# On Best Rational Approximations Using Large Integers

David T. Ashley, Joseph P. DeVoe, Karl Perttunen, Cory Pratt
Visteon Corporation
and
Anatoly Zhigljavsky
University Of Cardiff

Computer processors are equipped with instructions to multiply and divide very large integers. These instructions can be used to economically implement linear scalings from an integer domain to an integer range by choosing a rational number $r_A = h/k$, calculating the product $hx$ using an integer multiplication instruction, and applying an integer division instruction to form $\lfloor hx/k \rfloor$. This paper presents a novel $O(log\ max(h_{MAX}, k_{MAX}))$ algorithm based on continued fractions for finding the closest rational number $r_A = h/k$ to an arbitrary real number $r_I$ subject to constraints $h \leq h_{MAX}$ and $k \leq k_{MAX}$. Novel results are presented bounding the maximum distance between available choices of $r_A$ when $r_A$ will be chosen only in an interval $[l, r]$, utilizing a second novel $O(log\ max(h_{MAX}, k_{MAX}))$ continued fraction algorithm. Novel results bounding the error due to the necessity of an integer domain and range are presented. The results and techniques presented have relevance to scientific computing (where integer operations may execute much more quickly than floating point operations), to consumer electronics and embedded real-time systems (where the processor may have integer multiplication and division instructions, but no floating-point capability), to the design of special-purpose digital logic (which may implement multiplication and division in hardware), and to the design of mechanical systems (two gears meshed together mechanically implement a ratio which is the ratio of their numbers of teeth).

Categories and Subject Descriptors: G.1.2 [**Numerical Analysis**]: Approximation—*linear approximation*

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Rational approximation, Farey series, continued fraction, approximation error, integer lattice.

## 1. INTRODUCTION

Modern computer instruction sets contain instructions for multiplication and division of large integers. In many applications, the mainstay of efficient software design is the ability to phrase a computational problem in a form which is economically executed by the hardware available. In a very capable processor (such as a workstation or supercomputer), approximations involving only integers may be attractive because integer instructions execute more quickly than floating-point instructions, or because the processor design allows them to execute concurrently with floating-point instructions. In very inexpensive processors (such as those used in consumer electronics), approximations involving only integers may be attractive because the processor has no floating-point capability.

This paper presents results and techniques for making optimal use of integer multiplication and division instructions to approximate functions of the form $F(x) = r_I x$, $r_I \in \mathbb{R}^+$ using functions in the form of (1).[1]

$$J(x) = \lfloor r_A \lfloor x \rfloor \rfloor = \left\lfloor \frac{h\lfloor x \rfloor}{k} \right\rfloor ; h \in \mathbb{Z}^+, \leq h_{MAX}; k \in \mathbb{N}, \leq k_{MAX}. \qquad (1)$$

Because modern processors can multiply and divide *very* large integers (32- and 64-bit integers are typical), choosing $h$ and $k$ so as to place $r_A = h/k$ as close as possible to an arbitrary $r_I \in \mathbb{R}^+$ involves a very large search space, and an efficient algorithm is necessary for computational viability.

Section 2 presents a summary of important properties of the Farey series, and Section 3 presents a summary of important properties of the apparatus of continued fractions.[2]

Section 4 presents a novel $O(log\ k_{MAX})$ continued fraction algorithm for finding the best rational approximations $r_A = h/k$ to an arbitrary $r_I \in \mathbb{R}^+$ subject to the constraint $k \leq k_{MAX}$. Section 5 extends the algorithm of Section 4 to the case where both $h$ and $k$ are constrained, $h \leq h_{MAX} \wedge k \leq k_{MAX}$; and presents a novel $O(log\ max(h_{MAX}, k_{MAX}))$ continued fraction algorithm for finding the best rational approximations in the rectangular area of the integer lattice formed by the constraints.

Section 6 presents novel results bounding the distance between rational numbers in a rectangular area of the integer lattice when $r_I \in [l, r]$. Section 7 presents a method for bounding the end-to-end approximation error as a function of $r_A - r_I$.

Section 8 provides a practical design example illustrating the techniques.

## 2. THE FAREY SERIES OF ORDER $N$

The *Farey series of order* $N$, denoted $F_N$, is the ordered set of all irreducible rational numbers $h/k$ in the interval [0,1] with a denominator $k \leq N$. For example, the Farey series of order 7, $F_7$, is

$$F_7 = \left\{ \frac{0}{1}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{2}{7}, \frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{1}{2}, \frac{4}{7}, \frac{3}{5}, \frac{2}{3}, \frac{5}{7}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{1}{1} \right\}. \qquad (2)$$

The distribution of Farey rational numbers in [0,1] is repeated in any $[n, n+1]$, $n \in \mathbb{Z}^+$; so the distribution of Farey rationals in [0,1] supplies complete information

[1] Mnemonic for $r_I$ and $r_A$: $I$=ideal, $A$=actual. In this paper, $\mathbb{R}^+$, $\mathbb{Z}^+$, and $\mathbb{N}$ are the sets of non-negative real numbers, non-negative integers, and positive integers, respectively.

[2] The algorithms presented are based on the properties of the Farey series and the apparatus of continued fractions—because these are topics from number theory that seldom find application in practical computer arithmetic, a summary is necessary for readability.

about the distribution in all of $\mathbb{R}^+$.[3]

### 2.1 Properties Of Sequential Elements

Theorem 1. *If $H/K$ and $h/k$ are two successive terms of $F_N$, then*

$$Kh - Hk = 1. \tag{3}$$

*Note:* This condition is necessary but not sufficient for $h/k$ to be the Farey successor of $H/K$. In general, there is more than one $h/k$ with $k \leq N$ such that $Kh - Hk = 1$.

Proof. See [1] p.23, [6] p.222. □

Theorem 2. *If $H/K$ is a term of $F_N$, the successor of $H/K$ in $F_N$ is the $h/k$ satisfying $Kh - Hk = 1$ with the largest denominator $k \leq N$.*

Proof. Any potential successor of $H/K$ which meets $Kh - Hk = 1$ can be formed by adding $1/Kk$ to $H/K$ (4).

$$Kh - Hk = 1 \rightarrow \frac{h}{k} = \frac{1 + Hk}{Kk} = \frac{H}{K} + \frac{1}{Kk} \tag{4}$$

If $h/k$ and $h'/k'$ both satisfy $Kh - Hk = 1$ with $k' < k \leq N$, then $H/K < h/k < h'/k'$. Thus the $h/k$ with the largest $k \leq N$ that meets $Kh - Hk = 1$ is the successor in $F_N$ to $H/K$. □

Theorem 3. *If $H/K$ and $h/k$ are two successive terms of $F_N$, then*

$$K + k > N. \tag{5}$$

*Note:* This condition is necessary but not sufficient for $h/k$ to be the Farey successor of $H/K$.

Proof. See [1] p.23. □

Theorem 4. *If $h_{j-2}/k_{j-2}$, $h_{j-1}/k_{j-1}$, and $h_j/k_j$ are three consecutive terms of $F_N$, then:*

$$h_j = \left\lfloor \frac{k_{j-2} + N}{k_{j-1}} \right\rfloor h_{j-1} - h_{j-2} \tag{6}$$

$$k_j = \left\lfloor \frac{k_{j-2} + N}{k_{j-1}} \right\rfloor k_{j-1} - k_{j-2} \tag{7}$$

[3] We stretch the proper nomenclature by referring to sequential rational numbers outside the interval $[0, 1]$ as Farey terms or as part of $F_N$, which, in the strictest sense, they are not. All of the results presented in this paper (except Sections 2.2 and 2.3) apply everywhere in $\mathbb{R}^+$, and this abuse is not harmful.

*Notes:* (1)Theorem 4 gives recursive formulas for generating successive terms in $F_N$ if two consecutive terms are known. (2)Equations (6) and (7) can be solved to allow generation of terms in the decreasing direction (8, 9).

$$h_j = \left\lfloor \frac{k_{j+2} + N}{k_{j+1}} \right\rfloor h_{j+1} - h_{j+2} \tag{8}$$

$$k_j = \left\lfloor \frac{k_{j+2} + N}{k_{j+1}} \right\rfloor k_{j+1} - k_{j+2} \tag{9}$$

Proof. See [9] p.83. □

In general, given only a single irreducible rational number $h/k$, there is no method to find the immediate predecessor or successor in $F_N$ without some iteration (Equations 6, 7, 8, and 9 require two successive elements).

### 2.2 Number Of Elements

The number of elements in $F_N$ is approximately $3N^2/\pi^2$.[4] $F_{255=2^8-1}$ contains about 20,000 elements, $F_{65,535=2^{16}-1}$ contains about 1.3 billion elements, $F_{2^{32}-1}$ contains about $5.6 \times 10^{18}$ elements, and $F_{2^{64}-1}$ contains about $1.0 \times 10^{38}$ elements.

The large numbers of elements in the Farey series of the orders used in practice make it impractical to linearly search the Farey series to find the best rational approximations.[5]

### 2.3 Probabilistic Results On $|r_I - r_A|$

If rational numbers of the form $r_A = h/k$, subject to the constraint $k \leq k_{MAX}$, are used to approximate arbitrary real numbers $r_I$, it might not be clear how close we can "typically" choose $r_A$ to an aribtrary $r_I$ under the constraint. We consider different asymptotics for the precision of the approximation of an arbitrary $r_I$ by a rational number $r_A = h/k$ with $k \leq k_{MAX}$. For simplicity of notation we denote $\alpha = r_I$ and $N = k_{\mathrm{MAX}}$ and assume, without loss of generality, that $\alpha \in [0, 1]$.

We are thus interested in the asymptotic behaviour, when $N \to \infty$, of the quantity

$$\rho_N(\alpha) = \min_{h/k \in F_N} |\alpha - h/k| \ ,$$

which is the distance between $\alpha$ and $F_N$, the Farey series of order $N$.

The worst–case scenario is not very interesting: from the construction of the Farey series we observe that for a fixed $N$ the longest intervals between the neighbours of $F_N$ are $[0, 1/N]$ and $[1 - 1/N, 1]$ and therefore for all $N$

$$\max_{\alpha \in [0,1]} \rho_N(\alpha) = \frac{1}{2N} \, . \tag{10}$$

[4] This is a classic result from number theory, and its basis isn't discussed here. In this instance we mean $F_N$ strictly in the interval $[0, 1]$.

[5] For example, a particularly naive approach might be to start at an integer $i$, where three successive terms in $F_N$ are $(Ni - 1)/N$, $i/1$, and $(Ni + 1)/N$, and to use (6) through (9) to linearly search upward or downward until the real number of interest is enclosed. Even in $F_{2^{32}-1}$, searching 1,000,000 rational numbers per second, such a search would require up to about 90,000 years.

(This supremum is achieved at the points $1/(2N)$ and $1-1/(2N)$.)

However, such behaviour of $\rho_N(\alpha)$ is not typical: as is shown below, typical values of the approximation error $\rho_N(\alpha)$ are much smaller.

First consider the behaviour of $\rho_N(\alpha)$ for almost all $\alpha \in [0,1]$.[6] We then have (see [3], [2]) that for almost all $\alpha \in [0,1]$ and any $\varepsilon > 0$, (11) and (12) hold.

$$\lim_{N\to\infty} \rho_N(\alpha) N^2 \log^{1+\varepsilon} N = +\infty, \quad \liminf_{N\to\infty} \rho_N(\alpha) N^2 \log N = 0 \tag{11}$$

$$\limsup_{N\to\infty} \frac{\rho_N(\alpha)N^2}{\log N} = +\infty, \quad \lim_{N\to\infty} \frac{\rho_N(\alpha)N^2}{\log^{1+\varepsilon} N} = 0 \tag{12}$$

Even more is true: in (11) and (12) one can replace $\log N$ by $\log N \log\log N$, $\log N \log\log N \log\log\log N$, and so on. Analogously, $\log^{1+\varepsilon} N$ could be replaced by $\log N (\log\log N)^{1+\varepsilon}$, $\log N \log\log N (\log\log\log N)^{1+\varepsilon}$, and so on.

These statements are analogues of Khinchin's metric theorem, the classic result in metric number theory, see e.g. [2].

The asymptotic distribution of the suitably normalised $\rho_N(\alpha)$ was derived in [4]. A main result of this paper is that the sequence of functions $N^2\rho_N(\alpha)$ converges in distribution, when $N \to \infty$, to the probability measure on $[0,\infty)$ with the density given by (13).

$$p(\tau) = \begin{cases} 6/\pi^2, & \text{if } 0 \le \tau \le \frac{1}{2} \\ \frac{6}{\pi^2\tau}\left(1+\log\tau-\tau\right), & \text{if } \frac{1}{2} \le \tau \le 2 \\ \frac{3}{\pi^2\tau}\left(2\log(2\tau) - 4\log(\sqrt{\tau}+\sqrt{\tau-2}) - (\sqrt{\tau}-\sqrt{\tau-2})^2\right), & \text{if } 2 \le \tau < \infty \end{cases} \tag{13}$$

This means that for all $a, A$ such that $0 < a < A < \infty$, (14) applies, where 'meas' denotes for the standard Lebesgue measure on $[0,1]$.

$$\text{meas}\{\alpha \in [0,1] :\; a < N^2\rho_N(\alpha) \le A\} \to \int_a^A p(\tau)d\tau \quad \text{as } N \to \infty \tag{14}$$

Another result in [4] concerns the asymptotic behavior of the moments of the approximation error $\rho_N(\alpha)$. It says that for any $\delta \ne 0$ and $N \to \infty$, (15) applies, where $\zeta(\cdot)$ and $\mathrm{B}(\cdot,\cdot)$ are the Riemann zeta–function and the Beta–function, respectively.

[6] A statement is true for almost all $\alpha \in [0,1]$ if the measure of the set where this statement is wrong has measure zero.

$$\frac{\delta+1}{2}\int_0^1 \rho_N^\delta(\alpha)d\alpha = \begin{cases} \infty, & \text{if } \delta \le -1 \\ \frac{3}{\delta^2\pi^2}\left(2^{-\delta}+\delta 2^{\delta+2}\mathrm{B}(-\delta,\frac{1}{2})\right)N^{-2\delta}\,(1+o(1))\,, & \text{if } -1<\delta<1, \delta\neq 0 \\ \frac{3}{\pi^2}N^{-2}\log N + O\left(N^{-2}\right), & \text{if } \delta = 1 \\ 2^{-\delta}\frac{\zeta(\delta)}{\zeta(\delta+1)}N^{-\delta-1}+O\left(N^{-2\delta}\right), & \text{if } \delta > 1 \end{cases} \tag{15}$$

In particular, the average of the approximation error $\rho_N(\alpha)$ asymptotically equals

$$\int_0^1 \rho_N(\alpha)d\alpha = \frac{3}{\pi^2}\frac{\log N}{N^2} + O\left(\frac{1}{N^2}\right), \quad N \to \infty\,. \tag{16}$$

Comparison of (16) with (12) shows that the asymptotic behavior of the average approximation error $\int \rho_N(\alpha)d\alpha$ resembles the behavior of the superior limit of $\rho_N(\alpha)$. Even this limit decreases much faster than the maximum error $\max_\alpha \rho_N(\alpha)$, see (10): for typical $\alpha$ the rate of decrease of $\rho_N(\alpha)$, when $N \to \infty$, is, roughly speaking, $1/N^2$ rather than $1/N$, the error for the worst combination of $\alpha$ and $N$.

These results show that there is a significant advantage to using the Farey series as the set from which to choose rational approximations, rather than more naively using only rational numbers with the maximum denominator $k_{MAX}$ (as is often done in practice).

## 3. THE APPARATUS OF CONTINUED FRACTIONS

An *n-th order finite simple continued fraction* is a fraction in the form of (17), where $a_0 \in \mathbb{Z}^+$ and $a_k \in \mathbb{N}$ for $k > 0$. To ensure that two continued fractions which represent the same number can't be written differently, we also require that the final element $a_n$ not be equal to 1 (except when representing the integer 1).[7] A continued fraction in the form of (17) is denoted $[a_0; a_1, a_2, \ldots, a_n]$, and each $a_k$ is called an *element* or *partial quotient.*

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ldots + \cfrac{1}{a_n}}}} = [a_0; a_1, a_2, \ldots, a_n] \tag{17}$$

Continued fractions provide an alternate apparatus for representing real numbers. The form of (17) has important properties which are presented without proof.

- Every rational number can be represented by a finite simple continued fraction $[a_0; a_1, a_2, \ldots, a_n]$.

[7] If $a_n = 1$, the continued fraction can be reduced in order by one and $a_{n-1}$ can be increased by one while still preserving the value of the continued fraction.

- Each unique $[a_0; a_1, a_2, \ldots, a_n]$ corresponds to a uniquely valued rational number.

Without proof, we present the following algorithm for finding partial quotients $a_k$ of an arbitrary non-negative rational number $a/b$.

Algorithm 1.

```
• $k := -1$.
• $divisor_{-1} := a$.
• $remainder_{-1} := b$.
• Repeat
    • $k := k + 1$.
    • $dividend_k := divisor_{k-1}$.
    • $divisor_k := remainder_{k-1}$.
    • $a_k := dividend_k \; div \; divisor_k$.
    • $remainder_k := dividend_k \; mod \; divisor_k$.
• Until $(remainder_k = 0)$.
```

Without proof, we present the following properties of Algorithm 1.

- The algorithm will produce the same $[a_0; a_1, a_2, \ldots, a_n]$ for any $(ia)/(ib)$, $i \in \mathbb{N}$, i.e. the rational number $a/b$ need not be reduced before applying the algorithm.
- The algorithm will always terminate (i.e. the continued fraction representation $[a_0; a_1, a_2, \ldots, a_n]$ will be finite).

The apparatus of continued fractions is best viewed as an alternate apparatus for representing real numbers, and Algorithm 1 is best viewed as an algorithm for determining in which partition a rational number lies, in the same sense that long division successively partitions a rational number as each successive decimal digit is obtained. To say that the first three digits of a real number $x$ are "3.14" is logically equivalent to saying that $3.14 \leq x < 3.15$ (i.e. that $x$ lies in a certain partition). In the same sense, (18), (19), and (20) are valid equivalences.

$$(x = [a_0] \vee x = [a_0; \ldots]) \leftrightarrow (a_0 \leq x < a_0 + 1) \tag{18}$$

$$(x = [a_0; a_1] \vee x = [a_0; a_1, \ldots]) \leftrightarrow \left( a_0 + \frac{1}{a_1 + 1} < x \leq a_0 + \frac{1}{a_1} \right) \tag{19}$$

$$\begin{gathered} (x = [a_0; a_1, a_2] \vee x = [a_0; a_1, a_2, \ldots]) \\ \updownarrow \\ \left( a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2}} \leq x < a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + 1}} \right) \end{gathered} \tag{20}$$

The form of (18), (19), and (20) could be continued indefinitely to show the defining inequality for higher-order partitions. From the form of (18), (19), and

(20) it can be readily seen that irrational numbers have a non-terminating continued fraction representation, and that the algorithm for finding that representation would be symbolic and involve successively determining higher order partial quotients (i.e. at each step, in which partition the irrational number lies). The algorithm for determining the partial quotients of an irrational number isn't discussed in this paper. In most practical applications, $r_I$ is known empirically to at least several decimal places, and the most practical technique is to use the best known decimal approximation as the starting point to apply Algorithm 1.

The *kth convergent* of a finite simple continued fraction $[a_0; a_1, a_2, \ldots, a_n]$, denoted $s_k = p_k/q_k$, is the rational number corresponding to the continued fraction $[a_0; a_1, a_2, \ldots, a_k]$, $k \leq n$. Equations (21) through (26) define the canonical way to construct all $s_k = p_k/q_k$ from all $a_k$.

$$p_{-1} = 1 \tag{21}$$

$$q_{-1} = 0 \tag{22}$$

$$p_0 = a_0 = \lfloor r_I \rfloor \tag{23}$$

$$q_0 = 1 \tag{24}$$

$$p_k = a_k p_{k-1} + p_{k-2} \tag{25}$$

$$q_k = a_k q_{k-1} + q_{k-2} \tag{26}$$

When $p_k$ and $q_k$ (the numerator and denominator of the $k$th convergent $s_k$) are formed as specified by (21) through (26), convergents $s_k = p_k/q_k$ have the following properties, which are presented without proof.

- Each even-ordered convergent $s_k = p_k/q_k = [a_0; a_1, a_2, \ldots, a_k]$ is less than $[a_0; a_1, a_2, \ldots, a_n]$, and each odd-ordered convergent $s_k$ is greater than $[a_0;\ a_1,\ a_2,\ \ldots,\ a_n]$, with the exception of the final convergent $s_k$, $k = n$, which is equal to $[a_0; a_1, a_2, \ldots, a_n]$.
- Each convergent is irreducible; that is, $p_k$ and $q_k$ are coprime.
- Each $q_k$ is greater than $q_{k-1}$; that is, the denominators of convergents are ever-increasing. Furthermore, the denominators of convergents increase at a minimum rate that is exponential (Eq. 27), [5] Theorem 12.

$$q_k \geq 2^{\frac{k-1}{2}},\ k \geq 2 \tag{27}$$

An *intermediate fraction* is a fraction of the form

$$\frac{ip_k + p_{k-1}}{iq_k + q_{k-1}},\ i < a_{k+1}. \tag{28}$$

It can be seen by comparing (28) with (25) and (26) that an intermediate fraction can be denoted compactly by the continued fraction representation of a convergent, with the final element adjusted downward. For example, if $[a_0; a_1, a_2, \ldots, a_{k-1}]$ and $[a_0; a_1, a_2, \ldots, a_{k-1}, a_k]$, $k \leq n$, are convergents; $[a_0; a_1, a_2, \ldots, a_{k-1}, 1]$, $[a_0; a_1, a_2, \ldots, a_{k-1}, 2]$, $\ldots$, and $[a_0; a_1, a_2, \ldots, a_{k-1}, a_k - 1]$ are intermediate fractions.

## 4. CHOOSING $R_A = H/K$ SUBJECT TO $K \leq K_{MAX}$

Finding the best rational approximation $r_A = h/k$ to an arbitrary $r_I \in \mathbb{R}^+$ subject only to the constraint $k \leq k_{MAX}$ is equivalent to the problem of finding the two members of $F_{k_{MAX}}$ which enclose $r_I$. Potential naive algorithms include building $F_{k_{MAX}}$ starting at an integer $[O(k_{MAX}^2)]$, building $F_{k_{MAX}}$ starting at a rational number with a large prime denominator $[O(k_{MAX})]$, and building the Stern-Brocot tree $[O(k_{MAX})]$. For $k_{MAX}$ of a few hundred or less, any of these algorithms are satisfactory, and they can be carried out even with ordinary spreadsheet software, such as *Microsoft Excel*.

However, for $k_{MAX}$ typical of the more powerful microcontrollers used in consumer electronics ($2^{16}$ or $2^{32}$), and particularly for $k_{MAX}$ reflecting the integer arithmetic capability of workstations and supercomputers ($2^{32}$, $2^{64}$, or larger), $O(k_{MAX}^2)$ and $O(k_{MAX})$ algorithms are not computationally viable. This section presents a novel $O(log\ k_{MAX})$ algorithm which is suitable for finding best rational approximations even in Farey series of very large order, based on the apparatus of continued fractions.

### 4.1 Finding Best Rational Approximations With $r_I \notin F_{k_{MAX}}$

THEOREM 5. *For a non-negative rational*[8] *number $a/b$ not in $F_N$ which has a continued fraction representation $[a_0; a_1, a_2, \ldots, a_n]$, the highest-order convergent $s_k = p_k/q_k$ with $q_k \leq N$ is one neighbor*[9] *to $a/b$ in $F_N$, and the other neighbor in $F_N$ is*[10]

[8] Although it isn't discussed in this paper, it isn't required that a number be rational in order to apply this theorem. As emphasized by (18), (19), and (20), the process of obtaining continued fraction partial quotients is essentially a process of determining in which partition a number lies. All numbers in the same partition—rational or irrational—have the same Farey neighbors in all Farey series up to a certain order. If the partial quotients of an irrational number can be obtained up through $a_k$ s.t. $s_k = p_k/q_k$ is the highest-order convergent with $q_k \leq N$, then this theorem can be applied. Knowledge of all $a_0 \ldots a_k$ is equivalent to the knowledge that the number is in a partition where all numbers in that partition have the same Farey neighbors in all Farey series up through order $q_{k+1} - 1$.

[9] By neighbors in $F_N$ we mean the rational numbers in $F_N$ immediately to the left and immediately to the right of $a/b$.

[10] Theorem 5 is a somewhat stronger statement about best approximations than Khinchin makes in [5], Theorem 15. We were not able to locate this theorem or a proof in print, but this theorem is understood within the number theory community. It appears on the Web page of David Eppstein in the form of a 'C'-language computer program, `http://www.ics.uci.edu/~eppstein/numth/frap.c`. Although Dr. Eppstein phrases the solution in terms of modifying a partial quotient, his approach is equivalent to (29).

$$\frac{\left\lfloor \dfrac{N - q_{k-1}}{q_k} \right\rfloor p_k + p_{k-1}}{\left\lfloor \dfrac{N - q_{k-1}}{q_k} \right\rfloor q_k + q_{k-1}}. \tag{29}$$

PROOF. First, it is proved that the highest-order convergent $s_k = p_k/q_k$ with $q_k \leq N$ is one of the two neighbors to $a/b$ in $F_N$. $s_k \in F_N$, since $q_k \leq N$. By [5], Theorem 9, the upper bound on the difference between $a/b$ and arbitrary $s_k$ is given by

$$\left| \frac{a}{b} - \frac{p_k}{q_k} \right| < \frac{1}{q_k q_{k+1}}. \tag{30}$$

For two consecutive terms in $F_N$, $Kh - Hk = 1$. For a Farey neighbor $H/K$ to $s_k$ in $F_N$, (31) must hold.

$$\frac{1}{q_k N} \leq \left| \frac{H}{K} - \frac{p_k}{q_k} \right| \tag{31}$$

$q_{k+1} > N$, because $q_{k+1} > q_k$ and $p_k/q_k$ was chosen to be the highest-order convergent with $q_k \leq N$. Using this knowledge and combining (30) and (31) leads to (32).

$$\left| \frac{a}{b} - \frac{p_k}{q_k} \right| < \frac{1}{q_k q_{k+1}} < \frac{1}{q_k N} \leq \left| \frac{H}{K} - \frac{p_k}{q_k} \right| \tag{32}$$

This proves that $s_k$ is one neighbor to $a/b$ in $F_N$. The apparatus of continued fractions ensures that the highest order convergent $s_k$ with $q_k \leq N$ is closer to $a/b$ than to any neighboring term in $F_N$. Thus, there is no intervening term of $F_N$ between $s_k$ and $a/b$. If $k$ is even, $s_k < a/b$, and if $k$ is odd, $s_k > a/b$.

It must be proved that (29) is the other Farey neighbor. For brevity, only the case of $k$ even is proved: the case of $k$ odd is symmetrical. (29) is of the form (33), where $i \in \mathbb{Z}^+$.

$$\frac{ip_k + p_{k-1}}{iq_k + q_{k-1}} \tag{33}$$

$k$ is even, $s_k < a/b$, and the two Farey terms enclosing $a/b$, in order, are

$$\frac{p_k}{q_k}, \frac{ip_k + p_{k-1}}{iq_k + q_{k-1}}. \tag{34}$$

Applying the $Kh - Hk = 1$ test, (35), gives the result of 1, since by theorem ([5], Theorem 2), $q_k p_{k-1} - p_k q_{k-1} = (-1)^k$.

$$(q_k)(ip_k + p_{k-1}) - (p_k)(iq_k + q_{k-1}) = 1 \tag{35}$$

Thus, every potential Farey neighbor of the form (33) meets the $Kh - Hk = 1$ test. It is also straightforward to show that *only* potential Farey neighbors of the form (33) can meet the $Kh - Hk = 1$ test, using the property that $p_k$ and $q_k$ are coprime.

It must be established that a rational number of the form (33) is irreducible. This result comes directly from (35), since if the numerator and denominator of (29) or (33) are not coprime, the difference of 1 is not possible.

The denominator of (29) can be rewritten as

$$N - [(N - q_{k-1}) \bmod q_k] \in \{N - q_k + 1, ..., N\}\,. \tag{36}$$

It must be shown that if one irreducible rational number—namely, the rational number given by (29)—with a denominator $\in \{N - q_k + 1, \ldots, N\}$ meets the $Kh - Hk = 1$ test, there can be no other irreducible rational number in $F_N$ with a larger denominator which also meets this test.

Given (36), and given that *only* rational numbers of the form (33) can meet the $Kh - Hk = 1$ test, and given that any number of the form (33) is irreducible, the irreducible number meeting the $Kh - Hk = 1$ test with the next larger denominator after the denominator of (29) will have a denominator $\in \{N+1, \ldots, N+q_k\}$. Thus, no other irreducible rational number in $F_N$ besides that given by (29) with a larger denominator $\leq N$ and which meets the $Kh - Hk = 1$ test can exist; therefore (29) is the other enclosing Farey neighbor to $a/b$ in $F_N$. □

Theorem 5 suggests an algorithm for determining best approximations to a rational $r_I = a/b \notin F_{k_{MAX}}$ subject to the constraint $k \leq k_{MAX}$.

Algorithm 2.

```
• $k := -1$.
• $divisor_{-1} := a$.
• $remainder_{-1} := b$.
• $p_{-1} := 1$.
• $q_{-1} := 0$.
• Repeat
    • $k := k + 1$.
    • $dividend_k := divisor_{k-1}$.
    • $divisor_k := remainder_{k-1}$.
    • $a_k := dividend_k \; div \; divisor_k$.
    • $remainder_k := dividend_k \; mod \; divisor_k$.
    • If $k = 0$ then $p_k := a_k$ else $p_k := a_k p_{k-1} + p_{k-2}$.
    • If $k = 0$ then $q_k := 1$ else $q_k := a_k q_{k-1} + q_{k-2}$.
• Until $(q_k > k_{MAX})$.
• $s_{k-1} = p_{k-1}/q_{k-1}$ will be one Farey neighbor to $a/b$ in $F_{k_{MAX}}$. Apply (29) to
  obtain the other Farey neighbor.
```

Algorithm 2 builds the partial quotients $a_k$ and convergents $s_k = p_k/q_k$ of $a/b$ only as far as required to obtain the highest-order convergent with $q_k \leq N$; thus the number of iterations required is tied to $k_{MAX}$, rather than to the precision of

$a/b$. It is easy to see that Algorithm 2 is $O(log\ k_{MAX})$, since the the denominators of convergents $q_k$ have a minimum exponential rate of increase (27).[11]

### 4.2 Finding Best Rational Approximations With $r_I \in F_{k_{MAX}}$

The case where $r_I \in F_{k_{MAX}}$ corresponds to the case where $r_I$ is at the edge of a partition, in the sense suggested by (18), (19), and (20). In this case, the highest-order convergent $s_n = p_n/q_n = r_I \in F_{k_{MAX}}$, and (29) supplies the right Farey neighbor to $r_I$ if $n$ is even, or the left Farey neighbor to $r_I$ if $n$ is odd. In the former case (8) and (9) can be used to obtain the left Farey neighbor, and in the latter case (6) and (7) can be used to obtain the right Farey neighbor. The second half of the proof of Theorem 5 applies.

Thus, finding the neighbors in $F_{k_{MAX}}$ to an arbitrary $r_I = a/b \in F_{k_{MAX}}$ is also an $O(log\ k_{MAX})$ procedure, and easily accomplished using the apparatus of continued fractions.

## 5. CHOOSING $R_A = H/K$ SUBJECT TO $H \leq H_{MAX}$ AND $K \leq K_{MAX}$

Up to this point, only the case of constrained $k$ has been considered. However, in a practical application, $h$ is also typically constrained, usually by the size of the operands and results that machine multiplication and division instructions can accomodate.

When $h$ and $k$ are both constrained, $h \leq h_{MAX} \wedge k \leq k_{MAX}$, the set of rational numbers $h/k$ that can be formed has a convenient and intuitive graphical interpretation (Figure 1 illustrates this interpretation with $h_{MAX} = 3$ and $k_{MAX} = 5$). Each rational number $h/k$ that can be formed under the constraints corresponds to a point in the integer lattice.

From Figure 1, it is clear that:

- The angle $\theta$ of the ray from the origin to $(k, h)$ is monotonically increasing with respect to the value of $h/k$, and:
  - $h/k = tan\theta$.
  - $\theta = tan^{-1}h/k$.
- The smallest rational number that can be formed under the constraints is $0/1$, the smallest non-zero rational number is $1/k_{MAX}$, and the largest rational number is $h_{MAX}/1$.

[11] Although Algorithm 2 is the best known algorithm for finding Farey neighbors, it is an over-simplification to state that Algorithm 2 is $O(log\ k_{MAX})$. In the classical sense—speaking only in terms of numbers of operations and assuming that each type of operation takes the same amount of time regardless of the data—the algorithm is $O(log\ k_{MAX})$. However, when applying the algorithm for $a$, $b$, and $k_{MAX}$ much larger than the native data sizes of the computer used, one must use some sort of arbitrary-precision or long integer calculation package, and the calculation times of such packages are probably between $O(log\ N)$ and $O(N)$ with respect to the data values. Taking this into account, the algorithm may be as poor as $O(N\ log\ N)$ for data much larger than accomodated by the computer used. However, this is not an impediment to practical calculations. The rational approximation software packaged with this paper (submitted to CALGO) will find neighbors within the Farey series of order $2^{128}$ with a calculation time of just a few seconds on a personal computer, and will find neighbors within the Farey series of order $2^{1,000}$ with a calculation time of less than 60 seconds.

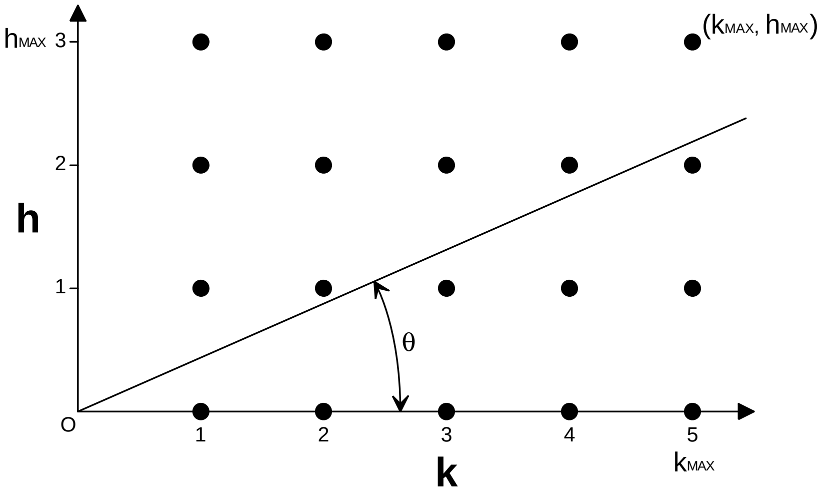

Fig. 1. Integer Lattice Interpretation Of Rational Numbers $h/k$ Formable Under Constraints $h \leq h_{MAX}$ And $k \leq k_{MAX}$

- Only irreducible rational numbers are directly "visible" from the origin (reducible numbers are hidden "behind" the irreducible numbers, when viewed from the origin).
- The ascending set of irreducible rational numbers that can be formed subject to the constraints can be constructed graphically by sweeping a line starting at $\theta = 0$ through the range $0 \leq \theta < \pi/2$, recording each integer lattice point that is directly "visible" from the origin.
- For $r_A = h/k \leq h_{MAX}/k_{MAX}$, the constraint $k \leq k_{MAX}$ is the dominant constraint, and the set of formable rational numbers $\leq h_{MAX}/k_{MAX}$ is simply $F_{k_{MAX}}$.

By symmetry in Figure 1, it can be seen that each formable rational number $\geq h_{MAX}/k_{MAX}$ is the reciprocal of an element of the Farey series of order $h_{MAX}$. Thus, it is clear that the set of formable rational numbers under the constraints $h \leq h_{MAX} \wedge k \leq k_{MAX}$ can be built by concatenating a portion of $F_{k_{MAX}}$ with a portion of $F_{h_{MAX}}$, but with the terms of $F_{h_{MAX}}$ inverted and reversed in order.

We denote the series formed from $F_N$ by inverting each element (except 0/1) and reversing the order as $F_{\overline{N}}$. For example, using this definition,

$$F_{\overline{3}} = \left\{ \ldots, \frac{3}{8}, \frac{2}{5}, \frac{3}{7}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}, \frac{3}{2}, \frac{2}{1}, \frac{3}{1} \right\}. \tag{37}$$

We denote the series formed by concatenating $F_{k_{MAX}}$ up through $h_{MAX}/k_{MAX}$ to $F_{\overline{h_{MAX}}}$ from $h_{MAX}/k_{MAX}$ through $h_{MAX}/1$ as $F_{k_{MAX},\overline{h_{MAX}}}$.

Note that $F_{k_{MAX},\overline{h_{MAX}}}$ is the ordered set of all irreducible rational numbers that can be formed subject to $h \leq h_{MAX} \wedge k \leq k_{MAX}$. For example, using this definition,

$$F_{5,\overline{3}} = \left\{ \frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}, \frac{3}{2}, \frac{2}{1}, \frac{3}{1} \right\}. \tag{38}$$

It can be verified that the result in (38) is the same as would be obtained in Figure 1 by sweeping a line from the origin counterclockwise through $0 \leq \theta < \pi/2$, recording each point of the integer lattice directly "visible" from the origin.

The following $O(log\ max(h_{MAX}, k_{MAX}))$ algorithm is presented for finding the neighbors in $F_{k_{MAX},\overline{h_{MAX}}}$ to an arbitrary irreducible rational number $a/b$.

ALGORITHM 3.

- If $a/b < h_{MAX}/k_{MAX}$, apply Algorithm 2 directly;
- Else if $a/b > h_{MAX}/k_{MAX}$, apply Algorithm 2 using $b/a$, rather than $a/b$ as $r_I$, and using $N = h_{MAX}$ rather than $N = k_{MAX}$, and invert and transpose the two Farey neighbors obtained;
- Else if $a/b = h_{MAX}/k_{MAX}$, apply both steps above: the first step to obtain the left neighbor in $F_{k_{MAX},\overline{h_{MAX}}}$ and the second step to obtain the right neighbor.

## 6. CHOOSING $R_A = H/K$ ONLY IN AN INTERVAL $[L, R]$

It is clear from the earlier discussion of the Farey series that the maximum distance between terms in $F_{k_{MAX}}$ is $1/k_{MAX}$, and that this maximum distance occurs only adjacent to an integer. It is also clear from the discussion of $F_{\overline{h_{MAX}}}$ that the maximum distance between terms is 1.

Thus, when we use $F_{k_{MAX},\overline{h_{MAX}}}$ to approximate real numbers, in general the worst-case distance between terms is 1.

In practical applications when rational approximation is used, the approximation tends to be used over a restricted interval $[l \gg 0, r \ll h_{MAX}]$ rather than over the full range of the rational numbers that can be formed, $[0, h_{MAX}]$. This section develops novel upper bounds on the distance between terms of $F_{k_{MAX},\overline{h_{MAX}}}$ in an interval $[l, r]$. For simplicity, assume $l, r \in F_{k_{MAX},\overline{h_{MAX}}}$.

Three distinct cases are developed (Figure 2). The upper bound developed from Case III is always larger than the upper bound developed from Case II, which is always larger than the upper bound developed from Case I; so if only the absolute maximum error over the interval $[l, r]$ is of interest, only the highest-numbered case which applies needs to be evaluated. However, some applications may have different error requirements in different regions of the interval $[l, r]$, and for these applications it may be beneficial to analyze more than one case.

### 6.1 Case I: $r_I < h_{MAX}/k_{MAX}$

With $r_I < h_{MAX}/k_{MAX}$, $k \leq k_{MAX}$ is the dominant constraint, and the neighbors available to $r_I$ are simply the terms of $F_{k_{MAX}}$. If $[l, r] \cap [0, h_{MAX}/k_{MAX}]$ includes an integer, clearly the maximum distance from $r_I$ to the nearest available term of $F_{k_{MAX},\overline{h_{MAX}}}$ is given by

$$\left| \frac{h}{k} - r_I \right| \leq \frac{1}{2k_{MAX}}. \tag{39}$$

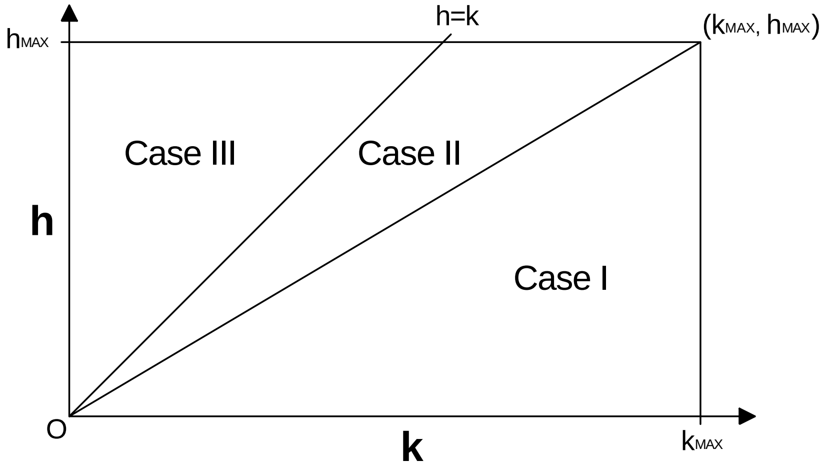

Fig. 2. Three Cases For Bounding Distance Between Terms In $F_{k_{MAX},\overline{h_{MAX}}}$

If $[l, r] \cap [0, h_{MAX}/k_{MAX}]$ does not include an integer, it can be shown that the maximum distance between Farey terms is driven by the rational number with the smallest denominator in the interval.

For two consecutive terms $p/q$ and $p'/q'$ in $F_{k_{MAX}}$, $p'q - pq' = 1$ (Theorem 1), so that

$$\frac{p'}{q'} - \frac{p}{q} = \frac{p'q - pq'}{qq'} = \frac{1}{qq'}. \tag{40}$$

By Theorem 3, $q + q' > k_{MAX}$, therefore

$$\frac{1}{qk_{MAX}} \leq \frac{1}{qq'} < \frac{1}{q(k_{MAX} - q)}. \tag{41}$$

Let $q_{MIN}$ be the smallest denominator of any rational number $\in F_{k_{MAX}}$ in the interval $[l, r]$. It is then easy to show that for any consecutive denominators $q, q'$ which occur in $F_{k_{MAX}}$ in the interval $[l, r]$,

$$\frac{1}{qq'} < \frac{1}{q_{MIN}\; max(q_{MIN}, k_{MAX} - q_{MIN})}. \tag{42}$$

Thus, the upper bound on the distance between consecutive terms of $F_{k_{MAX}}$ in an interval $[l, r]$ is tied to the minimum denominator of any rational number $\in F_{k_{MAX}}$ in $[l, r]$.

Note that clearly $q_{MIN} \leq 1/(r - l)$, so for most practical intervals $[l, r]$, the search for $q_{MIN}$ would not be computationally expensive. However, applications could arise where an approximation is used in an *extremely* narrow interval, and having an algorithm available that is computationally viable for such cases is advantageous. For example, locating the rational number $\in F_{2^{20,000}}$ with the smallest denominator in an interval of width $2^{-10,000}$ could be a serious computational problem.

To locate $q_{MIN}$ in $[l, r]$, note that at least one rational number with $q_{MIN}$ as a denominator in $[l, r]$ is the best approximation of order $q_{MIN}$ to the midpoint of the interval, $(l + r)/2$.[12] By theorem ([5], Theorem 15), every best approximation of a number is a convergent or intermediate fraction of the continued fraction representation of the number. We seek the convergent or intermediate fraction of $(l + r)/2$ with the smallest denominator that is in the interval $[l, r]$.

The convergents and intermediate fractions of $(l + r)/2$ are naturally arranged in order of increasing denominator. However, it would be inefficient to test *every* intermediate fraction for membership in $[l, r]$, as partial quotients $a_k$ are unlimited in size and such an algorithm may not be $O(log\ k_{MAX})$. Instead, since intermediate fractions are formed using the parameterized expression $(ip_k + p_{k-1})/(iq_k + q_{k-1})$, and since intermediate fractions are ever-increasing or ever-decreasing with respect to the parameter $i$, the smallest value of $i$ which will create an intermediate fraction potentially within $[l, r]$ can be directly calculated. Only the intermediate fraction formed with this calculated value of $i$ needs to be tested for membership in $[l, r]$.

Let $l_N$ and $l_D$ be the numerator and denominator of $l$, and let $r_N$ and $r_D$ be the numerator and denominator of $r$. In the case of $k$ even; $s_k < l < (l + r)/2$ (otherwise $s_k$ would have been identified as $\in [l, r]$, see Algorithm 4); $s_{k+1} \geq (l + r)/2$; with increasing $i$, $(ip_k + p_{k-1})/(iq_k + q_{k-1})$ forms a decreasing sequence; and the inequality we seek to solve is

$$\frac{ip_k + p_{k-1}}{iq_k + q_{k-1}} \leq \frac{r_N}{r_D}. \tag{43}$$

Solving (43), the smallest integral value of $i$ that will suffice is

$$i = \left\lceil \frac{r_N q_{k-1} - r_D p_{k-1}}{r_D p_k - r_N q_k} \right\rceil. \tag{44}$$

Similarly, for $k$ odd, the sequence is increasing, and the inequality and solution are

$$\frac{ip_k + p_{k-1}}{iq_k + q_{k-1}} \geq \frac{l_N}{l_D} \rightarrow i = \left\lceil \frac{l_N q_{k-1} - l_D p_{k-1}}{l_D p_k - l_N q_k} \right\rceil. \tag{45}$$

(43), (44), and (45) suggest the following continued fraction algorithm for finding a rational number with the smallest denominator in an interval $[l, r]$.

Algorithm 4.

- Calculate all partial quotients $a_k$ and all convergents $s_k = p_k/q_k$ of the midpoint of the interval, $(l + r)/2$.
- For each convergent $s_k = p_k/q_k$, in order of increasing $k$:
  - If $s_k = p_k/q_k \in [l, r]$, $s_k$ is a rational number with the lowest denominator, STOP.
  - If $k$ is even,

[12] Thanks to David M. Einstein and David Eppstein for this observation, contributed via the `sci.math` newsgroup, which is the linchpin of Algorithm 4.

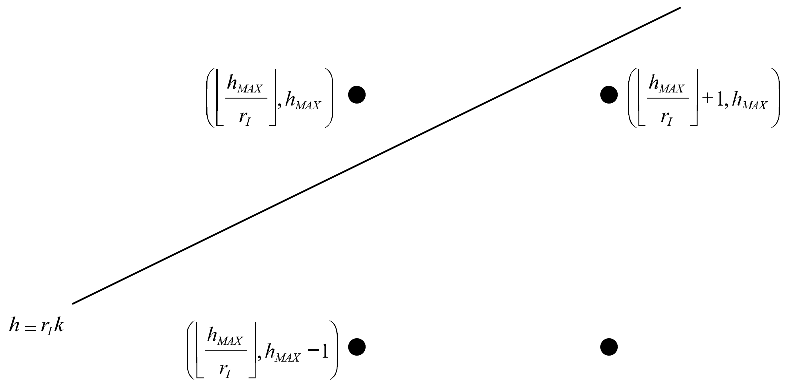
$\left(\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor, h_{MAX}\right)$
$\left(\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor + 1, h_{MAX}\right)$
$h = r_I k$
$\left(\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor, h_{MAX} - 1\right)$

Fig. 3. Graphical Interpretation Of Case II: $h_{MAX}/k_{MAX} < r_I < 1$

  - Calculate $i$ according to (44). If $i < a_{k+1}$ and the intermediate fraction $(ip_k + p_{k-1}) \ / \ (iq_k + q_{k-1}) \geq l$, this intermediate fraction is a rational number with the lowest denominator, STOP.
- Else if $k$ is odd,
  - Calculate $i$ according to (45). If $i < a_{k+1}$ and the intermediate fraction $(ip_k + p_{k-1}) \ / \ (iq_k + q_{k-1}) \leq r$, this intermediate fraction is a rational number with the lowest denominator, STOP.

Algorithm 4 is approximately $O(log\ k_{MAX})$, since there are a fixed number of steps per convergent, and the maximum number of convergents is $O(log\ k_{MAX})$. Once a rational number with the smallest denominator $q_{MIN}$ is located, (41) can be applied to bound $|r_A - r_I|$; namely,

$$\left| \frac{h}{k} - r_I \right| < \frac{1}{2q_{MIN}\ max(q_{MIN}, k_{MAX} - q_{MIN})}. \tag{46}$$

### 6.2 Case II: $h_{MAX}/k_{MAX} < r_I < 1$

If $h_{MAX}/k_{MAX} < r_I < 1$, a graphical argument (Figure 3) can be used to more tightly bound the maximum distance between terms of $F_{k_{MAX},\overline{h_{MAX}}}$.

In this case, a formable term at or to the left[13] of $r_I$ is represented by the point $(\lfloor h_{MAX}/r_I \rfloor + 1, h_{MAX})$ in the integer lattice, and a formable term at or to the right of $r_I$ is represented by the point $(\lfloor h_{MAX}/r_I \rfloor, h_{MAX})$ in the integer lattice. Thus, the maximum distance between neighboring terms in $F_{k_{MAX},\overline{h_{MAX}}}$ is given by the difference of these two terms,

$$\frac{h_{MAX}}{\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor} - \frac{h_{MAX}}{\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor + 1} = \frac{h_{MAX}}{\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor^2 + \left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor}, \tag{47}$$

and the maximum distance from $r_I$ to a neighboring term is given by

[13] To the left on the number line, but to the right in Figure 3.

$$\left|\frac{h}{k} - r_I\right| \leq \frac{h_{MAX}}{2\left(\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor^2 + \left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor\right)}. \tag{48}$$

Note that Case II will exist only if $h_{MAX}/k_{MAX} < 1$.

### 6.3 Case III: $1 < h_{MAX}/k_{MAX} < r_I$

It can be established graphically, using the coordinate system of Figure 1 or Figure 2, that the line $h = r_I k$ intercepts the line $h = h_{MAX}$ at the point $(h_{MAX}/r_I, h_{MAX})$. It is clear from a graphical argument that all of the terms of the Farey series of order $\lfloor h_{MAX}/r_I \rfloor$ are available as neighbors of $r_I$. Therefore,

$$\left|\frac{h}{k} - r_I\right| \leq \frac{1}{2\left\lfloor \frac{h_{MAX}}{r_I} \right\rfloor}. \tag{49}$$

## 7. END-TO-END APPROXIMATION ERROR

A rational approximation requires an integer domain and range, and there are three sources of error inherent in such an approximation:

- Input quantization error, as the input to the approximation is restricted to $\mathbb{Z}^+$.
- Error in selecting $r_A = h/k$, as in general it isn't possible to choose $r_A = r_I$.
- Output quantization error, as the remainder of the division of $hx$ by $k$ is discarded, and the output must be $\in \mathbb{Z}^+$.

To model the end-to-end approximation error, a model function is introduced which represents the function we wish to approximate,

$$F(x) = r_I x. \tag{50}$$

However, the approximation necessarily involves quantizing the input $x$, as well as the result of the integer division:

$$J(x) = \lfloor r_A \lfloor x \rfloor \rfloor = \left\lfloor \frac{h \lfloor x \rfloor}{k} \right\rfloor. \tag{51}$$

Quantization of a real argument $x$ which is not necessarily rational is treated by noting that quntization introduces an error $\varepsilon \in [0, 1)$:

$$\lfloor x \rfloor = x - \varepsilon;\ \varepsilon \in [0, 1). \tag{52}$$

Quantization of a rational argument $a/b$ is treated by noting that the largest quantization error $\varepsilon$ occurs when $a$ is one less than an integral multiple of $b$:

$$\left\lfloor \frac{a}{b} \right\rfloor = \frac{a}{b} - \varepsilon;\ \varepsilon \in \left[0, \frac{b-1}{b}\right]. \tag{53}$$

The difference function $J(x) - F(x)$, can be stated as in (54) or (55). The two quantizations in (54) can be treated by introducing $\varepsilon_1$ and $\varepsilon_2$ to yield (55). Note that $\varepsilon_1$ and $\varepsilon_2$ are independent, meaning for this application that in general $r_I$, $r_A = h/k$, and $x$ can be chosen so as to force any combination of $\varepsilon_1$ and $\varepsilon_2$, so that no combinations of $\varepsilon_1$ and $\varepsilon_2$ can be excluded.

$$J(x) - F(x) = \left\lfloor \frac{h \lfloor x \rfloor}{k} \right\rfloor - r_I x \tag{54}$$

$$J(x) - F(x) = \frac{h(x - \varepsilon_1)}{k} - \varepsilon_2 - r_I x;\ \varepsilon_1 \in [0, 1);\ \varepsilon_2 \in \left[0, \frac{k-1}{k}\right] \tag{55}$$

Choosing the extremes of $\varepsilon_1$ and $\varepsilon_2$ so as to minimize and maximize the difference function bounds the approximation error (56).

$$J(x) - F(x) \in \left( (r_A - r_I)x - r_A - \frac{k-1}{k}, (r_A - r_I)x \right] \tag{56}$$

Minimizing and maximizing (56) over a domain of $[0, x_{MAX}]$ leads to (57).

$$J(x) - F(x)|_{x \in [0, x_{MAX}]} \in \begin{cases} \left( (r_A - r_I)x_{MAX} - r_A - \frac{k-1}{k}, 0 \right], & r_A < r_I \\ \left( -r_A - \frac{k-1}{k}, 0 \right], & r_A = r_I \\ \left( -r_A - \frac{k-1}{k}, (r_A - r_I)x_{MAX} \right], & r_A > r_I \end{cases} \tag{57}$$

Thus, given an $r_I \in \mathbb{R}^+$ and a rational approximation to $r_I$, $r_A = h/k$, the error introduced by this rational approximation used over a domain $[0, x_{MAX}]$ can be bounded.

## 8. DESIGN EXAMPLE

A design example is presented to illustrate the methods presented. An $r_I$ specified with only modest precision and an $h_{MAX}$ and $k_{MAX}$ of only modest size are used to avoid a large number of partial quotients or large integers.[14]

**Design Example:** Assume that real numbers are to be approximated by rational numbers in the interval $[0.385, 2.160]$, subject to $h_{MAX} = 193 \wedge k_{MAX} = 500$. Bound $|r_A - r_I|$ under these constraints. Find the best rational approximations to $1/\pi \approx 0.31830989$ and $2/\pi \approx 0.63661977$ under the same constraints.

**Solution:** In this example, *Case I*, *Case II*, and *Case III* (Sections 6.1, 6.2, and 6.3) apply. Case III will dominate the upper bound on the error in selecting $r_A$, but it is instructive to work through Case I and Case II.

[14] The rational approximation software submitted with this paper will handle rational approximations involving hundreds of digits and hundreds of partial quotients. However, such approximations make unsuitable examples because of the length, the difficulty in typesetting huge integers and rational numbers, and because the examples can't be carried out manually by a reader.

Table 1. Partial Quotients And Convergents Of 0.3855 (Midpoint Of The Interval $[0.385, 0.386]$)

| Index ($k$) | $dividend_k$ | $divisor_k$ | $a_k$ | $remainder_k$ | $p_k$ | $q_k$ |
|---|---|---|---|---|---|---|
| -1 | N/A | 771 | N/A | 2000 | 1 | 0 |
| 0 | 771 | 2000 | 0 | 771 | 0 | 1 |
| 1 | 2000 | 771 | 2 | 458 | 1 | 2 |
| 2 | 771 | 458 | 1 | 313 | 1 | 3 |
| 3 | 458 | 313 | 1 | 145 | 2 | 5 |
| 4 | 313 | 145 | 2 | 23 | 5 | 13 |
| 5 | 145 | 23 | 6 | 7 | 32 | 83 |
| 6 | 23 | 7 | 3 | 2 | 101 | 262 |
| 7 | 7 | 2 | 3 | 1 | 335 | 869 |
| 8 | 2 | 1 | 2 | 0 | 771 | 2000 |

To apply the results from *Case I*, it is necessary to find a rational number with the smallest denominator in the interval $[l = 0.385, r = 193/500 = 0.386]$. The midpoint of the interval is $(l + r)/2 = 0.3855 = 771/2000$.

Table 1 shows the generation of the partial quotients and convergents of the midpoint, 771/2000, using Algorithm 2.

Algorithm 4 can be applied to locate the fraction in $[l, r]$ with the smallest denominator. $s_0 = 0/1 \notin [l, r]$. The intermediate fraction $(p_0 + p_{-1})/(q_0 + q_{-1}) = 1/1 \notin [l, r]$. $s_1 = 1/2 \notin [l, r]$. $s_2 = 1/3 \notin [l, r]$. $s_3 = 2/5 \notin [l, r]$. The intermediate fraction $(p_3 + p_2)/(q_3 + q_2) = 3/8 \notin [l, r]$. $s_4 = 5/13 \notin [l, r]$. (44) can be applied to determine the lowest value of the parameter $i$ for which an intermediate fraction *may* be in $[l, r]$:

$$i = \left\lceil \frac{(r_N = 193)(q_3 = 5) - (r_D = 500)(p_3 = 2)}{(r_D = 500)(p_4 = 5) - (r_N = 193)(q_4 = 13)} \right\rceil = \left\lceil \frac{-35}{-9} \right\rceil = 4. \tag{58}$$

Thus, it is only necessary to examine the intermediate fraction $(4p_4 + p_3)/(4q_4 + q_3)$ for potential membership in $[l, r]$. This intermediate fraction, $22/57 \approx 0.385965 \in [l, r]$. Thus, the fraction with the lowest denominator in the interval is 22/57, and $q_{min} = 57$.

Application of (46) yields

$$\left| \frac{h}{k} - r_I \right| < \left( \frac{1}{2q_{MIN}\ max(q_{MIN}, k_{MAX} - q_{MIN})} = \frac{1}{(2)(57)(500 - 57)} = \frac{1}{50,502} \right). \tag{59}$$

Note that the 1/50,502 maximum error in placing $r_A$ is much better than the 1/1,000 worst-case error for $F_{500}$ in general without restrictions on the interval.

Case II and Case III aren't as complicated as Case I—applying these cases is a simple matter of substitution into (48) or (49). Case II and (48) apply, but the error bounds from Case III will be larger, so Case II is not evaluated. Case III applies: the line $h = r_I k$ intersects the line $h = h_{MAX}$ at the point $(h_{MAX}/r_I, h_{MAX}) =$

Table 2. Partial Quotients And Convergents Of 31,830,989/100,000,000 (A Rational Approximation To $1/\pi$)

| Index ($k$) | $dividend_k$ | $divisor_k$ | $a_k$ | $remainder_k$ | $p_k$ | $q_k$ |
|---|---|---|---|---|---|---|
| -1 | N/A | 31,830,989 | N/A | 100,000,000 | 1 | 0 |
| 0 | 31,830,989 | 100,000,000 | 0 | 31,830,989 | 0 | 1 |
| 1 | 100,000,000 | 31,830,989 | 3 | 4,507,033 | 1 | 3 |
| 2 | 31,830,989 | 4,507,033 | 7 | 281,758 | 7 | 22 |
| 3 | 4,507,033 | 281,758 | 15 | 280,663 | 106 | 333 |
| 4 | 281,758 | 280,663 | 1 | 1,095 | 113 | 355 |
| 5 | 280,663 | 1,095 | 256 | 343 | 29,034 | 91,213 |
| 6 | 1,095 | 343 | 3 | 66 | 87,215 | 273,994 |
| 7 | 343 | 66 | 5 | 13 | 465,109 | 1,461,183 |
| 8 | 66 | 13 | 5 | 1 | 2,412,760 | 7,579,909 |
| 9 | 13 | 1 | 13 | 0 | 31,830,989 | 100,000,000 |

$(193/2.160, 193)$, thus all terms of the Farey series of order $\lfloor 193/2.160 \rfloor = 89$ are available for selection. Therefore, applying (49),

$$\left| \frac{h}{k} - r_I \right| \leq \frac{1}{2 \times 89} \approx 0.0056. \tag{60}$$

Thus, if $F_{500,\overline{193}}$ is used to approximate real numbers over the interval [0.385, 2.160], an upper bound on $|r_A - r_I|$ is $1/178 \approx 0.0056$. Note that Case III dominates, and that the upper bound on $|r_A - r_I|$ varies within the interval.

To find the best rational approximations to $1/\pi$ in $F_{500,\overline{193}}$, note that $1/\pi < 193/500$, so all of the terms in $F_{500}$ are available. Table 2 shows the partial quotients and convergents of $1/\pi$, using 0.31830989 as a rational approximation of $1/\pi$. $s_4$ is the highest-order convergent with $q_k \leq 500$, so $s_4 = 113/355$ is one Farey neighbor to $1/\pi$ in $F_{500}$. Applying (29) to generate the other neighbor in $F_{500}$ yields 106/333. Note that $113/355 - 1/\pi \approx -3 \times 10^{-8}$ and $106/333 - 1/\pi \approx 8 \times 10^{-6}$ (the errors are quite small).

To find the best rational approximations to $2/\pi$ in $F_{500,\overline{193}}$, note that $2/\pi > 193/500$, so the second clause of Algorithm 3 applies. Table 3 shows the partial quotients and convergents of $\pi/2$, using $1/0.63661977$ as a rational approximation of $\pi/2$. $s_3$ is the highest-order convergent with $q_k \leq 193$, so $s_3^{-1} = (11/7)^{-1}$ is one neighbor to $2/\pi$ in $F_{\overline{193}}$. Applying (29) to generate the reciprocal of the other neighbor in $F_{\overline{193}}$ yields 300/191, implying that 191/300 is the other neighbor. $7/11 - 2/\pi \approx -3 \times 10^{-4}$. $191/300 - 2/\pi \approx 5 \times 10^{-5}$.

## ACKNOWLEDGMENTS

We would like to gratefully acknowledge the assistance of Greg Bachelis, Robert Berman, Feng Lin, Nick Sahinidis, Adam Van Tuyl, Carl Schweiger, Ken Tindell, Steve Vestal, Bob Whitinger, and David B. Stewart in finding the areas of mathematics relevant to the rational number selection problem. We would also like to thank Johan Bengtsson, Michael J. Burke, Mark Endicott, David Eppstein, David M. Einstein, Mircea Munteanu, Adam Gibson, and Virgil (of the

Table 3. Partial Quotients And Convergents Of 100,000,000/63,661,977 (A Rational Approximation To $\pi/2$)

| Index ($k$) | $dividend_k$ | $divisor_k$ | $a_k$ | $remainder_k$ | $p_k$ | $q_k$ |
|---|---|---|---|---|---|---|
| -1 | N/A | 100,000,000 | N/A | 63,661,977 | 1 | 0 |
| 0 | 100,000,000 | 63,661,977 | 1 | 36,338,023 | 1 | 1 |
| 1 | 63,661,977 | 36,338,023 | 1 | 27,323,954 | 2 | 1 |
| 2 | 36,338,023 | 27,323,954 | 1 | 9,014,069 | 3 | 2 |
| 3 | 27,323,954 | 9,014,069 | 3 | 281,747 | 11 | 7 |
| 4 | 9,014,069 | 281,747 | 31 | 279,912 | 344 | 219 |
| 5 | 281,747 | 279,912 | 1 | 1,835 | 355 | 226 |
| 6 | 279,912 | 1,835 | 152 | 992 | 54,304 | 34,571 |
| 7 | 1,835 | 992 | 1 | 843 | 54,659 | 34,797 |
| 8 | 992 | 843 | 1 | 149 | 108,963 | 69,368 |
| 9 | 843 | 149 | 5 | 98 | 599,474 | 381,637 |
| 10 | 149 | 98 | 1 | 51 | 708,437 | 451,005 |
| 11 | 98 | 51 | 1 | 47 | 1,307,911 | 832,642 |
| 12 | 51 | 47 | 1 | 4 | 2,016,348 | 1,283,647 |
| 13 | 47 | 4 | 11 | 3 | 23,487,739 | 14,952,759 |
| 14 | 4 | 3 | 1 | 1 | 25,504,087 | 16,236,406 |
| 15 | 3 | 1 | 3 | 0 | 100,000,000 | 63,661,977 |

`sci.math.num-analysis` newsgroup) for insight into this problem; Cliff Stallings and Robert Kakos for support from Wayne State University's College Of Engineering; Paulette Groen and Paula Smith for support from Visteon; Bob Crosby for support from Texas Instruments; Klaus-Peter Zauner, Andrea Blome, Una Smith, Karsten Tinnefeld, and Axel Franke for other tool and logistical support; and the management team at Visteon for allowing us to pursue this effort in the workplace.

## REFERENCES

[1] G.H. Hardy, E.M. Wright, *An Introduction To The Theory Of Numbers*, ISBN 0-19-853171-0.

[2] G. Harman (1998), *Metric Number Theory*, Oxford University Press.

[3] P. Kargaev, A. Zhigljavsky (1966), *Approximation Of Real Numbers By Rationals: Some Metric Theorems*, Journal of Number Theory, 61, 209-225.

[4] P. Kargaev, A. Zhigljavsky (1967), *Asymptotic Distribution Of The Distance Function To The Farey Points*, Journal of Number Theory, 65, 130-149.

[5] A. Ya. Khinchin, *Continued Fractions*, University Of Chicago Press, 1964; Library Of Congress Catalog Card Number 64-15819.

[6] William J. LeVeque, *Fundamentals Of Number Theory*, Dover Publications, 1977, ISBN 0-486-68906-9.

[7] C. D. Olds, *Continued Fractions*, Randam House, 1963, Library Of Congress Catalog Card Number 61-12185.

[8] Oystein Ore, *Number Theory And Its History*, ISBN 0-486-65620-9.

[9] M. R. Schroeder, *Number Theory In Science And Communication*, ISBN 3-540-62006-0.

Fig. 4. Version Control Information (For Reference Only—Will Be Removed Before Submission Of Paper)

```
LaTeX compile date:  April 7, 2003.

CVS Version Control Information:
$Header: /cvs_root/dtaipubs/dtaipubs/acm0010/paper/acm_paper.tex,v 1.4 2003/04/08 03:49:14 dtashley Exp $.
```