IMPROVED PRECISION OF FIXED-POINT ALGORITHMS BY MEANS OF COMMON FACTORS

Yuriy A. Reznik[*]

Information Systems Laboratory
Stanford University
353 Serra Mall, Stanford, CA 94305
Email: yreznik@stanford.edu

Arianne T. Hinds, and Joan L. Mitchell

Ricoh | IBM
InfoPrint Solutions Company
6300 Diagonal Hwy, Boulder, CO, 80301
Email: {arianne, joanm}@us.ibm.com

ABSTRACT

We describe a general techniques for improving precision of fixed-point implementations of signal processing algorithms (such as filters, transforms, etc.) by introducing "common factors". These factors are applied to groups of real constants that need to be approximated by dyadic rational numbers, and we show that by carefully choosing values of common factors, the errors of final dyadic rational approximations can be significantly reduced. We show that the problem of design of such approximations is related to the classic Diophantine approximation problem, and include examples explaining how it can be solved, and used for improving practical designs.

Index Terms— Signal processing, fixed point algorithms, Diophantine approximations.

1. INTRODUCTION

One of the most basic tasks that arises in the design of fixed-point signal processing algorithms is that of approximating a given set of real (and possibly irrational) constants $\theta_1, \dots, \theta_m$, $(m \geqslant 2)$ with a set of rational numbers with common dyadic denominator:

$$\theta_1 \approx p_1/2^k \;,\; \dots \;,\; \theta_m \approx p_m/2^k \;, \qquad (1)$$

where $p_1, \dots, p_m$, and k are integers.

This way the numbers $\theta_1, \dots, \theta_m$ can be approximately represented in computer's memory by integers $p_1, \dots, p_m$, which, in turn, can be used for execution of basic arithmetic operations. For example, multiplications of an input variable x by $\theta_1, \dots, \theta_m$ can be conveniently mapped into integer instructions as follows:

$$x\theta_i \approx x p_i/2^k \rightsquigarrow (x * p_i) \gg k \;, \;\; (i = 1, \dots, m)$$

where $*$ and $\gg$ denote integer multiplication and bit-wise right shift operations correspondingly.

The key parameter that influences the complexity of algorithms using dyadic rational approximations (1) is the number of "precision bits" k. In software designs, this parameter is often constrained by the width of registers (e.g. 8, 16 or 32), and failure to meet such a constraint can possibly result in doubling (or in some cases – quadrupling) of the execution time. In hardware designs parameter k directly affects the number of gates needed to implement adders and multipliers.

The precision of approximations (1) also depends on the parameter k. Thus, given k and θ_i, the best choice of p_i yields

$$\left|\theta_i - p_i/2^k\right| = 2^{-k}\left|2^k\theta_i - p_i\right| = 2^{-k}\min_{z \in \mathbb{Z}}\left|2^k\theta_i - z\right| \leqslant 2^{-k-1},$$

which means, that minimum worst case magnitude of error

$$\Delta(k) = \min_{p_1,\dots,p_m} \max_i \left\{\left|\theta_i - p_i/2^k\right|\right\} \,. \qquad (2)$$

is also bounded by

$$\Delta(k) \leqslant 2^{-k-1} \,. \qquad (3)$$

In simple terms, this means, that on average, each bit of precision in dyadic approximations (1) reduces their worst case error at least by half.

This last observation is crucial for understanding precision-complexity tradeoffs in conventional fixed-point designs. It also underscores the importance of finding more efficient (with faster decaying errors) techniques for "importing" of real (and in particular, irrational) numbers into fixed-point algorithms.

In this paper, we study one such possible technique, involving the use of a "common factor". The main idea of this technique is to introduce an additional parameter ξ and use it for minimization of errors in approximations:

$$\theta_1\xi \approx p_1/2^k \;,\; \dots \;,\; \theta_m\xi \approx p_m/2^k \,. \qquad (4)$$

where $p_1, \dots, p_m$, and k are integers.

We note, that in many practical situations, the uniform scale of the original values $\theta_1, \dots, \theta_m$ by ξ can be either ignored (e.g. when it has no effect on the output), or "neutralized" by applying the inverse factor $1/\xi$ to constants in adjacent stages of the algorithm. In other words, we assume that

[*] On leave from Qualcomm Inc., 5775 Morehouse Dr., San Diego, CA.

the use of our modified approximations (4) instead of direct ones (1) will be feasible in practical designs.

We show, that for infinitely many k, by carefully choosing the value of a common factor ξ the equivalent (scaled by $1/\xi$) worst case error of approximations (4):

$$\Delta_\xi(k) = \tfrac{1}{\xi} \min_{p_1,\ldots,p_m} \max_i \left\{ \left| \theta_i \xi - p_i/2^k \right| \right\} \qquad (5)$$

can be made as small as

$$\Delta_\xi(k) \lesssim 2^{-k\left(1+\frac{1}{m-1}\right)} . \qquad (6)$$

In other words, we show that common-factor-based approximations can be significantly more precise than direct ones. We note that the magnitude of the achievable gain is particularly striking for small m. For example, when $m = 2$, the right side in (6) turns into 2^{-2k}, which implies, that the use of a common-factor might reduce the number of required precision bits by half!

The rest of this paper is organized as follows. Section 2 contains analysis of common-factor-based approximations and formulation of our main results. Practical examples of using this technique are given in Section 3.

2. PRECISION OF APPROXIMATIONS WITH COMMON FACTORS

2.1. Minimizing errors of pairs of approximations.

Consider first a special case when $m = 2$. By $\delta_1(\xi)$ and $\delta_2(\xi)$ we denote individual errors of approximations (4):

$$\delta_1(\xi) = \theta_1 \xi - p_1/2^k , \quad \delta_2(\xi) = \theta_2 \xi - p_2/2^k , \qquad (7)$$

and our first task would be to see if $\max\{|\delta_1(\xi)|, |\delta_2(\xi)|\}$ can be minimized by adjusting ξ.

We claim the following.

Lemma 1. *Let θ_1, θ_2 be real numbers, such that $\theta_1\theta_2 > 0$, and let k, p_1, and p_2 be integers. Then, there exist values ξ^* and δ^*, such that*

$$\delta^* = \max\{|\delta_1(\xi^*)|, |\delta_2(\xi^*)|\} = \min_\xi \max\{|\delta_1(\xi)|, |\delta_2(\xi)|\} .$$

These values are:

$$\xi^* = \tfrac{1}{2^k} \tfrac{p_1+p_2}{\theta_1+\theta_2} , \qquad (8)$$

and

$$\delta^* = \tfrac{1}{2^k} \left| \theta_1 \tfrac{p_1+p_2}{\theta_1+\theta_2} - p_1 \right| = \tfrac{1}{2^k} \left| \theta_2 \tfrac{p_1+p_2}{\theta_1+\theta_2} - p_2 \right| . \qquad (9)$$

Proof. Condition $\theta_1\theta_2 > 0$ implies that both $\delta_1(\xi)$ and $\delta_2(\xi)$ are non-constant and have the same direction of growth with ξ.

If $\delta_1(\xi)$ and $\delta_2(\xi)$ intersect 0 at the same location, then there exists point ξ^* such that $\delta_1(\xi^*) = \delta_2(\xi^*) = 0$. This implies that

$$\xi^* = \tfrac{1}{2^k} \tfrac{p_1}{\theta_1} = \tfrac{1}{2^k} \tfrac{p_2}{\theta_2} ,$$

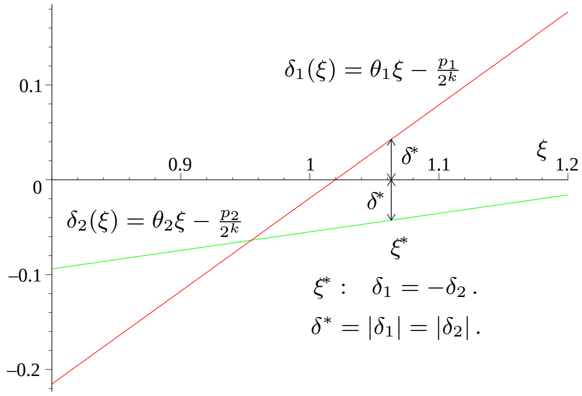

Fig. 1. Finding $\min_\xi \max \left\{ \left| \theta_1\xi - \frac{p_1}{2^k} \right|, \left| \theta_2\xi - \frac{p_2}{2^k} \right| \right\}$.

which is a special case of (8).

If $\delta_1(\xi)$ and $\delta_2(\xi)$ intersect 0 at different locations, then there exists ξ^* such that (see Fig. 1):

$$\delta_1(\xi^*) = -\delta_2(\xi^*) . \qquad (10)$$

Moreover, since both $\delta_1(\xi)$ and $\delta_2(\xi)$ have same direction of growth, moving ξ away from ξ^* will lead to asymmetric changes in absolute values of $\delta_1(\xi)$ or $\delta_2(\xi)$. That is, one of them will increase. Therefore, ξ^* is the point of minimum of $\max\{|\delta_1(\xi)|, |\delta_2(\xi)|\}$.

By solving (10) with respect to ξ^* we arrive at formula (8), and by plugging (8) in (7), and using (10) we arrive at (9). □

2.2. Associated Diophantine approximation

Let us now further assume that p_1, p_2 have same signs as θ_1, and θ_2. Then, by denoting $p = p_1$, $q = p_1 + p_2$, and

$$\theta^* = \tfrac{\theta_1}{\theta_1+\theta_2} . \qquad (11)$$

we observe that both parts of (9) turn into

$$\delta^* = \tfrac{|q|}{2^k} \left| \theta^* - p/q \right| .$$

By further de-scaling this quantity by ξ^* we arrive at

$$\delta^*/\xi^* = |\theta_1 + \theta_2| \, |\theta^* - p/q| , \qquad (12)$$

which means, that by plugging $\xi = \xi^*$, the problem of finding minimum of the worst case error of a pair of scaled dyadic rational approximations

$$\Delta_\xi(k) = \tfrac{1}{\xi} \min_{p_1,p_2} \max\{|\delta_1(\xi)|, |\delta_2(\xi)|\} .$$

becomes equivalent to the problem of finding rational approximations of a single number θ^*

$$\theta^* \approx p/q . \qquad (13)$$

Furthermore, if θ^* is irrational, then (13) turns into a classic Diophantine approximation problem [1].

The following result from Diophantine approximation theory (cf. [1, p. 11, Theorem V]) will be useful in our context.

Fact 1. *Let θ be irrational. Then there exist infinitely many integers q and p such that*

$$|\theta - p/q| < \kappa(\theta) q^{-2}, \tag{14}$$

where:

$$\kappa(\theta) = \left[\begin{array}{ll} 5^{-1/2}, & \textit{if } \theta = \frac{r\psi+s}{u\psi+v}, \textit{ where:} \\ & \psi = \frac{\sqrt{5}-1}{2}\,;\ r,s,u,v \in \mathbb{Z}\,, \\ & \textit{such that } rv - us = \pm 1\,, \\ 2^{-3/2}, & \textit{otherwise}\,. \end{array} \right. \tag{15}$$

2.3. Main result for approximations of pairs of constants

We state the following.

Theorem 1. *Let θ_1, θ_2 be irrational numbers of the same sign. Then, there exist infinitely many integers k and real numbers ξ, such that*

$$\begin{aligned} \Delta_\xi(k) &= \tfrac{1}{\xi} \min_{p_1,p_2} \max \left\{ \left|\theta_1 \xi - p_1/2^k\right|, \left|\theta_2 \xi - p_2/2^k\right| \right\} \\ &< \kappa\left(\tfrac{\theta_1}{\theta_1+\theta_2}\right) \tfrac{4}{|\theta_1+\theta_2|}\, 2^{-2k} = O\left(2^{-2k}\right). \end{aligned} \tag{16}$$

Proof. We use the following construction.

By assuming that $\xi = \xi^*$, and solving the associated Diophantine approximation problem (13), we find integers p, q satisfying precision constraint (14) of Fact 1. This also gives us integer factors $p_1 = p$ and $p_2 = q - p$ for our dyadic approximations. In order to select k, we can use some additional constraints. For example, we can require

$$1/2 < \xi^* \leqslant 1\,, \tag{17}$$

which is satisfied by choosing $k = \lceil \log_2 \left(q/(\theta_1 + \theta_2) \right) \rceil$.

Then, by plugging Diophantine precision bound (14) in (12), using lower bound for ξ^* from (17), and some simple algebra, we arrive at expression (16) claimed by the theorem. □

2.4. Extension of analysis to m-ary case

We now turn our attention to a problem of finding dyadic rational approximations for larger ($m > 2$) sets of numbers:

$$\theta_1 \xi \approx p_1/2^k\,,\ \ldots\,,\ \theta_m \xi \approx p_m/2^k\,. \tag{18}$$

For simplicity, we assume that all numbers $\theta_1, \ldots, \theta_m$ and $p_1, \ldots, p_m$ are either positive or negative.

From Lemma 1, we know that for any pair of numbers θ_i, θ_j, $i \neq j$, we can compute factor

$$\xi^*_{ij} = \tfrac{1}{2^k} \tfrac{p_i + p_j}{\theta_i + \theta_j}\,, \tag{19}$$

which will “symmetrize” errors of approximations:

$$\delta^*_{ij} = \tfrac{1}{2^k} \left| \theta_i \tfrac{p_i+p_j}{\theta_i+\theta_j} - p_i \right| = \tfrac{1}{2^k} \left| \theta_j \tfrac{p_i+p_j}{\theta_i+\theta_j} - p_j \right|. \tag{20}$$

and which will turn them into a Diophantine approximation:

$$\delta^*_{ij} = \tfrac{|q_{ij}|}{2^k} \left| \theta^*_{ij} - p_{ij}/q_{ij} \right|. \tag{21}$$

where $p_{ij} = p_i$, $q_{ij} = p_i + p_j$, and

$$\theta^*_{ij} = \tfrac{\theta_i}{\theta_i+\theta_j}. \tag{22}$$

By applying ξ^*_{ij} to the remaining constants $\{\theta_k, k \neq i,j\}$, we note that their approximations also turn into Diophantines

$$\left|\theta_k \xi^*_{ij} - p_k/2^k\right| = \tfrac{1}{2^k} \left|\theta_k \tfrac{p_i+p_j}{\theta_i+\theta_j} - p_k\right| = \tfrac{|q_{ij}|}{2^k} \left| \theta^*_{k|ij} - p_k/q_{ij} \right|,$$

where, however, the resulting constants

$$\theta^*_{k|ij} = \tfrac{\theta_k}{\theta_i+\theta_j}\,, \tag{23}$$

and errors of their approximations are different.

This means that by using factor ξ^*_{ij} we can reduce the problem of finding m dyadic rational approximations (18) to one of finding $m-1$ simultaneous Diophantine approximations:

$$\theta^*_{ij} \approx p_{ij}/q_{ij}, \left\{ \theta^*_{k|ij} \approx p_k/q_{ij}, k \neq i,j \right\}. \tag{24}$$

The relevant result from Diophantine approximation theory is given below (cf. [1, p. 14, Theorem III], [2, p.138]):

Fact 2. *Let $\theta_1, \ldots, \theta_m$, $(m \geqslant 2)$ be irrationals. Then, there are infinitely many integers q and $p_1, \ldots, p_m$, such that*

$$\max_i \left\{ |\theta_i - p_i/q| \right\} < \tfrac{m}{m+1}\, q^{-1-1/m}\,. \tag{25}$$

We are now ready to formulate and prove our main result.

Theorem 2. *Let $\theta_1, \ldots, \theta_2$ be $m > 2$ irrational numbers of the same sign. Then, there exist infinitely many integers k and real values ξ, such that*

$$\begin{aligned} \Delta_\xi(k) &= \tfrac{1}{\xi} \min_{p_1,\ldots,p_m} \max_i \left\{ \left|\theta_i \xi - p_i/2^k\right| \right\} \\ &< \tfrac{m-1}{m} \left(\min_{ij} \{|\theta_i + \theta_j|\} \right)^{-\frac{1}{m-1}} 2^{-(k-1)\left(1+\frac{1}{m-1}\right)} \\ &= O\left(2^{-k\left(1+\frac{1}{m-1}\right)} \right). \end{aligned} \tag{26}$$

Proof. We use the following construction.

We scan all $\binom{m}{2}$ pairs of indices i, j, and find a pair, for which the normalized (by $1/\xi^*_{ij}$) worst case error:

$$\begin{aligned} &\tfrac{1}{\xi^*_{ij}} \min_{p_{ij},p_k} \tfrac{|q_{ij}|}{2^k} \max \left\{ \left|\theta^*_{ij} - \tfrac{p_{ij}}{q_{ij}}\right|, \left|\theta^*_{k|ij} - \tfrac{p_k}{q_{ij}}\right|, k \neq i,j \right\} \\ &\quad = |\theta_i + \theta_j| \min_{p_{ij},p_k} \max \left\{ \left|\theta^*_{ij} - \tfrac{p_{ij}}{q_{ij}}\right|, \left|\theta^*_{k|ij} - \tfrac{p_k}{q_{ij}}\right|, k \neq i,j \right\} \end{aligned}$$

is the smallest one.

Then, by applying Fact 2, using (19) to replace q_{ij} with 2^k and ξ^*_{ij}, and subsequently, bounds $1/2 < \xi^*_{ij} \leqslant 1$ (which is attainable by choice of k), and $|\theta_i + \theta_j| \geqslant \min_{ij}\{|\theta_i + \theta_j|\}$, we arrive at estimate (26) claimed by the theorem. □

Table 1. Approximations of a pair of constants $\theta_1 = \cos\left(\frac{\pi}{16}\right)$, and $\theta_2 = \cos\left(\frac{7\pi}{16}\right)$.

Direct dyadic approximations: $\theta_1 \approx p_1/2^k, \quad \theta_2 \approx p_2/2^k$				Associated Diophantine approximation: $\theta^* = \theta_1/(\theta_1{+}\theta_2) \approx p/q$			Dyadic approximations with common factor ξ^*: $\theta_1\xi^* \approx p_1/2^k, \quad \theta_2\xi^* \approx p_2/2^k$			
k	p_1	p_2	$\max_i \left\|\theta_i - p_i/2^k\right\|$	q	p	$\|\theta^* - p/q\|$	$\xi^* = \frac{1}{2^k}\frac{q}{\theta_1+\theta_2}$	p_1	p_2	$\frac{1}{\xi^*}\max_i \left\|\theta_i\xi^* - p_i/2^k\right\|$
1	2	0	0.1950903220							
2	4	1	0.0549096780							
3	8	2	0.0549096780	6	5	0.0007559856	0.6378225711	5	1	0.0008889451
4	16	3	0.0192147196							
5	31	6	0.0120352804							
6	63	12	0.0075903220							
7	126	25	0.0035897196							
8	251	50	0.0003165304							
9	502	100	0.0003165304	440	367	0.0000015901	0.7308383627	367	73	0.0000018698
10	1004	200	0.0003165304							
11	2009	400	0.0002221780	1543	1287	0.0000001172	0.6407293146	1287	256	0.0000001378

Table 2. Approximations of a pair of constants $\theta_1 = \cos\left(\frac{3\pi}{16}\right)$, and $\theta_2 = \cos\left(\frac{5\pi}{16}\right)$.

Direct dyadic approximations: $\theta_1 \approx p_1/2^k, \quad \theta_2 \approx p_2/2^k$				Associated Diophantine approximation: $\theta^* = \theta_1/(\theta_1{+}\theta_2) \approx p/q$			Dyadic approximations with common factor ξ^*: $\theta_1\xi^* \approx p_1/2^k, \quad \theta_2\xi^* \approx p_2/2^k$			
k	p_1	p_2	$\max_i \left\|\theta_i - p_i/2^k\right\|$	q	p	$\|\theta^* - p/q\|$	$\xi^* = \frac{1}{2^k}\frac{q}{\theta_1+\theta_2}$	p_1	p_2	$\frac{1}{\xi^*}\max_i \left\|\theta_i\xi^* - p_i/2^k\right\|$
1	2	1	0.1685303877							
2	3	2	0.0814696123	5	3	0.0005438163	0.9011997775	3	2	0.0007542949
3	7	4	0.0555702330							
4	13	9	0.0189696123							
5	27	18	0.0122803877							
6	53	36	0.0069297670							
7	106	71	0.0033446123							
8	213	142	0.0008827330	367	220	0.0000011428	1.0335634948	220	147	0.0000015851
9	426	284	0.0008827330							
10	851	569	0.0004149248							
11	1703	1138	0.0000938295	2207	1323	0.0000000918	0.7769327769	1323	884	0.0000001273

3. EXAMPLE OF DESIGN USING COMMON FACTORS - BASED APPROXIMATIONS

As an example, let us consider a design of an 8x8 Discrete Cosine Transform (DCT) proposed by M. Vetterli and A. Ligtenberg [3]. We notice, that this is a scaled transform (that is, we can move leading factors outside 1D transforms), and that its 1-stage odd butterflies involve multiplications by pairs: (C_1, C_7), and (C_3, C_5), where $C_k = \cos\left(\frac{k\pi}{16}\right)$. Hence, we can introduce common factors for each of these pairs.

We illustrate the process of deriving scaled approximations for these pairs in Tables 1, and 2. First large columns in these tables illustrate the use of direct dyadic rational approximations. It is shown, that their worst case errors decay relatively slowly. Second large columns show solutions for the associated Diophantine approximations. We note that these results are sparse (not every denominator q leads to a precision stated by Fact 1), but their errors decay very rapidly. Finally, in third columns, we show parameters of approximations with common factors, derived from Diophantine solutions.

It can be seen, that common-factor-based approximations are remarkably more precise than the original dyadic rational approximations. For example, the top 3-bit solution from Table 1, reaches precision which would be normally achievable by using 8 bits. Similar observation can also be made regarding the top (2-bit) common-factor-based solution in Table 2.

In passing, we should note that for some applications precision of those top solutions might already be sufficient. Thus, using these solutions we were able to design an inverse (IDCT) transform passing all compliance tests of the JPEG standard.

4. REFERENCES

[1] J. W. S. Cassels, *An Introduction to Diophantine Approximations*, Cambridge University Press, 1957.

[2] M. Gröetschel, L. Lovácz, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, Springer, 1988.

[3] M. Vetterli and A. Ligtenberg, ”A Discrete Fourier-Cosine Transform Chip”, *IEEE Journal on Selected Areas in Communications*, Vol. SAC-4, No. 1, pp. 49-61, Jan. 1986.