LIMITS OF FINITE WORDLENGTH FIR DIGITAL FILTER DESIGN

Dušan M. Kodek

Faculty of Computer and Information Science
University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
kodek@fri.uni-lj.si

ABSTRACT

It has been known for some time that it is not possible to meet arbitrarily severe FIR filter specifications with fixed b-bit wordlength by sufficiently increasing the filter length N. For any given non-trivial specification there is a nonzero lower bound on the approximation error, below which it is not possible to go, no matter how large the value of N. For practical purposes it is even more useful to know a lower bound for given N and b. This bound represents a finite wordlength FIR filter design limit which is of theoretical importance and has not been known so far. This paper presents a method for computing this limit. The method is based on a lower bound theorem and can be used to estimate the approximation error limit in practical finite wordlength FIR design cases. It is also useful in the algorithm for the optimal finite wordlength design.

1. INTRODUCTION

There are many situations when it is not practical to use the optimal FIR digital filter coefficients obtained by some "infinite precision"[1] algorithm. One may, for example, wish to use a fixed point DSP processor which is usually cheaper and/or faster than a floating point one. The number of bits b that can be used to represent the filter coefficients will in general depend on the filter length N, processor properties, and on signal quantization. It is true in most cases that filter coefficients with as short as possible wordlength b are desirable. This gives more design freedom for other aspects of system design.

Replacing the optimal filter coefficients with the b-bit ones degrades filter's frequency response. The number of bits b must therefore not be too short or the filter will no longer be good enough. The designer faces the following question: given the filter specifications, what is the lowest number of bits b that will give an acceptable finite wordlength filter? The answer is a difficult one. It is clear that it cannot be answered by rounding the coefficients to b bits and computing the frequency response – rounding gives a suboptimal filter which can be up to 30 dB (or more) worse than the optimal filter. What is required is a frequency response of an optimal b-bit filter. Here lies the problem. Designing an optimal b-bit filter requires a solution of an NP-complete approximation problem that is very difficult to solve [1]. Most designers would prefer to solve it only if there is some assurance that the result is worth investing into a very long computation time.

This paper presents a method that gives such assurance. It is based on a lower bound for the increase of the approximation error that is caused by the b-bit constraint. This bound is a theoretical limit on the performance of a finite wordlength FIR digital filter of length N.

[1]The so called "infinite precision" coefficients are typically 32-bit floating point numbers. Though the 32-bit wordlength is hardly infinite, it is much longer than practical finite wordlengths that we are interested in.

2. THE APPROXIMATION PROBLEM

Let us start with the usual unconstrained design problem in which the coefficients a_j can be any real numbers. The optimal FIR digital filter is defined as a cosine polynomial $P(\omega)$ of order n

$$P(\omega) = \sum_{j=0}^{n} a_j \cos j\omega, \tag{1}$$

that minimizes the expression

$$\|D - P\|_\infty = \max_{a \le \omega \le b} |W(\omega)(D(\omega) - P(\omega))|. \tag{2}$$

The real function $D(\omega)$ is the desired frequency response, the weighting function $W(\omega)$ is by definition real and positive, and the interval $[a, b]$ is a subset (or a union of subsets) of the interval $[0, \pi]$. The polynomial order n is related to the filter length N; for positive symmetry $n = (N-1)/2$ for odd N or $n = N/2 - 1$ for even N. There are also simple formulas which give filter's impulse response from the coefficients a_j.

Let $P^*(\omega)$ be the optimal approximation to $D(\omega)$

$$\begin{aligned} P^*(\omega) &= \sum_{j=0}^{n} a_j^* \cos j\omega, \\ \|D - P^*\|_\infty &\le \|D - P\|_\infty, \quad \forall P(\omega). \end{aligned} \tag{3}$$

Several algorithms, from linear programming to various versions of the *exchange algorithm*, can be used to find $P^*(\omega)$ (ref. [2], pp.85-110). Design of optimal FIR digital filters has been considered a relatively easy problem ever since Parks and McClellan showed how to apply the Remez exchange algorithm. The main reason for this is the following well-known property of the optimal minimax approximation: there are exactly $(n+2)$ so called extremal points in $[a, b]$ at which the approximation error achieves its maximum. Let $\{\omega_i; i = 0, 1, \cdots, n+1\}$ be these extremal points. The following equations hold

$$W(\omega_i)(D(\omega_i) - \sum_{j=0}^{n} a_j^* \cos j\omega_i) = (-1)^i h^*, \; i = 0, 1, \cdots, n+1, \tag{4}$$

and $|h^*|$ denotes the optimal approximation error. Equations (4) also imply

$$\|D - P^*\|_\infty = |h^*|. \tag{5}$$

Things change dramatically when a finite wordlength constraint is introduced. We can, without loss of generality, make this constraint equivalent to forcing the coefficients $\{a_j; j = 0, 1, \cdots, n\}$ to be b-bit integers[2]. This will in most practical cases also require a multiplication of $D(\omega)$ and division of $W(\omega)$ by a suitable scaling factor. The scaling is simple and will be ignored in this paper. In other words, $D(\omega)$ and $W(\omega)$ are left unchanged, and the approximating polynomial $P(\omega)$ is replaced by

$$P(\omega) = \sum_{j=0}^{n} a_j \cos j\omega, \quad a_j \in I_b, \tag{6}$$

where I_b denotes a finite set of signed integers $I_b = \{-2^{b-1}, \cdots, -1, 0, 1, \cdots, 2^{b-1}\}$. The problem of finding the optimal integer polynomial $P(\omega)$ is much more difficult than the unconstrained case, although it may not appear so at first.

Notation $P(\omega)$ will from here on denote a polynomial with b-bit integer coefficients, while $P^*(\omega)$ remains the optimal unconstrained polynomial. Since $P^*(\omega)$ is unique the approximation error increases if $P(\omega) \neq P^*(\omega)$. We have

$$\epsilon = \|D - P\|_\infty - \|D - P^*\|_\infty = \|D - P\|_\infty - |h^*|, \tag{7}$$

where $\epsilon > 0$.

The problem we wish to solve can be stated like this: what is the minimum ϵ given the best possible integer coefficients $\{a_j; j = 0, 1, \cdots, n\}$? Or stated differently, how much will $\|D - P\|_\infty$ increase relative to $\|D - P^*\|_\infty$ because of the b-bit finite wordlength restriction. The lowest possible ϵ is needed to answer this question. This lowest ϵ is a theoretical limit on the performance of a given b-bit finite wordlength FIR digital filter of length N. Let us denote it as δ and define it formally as

$$\delta = \min_{P(\omega) \in I_b} (\epsilon) = \min_{P(\omega) \in I_b} (\|D - P\|_\infty) - |h^*|. \tag{8}$$

Note that we are looking over all b-bit integer polynomials, not just one particular $P(\omega)$. The optimal b-bit $P(\omega)$ and its coefficients a_j must be known in order to compute δ. These are, as mentioned in the introduction, difficult to compute so we wish to avoid this computation. An easily computed lower bound for δ will be derived instead.

3. LOWER BOUND THEOREM

The first step in our search for a lower bound of δ is a derivation of a lower bound for the approximation error increase ϵ for a single non-optimal polynomial $P(\omega)$. A special property of all functions that satisfy the Haar condition is useful here. It says that a $(n+1)(n+1)$ matrix with elements $\{\cos j\omega_i; j = 0, 1, \cdots, n; i = 0, 1, \cdots, n\}$, where $\{\omega_i; i = 0, 1, \cdots, n\}$ can be any set of $(n+1)$ distinct points from $[a, b]$, is always non-singular (see [2], pp.97-99). This means that there exist multipliers $\{\sigma_i; i = 0, 1, \cdots, n+1\}$, not all zero, that satisfy the conditions

$$\sum_{i=0}^{n+1} \sigma_i \cos j\omega_i = 0, \quad j = 0, 1, \cdots, n \tag{9}$$

[2]The integers are chosen for convenience. Any other finite set of numbers can be used instead.

for any $(n+2)$ points ω_i from interval $[a, b]$. It is easy to see that equations (1) and (9) imply

$$\sum_{i=0}^{n+1} \sigma_i P(\omega_i) = 0 \tag{10}$$

for any $P(\omega)$. The numbers $\{\sigma_i; i = 0, 1, \cdots, n+1\}$ have a very important property. All are nonzero and their signs alternate. That is

$$\text{sign}(\sigma_{i+1}) = -\text{sign}(\sigma_i), \quad i = 0, 1, \cdots, n. \tag{11}$$

The multipliers σ_i are available as a byproduct of the solution that gives optimal infinite precision coefficients a_j^*.

The lower bound for ϵ when a $P(\omega)$ is known is given by the following theorem:

Theorem *Let $P^*(\omega)$ be the optimal weighted minimax approximation to a real function $D(\omega)$ on the interval $[a, b]$ and let $P(\omega)$ be any other cosine polynomial. Then the increase in approximation error is bounded by*

$$\epsilon \geq \max_{0 \leq i \leq n+1} |c_i W(\omega_i)(P^*(\omega_i) - P(\omega_i))|, \tag{12}$$

where ω_i are extremal points and multipliers c_i are defined as

$$c_i = \begin{cases} \dfrac{|\frac{\sigma_i}{W(\omega_i)}|}{\sum_{\substack{k=0 \\ k \neq i}}^{n+1} |\frac{\sigma_k}{W(\omega_k)}|} & \text{if } (-1)^i h^*(P^*(\omega_i) - P(\omega_i)) < 0, \\ 1 \quad \text{otherwise}, & i = 0, 1, \cdots, n+1. \end{cases} \tag{13}$$

The proof is given in [3] and will not be repeated here. The theorem can be used to compute how much the approximation error increases if $P^*(\omega)$ is replaced by $P(\omega)$. Its application is straightforward – the signs of $(-1)^i h^*(P(\omega_i) - P^*(\omega_i))$ for a given $P(\omega)$ are known and c_is are easily obtained from (13). Doing this may hardly seem necessary, since one can simply use equation (2) and compute $\|D - P\|_\infty$ in extremal frequencies ω_i getting an exact increase. But we are not really interested in the case of a single $P(\omega)$. Instead, we need a lower bound that holds over all $P(\omega)$ with integer coefficients a_i. It is here that the theorem becomes quite useful.

4. DERIVATION OF LOWER BOUND

To get a lower bound for δ, we must be able to express it as a function of differences $\{a_j^* - a_j; j = 0, 1, \cdots, n\}$. This will be done following an approach similar to the one used in [4]. Let us first express the approximation error $e(\omega_i)$ that corresponds to $P(\omega)$

$$W(\omega_i)(D(\omega_i) - \sum_{j=0}^{n} a_j \cos j\omega_i) = e(\omega_i), \quad i = 0, 1, \cdots, n+1. \tag{14}$$

The following system of $n+2$ equations with $n+2$ unknowns can now be written using equations (4) and (14)

$$\frac{e(\omega_i)}{W(\omega_i)} = \sum_{j=0}^{n} (a_j^* - a_j) \cos j\omega_i + \frac{(-1)^i}{W(\omega_i)} h^*, \; i = 0, 1, \cdots, n+1. \tag{15}$$

The unknowns are $a_j^* - a_j$ and h^*. Note that the system's matrix is identical to the one in (4). Since (4) is already solved, (to get a_j^*) it is clear that (15) is always invertible. In addition, the inverse can be rather easily obtained as a byproduct of solving (4). It can be written as

$$\begin{aligned} a_j^* - a_j &= \sum_{i=0}^{n+1} g_{ji} \frac{e(\omega_i)}{W(\omega_i)} \quad j = 0, 1, \cdots, n, \\ h^* &= \sum_{i=0}^{n+1} g_{n+1i} \frac{e(\omega_i)}{W(\omega_i)}, \end{aligned} \tag{16}$$

where g_{ji} are the elements of the inverted matrix. To express the differences $a_j^* - a_j$ in terms of $P^*(\omega_i) - P(\omega_i)$, we note that

$$\frac{e(\omega_i)}{W(\omega_i)} = P^*(\omega_i) - P(\omega_i) + \frac{(-1)^i}{W(\omega_i)} h^*. \tag{17}$$

Inserting (17) into (16) gives

$$\begin{aligned} a_j^* - a_j &= \sum_{i=0}^{n+1} g_{ji} (P^*(\omega_i) - P(\omega_i) + \frac{(-1)^i}{W(\omega_i)} h^*), \\ h^* &= \sum_{i=0}^{n+1} g_{n+1i} (P^*(\omega_i) - P(\omega_i) + \frac{(-1)^i}{W(\omega_i)} h^*). \end{aligned} \tag{18}$$

It is easy to see that the h^* term in (18) amounts to zero. Setting $a_j = a_j^*$ for all j gives $P(\omega) = P^*(\omega)$ for all ω and the following property of matrix g_{ij} is revealed

$$\sum_{i=0}^{n+1} g_{ji} \frac{(-1)^i}{W(\omega_i)} = 0, \quad j = 0, 1, \cdots, n, \tag{19}$$

$$\sum_{i=0}^{n+1} g_{n+1i} \frac{(-1)^i}{W(\omega_i)} = 1. \tag{20}$$

Equations (16) and (17) can be rewritten as

$$a_j^* - a_j = \sum_{i=0}^{n+1} g_{ji} (P^*(\omega_i) - P(\omega_i)), \quad j = 0, 1, \cdots, n, \tag{21}$$

$$0 = \sum_{i=0}^{n+1} g_{n+1i} (P^*(\omega_i) - P(\omega_i)). \tag{22}$$

Let us now multiply and divide each term in equations (21)with $(-1)^i c_i W(\omega_i)$

$$a_j^* - a_j = \sum_{i=0}^{n+1} \frac{(-1)^i g_{ji}}{c_i W(\omega_i)} c_i W(\omega_i) (-1)^i (P^*(\omega_i) - P(\omega_i)), \tag{23}$$

$$0 = \sum_{i=0}^{n+1} \frac{(-1)^i g_{n+1i}}{c_i W(\omega_i)} c_i W(\omega_i) (-1)^i (P^*(\omega_i) - P(\omega_i)). \tag{24}$$

These equations contain the terms $c_i W(\omega_i)(P^*(\omega_i) - P(\omega_i))$ which appear in the theorem. Assume for a while that the differences $a_j^* - a_j$ are known. Obviously, the $\max_{0 \le i \le n+1} |c_i W(\omega_i)(P^*(\omega_i) - P(\omega_i))|$ will be minimal if the signs of all the terms in (23) are equal. This gives

$$\begin{aligned} &\max_{0 \le i \le n+1} |c_i W(\omega_i)(P^*(\omega_i) - P(\omega_i))| \\ &\ge \max_{0 \le j \le n} \left(\frac{|a_j^* - a_j|}{\sum_{i=0}^{n+1} |\frac{g_{ji}}{c_i W(\omega_i)}|} \right), \end{aligned} \tag{25}$$

which is exactly what is needed by the theorem in (12). There is a small problem here because the sign of $P^*(\omega) - P(\omega)$ is required by (13) in order to compute c_is. But this is easily solved since (25) assumes that the signs of all the terms in (23) are equal. Or formally

$$\text{sign}(a_j^* - a_j) = \text{sign}(g_{ji}(P^*(\omega_i) - P(\omega_i))), \tag{26}$$

for all i and j ($c_i W(\omega_i)$ are by definition positive). By multiplying both sides of (26) with $(-1)^i h^* g_{ji}$ we see that the $(-1)^i h^* (P^*(\omega_i) - P(\omega_i)) < 0$ criterion in (13) can be replaced by $(-1)^i h^* g_{ji}(a_j^* - a_j) < 0$. For the purpose of computation it is convenient to divide the indices i appearing in (13) into two subsets.

$$i \in \begin{cases} I_{Pj} & \text{if } (-1)^i h^* g_{ji} > 0, \\ I_{Mj} & \text{otherwise}, \end{cases} \quad i = 0, 1, \cdots, n+1. \tag{27}$$

Note that $c_i = 1$ for $i \in I_{Pj}$ if $a_j^* - a_j > 0$ and for $i \in I_{Mj}$ if $a_j^* - a_j < 0$. For all other cases

$$c_i = \frac{|\frac{\sigma_i}{W(\omega_i)}|}{\sum_{\substack{k=0 \\ k \ne i}}^{n+1} |\frac{\sigma_k}{W(\omega_k)}|}, \qquad i = 0, 1 \cdots, n+1. \tag{28}$$

Let us now remove the assumption about knowing the differences $a_j^* - a_j$. This is necessary to get the lower bound for δ (equation (8)) which is valid over all integer polynomials $P(\omega)$. For any set of optimal coefficients a_j^* there exist integers a_{j+} and a_{j-} that are the nearest upper and lower neighbors of a_j^*. In other words, a_{j+} is an element of I_b that gives the smallest (in an absolute sense) negative difference $a_j^* - a_j$ and a_{j-} is an element of I_b that gives the smallest positive difference $a_j^* - a_j$. Having a_{j+} and a_{j-} and using (25), the lower bounds δ_+ for $a_j^* - a_j < 0$ and δ_- for $a_j^* - a_j > 0$ can be written as

$$\delta_+ = \max_{0 \le j \le n} \left(\frac{|a_j^* - a_{j+}|}{\sum_{i \in I_{Pj}} |\frac{g_{ji}}{c_i W(\omega_i)}| + \sum_{i \in I_{Mj}} |\frac{g_{ji}}{W(\omega_i)}|} \right), \tag{29}$$

$$\delta_- = \max_{0 \le j \le n} \left(\frac{|a_j^* - a_{j-}|}{\sum_{i \in I_{Pj}} |\frac{g_{ji}}{W(\omega_i)}| + \sum_{i \in I_{Mj}} |\frac{g_{ji}}{c_i W(\omega_i)}|} \right). \tag{30}$$

It is obvious that there are no integer coefficients a_j that could possibly give lower deviation increase than δ_+ or δ_-. We have thus derived the lower bound

$$\delta \ge \min(\delta_+, \delta_-). \tag{31}$$

5. IMPROVEMENTS AND EXPERIMENTAL RESULTS

The lower bound (31) can be improved if we note that it is possible to decrease denominators in (29) and (30). The denominators decrease if there are as many as possible indices i in the set I_{Pj} for δ_+ and in the set I_{Mj} for δ_-. This eliminates the terms that include c_is and thereby decreases denominators because c_is are almost always significantly lower than 1. To see this compare (22) with (10) which shows

$$g_{n+1i} = k\sigma_i, \quad i = 0, 1, \cdots, n+1, \tag{32}$$

where k is an arbitrary real number. Because of (11) the signs of g_{n+1i} alternate: $\text{sign}(g_{n+1i+1}) = -\text{sign}(g_{n+1i})$. Equation (20) becomes

$$\sum_{i=0}^{n+1} |\frac{g_{n+1i}}{W(\omega_i)}| = 1, \tag{33}$$

and from (28)

$$c_i = \frac{|\frac{\sigma_i}{W(\omega_i)}|}{\sum_{\substack{k=0 \\ k\neq i}}^{n+1} |\frac{\sigma_k}{W(\omega_k)}|} = \frac{|\frac{g_{n+1i}}{W(\omega_i)}|}{\sum_{\substack{k=0 \\ k\neq i}}^{n+1} |\frac{g_{n+1k}}{W(\omega_k)}|} = \frac{|\frac{g_{n+1i}}{W(\omega_i)}|}{1 - |\frac{g_{n+1i}}{W(\omega_i)}|}. \tag{34}$$

This, together with (33), gives

$$\sum_{i=0}^{n+1} \frac{1}{1+c_i} = n+1. \tag{35}$$

Obviously, c_is are small and the lower bound will improve if they are eliminated from (29) and (30). This can be done with the help of equation (24). By multiplying (24) with a suitable factor f and subtracting it from equations (23) we get new coefficients g^*_{ji} which replace $(-1)^i g_{ji}$

$$g^*_{ji} = \frac{(-1)^i g_{ji} - f(-1)^i g_{n+1i}}{c_i W(\omega_i)}, \tag{36}$$

$$a^*_j - a_j = \sum_{i=0}^{n+1} g^*_{ji} c_i W(\omega_i)(-1)^i (P^*(\omega_i) - P(\omega_i)). \tag{37}$$

Note that all terms $(-1)^i g_{n+1i}$ have the same sign (because of (32)). A factor f_{mj} that makes all $h^* g^*_{ji} \leq 0$ therefore always exists. The set I_{Pj} is now empty and using (19) and (20) the denominator of (29) becomes

$$\sum_{i=0}^{n+1} |\frac{g^*_{ji}}{W(\omega_i)}| = \sum_{i=0}^{n+1} |\frac{(-1)^i g_{ji} - f_{mj}(-1)^i g_{n+1i}}{W(\omega_i)}| = |f_{mj}|. \tag{38}$$

The factor $|f_{mj}|$ should be as small as possible. The smallest f_{mj} that makes all $h^* g^*_{ji} \leq 0$ is given by

$$f_{mj} = \max_{i \in I_{Pj}} \left(\frac{g_{ji}\,\text{sign}(h^*)}{g_{n+1i}} \right). \tag{39}$$

A similar factor f_{pj} that makes all $h^* g^*_{ji} \geq 0$ and causes set I_{Mj} to be empty exists

$$f_{pj} = \min_{i \in I_{Mj}} \left(\frac{g_{ji}\,\text{sign}(h^*)}{g_{n+1i}} \right). \tag{40}$$

Table 1. Experimental results of lower bound effectiveness for 12 cases.

Filter	h^*	Opt. $\|D-P\|_\infty$	L.b. $\|D-P\|_\infty$
A15/5	0.119397	0.155174	0.142381
A25/5	0.039716	0.101226	0.061517
A35/7	0.015946	0.029838	0.020663
B15/7	0.279315	0.306864	0.292619
B25/7	0.122889	0.154227	0.132170
B35/7	0.052719	0.117187	0.065954
C15/5	0.051462	0.166736	0.120264
C25/5	0.012831	0.126398	0.082126
C35/7	0.002629	0.037575	0.012102
D15/7	0.189748	0.248478	0.219802
D25/7	0.048086	0.130607	0.078998
E25/6	0.033284	0.087922	0.046268

Using f_{mj} and f_{pj} the improved lower bound is

$$\delta \geq \min \left(\max_{0 \leq j \leq n} \frac{|a^*_j - a_{j+}|}{|f_{mj}|}, \max_{0 \leq j \leq n} \frac{|a^*_j - a_{j-}|}{|f_{pj}|} \right). \tag{41}$$

Additional improvements are possible; note that (41) uses at most two of the n equations in (23) – the ones that give maximum δ_- and δ_+. The other $n-2$ equations play no role. This is identical to saying that for these equations the differences $a^*_j - a_j$ are equal to zero. Since this is not the case, an improved lower bound can be obtained by adding or subtracting equations (23).

Twelve filters with five different sets of frequency-domain specifications, denoted A through E, were used for testing. The frequency specifications are identical to those that were used in [5]. We denote by A15/5 the filter design problem for specification A, length 15 (8 independent coefficients), and $b = 5$ bits (sign included); similarly for A25/5, B15/7, and so on. Table 1 shows a summary of the results, comparing the infinite precision deviation h^*, the optimal b-bit deviation $\|D-P\|_\infty$, and the lower bound number for $\|D-P\|_\infty$ that is computed with the help of (41).

The lower bound (41) was also implemented in a program for optimal finite wordlength FIR filter design. The computing time was, depending on the filter specifications, between 3 and 4 times lower than in the otherwise identical program which does not use the lower bound.

REFERENCES

[1] D.M.Kodek, "Design of optimal finite word- length FIR digital filters using integer programming techniques," IEEE Trans. on Acoustics, Speech and Signal Processing, vol.ASSP-28, pp.304-308, June 1980.

[2] M.J.D.Powell, Approximation theory and methods. Cambridge University Press, Cambridge, 1981.

[3] D.M.Kodek, "A theoretical performance bound for finite wordlength FIR digital filters," Proc. of the 1996 CISS Conference, pp.487- 492, Princeton, March 20-22, 1996.

[4] W.P.Niedringhaus, K.Steiglitz, D.M.Kodek, "An easily computed performance bound for finite wordlength direct-form FIR digital filters," IEEE Trans. Circuits Syst., vol.CAS-29, pp.191-193, Mar. 1982.

[5] D.M.Kodek, K.Steiglitz, "Comparison of optimal and local search methods for designing finite wordlength FIR digital filters," IEEE Trans. on Circuits and Systems, vol.CAS-28, pp.28-32, January 1981.