December 8, 2000: This paper is distributed with the *IjuTools* suite. The paper contained on the following pages was rejected by "*IEEE Transactions On Computers*" in 2000. Therefore, the five authors listed hold the copyright. LaTeX source code is also included in this distribution. The five authors hereby waive all rights under copyright law and grant permission to reproduce or use this paper and the accompanying LaTeX source code, in whole or in part, for any purpose whatsoever, commercial or non-commercial, with no requirement for renumeration of the authors, and no requirement to cite or otherwise supply the identity of the authors.

Under Adobe Arobat 3, some subscripts in this paper disappear when printed. Adobe Acrobat 4 or its successors are recommended for printing.

Economical Implementation Of Linear Scaling Functions In Microcontroller Software Using Rational Number Approximation Techniques

David T. Ashley, Joseph P. DeVoe, Cory Pratt, Karl Perttunen, Anatoly Zhigljavsky

Abstract— **Inexpensive microcontrollers are widely used in vehicles, consumer electronics, laboratory equipment, and medical equipment. A significant problem in the design of software for these devices is the efficient and reliable implementation of arithmetic. In this paper, techniques are developed to use an instruction set typical of an inexpensive microcontroller to economically approximate linear scaling functions of the form $f(x) = r_I x$ (r_I non-negative and real but not necessarily rational) in the first quadrant using a rational approximation factor $r_A = h/k$. Methods and considerations in choosing h and k are developed, and error terms are derived which bound the error introduced by the rational approximation.**

Keywords— **Microcontroller arithmetic, linear scaling, rational approximation, Farey series, continued fractions.**

I. Introduction

LOW-COST microcontrollers provide weak instruction sets, and a significant design challenge in the development of software for these devices is the economical implementation of arithmetic. This paper presents techniques which are suitable for economically implementing high-precision linear scalings using instructions which are characteristic of inexpensive 4-bit and 8-bit microcontrollers.

This paper is confined to the approximation of linear scalings of the form $y = r_I x$ ($r_I \in \mathbb{R}^+$, not necessarily $\in \mathbb{Q}^+$) by linear scalings of the form $y = hx/k$ ($h \in \mathbb{Z}^+$, $k \in \mathbb{N}$). In Section II, error terms for rational approximations are derived without regard for how integers h and k are chosen. In Section III, results from number theory are presented which give insight into how to choose rational numbers h/k for use as scaling factors. In Section IV, probabilistic results from number theory are presented outlining how close to a real r_I a rational $r_A = h/k$ can typically be chosen. In Section V, a special case in which k is chosen as an integral power of two is developed with the aim of providing a framework for tabulating scaling factors. In Section VI, practical techniques for economically implementing rational scaling functions using inexpensive microcontroller instruction sets are presented. In Section VII, design examples which illustrate the techniques are presented.

D.T. Ashley, J.P. DeVoe, C. Pratt, and K. Perttunen are with Visteon Automotive Systems in Dearborn, Michigan, USA. E-mail: {dashley1, jdevoe, cpratt2, kperttun}@visteon.com. Anatoly Zhigljavsky is with Cardiff University, Cardiff, UK. E-mail: zhigljavskyaa@cardiff.ac.uk.

II. Analysis Of Approximation Error

A function $y = r_I x$ ($r_I \in \mathbb{R}^+$, not necessarily $\in \mathbb{Q}^+$) is to be approximated by a function $y = r_A x$; $r_A = h/k$, $\in \mathbb{Q}^+$.[1] In this section, error terms are developed which bound the error introduced when $f(x) = r_I x$ is approximated by $f(x) = r_A x$ in the first quadrant only ($x \in \mathbb{R}^+$).

A. Model Functions

(1) through (4) provide models of the function to be approximated which vary in whether the domain is real or integral, and in whether the range is real or integral. The *floor(·)* function, denoted $\lfloor \cdot \rfloor$, is used to model the effect of quantization, such as occurs when a real argument is quantized for implementation using integer arithmetic, or when the fractional part of a quotient is discarded.[2]

In practical problems, the domain and range may be integral rather than real for either practical or conceptual reasons. As an example of a domain which is integral for *practical* reasons, consider an embedded software algorithm which has access only to integral data (as might happen with integral vehicle speed reported to an embedded software algorithm over a network). In this case, the behavior of the software may be specified only over $\mathbb{Z}^+$, as it is impossible to excite the software with non-integral values. It may also happen that the domain is *conceptually* integral, as occurs when the input argument is a count or other inherently integral quantity. The range may also be integral for similar practical or conceptual reasons.

(1) provides a model of the ideal function to be approximated when both domain and range are the non-negative real numbers.

$$F(x) = r_I x \tag{1}$$

(2) provides a model of the ideal function to be approximated when the function is to be evaluated on a domain of non-negative integers only.

$$G(x) = r_I \lfloor x \rfloor \tag{2}$$

(3) provides a model of the ideal function to be approximated when only the range is integral.

[1] Mnemonic for r_I and r_A: I=ideal, A=actual.

[2] The *ceiling(·)* function, denoted $\lceil \cdot \rceil$, is also used and appears in many algebraic results throughout the paper. There is often ambiguity in how the *floor(·)* and *ceiling(·)* functions are defined for negative arguments. Here, $\lfloor -1.1 \rfloor = \lceil -2.1 \rceil = -2$.

$$H(x) = \lfloor r_I x \rfloor \tag{3}$$

(4) provides a model of the ideal function to be aproximated when both the domain and range are the non-negative integers.

$$I(x) = \lfloor r_I \lfloor x \rfloor \rfloor \tag{4}$$

(5) defines r_A, the rational number used to approximate r_I. h/k is always assumed reduced.

$$r_A = \frac{h}{k};\ h \in \mathbb{Z}^+;\ k \in \mathbb{N} \tag{5}$$

(6) provides a model of the function which is used to approximate (1), (2), (3), or (4). The approximation always has an integral domain and range.

$$J(x) = \lfloor r_A \lfloor x \rfloor \rfloor = \left\lfloor \frac{h \lfloor x \rfloor}{k} \right\rfloor \tag{6}$$

(7) defines an enhancement to (6). The approximation error introduced can be shifted using integral parameter z.

$$K(x) = \left\lfloor \frac{h \lfloor x \rfloor + z}{k} \right\rfloor;\ z \in \mathbb{Z} \tag{7}$$

(8) and (9) are special cases of (6) and (7) which are useful in microcontroller work, since division by a power of two can be achieved very economically using right-shift instructions.

$$L(x) = \left\lfloor \frac{h \lfloor x \rfloor}{2^q} \right\rfloor;\ k = 2^q;\ r_A = \frac{h}{2^q} \tag{8}$$

$$M(x) = \left\lfloor \frac{h \lfloor x \rfloor + z}{2^q} \right\rfloor;\ k = 2^q;\ r_A = \frac{h}{2^q} \tag{9}$$

B. Methods Of Error Analysis

Quantization of a real argument which is not necessarily rational is treated by noting that quantization introduces an error $\varepsilon \in [0, 1)$ (Eq. 10).

$$\lfloor x \rfloor = x - \varepsilon;\ \varepsilon \in [0, 1) \tag{10}$$

Quantization of a rational argument a/b is treated by noting that the largest quantization error ε occurs when a is one less than an integral multiple of b (Eq. 11).[3]

$$\left\lfloor \frac{a}{b} \right\rfloor = \frac{a}{b} - \varepsilon;\ \ \varepsilon \in \left[0, \frac{b-1}{b}\right] \tag{11}$$

[3]Strictly speaking, $\varepsilon \in \left\{0, \frac{1}{b}, \ldots, \frac{b-2}{b}, \frac{b-1}{b}\right\}$; however, since only the smallest and largest values are of interest, (11) is used.

Since a difference of integers must also be an integer, results on differences of quantized values are constrained further by intersection with the set of integers. Care must be taken in the intersection of an interval with the set of integers, as the distinction between an open interval and a closed interval is significant. The identities in (12) through (15) are employed.[4]

$$[x, y] \cap \mathbb{Z} = [\lceil x \rceil, \lfloor y \rfloor]_{\mathbb{Z}} \tag{12}$$

$$[x, y) \cap \mathbb{Z} = [\lceil x \rceil, \lceil y - 1 \rceil]_{\mathbb{Z}} \tag{13}$$

$$(x, y] \cap \mathbb{Z} = [\lfloor x + 1 \rfloor, \lfloor y \rfloor]_{\mathbb{Z}} \tag{14}$$

$$(x, y) \cap \mathbb{Z} = [\lfloor x + 1 \rfloor, \lceil y - 1 \rceil]_{\mathbb{Z}} \tag{15}$$

C. Error Analysis Of $\{J(x), K(x)\} - I(x)$

(6) is a special case of (7), so the difference function $K(x) - I(x)$ with $z = 0$ is $J(x) - I(x)$. For this reason it is not necessary to derive $J(x) - I(x)$ separately. The difference function (or error function) is the difference between the ideal model function with integral domain and range $[I(x)]$, and the approximation function with integral domain and range $[J(x)$ or $K(x)]$ (Eq. 16).

$$K(x) - I(x) = \left\lfloor \frac{h \lfloor x \rfloor + z}{k} \right\rfloor - \lfloor r_I \lfloor x \rfloor \rfloor \tag{16}$$

The inner *floor(·)* function can be removed with the understanding that the difference function will be evaluated on a domain of integers only (17).

$$K(x) - I(x) = \left\lfloor \frac{hx + z}{k} \right\rfloor - \lfloor r_I x \rfloor;\quad x \in \mathbb{Z}^+ \tag{17}$$

Quantization (two occurrences of the *floor(·)* function) can be modeled as introducing errors ε_1 and ε_2 (18). Because the domain is integral, the largest quantization error in ε_1 occurs when $hx + z$ is one less than an integral multiple of k, hence $\varepsilon_1 \in \left[0, \frac{k-1}{k}\right]$. Because r_I may be irrational, $\varepsilon_2 \in [0, 1)$.

Choosing the extremes of ε_1 and ε_2 so as to minimize and maximize the difference function bounds the approximation error (19).

(19) may be intersected with the set of integers, because the result, a difference of integers, must also be an integer (20).

With $z = 0$, (20) supplies $J(x) - I(x)$ (21).

r_A must almost always be chosen unequal to r_I, and as a result the approximation error is larger for larger x. Most

[4]In (12) through (15) and throughout the paper, a subscript of $\mathbb{Z}$ is used to denote that a set may contain only integers.

$$K(x) - I(x) = \frac{hx+z}{k} - \varepsilon_1 - r_I x + \varepsilon_2;\ x \in \mathbb{Z}^+;\ \ \varepsilon_1 \in \left[0, \frac{k-1}{k}\right];\ \ \varepsilon_2 \in [0,1) \tag{18}$$

$$K(x) - I(x) \in \left[(r_A - r_I)x + \frac{z}{k} - \frac{k-1}{k}, (r_A - r_I)x + \frac{z}{k} + 1\right) \tag{19}$$

$$K(x) - I(x) \in \left[\left\lceil (r_A - r_I)x + \frac{z}{k} - \frac{k-1}{k} \right\rceil, \left\lceil (r_A - r_I)x + \frac{z}{k} \right\rceil\right]_{\mathbb{Z}} \tag{20}$$

$$J(x) - I(x) \in \left[\left\lceil (r_A - r_I)x - \frac{k-1}{k} \right\rceil, \lceil (r_A - r_I)x \rceil\right]_{\mathbb{Z}} \tag{21}$$

approximations are used only in a restricted domain, and it is useful to know the upper bound on approximation error when the approximation is used only in an interval $[0, x_{MAX}]_{\mathbb{Z}}$, $x_{MAX} \in \mathbb{N}$.[5] These upper bounds can be obtained by substitution into (20), and are presented as (22). Note that the second case of (22) may not be distinct, depending on the choice of z.

With $z = 0$, analogous results for $J(x) - I(x)$ are obtained (23).

In practice, there are three useful choices of the parameter z. (24) supplies the choice of z which will assure that the approximation error is never negative. (25) supplies the choice of z which will assure that the approximation error is never positive. (26) supplies the choice of z which will center the approximation error about zero.

$$z_{NONEG} = \begin{cases} \lceil (r_I - r_A)x_{MAX}k \rceil, & r_A < r_I \\ 0, & r_A \geq r_I \end{cases} \tag{24}$$

$$z_{NOPOS} = \begin{cases} 0, & r_A \leq r_I \\ \lfloor (r_I - r_A)x_{MAX}k \rfloor, & r_A > r_I \end{cases} \tag{25}$$

$$z_{CENTER} = \left\lfloor \frac{(r_I - r_A)x_{MAX}k}{2} \right\rfloor \tag{26}$$

Example 1: In a vehicle software load, vehicle speed is received in network messages as integral KPH, and is to be converted to MPH and retransmitted over a second network as integral MPH. If r_A=59/95≈0.62105263 is used to approximate $r_I \approx 0.6214$ (the ideal conversion factor from KPH to MPH), how much error might be introduced by this rational approximation (vs. retransmitting the quantized product of r_I and the received vehicle speed) for received vehicle speeds up to 255 KPH?

[5]Throughout the paper, it is assumed that $x_{MAX} \in \mathbb{N}$.

Solution: In this example, the domain is integral and the range is integral, so (23) applies with $x_{MAX} = 255$. The first row of (23) predicts that the error will always be in $[-1, 0]_{\mathbb{Z}}$={-1,0}, so that the calculated integral MPH may be up to one count less than implied by $I(x)$ with $r_I = 0.6214$.

D. Error Analysis Of $\{J(x), K(x)\} - G(x)$

Because (6) is a special case of (7), the difference function $K(x) - G(x)$ with $z = 0$ is $J(x) - G(x)$; hence there is no need to derive $J(x) - G(x)$ explicitly.

$$K(x) - G(x) = \left\lfloor \frac{h\lfloor x \rfloor + z}{k} \right\rfloor - r_I \lfloor x \rfloor \tag{27}$$

The inner *floor(·)* function of (27) can be removed with the understanding that the difference function will be evaluated on a domain of integers only (28). The difference function (28) is real (not required to be rational or integral).

$$K(x) - G(x) = \left\lfloor \frac{hx+z}{k} \right\rfloor - r_I x; \quad x \in \mathbb{Z}^+ \tag{28}$$

The arguments which support the derivation of (17) through (26) also support the derivation of (28) through (35). (33), (34), and (35) supply choices of the parameter z which ensure that the difference function is never negative, never positive, and centered about zero, respectively.

$$z_{NOPOS} = \begin{cases} 0, & r_A \leq r_I \\ \lfloor (r_I - r_A)x_{MAX}k \rfloor, & r_A > r_I \end{cases} \tag{34}$$

$$z_{CENTER} = \left\lfloor \frac{(r_I - r_A)x_{MAX}k + k}{2} \right\rfloor \tag{35}$$

$$K(x)-I(x)|_{x\in[0,x_{MAX}]_{\mathbb{Z}}} \in \begin{cases} \left[\left\lceil (r_A-r_I)x_{MAX}+\frac{z}{k}-\frac{k-1}{k}\right\rceil, \left\lceil \frac{z}{k}\right\rceil\right]_{\mathbb{Z}}, & r_A \le r_I-\frac{z+1}{x_{MAX}k} \\ \left[0, \left\lceil \frac{z}{k}\right\rceil\right]_{\mathbb{Z}}, & r_I-\frac{z+1}{x_{MAX}k} < r_A \le r_I \\ \left[\left\lceil \frac{z}{k}-\frac{k-1}{k}\right\rceil, \left\lceil \frac{z}{k}\right\rceil\right]_{\mathbb{Z}}, & r_A = r_I \\ \left[\left\lceil \frac{z}{k}-\frac{k-1}{k}\right\rceil, \left\lceil (r_A-r_I)x_{MAX}+\frac{z}{k}\right\rceil\right]_{\mathbb{Z}}, & r_A > r_I \end{cases} \tag{22}$$

$$J(x)-I(x)|_{x\in[0,x_{MAX}]_{\mathbb{Z}}} \in \begin{cases} \left[\left\lceil (r_A-r_I)x_{MAX}-\frac{k-1}{k}\right\rceil, 0\right]_{\mathbb{Z}}, & r_A \le r_I-\frac{1}{x_{MAX}k} \\ \{0\}, & r_I-\frac{1}{x_{MAX}k} < r_A \le r_I \\ [0, \lceil (r_A-r_I)x_{MAX}\rceil]_{\mathbb{Z}}, & r_A > r_I \end{cases} \tag{23}$$

$$K(x)-G(x) \in \left[(r_A-r_I)x+\frac{z}{k}-\frac{k-1}{k}, (r_A-r_I)x+\frac{z}{k}\right] \tag{29}$$

$$K(x)-G(x)|_{x\in[0,x_{MAX}]_{\mathbb{Z}}} \in \begin{cases} \left[(r_A-r_I)x_{MAX}+\frac{z}{k}-\frac{k-1}{k}, \frac{z}{k}\right], & r_A < r_I \\ \left[\frac{z}{k}-\frac{k-1}{k}, \frac{z}{k}\right], & r_A = r_I \\ \left[\frac{z}{k}-\frac{k-1}{k}, (r_A-r_I)x_{MAX}+\frac{z}{k}\right], & r_A > r_I \end{cases} \tag{30}$$

E. Error Analysis Of $\{J(x), K(x)\} - H(x)$

Because (6) is a special case of (7), the difference function $K(x)-H(x)$ with $z=0$ is $J(x)-H(x)$; hence there is no need to derive $J(x)-H(x)$ explicitly.

$$K(x)-H(x) = \left\lfloor \frac{h\lfloor x\rfloor + z}{k}\right\rfloor - \lfloor r_I x\rfloor \tag{36}$$

(36) corresponds to the approximation error introduced when a function with a real domain and integral range is approximated using a function with an integral domain and range. The error term supplied by (36) is integral.

The three quantizations in (36) can be treated by introducing ε_1, ε_2, and ε_3 (37).

Evaluating (37) at the extremes of ε_1, ε_2, and ε_3 leads to (38).

Intersection of (38) with the set of integers leads to (39).

Evaluating (39) at the extremes of $[0, x_{MAX}]$ yields (40). With $z=0$, (40) supplies $J(x)-H(x)$ (Eq. 41).

(42), (43), and (44) supply choices of the parameter z which ensure that the difference function is never negative, never positive, and centered about zero, respectively.

$$z_{NOPOS} = \begin{cases} 0, & r_A \le r_I \\ \lfloor (r_I-r_A)x_{MAX}k\rfloor, & r_A > r_I \end{cases} \tag{43}$$

$$z_{CENTER} = \left\lfloor \frac{(r_I-r_A)x_{MAX}k + r_A k}{2}\right\rfloor \tag{44}$$

F. Error Analysis Of $\{J(x), K(x)\} - F(x)$

Because (6) is a special case of (7), the difference function $K(x)-F(x)$ with $z=0$ is $J(x)-F(x)$; hence there is no need to derive $J(x)-F(x)$ explicitly.

$$K(x)-F(x) = \left\lfloor \frac{h\lfloor x\rfloor + z}{k}\right\rfloor - r_I x \tag{45}$$

(45) corresponds to the approximation error introduced when a function with a real domain and range is approximated using a function with an integral domain and range. The error term supplied by (45) is real rather than integral.

The two quantizations in (45) can be treated by introducing ε_1 and ε_2 (46).

Choosing the extremes of ε_1 and ε_2 so as to minimize and maximize the difference function bounds the approximation error (47).

Minimizing and maximizing (47) over a domain of $[0, x_{MAX}]$ gives (48). With $z=0$, (48) supplies $J(x)-F(x)$ (Eq. 49).

(50), (51), and (52) supply choices of the parameter z which ensure that the difference function is never negative, never positive, and centered about zero, respectively.

$$J(x) - G(x) \in \left[(r_A - r_I)x - \frac{k-1}{k}, (r_A - r_I)x\right] \tag{31}$$

$$J(x) - G(x)|_{x\in[0,x_{MAX}]_{\mathbb{Z}}} \in \begin{cases} \left[(r_A - r_I)x_{MAX} - \frac{k-1}{k}, 0\right], & r_A < r_I \\ \left[-\frac{k-1}{k}, 0\right], & r_A = r_I \\ \left[-\frac{k-1}{k}, (r_A - r_I)x_{MAX}\right], & r_A > r_I \end{cases} \tag{32}$$

$$z_{NONEG} = \begin{cases} \lceil (r_I - r_A)x_{MAX}k + k - 1 \rceil, & r_A < r_I \\ k-1, & r_A \geq r_I \end{cases} \tag{33}$$

$$K(x) - H(x) = \frac{h(x-\varepsilon_1)+z}{k} - \varepsilon_2 - r_I x + \varepsilon_3;\ \ \varepsilon_1 \in [0,1);\ \ \varepsilon_2 \in \left[0, \frac{k-1}{k}\right];\ \ \varepsilon_3 \in [0,1) \tag{37}$$

$$K(x) - H(x) \in \left((r_A - r_I)x - r_A + \frac{z}{k} - \frac{k-1}{k}, (r_A - r_I)x + \frac{z}{k} + 1\right) \tag{38}$$

$$K(x) - H(x) \in \left[\left\lfloor (r_A - r_I)x - r_A + \frac{z}{k} + \frac{1}{k}\right\rfloor, \left\lceil (r_A - r_I)x + \frac{z}{k}\right\rceil\right]_{\mathbb{Z}} \tag{39}$$

$$K(x) - H(x)|_{x\in[0,x_{MAX}]} \in \begin{cases} \left[\left\lfloor (r_A - r_I)x_{MAX} - r_A + \frac{z}{k} + \frac{1}{k}\right\rfloor, \left\lceil \frac{z}{k}\right\rceil\right]_{\mathbb{Z}}, & r_A < r_I \\ \left[\left\lfloor -r_A + \frac{z}{k} + \frac{1}{k}\right\rfloor, \left\lceil \frac{z}{k}\right\rceil\right]_{\mathbb{Z}}, & r_A = r_I \\ \left[\left\lfloor -r_A + \frac{z}{k} + \frac{1}{k}\right\rfloor, \left\lceil (r_A - r_I)x_{MAX} + \frac{z}{k}\right\rceil\right]_{\mathbb{Z}}, & r_A > r_I \end{cases} \tag{40}$$

$$J(x) - H(x)|_{x\in[0,x_{MAX}]} \in \begin{cases} \left[\left\lfloor (r_A - r_I)x_{MAX} - r_A + \frac{1}{k}\right\rfloor, 0\right]_{\mathbb{Z}}, & r_A < r_I \\ \left[\left\lfloor -r_A + \frac{1}{k}\right\rfloor, 0\right]_{\mathbb{Z}}, & r_A = r_I \\ \left[\left\lfloor -r_A + \frac{1}{k}\right\rfloor, \left\lceil (r_A - r_I)x_{MAX}\right\rceil\right]_{\mathbb{Z}}, & r_A > r_I \end{cases} \tag{41}$$

$$z_{NONEG} = \begin{cases} \lceil (r_I - r_A)x_{MAX}k + r_A k - 1 \rceil, & r_A < r_I \\ \lceil r_A k - 1 \rceil, & r_A \geq r_I \end{cases} \tag{42}$$

$$K(x) - F(x) = \frac{h(x-\varepsilon_1)+z}{k} - \varepsilon_2 - r_I x;\ \varepsilon_1 \in [0,1);\ \varepsilon_2 \in \left[0, \frac{k-1}{k}\right] \tag{46}$$

$$K(x) - F(x) \in \left((r_A - r_I)x - r_A + \frac{z}{k} - \frac{k-1}{k}, (r_A - r_I)x + \frac{z}{k} \right] \tag{47}$$

$$K(x) - F(x)|_{x\in[0,x_{MAX}]} \in \begin{cases} \left((r_A - r_I)x_{MAX} - r_A + \frac{z}{k} - \frac{k-1}{k}, \frac{z}{k}\right], & r_A < r_I \\ \left(-r_A + \frac{z}{k} - \frac{k-1}{k}, \frac{z}{k}\right], & r_A = r_I \\ \left(-r_A + \frac{z}{k} - \frac{k-1}{k}, (r_A - r_I)x_{MAX} + \frac{z}{k}\right], & r_A > r_I \end{cases} \tag{48}$$

$$J(x) - F(x)|_{x\in[0,x_{MAX}]} \in \begin{cases} \left((r_A - r_I)x_{MAX} - r_A - \frac{k-1}{k}, 0\right], & r_A < r_I \\ \left(-r_A - \frac{k-1}{k}, 0\right], & r_A = r_I \\ \left(-r_A - \frac{k-1}{k}, (r_A - r_I)x_{MAX}\right], & r_A > r_I \end{cases} \tag{49}$$

$$z_{NONEG} = \begin{cases} \lceil (r_I - r_A)x_{MAX}k + r_A k + k - 1 \rceil, & r_A < r_I \\ \lceil r_A k + k - 1 \rceil, & r_A \geq r_I \end{cases} \tag{50}$$

$$z_{NOPOS} = \begin{cases} 0, & r_A \leq r_I \\ \lfloor (r_I - r_A)x_{MAX}k \rfloor, & r_A > r_I \end{cases} \tag{51}$$

$$z_{CENTER} = \left\lfloor \frac{(r_I - r_A)x_{MAX}k + r_A k + k}{2} \right\rfloor \tag{52}$$

III. Methods Of Choosing h And k

In this section, algorithms for choosing h and k subject to the constraints $h \leq h_{MAX}$ and $k \leq k_{MAX}$ in order to place $r_A = h/k$ close to r_I are presented. The algorithms are presented in order of increasing sophistication (and efficiency).[6]

A. Farey Series Methods Of Choosing h And k

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 $k \leq k_{MAX}$. As an example, the Farey series of order 5, F_5, is shown in (53).

$$F_5 = \left\{ \frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1} \right\} \tag{53}$$

The distribution of Farey rational numbers in [0,1] is repeated in any $[n, n+1]$, $n \in \mathbb{Z}$; so that the distribution of Farey rationals in [0,1] supplies complete information about the distribution in all of $\mathbb{R}$.[7]

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

$$Kh - Hk = 1 \tag{54}$$

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 $k \leq k_{MAX}$ such that $Kh - Hk = 1$.

Proof: See [1] p.23, [6] p.222. ■

Theorem 2: If H/K and h/k are two successive terms of F_N, then:

$$K + k > N \tag{55}$$

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 3: 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:

$$h_j = \left\lfloor \frac{k_{j-2} + N}{k_{j-1}} \right\rfloor h_{j-1} - h_{j-2} \tag{56}$$

[6] Although it won't be shown in this paper, if $N = k_{MAX}$ is the maximum allowable denominator, Algorithm 1 is $O(N^2)$, Algorithm 2 is $O(N)$, and the continued fraction algorithm is $O(logN)$.

[7] We occasionally abuse the proper nomenclature by referring to sequential rational numbers outside the interval [0,1] as Farey terms or as part of F_N, which, technically, they are not. All of the results presented in this paper can be shown to apply everywhere in $\mathbb{R}$, so this abuse is not harmful.

$$k_j = \left\lfloor \frac{k_{j-2}+N}{k_{j-1}} \right\rfloor k_{j-1} - k_{j-2} \tag{57}$$

Notes: (1)Theorem 3 gives recursive formulas for generating successive terms in F_N if two consecutive terms are known. (2)Equations (56) and (57) can be solved to allow generation of terms in the decreasing direction (58, 59).

$$h_j = \left\lfloor \frac{k_{j+2}+N}{k_{j+1}} \right\rfloor h_{j+1} - h_{j+2} \tag{58}$$

$$k_j = \left\lfloor \frac{k_{j+2}+N}{k_{j+1}} \right\rfloor k_{j+1} - k_{j+2} \tag{59}$$

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 (Eqns. 56, 57, 58, and 59 require two successive elements). However, if the irreducible rational number is an integer $i = i/1$, the predecessor and successor in F_N are $(iN-1)/N$ and $(iN+1)/N$, so it is convenient to build F_N in either direction starting at an integer. This suggests an algorithm for finding the closest rational numbers in F_N to an r_I when N is small.

Algorithm 1:

- Choose an integer i as either $\lfloor r_I \rfloor$ or $\lceil r_I \rceil$.
- If $i = \lfloor r_I \rfloor$ is chosen, use $h_{j-2}/k_{j-2} = i/1, h_{j-1}/k_{j-1} = (iN+1)/N$ and use (56) and (57) to build successive increasing terms in F_N until the terms which enclose r_I are found. If $i = \lceil r_I \rceil$ is chosen, use $h_{j+2}/k_{j+2} = i/1$, $h_{j+1}/k_{j+1} = (iN-1)/N$ and use (58) and (59) to build successive decreasing terms in F_N until the terms which enclose r_I are found.[8]

The following additional theorem is presented, which can be useful in finding the next term of F_N given only a single term.

Theorem 4: If H/K is a term of F_N, the immediate successor of H/K in F_N is the h/k satisfying $Kh - Hk = 1$ with the largest denominator $k \le N$.

Proof: Any potential successor of H/K which meets $Kh - Hk = 1$ can be formed by adding $1/Kk$ to H/K (60, 61).

$$Kh - Hk = 1 \tag{60}$$

$$\Downarrow$$

$$\frac{h}{k} = \frac{1+Hk}{Kk} = \frac{H}{K} + \frac{1}{Kk} \tag{61}$$

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

[8]This procedure is easily carried out with spreadsheet software, such as *Microsoft Excel*.

Finding the Farey successor from a single Farey term $\notin \mathbb{Z}$ is labor-intensive and not easily done without a computer for even moderate N. Theorems 1, 2, and 4 outline a computationally tractable way (Algorithm 2) to use a computer to form the successor in F_N given only a single Farey term, even for large N. Once two successive Farey terms are known, Theorem 3 can be applied to generate additional terms at low cost. Algorithm 2 below outlines a method to economically find Farey terms on the left and right of a real number.

Algorithm 2:

- Choose a prime number $\alpha \ll N$. α is the number of denominators that a computer can test against $Kh - Hk = 1$ in a practical period of time.[9]
- Choose a rational number h'/α to the left of r_I ($h' = \lfloor r_I \alpha \rfloor$ is usually a good choice).
- Because α is prime, h'/α is not reducible unless h'/α is an integer.
- Denote the Farey term succeeding h'/α as h/k. Theorem 2 asserts that $\alpha + k > N$, implying that $k > N - \alpha$.
- Apply Theorems 1 and 4. Search downward from $k = N$ to $k = N - \alpha + 1$ for an h/k which satisfies Theorem 1. This will require at most α iterations.
- h'/α and h/k are now known to be successive Farey terms in F_N to the left of r_I. Theorem 3 can be employed to economically generate successive Farey terms until r_I is enclosed.

B. Continued Fraction Methods Of Choosing h And k

For selection of a suitable rational number from F_N when N is a few hundred or less, building F_N starting at an integer (Algorithm 1) or at a rational number with a large prime denominator (Algorithm 2) are practical techniques.[10] However, the number of elements of F_N is approximately $3N^2/\pi^2$; and so for large N, $\mathbb{R}$ is too dense with Farey rationals to economically search.

A more direct algorithm for locating the Farey neighbors of an arbitrary real r_I comes from the study of *continued fractions* (a topic in number theory).

A *finite simple continued fraction* is a fraction in the form of (62), where $a_0 \in \mathbb{Z}^+$ and $a_k \in \mathbb{N}$ for $k > 0$. A continued fraction in the form of (62) is denoted $[a_0; a_1, a_2, \ldots, a_n]$.

Continued fractions provide an alternate apparatus for representing real numbers. The form of (62) has important properties which are presented without proof.

- Every rational number can be represented by a finite simple continued fraction $[a_0; a_1, a_2, \ldots, a_n]$.
- Each unique $[a_0; a_1, a_2, \ldots, a_n]$ corresponds to a uniquely valued rational number, so long as $a_n \neq 1$.[11]

[9]Useful primes at each order of magnitude are 11; 101; 1,009; 10,007; 100,003; 1,000,003; and 10,000,019.

[10]*Microsoft Excel*, which maintains integers with 48 bits of precision, can be used to build F_N and select a suitable rational number for at least $N \approx 2^{24}$. (Multiplying two 24-bit numbers yields a 48-bit result, and so *Excel* should be usable until at least order $\approx 2^{24}$.)

[11]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. The restriction that the final element $a_n \neq 1$ is necessary to guarantee that each uniquely valued rational number has a unique finite simple continued fraction representation.

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ldots + \cfrac{1}{a_n}}}} = [a_0; a_1, a_2, \ldots, a_n] \tag{62}$$

TABLE I
CONTINUED FRACTION REPRESENTATION OF $3,362,997/2,924,082$

Index (k)	Dividend	Divisor	a_k	Remainder
0	3,362,997	2,924,082	1	438,915
1	2,924,082	438,915	6	290,592
2	438,915	290,592	1	148,323
3	290,592	148,323	1	142,269
4	148,323	142,269	1	6,054
5	142,269	6,054	23	3,027
6	6,054	3,027	2	0

Without proof, we present the following procedure for finding the continued fraction representation of an arbitrary non-negative rational number a/b.

Algorithm 3:

- Start with $k = 0$, dividend=a, divisor=b.
- Repeat
  - Carry out the division of dividend/divisor to form an integer quotient a_k and an integer remainder.
  - The divisor from the current iteration becomes the dividend for the next iteration, and the remainder from the current iteration becomes the divisor for the next iteration.
  - Increment k.
- Until (remainder is zero).

Without proof, we present the following properties of Algorithm 3.

- The algorithm will produce the same continued fraction representation $[a_0; a_1, a_2, \ldots, a_n]$ 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 last non-zero remainder will be the greatest common divisor of a and b.

Example 2: Find the continued fraction representation of a/b=3,362,997/2,924,082.

Solution: Table I shows the application of Algorithm 3 to form the continued fraction representation of a/b.

Table I implies that the continued fraction representation of $a/b = 3,362,997/2,924,082$ is $[1; 6, 1, 1, 1, 23, 2]$.

Note in Table I that the final non-zero remainder is 3,027, the g.c.d. of 3,362,997 and 2,924,082.

Irrational numbers also have a continued fraction representation, but this representation is necessarily infinite (non-terminating).

An algorithm does exist for obtaining the continued fraction representation of an irrational number; but in practice the algorithm must be carried out symbolically (which can be difficult and not amenable to automation). For this reason, only the algorithm for obtaining the continued fraction representation of *rational* numbers is presented here. Using a rational number as close as practical[12] to the irrational number to be approximated (such as using 3141592654/1000000000 for π, as is done in Example 4) is the recommended technique.

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 $[a_0; a_1, a_2, \ldots, a_k]$, $k \leq n$.

Each convergent s_k is a rational number with a numerator p_k and denominator q_k. Eqns. (64) through (69) define the canonical way to construct all $s_k = p_k/q_k$ from all a_k.

$$p_{-1} = 1 \tag{64}$$

$$q_{-1} = 0 \tag{65}$$

$$p_0 = a_0 = \lfloor r_I \rfloor \tag{66}$$

$$q_0 = 1 \tag{67}$$

$$p_k = a_k p_{k-1} + p_{k-2} \tag{68}$$

$$q_k = a_k q_{k-1} + q_{k-2} \tag{69}$$

When p_k and q_k (the numerator and denominator of the kth convergent s_k) are formed as specified by (64) through

[12]Let α be the irrational number to be approximated, let a/b be the rational number used as an approximation of α when applying Algorithm 3, and let H/K and h/k be the two Farey neighbors identified through Theorem 5. In a worst case, $\alpha < H/K < a/b < h/k$ or $H/K < a/b < h/k < \alpha$. In such cases, the misidentification of the two Farey neighbors is not detectable (because α is not known precisely enough, otherwise a more precise a/b would have been used). In *all* cases, $H/K < a/b < h/k$.

$$\frac{3,362,997}{2,924,082} = 1 + \cfrac{1}{6 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{23 + \cfrac{1}{2}}}}}} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4 + \cfrac{1}{a_5 + \cfrac{1}{a_6}}}}}} = [1; 6, 1, 1, 1, 23, 2] \tag{63}$$

TABLE II
CONVERGENTS OF CONTINUED FRACTION REPRESENTATION OF $3,362,997/2,924,082$

Index (k)	a_k	p_k	q_k
-1	Not defined	1	0
0	1	1	1
1	6	7	6
2	1	8	7
3	1	15	13
4	1	23	20
5	23	544	473
6	2	1,111	966

(69), 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.

Example 3: Find all convergents of the continued fraction representation of a/b=3,362,997/2,924,082; shown in Example 2 to be $[a_0; a_1, a_2, a_3, a_4, a_5, a_6]$=[1;6,1,1,1,23,2]

Solution: Table II shows the results of the application of equations (64) through (69) to form all convergents.

Note that the final convergent, s_6=p_6/q_6=1,111/966 is the reduced form of a/b. Note also that all convergents $s_k = p_k/q_k$ are irreducible. It may also be verified that each even-ordered convergent is less than a/b, and that each odd-ordered convergent is greater than a/b, with the exception of the final convergent, which is equal to a/b.

Theorem 5: For a non-negative rational 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[13] in F_N to a/b, and the other neighbor in F_N is given by (70).[14]

$$\frac{\left\lfloor \dfrac{N - q_{k-1}}{q_k} \right\rfloor p_k + p_{k-1}}{\left\lfloor \dfrac{N - q_{k-1}}{q_k} \right\rfloor q_k + q_{k-1}} \tag{70}$$

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. Note that $s_k \in F_N$, since s_k is rational and reduced with denominator not exceeding N. By theorem ([5], Theorem 9, p. 9), the upper bound on the difference between a/b and s_k is given by (71).

$$\left| \frac{a}{b} - \frac{p_k}{q_k} \right| < \frac{1}{q_k q_{k+1}} \tag{71}$$

For two consecutive terms in F_N, $Kh - Hk = 1$. For a Farey neighbor H/K to s_k in F_N, (72) must hold.

$$\frac{1}{q_k N} \leq \left| \frac{H}{K} - \frac{p_k}{q_k} \right| \tag{72}$$

$q_{k+1} > N$, 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 (71) and (72) leads to (73).

$$\left| \frac{a}{b} - \frac{p_k}{q_k} \right| < \frac{1}{q_k q_{k+1}} < \frac{1}{q_k N} \leq \left| \frac{H}{K} - \frac{p_k}{q_k} \right| \tag{73}$$

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 (70) is the other Farey neighbor. (70) is of the form (74), where $i \in \mathbb{Z}^+$.

$$\frac{ip_k + p_{k-1}}{iq_k + q_{k-1}} \tag{74}$$

[13]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.

[14]We were not able to locate Theorem 5 or a proof in print, but this theorem is known 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`.

If k is even, $s_k < a/b$, and the two Farey terms enclosing a/b, in order, are given in the first clause of (75). If k is odd, $s_k > a/b$, and the two Farey terms enclosing a/b, in order, are given in the second clause of (75).

$$\begin{cases} \dfrac{p_k}{q_k}, \dfrac{ip_k + p_{k-1}}{iq_k + q_{k-1}} \end{cases} \quad (k\ even)$$
$$\begin{cases} \dfrac{ip_k + p_{k-1}}{iq_k + q_{k-1}}, \dfrac{p_k}{q_k} \end{cases} \quad (k\ odd) \tag{75}$$

In either clause of (75), applying the $Kh - Hk = 1$ test, (76), gives the result of 1, since by theorem ([5], Theorem 2, p. 5), $q_k p_{k-1} - p_k q_{k-1} = (-1)^k$, with the exponent of k compensating for the ordering difference between the two clauses of (75), as shown in (77).

$$\begin{aligned} q_k p_{k-1} - p_k q_{k-1} &= (-1)^k = 1, \quad \text{(k even)} \\ q_{k-1} p_k - p_{k-1} q_k &= -(-1)^k = 1, \quad \text{(k odd)} \end{aligned} \tag{77}$$

Thus, every potential Farey neighbor of the form (74) meets the $Kh - Hk = 1$ test. In order to show that (70) is the companion Farey neighbor to p_k/q_k, it is only necessary to show that a term meeting the $Hk - Hk = 1$ test with a larger denominator still not greater than N cannot exist (Theorem 4).

It must first be established that a rational number of the form (74) is irreducible. This result comes directly from (76) and (77), since if the numerator and denominator of (70) or (74) are not coprime, the difference of 1 is not possible.

The denominator of (70) can be rewritten as (78).

$$N - [(N - q_{k-1}) \bmod q_k] \in \{N - q_k + 1, ..., N\} \tag{78}$$

Finally, it must be shown that if one irreducible rational number—namely, the rational number given by (70)—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.

Let c/d be the irreducible rational number given by (70), with d already shown above to be $\in \{N - q_k + 1, ..., N\}$. Since c/d and $s_k = p_k/q_k$ meet the $Kh - Hk = 1$ test, (79) follows.

$$c = \frac{1}{q_k} + \frac{p_k d}{q_k};\ c \in \mathbb{Z} \tag{79}$$

c as shown in (79) is necessarily an integer. Assume that $d \in \mathbb{Z}$ is to be perturbed by some amount $\Delta \in \mathbb{Z}$ to form a different integer $d + \Delta \in \mathbb{Z}$. In order for the $Kh - Hk = 1$ test to be met with the new choice of denominator $d + \Delta$, (80) is required.

$$\frac{1}{q_k} + \frac{p_k d}{q_k} + \frac{p_k \Delta}{q_k} \in \mathbb{Z} \tag{80}$$

TABLE III
CONTINUED FRACTION REPRESENTATION OF
3,141,592,654/1,000,000,000 (A RATIONAL APPROXIMATION TO π)

Index (k)	Dividend	Divisor	a_k	Remainder
0	3,141,592,654	1,000,000,000	3	141,592,654
1	1,000,000,000	141,592,654	7	8,851,422
2	141,592,654	8,851,422	15	8,821,324
3	8,851,422	8,821,324	1	30,098
4	8,821,324	30,098	293	2,610
5	30,098	2,610	11	1,388
6	2,610	1,388	1	1,222
7	1,388	1,222	1	166
8	1,222	166	7	60
9	166	60	2	46
10	60	46	1	14
11	46	14	3	4
12	14	4	3	2
13	4	2	2	0

Comparing (79) with (80), it can be seen that since the first two terms of (80) sum to an integer, (80) implies that $p_k\Delta/q_k \in \mathbb{Z}$. p_k and q_k are coprime, and so in order for q_k to divide $p_k\Delta$ with no remainder, Δ must contain at least every prime factor of q_k, which implies that $\Delta \geq q_k$. Noting that the denominator of (70) is necessarily $d \in \{N - q_k + 1, ..., N\}$, any positive perturbation $\Delta \geq q_k$ will form a $d + \Delta > N$. Thus, no other irreducible rational number in F_N besides that given by (70) with a larger denominator $\leq N$ and which meets the $Kh - Hk = 1$ test can exist; therefore (70) is the other enclosing Farey neighbor to a/b in F_N. ■

Example 4: Find the members of F_{65535} immediately before and immediately after π.

Solution: π is transcendental and cannot be expressed as a rational number. Using 3141592654/1000000000 as a rational approximation to π and applying Algorithm 3 yields Table III.

Table IV shows the formation of the convergents of the continued fraction representation of the rational approximation to π using (64) through (69).

By Theorem 5, one Farey neighbor is the convergent with the largest denominator not greater than 65,535. From Table IV, this convergent is $s_4 = p_4/q_4 = 104{,}348/33{,}215$ (and note that since this is an even-ordered convergent, it will be less than a/b). Also by Theorem 5, applying equation (70), the other Farey neighbor is 104,703/33,328.

C. Case Of Constrained h

In a practical design problem, a rational approximation will typically be implemented by multiplying the input argument x by h, adding an offset z, then dividing by k. Efficiency will often depend on being able to implement multiplication, addition, or division using single machine instructions, which are constrained in the size of the operands

$$\begin{array}{l}(q_k)(ip_k+p_{k-1})-(p_k)(iq_k+q_{k-1})=1, \quad \text{(k even)}\\ (iq_k+q_{k-1})(p_k)-(ip_k+p_{k-1})(q_k)=1, \quad \text{(k odd)}\end{array} \tag{76}$$

TABLE IV
CONVERGENTS OF CONTINUED FRACTION REPRESENTATION OF 3,141,592,654/1,000,000,000 (A RATIONAL APPROXIMATION TO π)

Index (k)	a_k	p_k	q_k
-1	Not defined	1	0
0	3	3	1
1	7	22	7
2	15	333	106
3	1	355	113
4	293	104,348	33,215
5	11	1,148,183	365,478
6	1	1,252,531	398,693
7	1	2,400,714	764,171
8	7	18,057,529	5,747,890
9	2	38,515,772	12,259,951
10	1	56,573,301	18,007,841
11	3	208,235,675	66,283,474
12	3	681,280,326	216,858,263
13	2	1,570,796,327	500,000,000

they can accomodate.

The results from number theory presented earlier are based only on the constraint $k \leq k_{MAX}$, i.e. only the constraint on the denominator is considered. However, in practical problems, the numerator is also typically constrained, usually by the size of operands that an integer multiplication instruction can accomodate.

r_I, h_{MAX}, and k_{MAX} can be specified so that the restriction on the numerator is the dominant constraint, which does not allow the Farey series of order $N = k_{MAX}$ to be economically used to find the best rational approximation to r_I, because $F_{k_{MAX}}$ near r_I will contain predominantly terms with numerators violating $h \leq h_{MAX}$.

To provide for a more economical search for the best rational approximations when the numerator is constrained, Theorem 6 is presented.

Theorem 6: Given a positive real number r_I and constraints on a rational approximation h/k to r_I, $0 \leq h \leq h_{MAX}$ and $0 < k \leq k_{MAX}$, the closest rational numbers to r_I on the left and right subject to the constraints lie in $F_{N'}$, with N' chosen as in (81).

$$N' = \left\lfloor \frac{h_{MAX}}{r_I} + 1 \right\rfloor \tag{81}$$

Proof: Note that by (81), $N' > h_{MAX}/r_I$, for all choices of h_{MAX} and r_I.

If $r_I \leq h_{MAX}/k_{MAX}$, $N' > k_{MAX}$; therefore $F_{N'} \supset F_{k_{MAX}}$, and the theorem is true.

If $r_I > h_{MAX}/k_{MAX}$, then $h_{MAX}/N' < r_I$. Note that h_{MAX}/N' or its reduced form if it is reducible is necessarily in $F_{N'}$. Any rational number $a/b > h_{MAX}/N'$ with $b > N'$ must also have $a > h_{MAX}$, which violates the constraints. Therefore, any a/b such that $h_{MAX}/N' < a/b < r_I$ must lie in $F_{N'}$, and the closest rational number to r_I on the right subject to the constraints must also lie in $F_{N'}$. ■

Example 5: Find the two best rational approximations to π subject to $h_{MAX} = 255$ and $k_{MAX} = 255$.

Solution: In this problem, both the numerator and denominator are constrained. The constrained numerator is not treated directly by the results from number theory. Applying (81) gives N'=82, thus it is only necessary to examine F_{82} for best approximations to π which meet both constraints. Building F_{82} yields 245/78 and 22/7 as the two best rational approximations to π under the constraints.

IV. PROBABILISTIC RESULTS ON $|R_I - R_A|$

In this section we consider different asymptotics for the precision of the approximation of an arbitrary r_I by a fraction $r_A = h/k$ with $k \leq k_{\mathrm{MAX}}$. 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 α 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

$$\max_{\alpha \in [0,1]} \rho_N(\alpha) = \frac{1}{2N} \,. \tag{82}$$

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

Such behaviour of $\rho_N(\alpha)$ is however not typical: as we shall see below, typical values of the approximation error $\rho_N(\alpha)$ are much smaller.

Let us first consider the behaviour of $\rho_N(\alpha)$ for almost all $\alpha \in [0, 1]$.[15] We then have, see [3] and also [2], that for almost all $\alpha \in [0, 1]$ and any $\varepsilon > 0$, (83) and (84) hold.

Even more is true: in (83) and (84) 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.

[15] 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.

$$\lim_{N\to\infty} \rho_N(\alpha) N^2 \log^{1+\varepsilon} N = +\infty, \quad \liminf_{N\to\infty} \rho_N(\alpha) N^2 \log N = 0 \tag{83}$$

$$\limsup_{N\to\infty} \frac{\rho_N(\alpha) N^2}{\log N} = +\infty, \quad \lim_{N\to\infty} \frac{\rho_N(\alpha) N^2}{\log^{1+\varepsilon} N} = 0 \tag{84}$$

These statements are analogues of Khinchin's metric theorem, the classic result in the 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 (85).

This means that for all a, A such that $0 < a < A < \infty$, (86) applies, where 'meas' denotes for the standard Lebesgue measure on $[0, 1]$.

Another result in [4] concerns the asymptotic behavior of the moments of the approximation error $\rho_N(\alpha)$. It says that for any $\delta \neq 0$ and $N \to \infty$, (87) applies, where $\zeta(\cdot)$ and $\mathrm{B}(\cdot,\cdot)$ are the Riemann zeta–function and the Beta–function, correspondingly.

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

$$\int_0^1 \rho_N(\alpha) d\alpha = \frac{3}{\pi^2} \frac{\log N}{N^2} + O\left(\frac{1}{N^2}\right), \quad N \to \infty\,. \tag{88}$$

Comparison of (88) with (84) 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 (82): for typical α the rate of decrease of $\rho_N(\alpha)$, when $N \to \infty$, is, roughly speaking, N^{-2} rather than N^{-1}, the error for the worst combination of α and N.

V. Tabulated Scaling Factors

Choosing $r_A = h/k$ using Farey series techniques as outlined in Section III is a suitable solution when r_I is invariant and known at the time the scaling function is designed. The error terms developed allow prediction of the maximum approximation error over a domain $[0, x_{MAX}]$ when a specific $r_A \approx r_I$ is chosen.

A different problem which occurs in practice is the need to tabulate scaling factors in ROM or EEPROM. These scaling factors (which represent r_A) may depend on sensor or actuator calibrations and are not known precisely in advance at the time the software is designed, and so the technique of choosing and evaluating a rational number r_A at design time as presented earlier cannot be applied.

A. Method Of Tabulating Scaling Factors

The previous sections have concentrated on scaling factors expressed as an arbitrary rational number h/k. However, it is usually not practical to tabulate scaling factors as rational numbers with an arbitrary denominator k for the following reasons.

- Not all processors have division instructions, and division in software (as would be required with an arbitrary tabulated denominator k) is expensive.
- Even for processors with division instructions, if the required maximum arbitrary denominator k exceeds the size which can be accomodated by the division instructions, there is no general way to perform a division of large operands using small division instructions (a solution involving arbitrary division is not scaleable).
- The rational elements of F_N are irregularly spaced in $\mathbb{R}$. The worst case occurs near integers, where the elements are $1/N$ apart. Allowing arbitrary tabulated rational numbers $r_A = h/k$ means that in some regions of $\mathbb{R}$, r_A can be placed very close to r_I, whereas in other regions, the maximum error may degrade to $|r_A - r_I| \leq 1/2N$. This irregular error is usually not useful in engineering endeavors, as the worst-case error must typically be assumed. The division by an arbitrary tabulated k carries computational cost but a limited engineering benefit.

The method of tabulating scaling factors presented in this paper is to create scaling factors of the form $h/2^q$, so that the denominator is an integral power of two. This approach has the following advantages.

- The required multiplication by h is scaleable, allowing large h when necessary.
- The division by 2^q is economically performed using right shift instructions, which every processor has, and which are scaleable.

B. Design Approaches For $h/2^q$ Tabulated Linear Scalings

Fig. 1 shows the three elements of design choice in engineering an $h/2^q$ tabulated linear scaling. If any two of these elements are fixed, the third can be derived. The *Maximum Approximation Error* (A) is the maximum approximation error that can be tolerated for any element of the domain and any tabulated r_A. The *Domain And Range Of Approximation* (B) are the domain over which the approximation will be used, and the range which must be reachable by the appropriate choice of tabulated scaling factor h. The *Data Size Of Tabulated h And Value Of q,* 2^q (C) are the number of bits which must be reserved for each tabulated h, and the number of bits by which the multiplication result hx must be right-shifted (this value is typically hard-coded into the scaling strategy).

The most typical case for beginning a design is that (A)

$$p(\tau) = \begin{cases} 6/\pi^2 & \text{if } 0 \le \tau \le \frac{1}{2} \\ \frac{6}{\pi^2 \tau}\left(1 + \log\tau - \tau\right) & \text{if } \frac{1}{2} \le \tau \le 2 \\ \frac{3}{\pi^2 \tau}\left(2\log(2\tau) - 4\log(\sqrt{\tau}+\sqrt{\tau-2}) - (\sqrt{\tau}-\sqrt{\tau-2})^2\right) & \text{if } 2 \le \tau < \infty \end{cases} \tag{85}$$

$$\text{meas}\{\alpha \in [0,1] :\ a < N^2 \rho_N(\alpha) \le A\} \to \int_a^A p(\tau) d\tau \quad \text{as } N \to \infty \tag{86}$$

$$\frac{\delta+1}{2}\int_0^1 \rho_N^{\delta}(\alpha) d\alpha = \begin{cases} \infty & \text{if } \delta \le -1 \\ \frac{3}{\delta^2\pi^2}\left(2^{-\delta} + \delta 2^{\delta+2}\mathrm{B}(-\delta, \frac{1}{2})\right) N^{-2\delta}\left(1+o(1)\right) & \text{if } -1<\delta<1, \delta\neq 0 \\ \frac{3}{\pi^2} N^{-2}\log N + O\left(N^{-2}\right) & \text{if } \delta = 1 \\ 2^{-\delta}\frac{\zeta(\delta)}{\zeta(\delta+1)} N^{-\delta-1} + O\left(N^{-2\delta}\right) & \text{if } \delta > 1 \end{cases} \tag{87}$$

Fig. 1. Three Elements Of Design Choice In Tabulated Linear $h/2^q$ Scaling Design

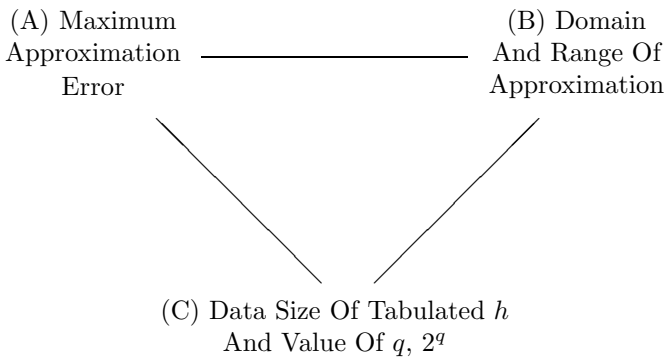
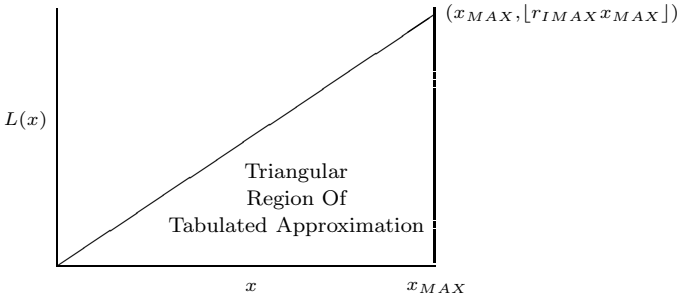


Fig. 2. Specification Of Domain And Possible Values Of r_I As Triangular Region



(Fig. 1) and (B) are known, so that (C) can be derived. A less typical case for beginning a design is that (B) and (C) are tentatively known, so that (A) can be derived (and if this error is unacceptable, the tentative design must be reevaluated). Although it is possible to make the necessary derivations, it never occurs in practice that a design begins with (A) and (C), with the intent of deriving (B). For this reason, only the first two cases are discussed in this paper.

C. Design By Placement Of r_A

A useful paradigm of design is to consider the problem of engineering a tabulated $h/2^q$ scaling in terms of what abilities we preserve for placing r_A with respect to r_I.

Assume that at design time, we know that the linear scaling will be used over the domain $[0, x_{MAX}]$, $x_{MAX} \in \mathbb{N}$, and that $0 \le r_I \le r_{IMAX}$ for all of the r_I we wish to tabulate. This establishes a triangular region in which the approximation will be used (Fig. 2).

The forms of $L(x)$ and $M(x)$ (Eqns. 8, 9) reveal that a specific choice of q allows the selection of r_A in steps of $1/2^q$. With q selected, there are four obvious choices for placement of r_A (89, 90, 91, 92), with almost no practical distinction between (90) and (91). For brevity, only (89) will be developed—i.e. we will consider only placing r_A at or to the left of r_I. Also for brevity, only $F(x)$ as a model function will be considered. All other cases can be developed using similar methods.

$$r_I - \frac{1}{2^q} < r_A \le r_I \tag{89}$$

$$r_I - \frac{1}{2^{q+1}} \le r_A < r_I + \frac{1}{2^{q+1}} \tag{90}$$

$$r_I - \frac{1}{2^{q+1}} < r_A \le r_I + \frac{1}{2^{q+1}} \tag{91}$$

$$r_I \le r_A < r_I + \frac{1}{2^q} \tag{92}$$

Placing r_A consistently at or to the left of r_I (89) will be accomplished if h is chosen by (93).

$$h = \lfloor r_I 2^q \rfloor \tag{93}$$

When h is selected using (93), $r_A \le r_I$ and (94) applies.

$$r_A - r_I \in \left(-\frac{1}{2^q}, 0 \right] \tag{94}$$

To obtain a relationship (B,C)→(A) (Fig. 1), (94) may be substituted into (48) to yield (95).

To obtain a relationship (A,B)→(C), define ε_{SPAN} as the maximum span of error with a fixed z and fixed r_I as x is allowed to vary throughout $[0, x_{MAX}]$ (Eq. 48). It can be shown by solving (48) that q must be chosen so as to meet (96) if the error span is not to exceed ε_{SPAN}. (97) supplies the smallest choice of q which satisfies (96).

$$q \ge \log_2 \left(\frac{x_{MAX} - 1}{\varepsilon_{SPAN} - r_{IMAX} - 1} \right) \tag{96}$$

q may be chosen larger than suggested by (97), but not smaller, while still meeting (96). In practice, this may be done because it is economical to choose q to be a multiple of eight so that the division by 2^q is accomplished by ignoring the least significant byte(s) of $hx + z$, rather than by shifting. (This technique, however, will eliminate shifting at the expense of h_{MAX}, and usually only makes sense when h_{MAX} can be increased without choosing different processor instructions to calculate $hx + z$.)

Once q is fixed, h_{MAX} can be calculated. Substituting $r_I = r_{IMAX}$ into (93) yields (98).

$$h_{MAX} = \lfloor r_{IMAX} 2^q \rfloor \tag{98}$$

D. Design By Placement Of $L(x_{MAX})$

A second useful paradigm of design is to consider the problem of engineering a tabulated $h/2^q$ scaling in terms of what abilities we preserve for placing the terminal point $L(x_{MAX})$ with respect to the terminal point of the model function $I(x)$, $I(x_{MAX})$, $\psi_L \le 0 \le \psi_U$, so that (99) is met (Fig 3).

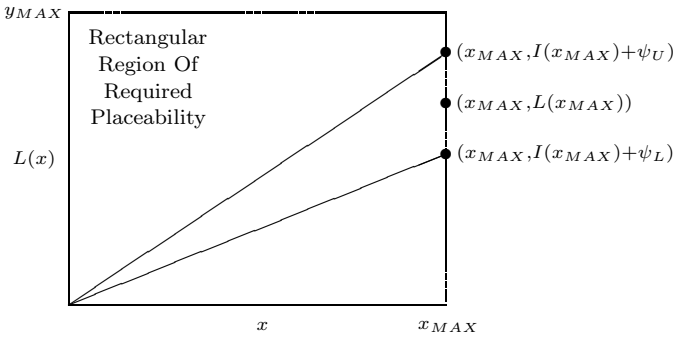
$$I(x_{MAX}) + \psi_L \le L(x_{MAX}) \le I(x_{MAX}) + \psi_U \tag{99}$$

$$\Downarrow$$

$$r_A - r_I \in \left(\frac{\psi_L - 1}{x_{MAX}}, \frac{\psi_U + 1}{x_{MAX}} \right) \tag{100}$$

(99) places restrictions on the relationship between r_A and r_I, and implies (100). This implication is not reversible.

Fig. 3. Specification Of Domain And Possible Values Of r_I As Rectangular Region



For brevity, only $F(x)$ will be considered as a model function and only $L(x)$ will be considered as an approximation. To obtain a relationship which shows how the choice of ψ_L and ψ_U affect the error function $L(x) - F(x)$, (100) may be substituted into (48) to yield (101).

The ability to place the target point $L(x_{MAX})$ in the interval indicated in (99) requires (102). Solving for q leads to the constraint in (103) and the smallest possible choice of q in (104).

$$\frac{1}{2^q} \le \frac{\psi_U - \psi_L + 1}{x_{MAX}} \tag{102}$$

$$q \ge \log_2 \left(\frac{x_{MAX}}{\psi_U - \psi_L + 1} \right) \tag{103}$$

With q chosen by (103) or (104), h can be chosen by (105) so as to meet (99).

$$h = \left\lceil \frac{2^q \left(\lfloor r_I x_{MAX} \rfloor + \psi_L \right)}{x_{MAX}} \right\rceil \tag{105}$$

(102) through (104) give useful rules of thumb for sizing q based on allowed variability in $L(x_{MAX})$. There are two additional useful practical applications for the paradigm of thought (*observation* implies *scaling factor*), and for the equations themselves.

The first additional practical application (for the paradigm of thought) is in the error analysis of self-calibrating systems. In Example 7, it is assumed that we precisely know the transfer characteristics of each bathroom scale transducer and can thereby choose h. A practical bathroom scale is more likely to be self-calibrating, so that at manufacture a known calibration weight $x_{CAL} \in \mathbb{R}^+$ can be placed on the scale and the scale itself will determine the transfer characteristics of the transducer and choose h.[16] For a self-calibrating scale, an important question is if a known calibration weight x_{CAL} is

[16]In this discussion and in (106) and (107), r_I is taken to be the transfer characteristic of of the transducer; whereas in Example 7, $1/r_I$ is the transfer characteristic of the transducer, and r_I is the desired transfer characteristic of the linear scaling in the software.

$$M(x) - F(x)\big|_{x\in[0,x_{MAX}],r_A-r_I\in\left(-\frac{1}{2^q},0\right]} \in \left(\frac{-x_{MAX}+z+1}{2^q} - r_{IMAX} - 1, \frac{z}{2^q}\right] \tag{95}$$

$$q = \left\lceil \log_2 \left(\frac{x_{MAX}-1}{\varepsilon_{SPAN} - r_{IMAX} - 1} \right) \right\rceil = \left\lceil \frac{\ln\left(\frac{x_{MAX}-1}{\varepsilon_{SPAN} - r_{IMAX} - 1} \right)}{\ln 2} \right\rceil \tag{97}$$

$$L(x) - F(x)\big|_{x\in[0,x_{MAX}],r_A-r_I\in\left(\frac{\psi_L-1}{x_{MAX}},\frac{\psi_U+1}{x_{MAX}}\right)} \in \left(\psi_L - 1 - r_A - \frac{2^q-1}{2^q}, \psi_L + 1\right) \tag{101}$$

$$q = \left\lceil \log_2 \left(\frac{x_{MAX}}{\psi_U - \psi_L + 1} \right) \right\rceil = \left\lceil \frac{\ln\left(\frac{x_{MAX}}{\psi_U - \psi_L + 1} \right)}{\ln 2} \right\rceil \tag{104}$$

placed on the scale and produces an A/D converter count $y_{CAL} = H(x_{CAL})$, how much can be inferred about the underlying r_I of the transducer? It can be shown that the implication relationship in (106) and (107) applies. This self-calibration uncertainty should not be neglected in error analyses. Note in (107) that the self-calibration uncertainty in r_I decreases with increasing x_{CAL}, which is consistent with intuition.

$$y_{CAL} = H(x_{CAL}) = \lfloor r_I x_{CAL} \rfloor \tag{106}$$

$$\Downarrow$$

$$r_I \in \left[\frac{y_{CAL}}{x_{CAL}}, \frac{y_{CAL}}{x_{CAL}} + \frac{1}{x_{CAL}} \right) \tag{107}$$

The second additional practical application (for the equations) is the special case of $\psi_L = \psi_U = 0$, which is useful in devising a tabulated linear scaling for piecewise linear functions when the linear segments must join neatly, or when the required accuracy of a linear scaling is not known precisely in advance and a reasonable default tabulated scaling strategy must be chosen. Theorem 7 supplies a choice of q and a result about h_{MAX} which is useful in such cases.

Theorem 7: Given a rational linear scaling of the form (8) with an m-bit domain and an n-bit range, choosing $q = m$ and $h_{MAX} = 2^{m+n}-1$ (i.e. choosing a data width of $m+n$ bits for h) will allow an h to be chosen so that $L(x') = y'$ for any $x' \in [1, x_{MAX} = 2^m - 1]_{\mathbb{Z}}$, $y' \in [0, y_{MAX} = 2^n - 1]_{\mathbb{Z}}$.

Proof: At $x' = x_{MAX}$, in order to be able to choose y', it is required that $x_{MAX}/2^q \le 1$, and $q = m$ is the smallest integral choice of q that will satisfy this constraint. With $q = m$, $L(1) = y_{MAX}$ requires $h_{MAX}/2^m \ge y_{MAX}$, and $h_{MAX} = 2^{m+n} - 1$ (a bit-width of at least $m + n$ for h) is required. ■

VI. Implementation Techniques

Practical microprocessors fall into the following categories, ranked from least capable to most capable.

- Processors with shift and addition instructions.
- Processors with shift, addition, and multiplication instructions.[17]
- Processors with shift, addition, multiplication, and division instructions.

Shift instructions, addition instructions, and multiplication instructions are always *scaleable*, meaning that operands of arbitrary size can be shifted, added, or multiplied by the repeated use of instructions which inherently accept smaller operands. Division instructions, however, are not scaleable. No general method exists to use processor division instructions to divide operands of arbitrary size.

For multiplication, certain values of h can lead to especially economical implementations. Multiplication by an h which is an integral power of two can be performed using shift instructions. Multiplication by an h whose bit pattern is very sparsely populated with 1's can also lead to an economical implementation. For example, multiplication by $h = 33_{10} = 100001_2$ can be performed using five left shifts and an addition. For division, a value of k which is an integral power will lead to a very economical implementation.

The following steps are recommended to economize an $(hx + z)/k$ linear scaling for implementation.

[17] To date, the authors have not encountered a processor with division instructions but no multiplication instructions.

TABLE V
F_{159} NEAR 1.6093 (EXAMPLE 6)

h	k	h/k	Error
214	133	1.60902256	-0.00027744
177	110	1.60909091	-0.00020909
140	87	1.60919540	-0.00010460
243	151	1.60927152	-0.00002848
103	64	1.60937500	+0.00007500
169	105	1.60952381	+0.00022381
235	146	1.60958904	+0.00028904
227	141	1.60992908	+0.00062903

- If the processor does not have a division instruction to directly support the integer division by k, use an $(hx+z)/2^q$ scaling rather than an $(hx+z)/k$ scaling (due to the high cost of division in software).
- If the bit pattern of h is sparsely populated with 1's, evaluate implementation of the multiplication via repeated left-shifting and addition.

VII. DESIGN EXAMPLES

Example 6: (CONVERSION FROM MPH TO KPH) Devise an economical and accurate linear scaling algorithm from integral MPH to integral KPH which operates over a domain of $[0,255]_{\mathbb{Z}}$ MPH (one unsigned byte) and delivers an unsigned byte in $[0,255]_{\mathbb{Z}}$ KPH, and bound the error introduced by the scaling, assuming that the input speed is quantized (already contains error). The error between actual speed (before input quantization) and the output of the algorithm must never be negative—i.e. the output of the algorithm must never *under*state the speed. In the event that the result is too large for one byte, the result should be 255. Implement the algorithm on a TMS370 CPU core, which is characterized by an 8×8=16 unsigned multiplication instruction and a 16×8=8 division instruction.

Solution: One mile is 1.6093 kilometers, thus $r_I = 1.6093$. Efficient implementation using the instruction set of the TMS370 is best done by a single multiplication instruction followed by a single division instruction, implying that $h_{MAX} = k_{MAX} = 255$. Applying Theorem 6 yields $N' = 159$, so it is only necessary to examine F_{159} for the two enclosing rational numbers which meet the constraints on numerator and denominator.[18] Building F_{159} near 1.6093 yields Table V.[19]

[18]Theorem 6 must be applied with caution, as it only guarantees that the *two enclosing* rational numbers are in F_{159}. Table V is included to show the construction of F_{159}—it should be noted that without further analysis there is no guarantee that there are not rational numbers which meet the constraints to the left of 243/151 with $k > 159$.

[19]Table V can be built using spreadsheet software to construct F_{159} starting with the rational numbers 1/1 and 160/159 and using (56) and (57) to build increasing Farey terms until r_I is enclosed.

Fig. 4. TMS370 Solution To Example 6 Using The Rational Number 243/151

```
  MOV input, A    ;12 cycles, load far input
                  ;into A
  MPY #243, A     ;45 cycles, multiply by 243,
                  ;result in (MSB:LSB)=(A:B)
  ADD #139, B     ;6 cycles, 139 = 395 mod 256
  ADC #1, A       ;6 cycles, 1 = 395 div 256
  MOV #151, R02   ;8 cycles, set up for divide
  DIV R02, A      ;Max 63 cycles, divide,
                  ;quotient in A, carry set if
                  ;overflow
  JNC NOOVERFLOW  ;5 cycles if jump not taken
                  ;7 cycles if jump taken
                  ;Jump if div without overflow
  MOV #255, A     ;6 cycles, replace div result
                  ;if too large and won't fit in
                  ;one byte.  DIV instruction
                  ;set C on overflow
NOOVERFLOW:       ;Label, no code generated
  MOV A, output   ;10 cycles, move result to
                  ;far output var

(Total:  Max. 161 clocks, 53.7us with 12 Mhz
crystal.)
```

From Table V, the two rational numbers which enclose 1.6093 are 243/151 and 103/64. Choosing $r_A = 243/151$[20] and using $x_{MAX} = 256$[21] with $z = 395$ by (50) leads to the assembly-language shown in Fig. 4.

From the problem statement, the input to the algorithm is assumed to be quantized (it already contains error), so (48) applies. Evaluating (48) with $r_A = 243/151$, $z = 395$, $r_I = 1.6093$ and $x_{MAX} = 256$ yields $K(x) - F(x) \in (0.0060, 2.6159]$ KPH over the domain [0,256).

Example 7: (BATHROOM SCALE) A manufacturer wishes to build a family of electronic bathroom scales using linear transducers which convert weight to voltage. The voltage from a transducer is measured using a 10-bit A/D converter and a custom combinational logic integrated circuit which will multiply the integer $x \in [0, 2^{10}-1]_{\mathbb{Z}}$ from the A/D converter by a programmable calibration constant h, neglect a number q of least significant bits of the product hx, and display the non-neglected bits as a weight, in integral pounds, for the user. In the event that the A/D converter is saturated ($x = 2^{10}-1 = 1,023$), an overflow indicator will be displayed. The transducers to be used always produce exactly zero volts with no weight applied, but vary from 0.25 lbs. per A/D count to 0.35 lbs. per A/D

[20]From Table V, 103/64 also appears to be an attractive rational number, because the division by 64 can be accomplished by shifting $hx+z$ right by 5 bits. However, with the TMS370, each shift of a 16-bit operand requires two `RRC` instructions, for a total of 10 instructions at 8 clock cycles each, or 80 clock cycles (more than the `DIV` instruction). Therefore, 243/151 is the more attractive rational number.

[21]A value of 256 rather than 255 must be used for x_{MAX} because it is assumed that $x \in [0, 256)$.

count in their transfer characteristic. Market research has shown that users strongly dislike bathroom scales which overstate their weight; but simultaneously prefer that their weight not be understated by more than 2 lbs. In the design of the custom integrated circuit for the family of bathroom scales, what value of q should be chosen? How accurate will each bathroom scale be? How many bits must be reserved for h, and what strategy should be used to choose h based on the transfer characterisic r_I of each transducer?

Solution: From the problem statement, $r_I \in [0.25, 0.35]$, $r_{IMAX} = 0.35$, $x_{MAX} = 1,023$[22], and it is required that $L(x) - F(x) \in [-2, 0]$. It is also required that $r_A \leq r_I$ (as in Eqs. 93, 94); otherwise it is possible that $\exists x \in [0, x_{MAX}]$, $L(x) > F(x)$, which contradicts the product requirements.

With $x_{MAX} = 1,023$, $\varepsilon_{SPAN} = 2$, and $r_{IMAX} = 0.35$, (97) yields $q = 11$ as the minimum choice for q to meet accuracy requirements.

With $x_{MAX} = 1,023$, $z = 0$, $q = 11$, $2^q = 2,048$, and $r_{IMAX} = 0.35$, (95) predicts that $L(x) - F(x) \in (-1.84, 0]$.

With $q = 11$, $2^q = 2,048$, and $r_{IMAX} = 0.35$, (98) yields $h_{MAX} = 716$ and 10 bits must be reserved in the custom integrated circuit for the calibration factor h. For each r_I to be tabulated, h should be chosen by (93): $h = \lfloor r_I 2^q \rfloor = \lfloor 2,048 r_I \rfloor$.

VIII. Conclusion

The techniques presented in this paper demonstrate how linear scaling functions of the form $y = r_I x$ ($r_I \in \mathbb{R}^+$, not necessarily $\in \mathbb{Q}^+$) can be implemented on inexpensive 4-bit and 8-bit microcontrollers by approximating r_I with a rational scaling factor $r_A = h/k$ ($h \in \mathbb{Z}^+$, $k \in \mathbb{N}$). Several methods for choosing h and k and several implementation techniques were presented.

A detailed analysis of approximation error due to the inability to choose $r_A = r_I$ and due to the quantization inherent in digital systems and integer arithmetic was provided. It was shown that the approximation error can be bounded when it is known that the approximation will be used only over a domain of $[0, x_{MAX}]$, $x_{MAX} \in \mathbb{N}$. It was also shown that by introducing an integral parameter z, the error function could be adjusted to be never negative, never positive, or centered about zero.

For cases where r_I is known at design time, three algorithms were presented that allow h and k to be chosen subject to the constraints $h \leq h_{MAX}$ and $k \leq k_{MAX}$ so as to place $r_A = h/k$ as close as possible to r_I. For cases where r_I is not precisely known at design time, methods were presented for designing tabulated scalings and bounding the approximation error introduced.

Important results from number theory were presented which show that although the worst-case error in placing r_A decreases as $1/N$, the typical error decreases as $1/N^2$.

Techniques which allow linear approximations to be performed economically on microcontrollers of varying capability were also presented. It was shown how the form of $r_A = h/k$ could be modified to improve software efficiency, enable use of only scaleable (non-division) instructions, or facilitate tabulation of scaling factors.

[22] From the problem statement, $x = 1,023$ will generate an overflow display, so we need not consider $x \in [1,023,\ 1,024)$.

IX. Acknowledgements

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, Mircea Munteanu, Adam Gibson, and Virgil (of the `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; Yu-Tzu Tsai and William J. Hagen for support from the IEEE; 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.

X. Definitions, Acronyms, Abbreviations, Variables And Mathematical Notation

$\lfloor \cdot \rfloor$

Used to denote the *floor(·)* function. The *floor(·)* function is the largest integer not larger than the argument.

$\lceil \cdot \rceil$

Used to denote the *ceiling(·)* function. The *ceiling(·)* function is the smallest integer not smaller than the argument.

a/b

An arbitrary rational number.

coprime

Two integers that share no prime factors are *coprime*. *Example:* 6 and 7 are coprime, whereas 6 and 8 are not.

F_N

The Farey series of order N. The Farey series is the ordered set of all reduced rational numbers with a denominator not larger than N.

greatest common divisor (g.c.d.) The greatest common divisor of two integers is the largest integer which divides both integers without a remainder. *Example:* the g.c.d. of 30 and 42 is 6.

H/K*, *h/k*, *h'/k'*, *h''/k''*, *h_i/k_i

Terms in a Farey series of order N.

irreducible

A rational number p/q where p and q are coprime is said to be *irreducible*. Equivalently, it may be stated that p and q share no prime factors or that the greatest common divisor of p and q is 1.

KPH

Kilometers per hour.

MPH

Miles per hour.

$\mathbb{N}$

The set of positive integers (natural numbers).

$\mathbb{Q}$

The set of rational numbers.

$\mathbb{Q}^+$

The set of non-negative rational numbers.

$\boldsymbol{r_A}$

The rational number h/k used to approximate an arbitrary real number r_I.

$\boldsymbol{r_I}$

The real number, which may or may not be rational, which is to be approximated by a rational number $r_A = h/k$.

$\mathbb{R}$

The set of real numbers.

$\mathbb{R}^+$

The set of non-negative real numbers.

reduced

See *irreducible*.

$\boldsymbol{s_k = p_k/q_k}$

The kth convergent of a continued fraction.

$\boldsymbol{x_{MAX}}$

The largest element of the domain for which the behavior of an approximation must be guaranteed. In this paper, most derivations assume that $x \in [0, x_{MAX}]$, $x_{MAX} \in \mathbb{N}$.

$\mathbb{Z}$

The set of integers.

$\mathbb{Z}^+$

The set of non-negative integers.

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.

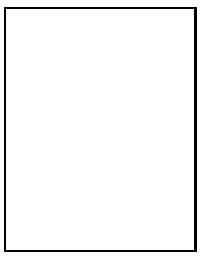
[5] A. Ya. Khinchin, *Continued Fractions*, University Of Chicago Press, 1964; Library Of Congress Catalog Card Number 64-15819.

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

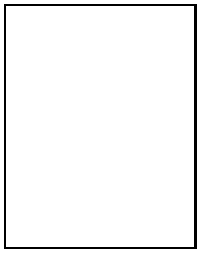
[7] C. D. Olds, *Continued Fractions*, Randam House, 1963, Library Of Congress Catalog Card Number 61-12185.

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

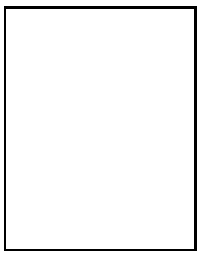
[9] M. R. Schroeder, *Number Theory In Science And Communication*, ISBN 3-540-62006-0.



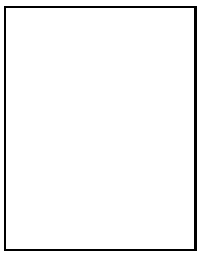
David T. Ashley (biography not yet included in document).



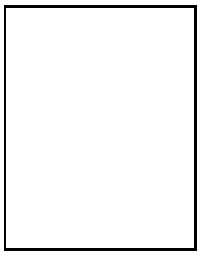
Joseph P. DeVoe (biography not yet included in document).



Cory Pratt (biography not yet included in document).



Karl Perttunen (biography not yet included in document).



Anatoly Zhigljavsky (biography not yet included in document).

Fig. 5. Version Control Information (For Reference Only—Will Be Removed Before Submission Of Paper)

```
LaTeX compile date:  December 8, 2000.

PVCS version control information:
$Header:   J:/arch/dashley1/misc_asn/misc_asn/pq_paper.tev   1.62   25 May 2000 14:05:22   dashley1  $.
```