frac{1}{\sqrt{NA}} \sum_{j=0}^{A-1} \omega_N^{jpy} \right|^2 = \frac{1}{NA} \left| \sum_{j=0}^{A-1} \omega_N^{jpy} \right|^2 \tag{28}$$

The argument of the modulus is a geometric sum $\sum q^j$ with $q = \omega_N^{py} = e^{\frac{2\pi i}{N}py}$; thus, in case $q \neq 1$,

$$P(y) = \frac{1}{NA} \left| \sum_{j=0}^{A-1} q^j \right|^2 = \frac{1}{NA} \left| \frac{1 - q^A}{1 - q} \right|^2 \tag{29}$$

With $q^A = e^{\frac{2\pi i}{N}Apy}$. In this section, in order to compute a lower bound for $P(y)$, we investigate some relations between the following parameters appearing in Equation (28):

- n: the number to be factorized;
- N: a power of 2 (e.g., $N = 2^m$) with $n^2 < N < 2n^2$;
    - the choice of N effectively determines the domain of numbers that can be represented in the $|a\rangle$-part of the quantum register (see Equation (1)).
- p: the period of the modular exponentiation function $f(x) = a^x\ mod\ n$;
- A: the number of arguments mapped to a given value of f.

We also estimate bounds of the argument $2\pi\frac{Apy}{N}$ of $q^A = e^{\frac{2\pi i}{N}Apy}$.

3.2.1. Basics from Number Theory

For convenience, we state the definition of the modulo function.

Definition 7. The *modulo function* is the following map:

$$\begin{aligned} mod : \mathbb{N}_0 \times \mathbb{N} &\to \mathbb{N} \\ (z, n) &\mapsto z - \lfloor \tfrac{z}{n} \rfloor n \overset{def}{=} z \; mod \; n \end{aligned} \tag{30}$$

$z \; mod \; n$ is, thus, the residue left when dividing z by n. I.e., if $r = z \; mod \; n$, then there is a number $k \in \mathbb{N}_0$ such that $z = kn + r$ with $0 \leq r < n$.

If $z \; mod \; n = \widetilde{z} \; mod \; n = r$, we find numbers k_1 and k_2 such that $z = k_1 n + r$ and $\widetilde{z} = k_2 n + r$ with $0 \leq r < n$. This implies that $z - \widetilde{z} = (k_1 - k_2)n =: kn$, i.e., n is a divisor of $z - \widetilde{z}$ (in symbols: $n|(z - \widetilde{z})$). We also obtain that $z \; mod \; n = \widetilde{z} \; mod \; n$ implies that $\widetilde{z} = z + kn$.

The equation $z \; mod \; n = \widetilde{z} \; mod \; n$ is abbreviated as $z \equiv \widetilde{z} \; (mod \; n)$; in words, z is *congruent* $\widetilde{z}$ modulo n. As shown just before, $z \equiv \widetilde{z} \; (modn \;)$ is equivalent to $n|(z - \widetilde{z})$ and to $\widetilde{z} = z + kn$. We keep this as

Note 6.
$z \equiv \widetilde{z} \; (mod \; n) \Leftrightarrow n|(z - \widetilde{z}) \Leftrightarrow \widetilde{z} = z + kn.$ □

Furthermore, we state the definition of modular exponentiation which turns out to play a key role in finding factors.

Definition 8.
For $0 < a < n$, the *modular exponentiation function* is the following map:

$$\begin{aligned} f : \mathbb{N}_0 &\to \mathbb{N}_0 \\ x &\mapsto a^x \; mod \; n \end{aligned} \tag{31}$$

The smallest number p that satisfies $f(x) = f(x + p)$ for all x is called the *period* of *f*. Especially, with $x = 0$, we get $f(0) = f(p)$ which means that $a^p \; mod \; n = a^0 \; modn = 1 \; mod \; n$, i.e., $a^p \equiv 1 \; (mod \; n)$ which in turn is equivalent to $n|(a^p - 1)$. Thus, we have proven:

Note 7.
$f(x) = a^x modn$ has period p $\Leftrightarrow a^p \equiv 1(modn) \Leftrightarrow n|(a^p - 1)$. □

Finding a factor of n can be achieved by finding the period p of the function $f(x) = a^x modn$. This is seen as follows: Let p be the period of f, then $n|(a^p - 1)$, i.e., $(a^p - 1) = kn$. Assume p is even (if p is odd, Shor's algorithm is repeated with a different a, until an even p is found). With such an even p, it is $(a^p - 1) = \left(a^{p/2} - 1\right)\left(a^{p/2} + 1\right) = kn$ which implies that $\left(a^{p/2} - 1\right)$ and $\left(a^{p/2} + 1\right)$ have a common divisor, which in turn means that $gcd\left(a^{p/2} - 1, n\right)$ or $gcd\left(a^{p/2} + 1, n\right)$ is a divisor of n. Thus, if an even period has been determined, classically efficient calculations can be used to compute a factor of n. If this factor is a prime number, we can finish. Otherwise, we continue determining a factor of the former factor, and so on, until we end up with a prime factor of n.

Next, we determine an upper bound of the period p of the modular exponentiation by using group theory. A simple calculation shows that "≡" is an equivalence relation on $\mathbb{Z}$. The equivalence class of $z \in \mathbb{Z}$ is denoted as $[z]$ and is referred to as the *residue class* of z modulo n. It is $[z] = \{\widetilde{z} \in \mathbb{Z} \,|\, \widetilde{z} \equiv z \; (mod \; n)\} = \{z + kn \mid k \in \mathbb{Z}\}$ (see Note 6), where the latter set is sometimes written as $z + n\mathbb{Z}$. The set of all residue classes modulo n is denoted as $\mathbb{Z}_n$, i.e., $\mathbb{Z}_n = \{[0], [1], \ldots, [n-1]\}$.

We can multiply two residue classes modulo n as follows: $[x] \cdot [y] = [x \cdot y]$. With this multiplication, $\mathbb{Z}_n^* = \{[z] \in \mathbb{Z}_n \mid gcd(z, n) = 1\}$ becomes a group. Because $\mathbb{Z}_n^* \subseteq \mathbb{Z}_n$, it is $\varphi(n) := card \; \mathbb{Z}_n^* \leq card \; \mathbb{Z}_n = n$. Since every integer is a divisor of itself, it is

$gcd(n,n) = n \neq 1$ (for $n \geq 2$), i.e., the cardinality of numbers co-prime to n is less than n: $n \geq 2 \Rightarrow \varphi(n) < n$.

The well-known Lagrange's theorem from group theory states that for a group G with $card\ G = m < \infty$ and for each $x \in G$, it is $x^m = e$ (e is the unit element of G)—see Lemma 3.2.5 in [3], for example. Thus, for $x \in \mathbb{Z}_n^*$, it is $x^{\varphi(n)} = 1$, i.e., $x^{\varphi(n)} \equiv 1\ (mod\ n)$. Since the period p is the smallest number with $x^p \equiv 1\ (mod\ n)$, it follows that $p \leq \varphi(n)$ and, thus, $p < n$.

Now, the assumption of Shor's algorithm is that $0 < a < n$ and that $gcd(a,n) = 1$, which ensures that $[a] \in \mathbb{Z}_n^*$; thus, $[a]^p \equiv 1\ (mod\ n)$ and $p < n$.

Lemma 8.

Let p be the period of $f(x) = a^x\ mod\ n$. Then, $p < n$. □

3.2.2. Intervals of Consecutive Multiples of the Period

The relation between N and p is depicted in Figure 19; multiples of N are always contained in closed intervals defined by consecutive multiples of p, i.e., it may happen that a multiple of N coincides with a multiple of p.

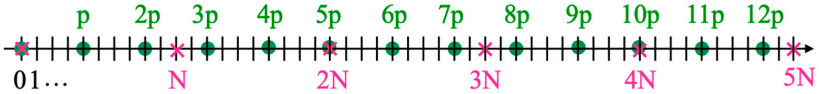


Figure 19. Multiples of N are enclosed by immediately succeeding multiples of p.

Note 8.

$$\forall k \in \mathbb{N} \exists t \in \mathbb{N} : (t-1)p \leq kN \leq tp.$$

Proof. Pick an arbitrary $k \in \mathbb{N}$, i.e., $kN \in \mathbb{N}$ is also given. Then, we find a $\tilde{t} \in \mathbb{N}$ such that $\tilde{t}p \geq kN$ (trivial because $p > 0$). Let t be the smallest of such $\tilde{t}$, i.e., $t \overset{def}{=} min\{\tilde{t} | \tilde{t}p \geq kN\}$. Thus, $(t-1)p \leq kN$ because otherwise $(t-1)p > kN$, which is a contradiction because t was chosen minimal.

Together, $(t-1)p \leq kN \leq tp$. The claim follows because k an arbitrary number. □

The situation we just discussed is shown in Figure 20.

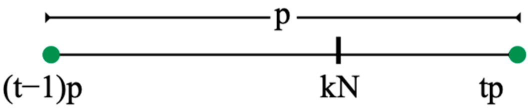


Figure 20. Determining the interval of succeeding multiples of p enclosing a multiple of N.

Furthermore, two different multiples of N are in different intervals defined by succeeding multiples of p. Otherwise, the situation of Figure 21 would imply that $N \leq p$, which is a contradiction as shown by the proof of Note 9 below.

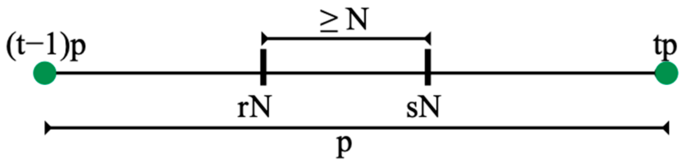


Figure 21. No two multiples of N can be enclosed by succeeding multiples of p.

We denote a t with $(t-1)p \leq kN \leq tp$ by t_k. This is justified by the next Note 9 which proves that such a t is uniquely determined by k. Especially, a multiple kN is contained in "its" interval:

$$\forall k \in \mathbb{N} \exists t_k \in \mathbb{N} : kN \in [(t_k-1)p, t_k p] \tag{32}$$

Note 9.

Let $k \in \mathbb{N}$ and $t_k \in \mathbb{N}$ with $(t_k-1)p \leq kN \leq t_k p$.
Then, $r \neq s \in \{0, \ldots, p-1\}$ implies $t_r \neq t_s$.

Proof (by contradiction). Assume $r \neq s$ but $t_r = t_s \overset{def}{=} t$ with $(t-1)p \leq rN \leq tp$ and $(t-1)p \leq sN \leq tp$ (see Figure 20). W.l.o.g. $r < s \Rightarrow r+1 \leq s \Rightarrow sN - rN \geq (r+1)N - rN = N$. Further, $sN - rN \leq tp - (t-1)p = p$. Together, it is $N \leq sN - rN \leq p$, i.e., $N < p$.

According to Lemma 8, we know $p < n \Rightarrow N < p < n < n^2$. However, by selection of N (see the bullet list at the beginning of Section 8), it is $n^2 < N$, which is a contradiction. □

The proof of Note 9 has shown especially:

Corollary 6.

$$r \neq s \in \{0, \ldots, p-1\} \;\Rightarrow\; rN \notin [(t_s-1)p, t_s p]\square$$

By Note 9, for $r \neq s \in \{0, \ldots, p-1\}$, the numbers t_r, t_s are different, i.e., for each $k \in \{0, \ldots, p-1\}$, such a unique t_k exists, i.e., the p numbers $t_0, t_1, \ldots, t_{p-1}$ are different. Thus:

Corollary 7.

There exist p different numbers t_k, $0 \leq k \leq p-1$, such that $(t_k-1)p \leq kN \leq t_k p$. □

These different numbers are strictly monotonically increasing.

Note 10.

Let $k \in \mathbb{N}$ and $t_k \in \mathbb{N}$ with $(t_k-1)p \leq kN \leq t_k p$. Then, $t_k < t_{k+1}$.
Thus, $t_0 < t_1 < \ldots < t_{p-1}$.

Proof (by contradiction). Assume $t_{k+1} \leq t_k$; thus, $t_{k+1} - 1 \leq t_k - 1$, which implies $t_{k+1}p \leq t_k p$ and $(t_{k+1}-1)p \leq (t_k-1)p$.

Now, $kN < (k+1)N$, $(t_k-1)p \leq kN \leq t_k p$, and $(t_{k+1}-1)p \leq (k+1)N \leq t_{k+1}p$. This implies $(k+1)N \leq t_{k+1}p \leq t_k p$ and $(t_k-1)p \leq kN < (k+1)N$, which finally results in $(t_k-1)p < (k+1)N \leq t_k p$—which is a contradiction to corollary 6 because this would imply that $(k+1)N \in [(t_k-1)p, t_k p]$. □

Each multiple kN of N is "close" to a multiple tp in the sense that kN is at most $p/2$ apart from $(t_k-1)p$ or $t_k p$ (see Figure 22).

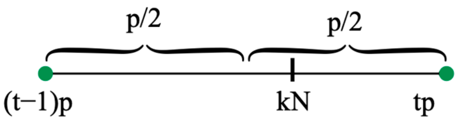


Figure 22. A multiple of N is always "close" to a multiple of p.

More precisely:

Note 11.

$$\forall k \in \mathbb{N} \exists t \in \mathbb{N} : |(t-1)p - kN| \leq \frac{p}{2} \vee |tp - kN| \leq \frac{p}{2}.$$

Proof. It is $(t-1)p \leq kN \leq tp$, i.e., by definition $kN \in [(t-1)p, tp]$. This implies $kN-(t-1)p \leq \frac{p}{2} \vee tp - kN \leq \frac{p}{2}$ (see Figure 22), otherwise:

$$kN-(t-1)p > \frac{p}{2} \wedge tp - kN > \frac{p}{2} \Leftrightarrow -(t-1)p > \frac{p}{2} - kN \wedge tp > \frac{p}{2} + kN$$

$\Rightarrow tp-(t-1)p > \frac{p}{2} - kN + \frac{p}{2} + kN = p$, but $tp-(t-1)p = p$, i.e., $p > p$, which is a contradiction! This proves the claim $|(t-1)p - kN| \leq \frac{p}{2} \vee |tp - kN| \leq \frac{p}{2}$. □

As before, from Note 9, it follows that for $k \in \{0, \ldots, p-1\}$, these numbers t are all different. Precisely:

Corollary 8.
Let $0 \leq k \leq p-1$ and $t_k \in \mathbb{N}$ such that $|(t_k-1)p - kN| \leq \frac{p}{2} \vee |t_k p - kN| \leq \frac{p}{2}$. If $r \neq s$, then $t_r \neq t_s$. □

The multiples of N are sparsely scattered across the intervals of consecutive multiples of p. More precisely, intervals of consecutive multiples of p, which contain a multiple of N, are not consecutive. This is the content of

Note 12.
Let $n > 2$, $k \in \{0, \ldots, p-1\}$, and $t_k \in \mathbb{N}$ with $(t_k-1)p \leq kN \leq t_k p$.
Then $t_{k+1} > t_k + 1$ as well as $t_{k-1} < t_k - 1$.

Proof. Because of Note 10, it is $t_{k+1} > t_k$; thus, $t_{k+1} \geq t_k + 1$.

Assumption. $t_{k+1} = t_k + 1$.
By definition, $(k+1)N \in [(t_{k+1}-1)p, t_{k+1}p]$, and by assumption, $t_{k+1} = t_k + 1$; thus, it is $(k+1)N \in [t_k p, (t_k+1)p]$. Furthermore, by definition, $kN \in [(t_k-1)p, t_k p]$ (see Figure 23).

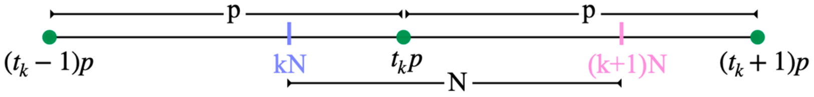


Figure 23. Situation in case kN and $(k+1)N$ lying within two consecutive intervals of consecutive multiples of p.

Now, $[(t_k-1)p, t_k p], [t_k p, (t_k+1)p] \subseteq [(t_k-1)p, (t_k+1)p]$ which implies $kN, (k+1)N \in [(t_k-1)p, (t_k+1)p]$.
Thus, $N = (k+1)N - kN \leq (t_k+1)p - (t_k-1)p = 2p$, i.e., $N \leq 2p$ (see Figure 23).
By Lemma 8, it is $p < n \Rightarrow 2p < 2n$. With $n > 2 \Rightarrow n^2 > 2n$ and by definition of N, it is $n^2 < N$; thus, $N > n^2 > 2n > 2p$: a contradiction!
Thus, the assumption is false, which implies $t_{k+1} > t_k + 1$. The claim $t_{k-1} < t_k - 1$ is proven similarly. □

The resulting geometric situation is depicted in Figure 24.

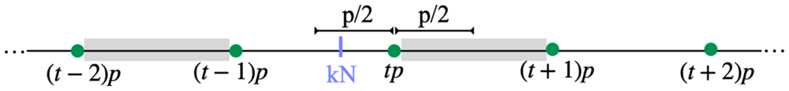


Figure 24. If kN is in an interval defined by two consecutive multiples of p, the preceding and succeeding intervals do not contain a multiple of N.

If $kN \in [(t_k-1)p, t_k p]$, it is $t_k p - kN \leq p/2$ or $kN - (t_k-1)p \leq p/2$ (see Figure 22 or Figure 24). In case $kN - (t_k-1)p \leq p/2$, we define $\hat{t} := t_k - 1$ and $kN - \hat{t}p \leq p/2$ results, and in case of $t_k p - kN \leq p/2$, we define $\hat{t} := t_k$ implying $\hat{t}p - kN \leq p/2$. According to Note 9, this $\hat{t}$ is uniquely defined. This proves

Note 13.
$\forall k \in \mathbb{N} \ \exists! \hat{t} \in \mathbb{N} : \left|\hat{t}p - kN\right| \leq \frac{p}{2}$, i.e., $\hat{t}$ is uniquely determined by k. □

This is next rewritten into a format more useful for what follows.

Note 14.
Let $k \in \{0, \ldots, p-1\}$ and $t_k \in \mathbb{N}$ with $t_k \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$.
If $r \neq s \in \{0, \ldots, p-1\}$, then $t_r \neq t_s$.

Proof. It is $t_k \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right] \Leftrightarrow k\frac{N}{p} - \frac{1}{2} \leq t_k \leq k\frac{N}{p} + \frac{1}{2} \Leftrightarrow kN - \frac{p}{2} \leq pt_k \leq kN + \frac{p}{2}$ $\Leftrightarrow -\frac{p}{2} \leq pt_k - kN \leq \frac{p}{2} \Leftrightarrow |pt_k - kN| \leq \frac{p}{2}$. Note 13 shows that t_k is uniquely determined by k. □

Finally, we can prove the following:

Corollary 9.
There exist p different numbers t_k, $0 \leq k \leq p-1$, such that

$$t_k \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$$

Proof. There exist p different numbers t_k, $0 \leq k \leq p-1$, such that $(t_k - 1)p \leq kN \leq t_k p$ (Corollary 7). The proof of Note 11 shows that this implies $|(t_k - 1)p - kN| \leq \frac{p}{2} \vee |t_k p - kN| \leq \frac{p}{2}$. The proof of Note 14 shows that this implies $t_k \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$. □

3.2.3. Cardinality of Pre-Images

First, we show that the parameter A is greater than 1, i.e., at least two numbers available in the $|a\rangle$ -part of the quantum register are mapped by f to the same value.

Note 15.

$$A > 1.$$

Proof. As reminded in the introduction, the quantum Fourier transform of Shor's algorithm produces the following state:

$$|a\rangle|b\rangle = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} |x\rangle|f(x)\rangle$$

After measurement of the $|b\rangle$ -part of the register, the $|a\rangle$ -part is in the state

$$|a\rangle = \frac{1}{\sqrt{A}}(|x\rangle + |x+p\rangle + |x+2p\rangle + \cdots + |x+(A-1)p\rangle) \tag{33}$$

i.e., $f^{-1}(|x\rangle) = \{|x\rangle, |x+p\rangle, |x+2p\rangle, \cdots, |x+(A-1)p\rangle\}$.

Choose $x < p$—such an x exists because otherwise it would be $p = 0$, but a period p satisfies $p > 0$. With $p < n$ (Lemma 8) and $n < n^2 < N$ (by choice of N), it is $x + p < 2p < N$ (see the proof of Note 12). Thus, $x + p$ is in the domain of f (in the sense that it is a value in the $|a\rangle$ -part of the quantum register available as an argument for f), i.e., $f(x+p)$ is available in the $|b\rangle$ -part of the register.

$A = 1$ would imply that $f^{-1}(|x\rangle) = \{|x\rangle\}$ and, thus, $|x+p\rangle \notin f^{-1}(|x\rangle)$, i.e., $f(|x\rangle) \neq f(|x+p\rangle)$ for $p \neq 0$. Since $p > 0$, the function f would not be periodic. □

Next, we prove tighter bounds for the parameter A.

Note 16.

$$(A-1)p < N < (A+1)p.$$

Proof. As in the proof of Note 15, we choose $x < p$. With

$$|a = \frac{1}{\sqrt{A}}(|x + |x+p + |x+2p + \cdots + |x+(A-1)p),$$

i.e., $\{|x\rangle, |x+p\rangle, |x+2p\rangle, \cdots, |x+(A-1)p\rangle\}$ are all values in the $|a\rangle$-part of the register being mapped to $f(x)$, i.e., A is the largest number satisfying $x + (A-1)p < N$. With $x \geq 0$, this implies $(A-1)p < N$—which is the first part of the claim.

Thus, $(A+1)p > N$. Otherwise, $(A+1)p \leq N$ and with $x < p$, it would be $x + Ap < p + Ap = (A+1)p \leq N$, i.e., $|x + Ap\rangle$ would also be in the $|a\rangle$-part of the register being mapped to $f(x)$, which is a contradiction to the definition of A. This proves the second part of the claim. □

The next estimation gives an approximation of N in terms of the product of A and p.

Note 17.

$$N \approx Ap.$$

Proof. Because of $(A-1)p < N < (A+1)p$, the geometric situation is as depicted in Figure 25, i.e., $N \in [(A-1)p, Ap]$ or $N \in [Ap, (A+1)p]$. Thus, $|N - Ap| \leq p$.

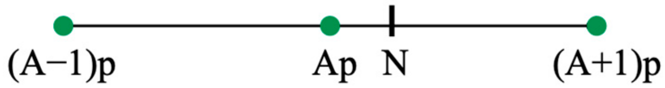


Figure 25. N is embraced by $(A-1)p$ and $(A+1)p$.

Now, $p < n$ (Lemma 8) and $n < n^2 < N$ by choice of N. In practice, n is a large number, i.e., n^2 is huge compared to n: $n \ll n^2 < N$. Together:

$$p \ll N \tag{34}$$

In this sense, p is a small number, i.e., $|N - Ap|$ is small too: $N \approx Ap$. □

3.2.4. Estimating Arguments of Amplitudes of Potential Measurement Results

The next Lemma is the main result of this section for what follows.

Lemma 9.

Let $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ and $k \in \{0, \ldots, p-1\}$. Then:

$$2\pi\frac{y(A-1)p}{N}, 2\pi\frac{py}{N} \in [-\pi, +\pi]$$

Proof. It is $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right] \Leftrightarrow k\frac{N}{p} - \frac{1}{2} \leq y \leq k\frac{N}{p} + \frac{1}{2} \Leftrightarrow$ (multiply with p) $kN - \frac{p}{2} \leq py \leq kN + \frac{p}{2} \Leftrightarrow -\frac{p}{2} \leq py - kN \leq \frac{p}{2} \Leftrightarrow$

$$\text{(i) } yp - kN \leq \frac{p}{2} \wedge kN - yp \leq \frac{p}{2}$$

By Note 16, it is

$$\text{(ii) } (A-1)p < N \Rightarrow \frac{(A-1)p}{N} < 1$$

Furthermore:

$$\text{(iii)}\ \frac{(A-1)p}{N} = \frac{Ap}{N} - \frac{p}{N} \overset{(A)}{\approx} \frac{N}{N} - \frac{p}{N} = 1 - \frac{p}{N} \overset{(B)}{\approx} 1$$

where (A) is because of Note 17 ($N \approx Ap$), and (B) is because of Equation (34) ($p \ll N$).

Next, we compute the lower bound for the first fraction of the claim:

$$\begin{aligned} 2\pi\tfrac{y(A-1)p}{N} &= 2\pi\tfrac{(A-1)}{N}yp \\ &\overset{(C)}{\geq} 2\pi\tfrac{(A-1)}{N}\left(kN - \tfrac{p}{2}\right) \\ &= 2\pi(A-1)k - \pi\tfrac{(A-1)p}{N} \\ &\overset{(D)}{\geq} -\pi\tfrac{(A-1)p}{N} \overset{(E)}{\approx} -\pi \end{aligned}$$

where (C) follows from the second inequation of (i) above, (D) is because of $2\pi(A-1)k \geq 0$, and (E) is implied by (iii) above.

The upper bound for the first fraction of the claim is computed next:

$$\begin{aligned} 2\pi\tfrac{y(A-1)p}{N} &= 2\pi\tfrac{(A-1)}{N}yp \overset{(F)}{\leq} 2\pi\tfrac{(A-1)}{N}\left(kN + \tfrac{p}{2}\right) \\ &= 2\pi(A-1)k + \pi\tfrac{(A-1)p}{N} \overset{(G)}{<} 2\pi(A-1)k + \pi \\ &\overset{(H)}{<} 2\pi(A-1)p + \pi \overset{(I)}{<} 2\pi N + \pi \overset{(J)}{\equiv} \pi \end{aligned}$$

where (F) is implied by the first inequation of (i) above, (G) is (ii) above, (H) follows from the prerequisite $k \in I\{0,\ldots,p-1\}$, i.e., $k < p$, and (I) is the first inequation of (ii) above. Finally, we will estimate $e^{i\varphi}$, and because of $e^{i2\pi N} = 1$, (J) is justified.

Together, $-\pi \leq 2\pi\frac{y(A-1)p}{N} \leq \pi$, which proves the first claim. $\square_{(firstfraction)}$

Next,

$$\begin{aligned} 2\pi\tfrac{py}{N} &= \tfrac{2\pi}{N}py \overset{(K)}{\leq} \tfrac{2\pi}{N}\left(kN + \tfrac{p}{2}\right) \\ &= 2\pi k + \pi\tfrac{p}{N} \overset{(L)}{<} 2\pi k + \pi \overset{(M)}{\equiv} \pi \end{aligned}$$

with (K) from the first inequation of (i) before, (L) because $p < N$, and (M) because we will estimate $e^{i\varphi}$. I.e., the upper bound of the second fraction is as claimed.

The correctness of the lower bound is seen as follows:

$$\begin{aligned} 2\pi\tfrac{py}{N} &= \tfrac{2\pi}{N}py \overset{(N)}{\geq} \tfrac{2\pi}{N}\left(kN - \tfrac{p}{2}\right) \\ &= 2\pi k - \pi\tfrac{p}{N} \overset{(O)}{>} -\pi\tfrac{p}{N} \overset{(Q)}{>} -\pi \end{aligned}$$

with the second inequation of (i) before giving (N), (O) is because of $2\pi k > 0$, and (Q) is true because $0 < p < N$; thus, $0 < p/N < 1$. $\square_{(secondfraction)}$ $\square$

3.3. Estimating Probabilities

We are now ready to compute the probability $P(y)$ that the state $|\mathrm{y}\rangle$, which is prepared by the quantum part of Shor's algorithm, is "close" (i.e., within a distance of $1/2$) to a multiple of p/N.

Lemma 10.

Assume $q = e^{i2\pi\frac{yp}{N}} \neq 1$ and let P be the probability that $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ for a $k \in \{0,\ldots,p-1\}$. Then, $P \approx \frac{4}{\pi^2}$.

Proof. According to Equation (29), the probability $P(y)$ to measure a particular $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ is

$$P(y) = \frac{1}{NA}\left|\frac{1-q^A}{1-q}\right|^2 \tag{35}$$

In case $q \neq 1$ (which is the assumption) where $q = e^{i2\pi\frac{yp}{N}}$ (the case $q = 1$ will be treated separately in Note 18). Thus, with $q^A = e^{i2\pi\frac{yAp}{N}}$, it is

$$\left|\frac{1-q^A}{1-q}\right| = \frac{\left|1-q^A\right|}{|1-q|} = \frac{\left|1-e^{i2\pi\frac{yAp}{N}}\right|}{\left|1-e^{i2\pi\frac{yp}{N}}\right|} \tag{36}$$

The structure of the numerator and denominator recommends the estimation of both by means of the Lemma 7 (Secant Length Estimation). However, applying Lemma 7 requires that $2\pi\frac{yAp}{N}, 2\pi\frac{yp}{N} \in [-\pi, +\pi]$. By Lemma 9, we know that under the prerequisite $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$, it is $2\pi\frac{py}{N} \in [-\pi, +\pi]$ as well as $2\pi\frac{y(A-1)p}{N} \in [-\pi, +\pi]$, but Lemma 9 does not imply $2\pi\frac{yAp}{N} \in [-\pi, +\pi]$.

Now, consider the following calculation:

$$\left|\frac{1-q^A}{1-q}\right| = \left|\frac{1-q^{A-1}}{1-q} + q^{A-1}\right| \overset{(A)}{\geq} \left|\frac{1-q^{A-1}}{1-q}\right| - \left|q^{A-1}\right| \overset{(B)}{=} \left|\frac{1-q^{A-1}}{1-q}\right| - 1 \tag{37}$$

where (A) holds because of $|a+b| \geq |a| - |b|$, and $\left|e^{i\varphi}\right| = 1$ implies (B): $\left|q^t\right| = \left|\left(e^{i2\pi\frac{yp}{N}}\right)^t\right| = \left|e^{i(2\pi\frac{ypt}{N})}\right| = 1$.

Equation (37) allows us to apply the secant length estimation (Lemma 7) because in

$$\left|\frac{1-q^{A-1}}{1-q}\right| = \frac{\left|1-q^{A-1}\right|}{|1-q|} = \frac{\left|1-e^{i2\pi\frac{y(A-1)p}{N}}\right|}{\left|1-e^{i2\pi\frac{yp}{N}}\right|} \tag{38}$$

it is now $2\pi\frac{y(A-1)p}{N}, 2\pi\frac{yp}{N} \in [-\pi, +\pi]$ according to Lemma 9.

First, we use the second inequation of $\frac{2|\varphi|}{\pi} \leq \left|1-e^{i\varphi}\right| \leq |\varphi|$ from Lemma 7 with $\varphi = 2\pi\frac{yp}{N} \in [-\pi, +\pi]$ and obtain

$$\left|1-e^{i2\pi\frac{yp}{N}}\right| \leq 2\pi\frac{yp}{N} \tag{39}$$

Then, we use the first inequation of $\frac{2|\varphi|}{\pi} \leq \left|1-e^{i\varphi}\right| \leq |\varphi|$ from Lemma 7 with $\varphi = 2\pi\frac{y(A-1)p}{N} \in [-\pi, +\pi]$ and obtain

$$\left|1-e^{i2\pi\frac{y(A-1)p}{N}}\right| \geq \frac{2}{\pi} \cdot 2\pi\frac{y(A-1)p}{N} \tag{40}$$

Using Equations (39) and (40) in Equation (38) results in

$$\left|\frac{1-q^{A-1}}{1-q}\right| = \frac{\left|1-e^{i2\pi\frac{y(A-1)p}{N}}\right|}{\left|1-e^{i2\pi\frac{yp}{N}}\right|} \geq \frac{2}{\pi} \cdot 2\pi y\frac{(A-1)p}{N} \cdot \frac{N}{2\pi yp} = \frac{2(A-1)}{\pi} \tag{41}$$

This result is now used in Equation (37) (step (C) below) and we obtain

$$\begin{aligned}\left|\frac{1-q^A}{1-q}\right| = \left|\frac{1-q^{A-1}}{1-q}\right| - 1 \qquad & \overset{(C)}{\geq} \frac{2(A-1)}{\pi} - 1 \\ & = \frac{2A}{\pi} - \frac{2}{\pi} - 1 = \frac{2A}{\pi} - \left(\frac{2}{\pi}+1\right)\end{aligned} \tag{42}$$

Using Equation (42) in Equation (35) (step (D) below) results in

$$\begin{aligned} P(y) = \frac{1}{NA}\left|\frac{1-q^A}{1-q}\right|^2 & \overset{(D)}{\geq} \frac{1}{NA}\left(\frac{2A}{\pi} - \left(\frac{2}{\pi}+1\right)\right)^2 \\ &= \frac{1}{NA}\left(\frac{4A^2}{\pi^2} - \frac{4A}{\pi}\left(\frac{2}{\pi}+1\right) + \left(\frac{2}{\pi}+1\right)^2\right) \\ &= \frac{1}{NA}\left(\frac{4A^2}{\pi^2} - \frac{8A}{\pi^2} - \frac{4A}{\pi} + \frac{4}{\pi^2} + \frac{4}{\pi} + 1\right) \\ &= \frac{4A}{\pi^2 N} - \frac{8}{\pi^2 N} - \frac{4}{\pi N} + \frac{4}{\pi^2 NA} + \frac{4}{\pi NA} + \frac{1}{NA} \\ &\geq \frac{4A}{\pi^2 N} - \frac{8}{\pi^2 N} - \frac{4}{\pi N} = \frac{4A}{\pi^2 N} - \frac{4}{\pi N}\left(1+\frac{2}{\pi}\right) \end{aligned}$$

Thus,

$$P(y) \geq \frac{4A}{\pi^2 N} - \frac{4}{\pi N}\left(1+\frac{2}{\pi}\right) \tag{43}$$

According to Note 17, we know $N \approx Ap \Rightarrow \frac{A}{N} \approx \frac{1}{p}$, i.e.,

$$\frac{4A}{\pi^2 N} \approx \frac{4}{\pi^2 p} \tag{44}$$

Furthermore, since N is a "huge" number, we know that the following is "small":

$$\frac{4}{\pi N}\left(1+\frac{2}{\pi}\right) \overset{def}{=} \varepsilon \tag{45}$$

Using Equations (44) and (45) in Equation (43) results in

$$P(y) \geq \frac{4}{\pi^2}\frac{1}{p} - \varepsilon \tag{46}$$

for each $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$. According to Corollary 9, there exist p different numbers y_k with $y_k \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ and for each of them $P(y_k) \geq \frac{4}{\pi^2}\frac{1}{p} - \varepsilon$. Since we are not interested in a particular y_k, but in any of them, we need to sum up all probabilities $P(y_k)$ to obtain the overall probability P:

$$P = \sum_{i=0}^{p-1} P(y_i) \geq \frac{4}{\pi^2} - p\varepsilon \approx \frac{4}{\pi^2}$$

This proves the claim. □

We still need to estimate the probability for the case $q = 1$.

Note 18.
Let $q = 1$. Then $P(y) = \frac{A}{N}$.

Proof. In case $q = 1$, the probability is

$$P(y) = \frac{1}{NA}\left|\sum_{j=0}^{A-1} q^A\right|^2 = \frac{1}{NA}\left|\sum_{j=0}^{A-1} 1\right|^2 = \frac{1}{NA}A^2 = \frac{A}{N} \ \square$$

3.4. Computing the Period

Let y be the result of the measurement produced by Shor's algorithm. Under the assumption that $q \neq 1$, the following holds:

Theorem 15.
With probability $P \approx \frac{4}{\pi^2}$, there exists a $k \in \{0, \ldots, p-1\}$, such that

$$\left|\frac{y}{N} - \frac{k}{p}\right| < \frac{1}{2p^2}$$

Proof. According to Lemma 10, the probability that $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ for a $k \in \{0, \ldots, p-1\}$ is $\approx 4/\pi^2$.

However, $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right] \Leftrightarrow -\frac{1}{2} \leq y - \frac{kN}{p} \leq +\frac{1}{2}$. Dividing the latter inequations by N yields: $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right] \Leftrightarrow -\frac{1}{2N} \leq \frac{y}{N} - \frac{k}{p} \leq +\frac{1}{2N}$.

Thus, $\left|\frac{y}{N} - \frac{k}{p}\right| \leq \frac{1}{2N}$. By choice of N, it is $n^2 < N$. Furthermore, $p < n \Rightarrow p^2 < n^2 \Rightarrow p^2 < N \Rightarrow \frac{1}{N} < \frac{1}{p^2}$. This results in $\left|\frac{y}{N} - \frac{k}{p}\right| \leq \frac{1}{2N} < \frac{1}{2p^2}$. $\square$

Legendre's Theorem (Theorem 14) proves immediately:

Theorem 16.
With probability $\approx 4/\pi^2$, k/p is a convergent of y/N. $\square$

3.4.1. Determining the Period by Convergents: q $\neq$ 1

The Algorithm 2 determines with probability of approximately $4/\pi^2$ the period p we are looking for; is is applicable in the case $q \neq 1$:

Algorithm 2 Determining with probability of approximately $4/\pi^2$ the period p we are looking for

```
1. Compute $\frac{y}{N} \in \mathbb{Q}_{>0}$;
   a. The result of the measurement is $y \in \mathbb{N}$ and $N \in \mathbb{N}$ has been chosen
      $\Rightarrow \frac{y}{N} \in \mathbb{Q}_{>0}$ can be computed.
2. Compute the continued fraction representation $[a_0; a_1, \ldots, a_m]$ of $\frac{y}{N} \in \mathbb{Q}$;
3. Compute the convergents $[a_0; a_1, \ldots, a_u] = \frac{g_u}{h_u}$, $1 \leq u \leq m$;
4. Determine $h_\omega$ with $h_\omega \geq h_u$ for $1 \leq u \leq m$ and $h_\omega < n$
   $\Rightarrow \frac{g_\omega}{h_\omega}$ is a very good approximation of $\frac{k}{p}$ because $\frac{1}{2h_\omega^2} \leq \frac{1}{2h_u^2}$;
5. Thus, $h_\omega \approx p$ is a candidate for the period p;
6. Check whether p is in fact the period.
```

3.4.2. Determining the Period by Convergents: q = 1

In case $q = 1$, the above algorithm is not applicable. However, $q = 1 \Leftrightarrow e^{\frac{2\pi i}{N}py} = 1 \Leftrightarrow \frac{py}{N} \in \mathbb{Z} \Leftrightarrow p = k\frac{N}{y}$ with $k \in \mathbb{Z}$. Thus, the Algorithm 3 can be used:

Algorithm 3 $q = 1 \Leftrightarrow e^{\frac{2\pi i}{N} py} = 1 \Leftrightarrow \frac{py}{N} \in \mathbb{Z} \Leftrightarrow p = k\frac{N}{y}$ with $k \in \mathbb{Z}$

```
1.  Compute $\frac{N}{y} \in \mathbb{Q}_{>0}$. The result of the measurement is $y \in \mathbb{N}$ and $N \in \mathbb{N}$ has been chosen $\Rightarrow$
    $\frac{N}{y} \in \mathbb{Q}_{>0}$ can be computed;
2.  Select $k \in \mathbb{N}$;
3.  Compute $k\frac{N}{y}$;
4.  If $k\frac{N}{y} \notin \mathbb{N}$, go back to step (2);
5.  If $k\frac{N}{y} \geq n$, go back to step (2);
6.  $p = k\frac{N}{y}$ is a candidate for the period p;
7.  Check whether p is in fact the period;
8.  If p is not the period:
    a.  If some predefined termination criterion is met: stop;
    b.  Go back to step (2).
```

This may yield the period p but does not guarantee it.

3.5. *How the Presented Results Relate*

The contribution contains several low-level details. In order to avoid getting lost in these details, this section sketches how the main details contribute to the proof of Shor's algorithm. The Figure 26 at the end of this section is a cartoon of these relations.

3.5.1. Applying the Results about Continued Factions

Determining a divisor and finally a prime factor of a natural number $n \in \mathbb{N}$ can be reduced to finding the period p of the modular exponentiation function $f(x) = a^x mod n$ for an a with $0 < a < n$—see Section 3.2.1 and Note 7.

The quantum part of Shor's algorithm produces the state $\frac{1}{\sqrt{NA}} \sum_{j=0}^{A-1} \omega_N^{jpy} |y\rangle$ from Equation (2). Measuring this state results in a natural number $y \in \mathbb{N}$.

The natural number N in Equation (2) must be chosen in advance based on the number n to be factorized: it is chosen as $N = 2^m$ with $n^2 < N < 2n^2$—see the introduction of Section 8. This ensures that the relevant arguments to compute the $f(x)$ by the quantum part of Shor's algorithm can be captured as quantum states.

Theorem 15 guarantees with probability $P \approx 4/\pi^2$ the existence of a $k \in \{0, \ldots, p-1\}$ such that $\left|\frac{y}{N} - \frac{k}{p}\right| < \frac{1}{2p^2}$.

Thus, according to the convergent criterion of Legendre's Theorem (Theorem 14), k/p is a convergent of y/N.

The proof of Legendre's convergent criterion (Theorem 14), in turn, is based on the fact that convergents are exactly the best approximations of the second kind (Theorem 11 and Lagrange's theorem (Theorem 12)).

The proof of Lagrange's theorem (Theorem 12—each convergent is a best approximation of the second kind) makes use of the recursion theorem (Theorem 1), the sign theorem (Theorem 2), the monotony property of denominators of convergents (Corollary 1), as well as the estimations of the upper bounds of convergents (Lemma 3) and their lower bounds (Lemma 5).

The proof of Theorem 11 (each best approximation of the second kind is a convergent) makes use of the recursion theorem (Theorem 1), the distance theorem (Theorem 8), the computation of the difference of convergents (Corollary 3), the distance of fractions (Note 4), and the estimation of the upper bounds of convergents (Lemma 3).

The estimations of the lower and upper bounds of convergents depend on the distance theorem (Theorem 8), the computation of the difference of convergents (Corollary 3), the monotony property of denominators of convergents (Corollary 1), and on estimations of the size of denominators of convergents (Lemma 1). The estimation of the lower bounds

(Lemma 5) makes use of semiconvergents (Definition 4) and their monotony property (Lemma 4) as well as the nesting theorem (Theorem 7).

Remark: Theorem 13, which proves that best approximations of the first kind are convergents or semiconvergents, is not immediately relevant to Shor′s algorithm and may be ignored when focusing on Shor′s algorithm.

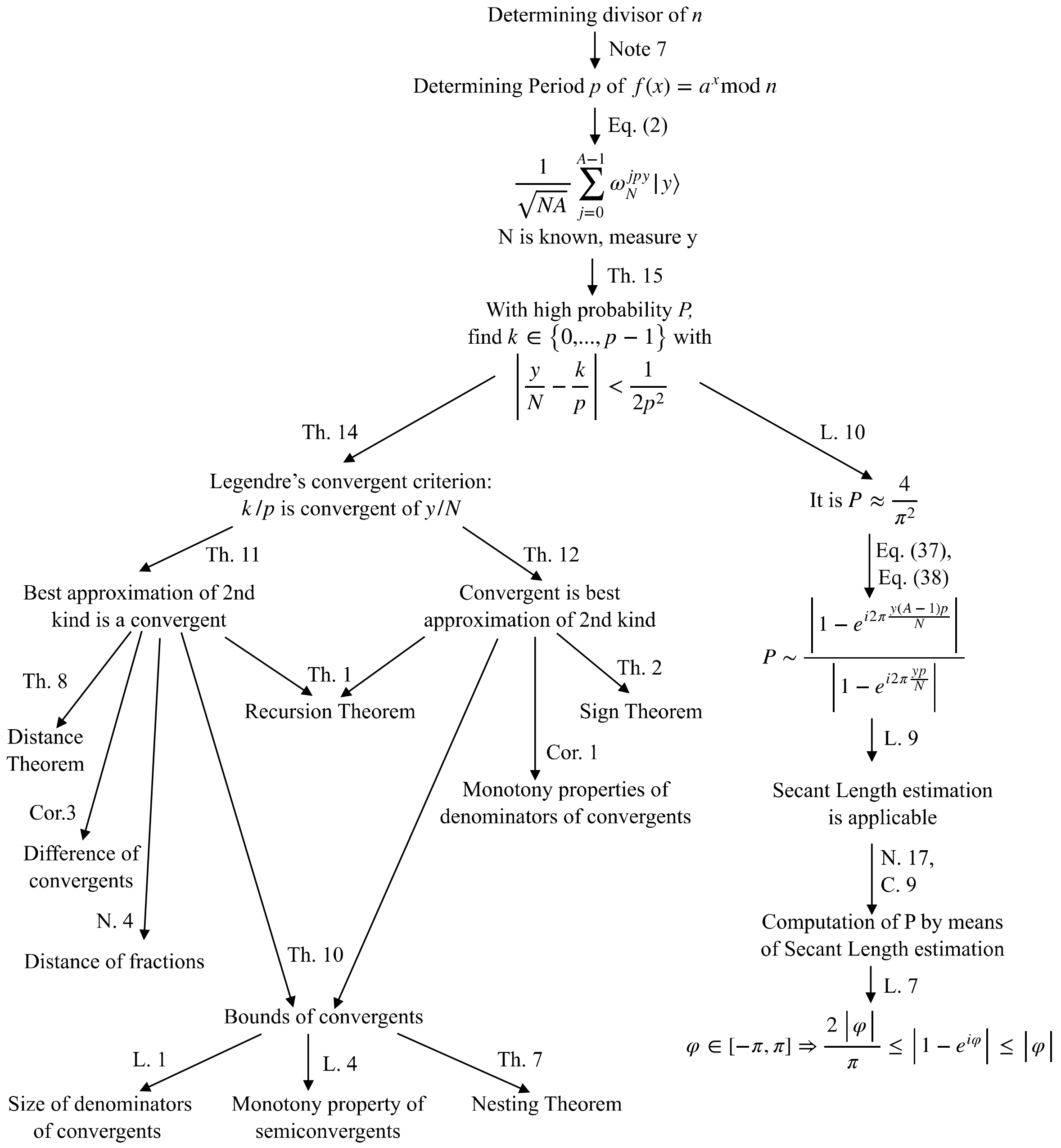

Determining divisor of n
Note 7
Determining Period p of $f(x) = a^x \bmod n$
Eq. (2)
$\frac{1}{\sqrt{NA}} \sum_{j=0}^{A-1} \omega_N^{jpy} |y\rangle$
N is known, measure y
Th. 15
With high probability P,
find $k \in \{0,...,p-1\}$ with
$\left|\frac{y}{N} - \frac{k}{p}\right| < \frac{1}{2p^2}$
Th. 14
L. 10
Legendre’s convergent criterion:
k/p is convergent of y/N
It is $P \approx \frac{4}{\pi^2}$
Th. 11
Th. 12
Eq. (37),
Eq. (38)
Best approximation of 2nd kind is a convergent
Convergent is best approximation of 2nd kind
$P \sim \frac{\left|1 - e^{i2\pi\frac{y(A-1)p}{N}}\right|}{\left|1 - e^{i2\pi\frac{yp}{N}}\right|}$
Th. 8
Th. 1
Recursion Theorem
Th. 2
Sign Theorem
Distance Theorem
Cor. 1
L. 9
Monotony properties of denominators of convergents
Secant Length estimation is applicable
Cor.3
Difference of convergents
N. 17,
C. 9
N. 4
Computation of P by means of Secant Length estimation
Distance of fractions
Th. 10
L. 7
Bounds of convergents
$\varphi \in [-\pi, \pi] \Rightarrow \frac{2|\varphi|}{\pi} \leq |1 - e^{i\varphi}| \leq |\varphi|$
L. 1
L. 4
Th. 7
Size of denominators of convergents
Monotony property of semiconvergents
Nesting Theorem

Figure 26. How the main results of the paper relate.

3.5.2. Applying Probability Estimations

According to Equation (29) (which is implied by the Born rule), the probability $P(y)$ to measure a particular y is $P(y) = \frac{1}{NA}\left|\frac{1-q^A}{1-q}\right|^2$ for $q = e^{i2\pi\frac{py}{N}} \neq 1$.

Equations (37) and (38) show that this probability can be estimated as $\left|\frac{1-q^A}{1-q}\right| \geq \left|\frac{1-q^{A-1}}{1-q}\right| - 1 = \frac{\left|1-e^{i2\pi\frac{y(A-1)p}{N}}\right|}{\left|1-e^{i2\pi\frac{yp}{N}}\right|} - 1$. The latter fraction, in turn, can be estimated by means of Lemma 7 (Secant Length Estimation) in case $2\pi\frac{y(A-1)p}{N}, 2\pi\frac{py}{N} \in [-\pi, +\pi]$.

Lemma 9 shows that the latter inclusion is satisfied in case of $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ and $k \in \{0, \ldots, p-1\}$.

Lemma 10 proves that with probability $P(y) \approx 4/\pi^2$ it is, in fact, $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ for $k \in \{0, \ldots, p-1\}$. The proof of this lemma is based on a proper estimation of N (Note 17) which in turn relies on Note 16. Further, it makes use of Corollary 9, which is the summary of the various results of Section 3.2.2.

A simple calculation in the proof of Theorem 15 finally shows that $y \in \left[k\frac{N}{p} - \frac{1}{2}, k\frac{N}{p} + \frac{1}{2}\right]$ implies $\left|\frac{y}{N} - \frac{k}{p}\right| \leq \frac{1}{2N} < \frac{1}{2p^2}$. Thus, with probability $P(y) \approx 4/\pi^2$, the convergent criterion of Legendre's Theorem (Theorem 14) is satisfied.

4. Conclusions and Related Work

The literature analyzing, discussing, and refining Shor's algorithm [1] is vast. Of course, most text books on quantaum computing explain the algorithm too (e.g., [4,5]). In doing so, all this literature puts a sharp focus on the quantum part of the algorithm and sketches its classical parts at various depths. However, the mathematical treatment of the classical aspects is sketchy, omitting most of the details and leaving them as an exercise for the reader with references to corresponding text books from mathematics such as [6] or [7]. The lecture notes by Preskill [2] go a bit deeper, especially on the estimation of probabilities, but still omit the low-level details; however, the authors of the contribution at hand benefited a lot by the treatment in [2]. It is noted that the genesis for the authors' treatment of probability estimations was inspired by unpublished, non-public work to which the authors had access to several years ago.

In doing so, the contribution at hand is very detailed on the probability estimation of being able to use Legendre's Theorem in Shor's algorithm. The authors are not aware of any other publication providing these low-level details.

Furthermore, the contribution at hand is a self-contained treatment on continued fractions up to Legendre's Theorem. All background that is needed to understand this theorem is presented, including all proofs with low-level details step by step.

The authors hope to foster the comprehension of the classical aspects of Shor's algorithm even at the level of beginners in quantum computing.

Author Contributions: Writing—original draft, F.L.; Writing—review & editing, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the BMWK project PlanQK (01MK20005N).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shor, P.W. Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer. *SIAM J. Sci. Stat. Comput.* **1997**, *26*, 1484–1509. [CrossRef]
2. Preskill, J. *Lecture on Quantum Information—Chapter 6. Quantum Algorithms*; California Institute of Technology: Pasadena, CA, USA, 2020. Available online: http://theory.caltech.edu/~{}preskill/ph219/chap6_20_6A.pdf (accessed on 11 July 2022).
3. Shult, E.; Surowski, D. *Algebra*; Springer: Berlin/Heidelberg, Germany, 2015.

4. Nielsen, M.A.; Chuang, I.L. *Quantum Computation and Quantum Information*; Cambridge University Press: Cambridge, UK, 2016.
5. Rieffel, E.; Polak, W. *Quantum Computing: A Gentle Introduction*; The MIT Press: Cambridge, MA, USA, 2011.
6. Hardy, G.H.; Wright, E.M. *An Introduction to the Theory of Numbers*, 4th ed.; Oxford University Press: New York, NY, USA, 1975.
7. Khinchin, A.Y. *Continued Fractions*, 3rd ed.; The University of Chicago Press: Chicago, IL, USA, 1964.