Digital Signal Processing 23 (2013) 1798–1805

Contents lists available at SciVerse ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Length limit of optimal finite wordlength FIR filters

Dušan M. Kodek

University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Available online 11 July 2013

Keywords:
FIR digital filters
Finite wordlength effects
Quantization
Minimax approximation
Combinatorial optimization

ABSTRACT

In practical FIR digital filter applications it is often necessary to represent the filter coefficients with a finite number of bits. The optimal finite wordlength coefficients have an interesting property. For all finite wordlength filters there exists a maximum filter length N_{max} and the corresponding cosine polynomial degree n_{max}. Increasing the filter length N beyond N_{max} gives additional coefficients that are all zero. A theoretical explanation for the existence of N_{max} is given in the paper. The influence of the filter specifications on N_{max} is investigated. In addition, a simple method that gives a reasonably accurate estimate of N_{max} is also given. Knowing N_{max} and its relationship to the filter specifications is important in the finite wordlength FIR design because it can reduce the time needed to compute the optimal coefficients.

1. Introduction

There are many practical situations in which the coefficients of an FIR digital filter must be represented with a finite number of bits. This requires that the "infinite precision" coefficients are replaced by the finite wordlength ones. If we wish to use a fixed point DSP processor, which is almost always cheaper and/or faster than a floating point one, we would like to meet the filter specifications with coefficients that are represented with a small number of bits b.

An interesting phenomenon was observed when optimal filters with b-bit coefficients were computed [1]. It was found that beyond a certain length all coefficients are always zero. This means that it is not possible to meet arbitrarily severe FIR filter specifications with a small number of bits b by increasing the filter length N. This phenomenon was further studied in [2,3]. For all finite wordlength filters there exists a maximum filter length N_{max}. If the filter length N is increased beyond N_{max}, the additional coefficients are all zero. Obviously, these filters do not get any better by using $N > N_{max}$. The time needed to compute the optimal coefficients, however, increases considerably.

The fact that a length limit exists was established experimentally and is perhaps somewhat surprising. Such a limit does not exist if the finite wordlength restriction is removed. The Weierstrass theorem [4] assures us that the minimax approximation error goes towards zero when $N \to \infty$. This is not the case if the filter coefficients are constrained to b bits.

Note that the simple idea of finding N_{max} by computing a large N filter and then rounding the coefficients to their nearest b-bit representation does not work. The rounded b-bit coefficients will certainly be zero from some N_x on. The problem is that this N_x is practically always much higher than the true N_{max}. In typical cases it was between two or three times higher than N_{max}. Obviously, we need something better.

No quick and simple method that finds N_{max} is known, and this remains an interesting open problem which has received little attention in the literature. This is not surprising. Approximation of functions by polynomials with coefficients which are integers is famous for being very hard. Ferguson [5] gives a sampling of difficulties and some of the more accessible results. These results have some bearing on our problem which is a special case of a general integer polynomial approximation. It is quite different from the general case. The polynomial is a cosine polynomial, the interval is a union of disjoint subsets, and there is an additional weight function which can be different in every subset. It appears that there are no mathematical papers that deal with this particular approximation problem.

The problem of finding N_{max}, and the corresponding cosine polynomial degree n_{max}, is important both for theoretical and for practical reasons. A theoretical explanation is needed to understand why it exists at all. Knowing N_{max} and n_{max}, even if only approximately, is important for practical reasons since it can reduce the time needed to compute the optimal finite wordlength coefficients. The designer would also often like to know how the filter specifications like the width of bands and transition or don't care bands, the weight function, and the number of bits affect n_{max}. It is the purpose of this paper to give some answers to these questions.

The outline of the paper is the following: In Section 2 we recall the basic facts about the finite wordlength design problem. We then show that the number of nonzero filter coefficients is finite in Section 3. In Section 4 we consider a cosine polynomial

E-mail address: duke@fri.uni-lj.si.

1051-2004/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.dsp.2013.07.004

with real coefficients of degree n to which an integer coefficient of degree n_x, $n_x > n$, is added. We prove that a degree $n_{f\,max}$ exists for which adding an integer coefficient always increases the approximation error if $n \geqslant n_{f\,max}$. For standard filters $n_{f\,max}$ can be investigated more precisely. This is done in Section 5. Practical search for n_{max} requires a bound on the maximum number of consecutive zero coefficients. A simple formula that gives an estimate n_{czer} for this bound is derived in Section 6. In Section 7 we demonstrate that $n_{f\,max}$ can be used to predict n_{max} and then use n_{czer} to verify the prediction by computing the true n_{max}.

2. The finite wordlength design problem

Let us start with the infinite precision design problem. We limit our attention to linear phase filters with real-valued coefficients although it will be shown that N_{max} almost certainly exists for nonlinear phase filters too. Details are more complicated which is also true for filters with complex coefficients. The frequency response $H^*(\omega)$ of a length N optimal infinite precision (i.e., filter coefficients can be any real number) linear phase FIR digital filter is equal to

$$H^*(\omega) = \sum_{k=0}^{N-1} h^*(k) e^{-j\omega k}$$
$$= e^{j(L\frac{\pi}{2} - \frac{N-1}{2}\omega)} Q(\omega) \sum_{k=0}^{n} a_k^* \cos k\omega \tag{1}$$

where $L = 0$ or 1. Depending on N and filter symmetry there are exactly four types of FIR filters and four real functions $Q(\omega)$. The degree n of the cosine polynomial

$$P_n^*(\omega) = \sum_{k=0}^{n} a_k^* \cos k\omega \tag{2}$$

is related to the filter length N and there are formulas which relate the optimal coefficients $h^*(k)$ and a_k^*. Function $Q(\omega)$ is irrelevant from the point of view of the approximation problem and will be ignored. To find $P_n^*(\omega)$ one must solve the following minimax approximation problem

$$\min_{P_n(\omega)} \max_{\omega\in\Omega} \left| W(\omega)\big(D(\omega) - P_n(\omega)\big)\right|. \tag{3}$$

The real function $D(\omega)$ is the desired frequency response, the weighting function $W(\omega)$ is by definition real and positive, and the set Ω is a subset, or a union of subsets, of the interval $[0, \pi]$.

The standard approach to solving (3) is to use the Remez algorithm in a way that was first described by Parks and McClellan [6]. The problem becomes much more complex when the finite wordlength constraint is introduced. Although there seems to be no formal proof that it is NP-hard, this is almost certainly so.

For the purpose of this paper we will make the finite wordlength constraint equal to requesting that the filter coefficients $h(k)$ are b-bit integers from the set I_b, where $I_b = \{-2^{b-1}, \ldots, -1, 0, 1, \ldots, 2^{b-1}\}$. The integer set I_b is chosen for convenience only — any other finite set of b-bit numbers (sums of a limited number of power-of-two terms, for example) can be used instead. Constraining the coefficients $h(k)$ to the set I_b requires a redefinition or scaling of the original infinite precision approximation problem. This is necessary to bring the coefficients $h(k)$ within the range of numbers in I_b and can be done with the help of a scaling factor s. Let us assume that s is known and denote as $D_u(\omega)$, $W_u(\omega)$, and $P_{nu}(\omega)$ the original (unscaled) problem. The approximation problem that gives the optimal b-bit approximation error of degree n can be written as

$$E_{min}(n) = \min_{P_{nu}(\omega)} \max_{\omega\in\Omega} \left| \frac{W_u(\omega)}{s}\big(sD_u(\omega) - sP_{nu}(\omega)\big)\right|$$
$$= \min_{P_n(\omega)} \max_{\omega\in\Omega} \left| W(\omega)\big(D(\omega) - P_n(\omega)\big)\right| \tag{4}$$

where the scaled functions $D(\omega)$ and $W(\omega)$ are defined as

$$D(\omega) = sD_u(\omega), \qquad W(\omega) = W_u(\omega)/s. \tag{5}$$

The finite wordlength polynomial $P_n(\omega)$ equals

$$P_n(\omega) = sP_{nu}(\omega) = \sum_{k=0}^{n} a_k \cos k\omega, \quad a_k \in I_b,\ a_0 \in I_b/2, \tag{6}$$

and $I_b/2$ denotes the set I_b in which all elements are divided by 2. Observe that $a_k \in I_b$ and $a_0 \in I_b/2$ in (6) which means that the coefficient a_0 is a special case. This follows from the requirement that all $h(k)$ must be in I_b and for type 1 (odd N, positive symmetry) FIR filters we have

$$h(n) = 2a_0, \quad n = (N-1)/2$$
$$h(n-k) = h(n+k) = a_k, \quad k = 1, 2, \ldots, n \tag{7}$$

where $2a_0$ and a_k are from I_b. It follows from (7) that $H(\omega)$ will be multiplied by $2s$. The formulas for type 2, 3, and 4 FIR filters are somewhat more complicated but the conclusions are similar.

The scaling factor s can be interpreted as the filter gain. Its choice is not trivial and is described in [7–9]. Two approaches are typically used in practice:

1. The scaling factor s is included in the approximation problem (4) as a variable. This gives the optimal scaling factor and the lowest approximation error but makes solving (4) significantly more difficult.
2. A constant scaling factor s determined by some ad hoc method is used.

Since we are only interested in determining the maximum filter length N_{max} and the corresponding polynomial degree n_{max}, we will assume that s is a known constant. Notation $P_n(\omega)$ will from here on denote a polynomial with b-bit coefficients $a_k \in I_b$, $k = 0, 1, \ldots, n$, whereas $D(\omega)$ and $W(\omega)$ are the scaled input functions. We begin by proving that the number of nonzero b-bit coefficients a_k is always finite.

3. The number of nonzero coefficients a_k is finite

For any real sequence $x(k)$ of length N the discrete version of the Parseval theorem [10] states

$$\sum_{k=0}^{N-1} x^2(k) = \frac{1}{N}\sum_{m=0}^{N-1} \left|X(m)\right|^2 \tag{8}$$

where

$$X(m) = \sum_{k=0}^{N-1} x(k) e^{-\frac{2\pi jkm}{N}}. \tag{9}$$

$X(m)$ is simply the frequency response of $x(k)$ computed at $\omega = 2\pi n/N$. If we use $h(k)$ instead of $x(k)$, (8) becomes

$$\sum_{k=0}^{N-1} h^2(k) = \frac{1}{N}\sum_{m=0}^{N-1} \left|P_n\left(\frac{2\pi m}{N}\right)\right|^2 \tag{10}$$

where $P_n(\omega)$ is defined by (6). It follows from (4) that $|P_n(\omega)|$ is bounded by

$$\left|P_n(\omega)\right| \leqslant \left|D(\omega)\right| + \frac{E_{min}(n)}{W(\omega)}, \quad \omega \in \Omega \tag{11}$$

where $E_{min}(n)/W(\omega)$ is by definition positive. Since Ω is a subset, or a union of subsets, of the interval $[0,\pi]$, some of the frequencies $\omega = 2\pi m/N$ may not belong to Ω.

The missing parts of the interval $[0,\pi]$ are called transition or don't care bands and the FIR filter specifications typically do not define $D(\omega)$ and $W(\omega)$ in these bands. The reason for this is that it is easier, and almost always better, to let the approximation algorithm find the $P_n(\omega)$ in the don't care bands. In this paper we assume that the filter specifications are such that there exists a positive number C that upper bounds $P_n(\omega)$ as

$$\left|P_n(\omega)\right| \leqslant C, \quad \omega \in [0,\pi]. \tag{12}$$

For filters that do not have peaks in don't care bands C is given by (11). For other filters C is larger but finite.

In what follows $D(\omega)$ and $W(\omega)$ are known continuous functions defined over the complete interval $[0,\pi]$. Obviously, their values in the don't care bands can always be computed exactly after the filter coefficients are known. The practicality of such $D(\omega)$ and $W(\omega)$ definitions is not important for our purpose. Inserting (12) into (10) gives

$$\sum_{k=0}^{N-1} h^2(k) \leqslant C^2, \quad h(k) \in I_b. \tag{13}$$

This also holds for nonlinear phase filters which is an indication that N_{max} exists for them too. Details, however, are different because (7) holds for linear phase only and gives

$$(2a_0)^2 + \sum_{k=1}^{n} a_k^2 \leqslant C^2, \quad a_k \in I_b,\ a_0 \in I_b/2 \tag{14}$$

where $n = (N-1)/2$. Assume now that $n \to \infty$. Since C does not increase with increasing n and since the coefficients a_k are integers from I_b, Eqs. (13) and (14) prove that there can be only a finite number of nonzero b-bit coefficients a_k and $h(k)$.

No such limit exists if a_k are not constrained to integer values. Eqs. (13) and (14) still hold, but the sum of a_k^2 does not become infinite when $n \to \infty$ because a_k converge to zero.

4. The number of consecutive zero coefficients a_k is small

Knowing that the number of finite wordlength nonzero coefficients a_k is finite is important. However, this only tells that a finite n_{max} exists. It could, in principle, be a very large number. In particular, it is easy to see that (14) does not exclude a situation where there exists a large number of consecutive zero a_k before the last nonzero one. To prove that the best integer polynomial $P_n(\omega)$ cannot have many zero coefficients before the last nonzero one, we need to look deeper into the properties of polynomial minimax approximation.

Function $D(\omega)$ is by definition even and can always be written as

$$D(\omega) = \sum_{k=0}^{\infty} c_k \cos k\omega \tag{15}$$

where c_k are Fourier coefficients

$$c_0 = \frac{1}{\pi}\int_0^{\pi} D(\omega)\,d\omega, \qquad c_k = \frac{2}{\pi}\int_0^{\pi} D(\omega)\cos k\omega\,d\omega, \quad k \geqslant 1. \tag{16}$$

If the function $D(\omega)$ satisfies the Dini–Lipschitz condition on $[0,\pi]$, which is practically always true in filter design cases, then the sum (15) converges uniformly and c_k converge to zero [11].

For the derivations that follow in this and the next section we will temporarily remove the integer constraint from $P_n(\omega)$. The cosine polynomial $P_n(\omega)$ of degree n that approximates (4) is not constrained to integer coefficients. The following lower bound on $E_{min}(n)$ can be found in [12]

$$E_{min}(n) \geqslant \frac{\pi}{4} \max_{k \geqslant n+1}\left(\frac{|c_k|}{B_k}\right). \tag{17}$$

For $W(\omega) = 1$ the constants B_k equal 1 for all k. Otherwise they can be computed with a formula[1]

$$B_k = \int_0^{\pi} \frac{|\cos k\omega|}{2W(\omega)}\,d\omega, \quad k \geqslant n+1. \tag{18}$$

For functions $W(\omega)$ that are typical in filter design the values of B_k are easy to compute. They are almost independent of k and it follows from (5) that they are a function of the scaling factor s which is in turn proportional to 2^{b-1}.

Since we are only interested in proving that there cannot be many consecutive zero coefficients a_k, the exact values of B_k are not very important. Let $k = n_x$ be the index of the nonzero integer coefficient and let a_k, $k = n+1, n+2, \ldots, n_x - 1$, be equal to zero. It is easy to see that this is equivalent to approximation of the function

$$F(\omega) = D(\omega) + a_{n_x}\cos n_x\omega, \quad a_{n_x} \in I_b,\ n_x \geqslant n+1 \tag{19}$$

where a_{n_x} is the nonzero integer coefficient. The corresponding approximation problem with $n_x - n - 1$ consecutive zeros is equivalent to the following problem of degree n, $n < n_x$,

$$E_{F\,min}(n) = \min_{P_n(\omega), a_{n_x}} \max_{\omega \in [0,\pi]} \left|W(\omega)\bigl(F(\omega) - P_n(\omega)\bigr)\right| \tag{20}$$

where $D(\omega)$ and $W(\omega)$ are scaled functions defined in (5).

Coefficient a_{n_x} in (20) can in principle increase or decrease approximation error $E_{F\,min}(n)$ relative to $E_{min}(n)$. We will prove that for every filter specification there exists a degree $n_{f\,max}$ such that $E_{F\,min}(n) > E_{min}(n)$ for all $n \geqslant n_{f\,max}$.

The function $F(\omega)$ is even and can be written as

$$F(\omega) = \sum_{k=0}^{\infty} d_k \cos k\omega \tag{21}$$

where d_k are Fourier coefficients. Fourier transform is linear and it follows from (16) and (19) that $d_k = c_k$ for all k with the exception of $k = n_x$ for which

$$d_{n_x} = c_{n_x} + a_{n_x}. \tag{22}$$

Using (17) the lower bound on $E_{F\,min}(n)$ equals

$$E_{F\,min}(n) \geqslant \frac{\pi}{4}\left(\frac{|d_{n_x}|}{B_{n_x}}\right) \geqslant \frac{\pi}{4}\left(\frac{|a_{n_x}| - |c_{n_x}|}{B_{n_x}}\right) \tag{23}$$

where an integer a_{n_x} that gives the lowest bound must be used. Since a_{n_x} is a nonzero integer and since the coefficients c_k converge to zero, $E_{F\,min}(n)$ converges to

$$L = \frac{\pi}{4}\left(\frac{1}{B_\infty}\right) \tag{24}$$

with increasing n_x.

[1] The bound in [12] does not use $W(\omega)$. Including an arbitrary positive weight function $W(\omega)$ is simple and follows from the derivation of the bound.

It is now easy to show that the optimal polynomial cannot have any nonzero integer coefficients after a_n if n is high enough. This follows from the fact that $E_{min}(n)$ decreases with n whereas $E_{F\,min}(n)$ stops decreasing when n_x is high enough. This in turn means that (23) ensures that there exists a degree $n = n_{f\,max}$ for which

$$E_{F\,min}(n) > E_{min}(n), \quad n \geqslant n_{f\,max} \tag{25}$$

for all $n_x \geqslant n_{f\,max} + 1$. In other words, adding an integer coefficient a_{n_x} always increases the approximation error if $n \geqslant n_{f\,max}$. The optimal polynomial $P_n(\omega), n \geqslant n_{f\,max}$, therefore cannot have one or more nonzero integer coefficients after $a_{n_{f\,max}}$. Even though the coefficients of $P_n(\omega)$ are real, this conclusion obviously also holds for integer $P_n(\omega)$ because $E_{F\,min}(n)$ can only be higher if the coefficients are constrained to integers. This also means that the limit L for the integer $P_n(\omega)$ is higher than the one given by (24). Another difference is that this limit and $n_{f\,max}$ are much harder to find for integer $P_n(\omega)$.

Finding $n_{f\,max}$ for non-integer $P_n(\omega)$ is a simple matter. However, the first n satisfying $E_{F\,min}(n) > E_{min}(n)$ does not automatically equal $n_{f\,max}$. The rate of convergence of Fourier coefficients c_k depends on the function $D(\omega)$, and the sequence of coefficients c_k is not necessarily monotone. Correspondingly, the convergence of $E_{F\,min}(n)$ to L also does not have to be monotone. For example, functions $D(\omega)$ which are common in the filter design cases often have $c_k = 0$ for even k. The condition $E_{F\,min}(n) > E_{min}(n)$ must therefore be examined for several values n before $n_{f\,max}$ is found. Since $E_{min}(n)$ typically decreases faster than $E_{F\,min}(n)$ the $n_{f\,max}$ is usually found when $E_{F\,min}(n)$ is still quite far away from its limit L.

5. Improved bound for piecewise constant $D(\omega)$ and $W(\omega)$

The bounds (23) and (24) hold for arbitrary functions $D(\omega)$ and $W(\omega)$. The only restriction is that the Fourier coefficients of $D(\omega)$ converge to zero which is practically never a problem. For standard frequency selective filters (lowpass, highpass, bandpass, bandstop) the functions $D(\omega)$ and $W(\omega)$ are piecewise constant. Such functions allow a more accurate estimate of limit L and also allow an insight into the influence of filter specifications on $n_{f\,max}$.

Let us define $G_n(\omega)$ as the cosine polynomial of degree n which is the best weighted minimax approximation of $F(\omega)$ on the $n_x + 1$ points $\omega_0, \omega_1, \ldots, \omega_{n_x}$ defined by

$$\omega_\ell = \frac{\ell\pi}{n_x}, \quad \ell = 0, 1, \ldots, n_x. \tag{26}$$

This particular set of points plays a special role in our derivation. We use $n_x = n + 1$ and by the alternation theorem $G_n(\omega)$ satisfies

$$\delta \frac{(-1)^i}{W(\omega_i)} = F(\omega_i) - G_n(\omega_i), \quad i = 0, 1, \ldots, n_x. \tag{27}$$

Since $G_n(\omega)$ is the best approximation on the ω_ℓ defined by (26) and not on $\omega \in [0, \pi]$, it is not equal to $P_n(\omega)$ from (20). But the set (26) is a subset of $[0, \pi]$ which means that the approximation error $E_{F\,min}(n)$ cannot be lower than δ. We have

$$E_{F\,min}(n) \geqslant |\delta| = \left|W(\omega_i)\big(F(\omega_i) - G_n(\omega_i)\big)\right|. \tag{28}$$

Because $E_{F\,min}(n)$ increases with decreasing n, (28) holds for all $n < n_x$. To get a bound for δ we follow an approach used in [12] and define the operator $\sum''$ where the primes indicate that the first and last terms in the sum are to be halved. We then multiply (27) by $\cos n_x\omega_i$ and apply the operator $\sum''$ to both sides

$$\delta \sum_{i=0}^{n_x}{}'' \frac{(-1)^i}{W(\omega_i)} \cos n_x\omega_i = \sum_{i=0}^{n_x}{}'' F(\omega_i) \cos n_x\omega_i - \sum_{i=0}^{n_x}{}'' G_n(\omega_i) \cos n_x\omega_i. \tag{29}$$

Now $G_n(\omega)$ is a cosine polynomial of degree n, $n < n_x$, and the term containing it can be written as

$$\sum_{i=0}^{n_x}{}'' G_n(\omega_i) \cos n_x\omega_i = \sum_{i=0}^{n_x}{}'' \sum_{k=0}^{n} a'_k \cos k\omega_i \cos n_x\omega_i \tag{30}$$

where a'_k are the coefficients of $G_n(\omega)$. The sum

$$\sum_{i=0}^{n_x}{}'' \cos k\omega_i \cos n_x\omega_i \tag{31}$$

equals zero for all $k < n_x$ [13]. Hence in (29) the term with $G_n(\omega_i)$ vanishes. For other terms observe that on the set (26) $\cos n_x\omega_i = (-1)^i$. Using the definition (19) for $F(\omega)$ makes (29) equal to

$$\delta \sum_{i=0}^{n_x}{}'' \frac{1}{W(\omega_i)} = \sum_{i=0}^{n_x}{}'' \big((-1)^i D(\omega_i) + a_{n_x}\big) \tag{32}$$

and

$$\delta = \frac{1}{\sum''^{n_x}_{i=0} \frac{1}{W(\omega_i)}} \left(n_x a_{n_x} + \sum_{i=0}^{n_x}{}'' (-1)^i D(\omega_i) \right). \tag{33}$$

Assume without loss of generality that $W_u(\omega) \geqslant 1$ which from (5) gives $W(\omega) \geqslant 1/s$. The sum containing $1/W(\omega_i)$ equals sn_x if $W_u(\omega) = 1$. For any other $W_u(\omega)$ the sum is always lower than sn_x. Since $D(\omega)$ is piecewise constant, the sum containing $D(\omega_i)$ is almost independent of n_x. This means that its contribution converges to zero with increasing n_x and δ converges to

$$\delta = \frac{n_x a_{n_x}}{\sum''^{n_x}_{i=0} \frac{1}{W(\omega_i)}}. \tag{34}$$

We see from (28) that $E_{F\,min}(n) \geqslant |\delta|$ and because a_{n_x} is a nonzero integer $E_{F\,min}(n)$ converges to

$$L = \frac{n_x}{\sum''^{n_x}_{i=0} \frac{1}{W(\omega_i)}} \tag{35}$$

with increasing n_x. There is always $L \geqslant 1/s$ and L becomes almost independent of n_x with increasing n_x. This bound was derived without the help of Fourier coefficients and also gives higher values than (24). As in (24) the convergence to L is typically not monotone because the sum containing $D(\omega_i)$ in (33) can change sign. The constant B_∞ from (24) is similar, but not identical, to the sum in (35). Still, the conclusions that follow are identical to those that were described in connection with (23). $E_{F\,min}(n)$ stops decreasing when n_x is high enough which confirms again that $n_{f\,max}$ exists. Observe that the limit L depends on the scaling factor s which is included in $W(\omega)$. The scaling factor is proportional to 2^{b-1} which means that L decreases with the number of bits b. As expected, degree $n_{f\,max}$ will increase with b.

6. Computing an estimate for the maximum number of consecutive zero coefficients a_k

We will demonstrate in Section 7 that $n_{f\,max}$ can be used to predict n_{max}. To see how good this prediction is requires that we know the true n_{max}. There is a problem here. The only known method that finds the true n_{max} is to solve the finite wordlength polynomial problem (4) for $n = 1, 2, \ldots$, and so on. From some n on the coefficients a_k, $k > n$, will start to be all zero. How do we know when we can stop the search and declare that $n = n_{max}$? It is clear that we need a bound on the maximum number of consecutive coefficients a_k before the last nonzero one.

The way to this bound follows from the observation derived in (24) and (35) that $E_{F\,min}(n)$ cannot decrease if a nonzero coefficient of degree n_x is added after a string of zero coefficients. The piecewise constant functions $D(\omega)$ and $W(\omega)$ simplify the search for bound. Consider the following approximation problem

$$\min_{a_k} \max_{[0,\pi]} \left| K + a_{n_x} \cos n_x\omega - \sum_{k=0}^{n} a_k \cos k\omega \right| \tag{36}$$

where $n_x > n$ and K is an arbitrary constant. The solution of this problem $a_0 = K$ and $a_k = 0$ for $k = 1, \ldots, n$ is well known [14]. It gives the error function $e(\omega) = a_{n_x} \cos n_x\omega$ which has $n_x + 1$ extremal points with alternating sign

$$e(\omega_i) = -e(\omega_{i+1}), \quad i = 0, 1, \ldots, n_x - 1 \tag{37}$$

where $\omega_i = i\pi/n_x$ and $|e(\omega_i)| = 1$. The alternation theorem states that the necessary and sufficient condition for the best polynomial approximation of degree n is that the error function $e(\omega)$ exhibit at least $n+2$ extremal points ω_i with the alternating property (37). Because $n_x > n$, this solution clearly fulfills this condition.

The problem (36) demonstrates the inability of lower degree cosines to approximate a higher degree cosine. It is indeed possible to state that this is the underlying reason for the existence of both n_{max} and $n_{f\,max}$. Our approximation problem (20) is of course more complicated, but there are similarities. If each of the piecewise constant sections of $D(\omega)$ is observed separately, they lead to approximation problems that are similar to (36). Let us examine the band with the maximum $W(\omega)$ and denote its weight as

$$W_{max} = \max_{\omega\in[0,\pi]} W(\omega). \tag{38}$$

The number of extremal points that the optimal solution has in this band will in general depend on its width and also on $D(\omega)$, $W(\omega)$, n, and n_x. If n_x is high enough at least one of the extremal points will be in this band. We again use the frequency points (26) and denote the set of frequencies ω_ℓ that belong to this band as Ω_{wx}. Because there are extremal points in Ω_{wx}, (28) can be rewritten as

$$E_{F\,min}(n) \geqslant |\delta| = W_{max} \max_{\omega_\ell\in\Omega_{wx}} \left| F(\omega_\ell) - G_n(\omega_\ell) \right|. \tag{39}$$

Using the definition (21) the term containing $F(\omega_\ell) - G_n(\omega_\ell)$ equals

$$\max_{\omega_\ell\in\Omega_{wx}} \left| D(\omega_\ell) + a_{n_x} \cos n_x\omega_\ell - \sum_{k=0}^{n} g_k \cos k\omega_\ell \right| \tag{40}$$

where g_k are the coefficients of polynomial $G_n(\omega)$. $D(\omega_\ell)$ is constant in Ω_{wx}, and we have a case that is similar to the one in (36). There are two differences, however. The first one is that only a subset Ω_{wx} of the interval $[0, \pi]$ is used. The second is that $G_n(\omega)$ must approximate $D(\omega)$ in other bands too. This means that only a part of the polynomial degree n can be used to minimize (40).

We wish to examine the effect of the maximum weight W_{max} on the polynomial $G_n(\omega)$ that gives the optimal $E_{F\,min}(n)$. Let $G_n^*(\omega)$ be a cosine polynomial with $g_0^* = D(\omega_\ell)$ and $g_k^* = 0$, $k = 1, \ldots, n$. We denote the number of frequencies ω_ℓ in Ω_{wx} as $n_{\Omega wx}$. For this $G_n^*(\omega)$ the error function is

$$\begin{aligned} e^*(\omega_\ell) &= W_{max}\bigl(D(\omega_\ell) + a_{n_x} \cos n_x\omega_\ell - G_n^*(\omega_\ell)\bigr) \\ &= W_{max} a_{n_x} (-1)^\ell, \quad \ell = \ell_{x1}, \ldots, \ell_{x2} \end{aligned} \tag{41}$$

where ℓ_{x1} and ℓ_{x2} are the indices of the first and last ω_ℓ in Ω_{wx}. Suppose that a polynomial which is better than $G_n^*(\omega)$ exists. Without loss of generality we let $G_n^*(\omega) + R_n(\omega)$ be the best polynomial. Hence the reduction of approximation error is obtained and there must be

$$\left| e^*(\omega_\ell) - R_n(\omega_\ell) \right| < |W_{max} a_{n_x}|, \quad \ell = \ell_{x1}, \ldots, \ell_{x2}. \tag{42}$$

It follows from (42) that the sign of $e^*(\omega_\ell)$ is the same as the sign of $R_n(\omega_\ell)$ for all $\omega_\ell \in \Omega_{wx}$. This means that $G_n^*(\omega)$ is the best approximation if there is no polynomial $R_n(\omega)$ that satisfies the condition

$$e^*(\omega_\ell) R_n(\omega_\ell) > 0, \quad \ell = \ell_{x1}, \ldots, \ell_{x2}. \tag{43}$$

Observe that $R_n(\omega)$ is a cosine polynomial of degree n and therefore cannot have more than n changes of sign on $[0, \pi]$. But $e^*(\omega)$ changes sign $n_{\Omega wx} - 1$ times in Ω_{wx} and this number is increasing with n_x. It follows that $R_n(\omega)$ satisfying (43) never exists if $n_{\Omega wx} - 1 > n$ and this in turn means that $G_n(\omega) = G_n^*(\omega)$ is the best polynomial. It now follows from (39) that for a given n there always exists n_x giving

$$E_{F\,min}(n) \geqslant W_{max}. \tag{44}$$

A pathological case occurs for filters that have a very large W_{max} in stopbands ($D(\omega) = 0$). For such filters the optimal polynomial is simply $G_n(\omega) = 0$ which gives error function $e(\omega) = W(\omega)D(\omega)$. The approximation error $E_{F\,min}(n) = \max_{\omega\in[0,\pi]}(W(\omega)D(\omega))$ is lower than W_{max} but filters with zero coefficients are of course not practical.

For practical filters we can use (44) to get a bound on the maximum number of consecutive zero coefficients a_k. For a given n_x the number of consecutive zero coefficients a_k equals $n_x - n - 1$ and (44) states that the approximation error $E_{F\,min}(n)$ cannot go below W_{max} with increasing n_x. Or in other words, the approximation error $E_{F\,min}(n)$ cannot get better once $n_x - n - 1$ is big enough.

To get an estimate when exactly this occurs note first that the optimal polynomial $G_n(\omega)$ must approximate $D(\omega)$ in all bands. The condition $n_{\Omega wx} - 1 > n$ is therefore pessimistic because the number of sign changes that $e(\omega)$ can have on $\omega \in \Omega_{wx}$ is typically lower than n. Let us denote this number as $z_{\Omega wx}$. We wish to find an easily computed estimate for n_x that satisfies the condition

$$n_{\Omega wx} - 1 > z_{\Omega wx} \tag{45}$$

which is sufficient for (44) to hold. Getting an estimate for $z_{\Omega wx}$ is not difficult. The Cauchy remainder theorem [4] tells us that the zeros and the extremes of $e(\omega)$ of the optimal cosine polynomial are equidistant on the interval $[0, \pi]$. In our case the don't care bands are excluded from $[0, \pi]$ which affects the equidistant property near the band edges. But the position of zeros is still almost equidistant in the rest of the interval.

We use this property to get an estimate for $z_{\Omega wx}$. Let $\Delta\omega_{tr}$ denote the total width of all don't care bands and let $\Delta\omega_{\Omega wx}$ denote the width of the band with the weight W_{max}. An approximate formula gives

$$z_{\Omega wx} \approx n \frac{\Delta\omega_{\Omega wx}}{\pi - \Delta\omega_{tr}}. \tag{46}$$

The frequencies ω_ℓ are equidistant by definition and their number in Ω_{wx} is approximately

$$n_{\Omega wx} \approx n_x \frac{\Delta\omega_{\Omega wx}}{\pi}. \tag{47}$$

Combining (46) and (47) with (45) gives an estimate for n_x that we are looking for

$$n_x \approx n + \left\lceil n\frac{\Delta\omega_{tr}}{\pi - \Delta\omega_{tr}} + \frac{2\pi}{\Delta\omega_{\Omega wx}} \right\rceil \tag{48}$$

where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Let us denote the maximum number of consecutive zero coefficients a_k as n_{czer}. It follows from (48) that it is equal to

$$n_{czer} = n_x - n - 1 \approx \left\lceil n\frac{\Delta\omega_{tr}}{\pi - \Delta\omega_{tr}} + \frac{2\pi}{\Delta\omega_{\Omega wx}} \right\rceil - 1. \tag{49}$$

This simple formula was tested on a large number of filter design cases. For filters with $\Delta\omega_{\Omega wx} > 0.05\pi$ and $sW_{max} < 100$ it was found that it gives an excellent prediction for n_{czer}. The n_{czer} from (49) gave approximation error $E_{F\,min}(n)$ that was never more than 5% below the bound (44). The formula becomes inaccurate for filters with a very narrow band $\Delta\omega_{\Omega wx}$ that also have a very high weight W_{max}. In such cases one can use the exact $z_{\Omega wx}$ by counting the number of ω_ℓ in Ω_{wx} and a much better estimate of $n_{\Omega wx}$ by using the number of extremal points in Ω_{wx} that is obtained from the Remez algorithm when (4) is solved.

Filters with the same weight in all bands are a special case because $\Delta\omega_{\Omega wx} = \pi - \Delta\omega_{tr}$ and (49) becomes

$$n_{czer} = n_x - n - 1 \approx \left\lceil \frac{n\Delta\omega_{tr} + 2\pi}{\pi - \Delta\omega_{tr}} \right\rceil - 1 \tag{50}$$

giving lower n_{czer} than (49). Eqs. (49) and (50) describe the influence of filter specifications on n_{czer}. It is inversely proportional to the width $\Delta\omega_{\Omega wx}$ of the band with the highest weight and is almost independent of n if the don't care bands are narrow. In addition, n_{czer} is almost independent of W_{max}.

Although we derived n_{czer} for real polynomial $P_n(\omega)$, it is easy to see that it is also valid for the optimal integer $P_n(\omega)$. This is true because the optimal integer $P_n(\omega)$ gives $E_{F\,min}(n)$ that cannot be lower than the one for the unconstrained $P_n(\omega)$. The maximum number of consecutive zeros n_{czer} for integer $P_n(\omega)$ therefore cannot be higher than the one given by (49) or (50). It will in fact almost always be lower which means that it is safe to use it to stop the search for true n_{max}.

Knowing n_{czer} also gives an indirect insight into the maximum integer polynomial degree n_{max}. Higher n_{czer} gives an indication that n_{max} will also be higher. We can expect with a reasonable certainty that the filter specifications affect the n_{max} in a way that is similar to the one described above. This can be useful because there is no formula that directly relates n_{max} to filter specifications. These conclusions are in excellent agreement with the results that were observed in the practical optimal finite wordlength design cases.

7. Results and conclusion

Let us now return to our starting problem which is how to find n_{max} and N_{max} for optimal filters with coefficients from I_b. This problem is related to (4) and is almost certainly NP-hard. It is computationally very demanding and we wish to find an easily computed estimate for n_{max}. In order to get this estimate we derived a bound $n_{f\,max}$ using polynomials $P_n(\omega)$ with real coefficients in which only the highest order coefficient a_{n_x} is from I_b. Finding $n_{f\,max}$ is quite easy and it is worth investigating if it can be used to compute a reasonably good estimate of n_{max}.

Both $n_{f\,max}$ and n_{max} are computed from the same filter specifications. To see if they are related let us make the following two observations:

1. It follows from (24) and (35) that for real $P_n(\omega)$ the approximation error $E_{F\,min}(n)$ converges to limit L with increasing n. Because of its higher approximation power the $E_{F\,min}(n)$ of real $P_n(\omega)$ is lower than $E_{min}(n)$ of the integer $P_n(\omega)$. Since it also converges to L faster, we might expect that this also means that $n_{f\,max}$ is lower than n_{max}. This, however, is not necessarily so.
2. The reason for this is that the limit L is not the same for real and integer $P_n(\omega)$. Even if no formula that gives L for integer $P_n(\omega)$ is known, it is easy to see that it is higher than for real $P_n(\omega)$. Convergence to a higher limit obviously also means lower n_{max} for the integer $P_n(\omega)$. The effects of higher approximation power of real $P_n(\omega)$ and higher L of integer $P_n(\omega)$ influence $n_{f\,max}$ and n_{max} in the same direction.

Table 1
The six sets of filter specifications. The frequency edges are divided by 2π.

Filter	Band 1	Band 2	Band 3
A			
edges	0–0.2	0.25–0.5	
$D(\omega)$	1	0	
$W(\omega)$	1	1	
B			
edges	0–0.2	0.25–0.5	
$D(\omega)$	1	0	
$W(\omega)$	1	10	
C			
edges	0–0.14	0.18–0.32	0.36–0.5
$D(\omega)$	1	0	1
$W(\omega)$	1	1	1
D			
edges	0–0.14	0.18–0.32	0.36–0.5
$D(\omega)$	1	0	1
$W(\omega)$	1	10	1
E			
edges	0.01–0.21	0.26–0.49	
$D(\omega)$	1	0	
$W(\omega)$	1	1	
F			
edges	0.01–0.21	0.26–0.49	
$D(\omega)$	1	0	
$W(\omega)$	1	10	

Although we have no formal proof that these effects lead to similar $n_{f\,max}$ and n_{max}, it is likely that they do. The idea of predicting n_{max} simply as

$$n_{max} = n_{f\,max} \tag{51}$$

seems promising and is worth investigating experimentally.

Thirty six filters with six different sets of frequency-domain specifications, denoted A through F, with wordlength $b = 7$ to 12 were used for testing. The frequency specifications are similar to those that were used in [15,16]. They were chosen because they are completely unrelated to the n_{max} problem. The frequency specifications are given in Table 1. A is a low-pass filter with unit weighting in both bands. B is the same, except that the stopband has a weighting of 10. C is a bandstop filter with unit weighting in all bands, while D has a weighting of 10 in stopband. E is a low-pass filter whose passband and stopbands do not include $\omega = 0$ or π. Again, F is the same, except that the stopband has a weighting of 10.

We denote by A/7 the filter design problem for specification A and $b = 7$ bits; similarly for A/8, A/9, A/10, A/11, A/12, B/7 and so on. All filters are type 1 (odd length, positive symmetry), which means that the maximum filter length is equal to $N_{max} = 2n_{max} - 1$. For each of the filters $n_{f\,max}$ for real $P_n(\omega)$ was computed as described in Section 4 giving a predicted $n_{max} = n_{f\,max}$.

Table 2
Predicted and true maximum polynomial degree n_{max} for filters from Table 1. The filter length N_{max} equals $2n_{max} - 1$.

Filter	Predicted n_{max}	True n_{max}	Predicted appr. error	True appr. error
A/7	15	15	0.0517669	0.0517669
A/8	20	20	0.0296271	0.0296271
A/9	24	22	0.0152829	0.0152829
A/10	29	30	0.0085716	0.0083668
A/11	33	31	0.0039812	0.0039812
A/12	38	42	0.0021621	0.0020973
B/7	15	17	0.2119286	0.1906503
B/8	19	19	0.1121105	0.1121105
B/9	24	22	0.0568042	0.0568042
B/10	28	26	0.0291078	0.0291078
B/11	30	30	0.0152946	0.0152946
B/12	37	35	0.0070114	0.0070114
C/7	21	21	0.0424776	0.0424776
C/8	27	27	0.0222873	0.0222873
C/9	31	33	0.0130958	0.0117153
C/10	37	39	0.0064574	0.0052595
C/11	43	43	0.0030568	0.0030568
C/12	49	49	0.0014760	0.0014760
D/7	23	23	0.0994322	0.0994322
D/8	29	27	0.0610920	0.0610920
D/9	35	35	0.0248223	0.0248223
D/10	39	43	0.0153054	0.0143894
D/11	45	45	0.0069477	0.0069477
D/12	51	53	0.0037637	0.0037085
E/7	17	15	0.0512550	0.0512550
E/8	21	19	0.0288770	0.0288770
E/9	24	26	0.0156386	0.0152211
E/10	30	26	0.0079545	0.0079545
E/11	34	34	0.0039878	0.0039878
E/12	38	40	0.0021529	0.0019435
F/7	16	16	0.1973700	0.1973700
F/8	18	19	0.1104889	0.1029697
F/9	25	25	0.0523973	0.0523973
F/10	29	31	0.0286058	0.0272422
F/11	33	31	0.0138546	0.0138546
F/12	38	36	0.0072031	0.0072031

The program for the optimal finite wordlength filter design [16] was used to find the true n_{max} for integer $P_n(\omega)$ for each of the filters. This is a computationally slow procedure which requires trying different filter lengths n until a solution with n_{czer} consecutive zeros is found. For example, to find true n_{max} for filter D/12 requires a solution of length $n = 59$ ($N = 117$) which took 2.6 h (on a 3.4 GHz Intel Core i7 3770). Increasing the number of bits b to 13 or more makes the computation of true n_{max} very difficult. Computing the predicted n_{max}, however, is easy for any b.

Table 2 shows a summary of the results, comparing the predicted n_{max} and true n_{max} together with the corresponding approximation errors. A constant scaling factor $s = 2^{b-2}$ was used in all design cases. This scaling was chosen for simplicity and also to allow easier comparison with the older results. As described by (7) it gives $H(\omega)$ that is multiplied by 2^{b-1}. An obvious alternative would be to use the optimal scaling factor. Unfortunately, this requires much longer computation times and makes the computations that are needed in Table 2 impractical. Although the scaling factor s affects the resulting approximation error, it does that in approximately the same direction for computation of both predicted and true n_{max}. This means that the relationship between the predicted and true n_{max} stays more or less the same for all reasonable s.

Results show that the predicted n_{max} is close to the true n_{max}. This is quite remarkable when we consider how much easier it is to compute the predicted n_{max}. In 14 of the 36 cases the predicted n_{max} equals true n_{max}. The largest difference is 4 (A/12, D/10, E/10) which in the case of E/10 gives a 15% mismatch. The difference is smaller for all other filters with an average mismatch of less than 5%. Filter E/10 is the worst and is also quite unusual in another respect. The true n_{max} for both E/9 and E/10 is 26 which is contrary to the expected increase of n_{max} when the number of bits b increases from 9 to 10. E/10 is the only such case, the other 35 cases behave as expected. Results also show that n_{max} is almost independent of weight which is in agreement with the predictions that follow from (49) or (50).

It is interesting to observe the filter performance when the predicted n_{max} is smaller than the true n_{max}. For filter E/12 the approximation error of predicted $n_{max} = 38$ equals 0.0021529 whereas the approximation error of true $n_{max} = 40$ is 0.0019435. The difference is 10.8% which corresponds to a stopband difference of 1.3 dB. Similarly, for C/10 (predicted $n_{max} = 31$, true $n_{max} = 33$) the numbers are 0.0064574 and 0.0052595 or 22.7% which corresponds to a stopband difference of 1.5 dB. These are the largest differences and are much smaller for all other filters. The performance of predicted n_{max} filter is close to performance of true n_{max} filter even in the worst examples.

The easily computed predicted maximum degree n_{max} is useful in the practical finite wordlength FIR design. Let us assume that we would like to find the b-bit filter with the lowest approximation error for a given set of specifications. Without the predicted n_{max} we must compute a large number of optimal finite wordlength solutions which is a very time consuming procedure. Using predicted n_{max} gives a much faster answer which is almost always equal or close to the correct answer. We conclude that this result is useful in the practical finite wordlength FIR filter design.

References

[1] D.M. Kodek, Design of optimal finite word-length FIR digital filters using integer programming techniques, IEEE Trans. Acoust. Speech Signal Process. 28 (June 1980) 304–308.

[2] D.M. Kodek, K. Steiglitz, Filter-length word-length tradeoffs in FIR digital filter design, IEEE Trans. Acoust. Speech Signal Process. 28 (Dec. 1980) 739–744.

[3] D.M. Kodek, M. Krisper, Telescopic rounding for suboptimal finite wordlength FIR digital filter design, Digital Signal Process. 15 (Nov. 2005) 522–535.

[4] P.J. Davis, Interpolation and Approximation, Dover, New York, 1975, pp. 107–118, pp. 56–64.

[5] L.B.O. Ferguson, What can be approximated by polynomials with integer coefficients, Amer. Math. Monthly 113 (May 2006) 403–414.

[6] T.W. Parks, J.H. McClellan, Chebyshev-approximation for nonrecursive digital filters with linear phase, IEEE Trans. Circuit Theory 19 (March 1972) 189–194.

[7] D.M. Kodek, Design of optimal finite wordlength FIR digital filters, in: Proc. European Conf. on Circuit Theory and Design ECCTD'99, vol. I, Stresa, Italy, Aug. 29–31, 1999, pp. 401–404.

[8] Y.C. Lim, Design of discrete-coefficient-value linear phase FIR filters with optimum normalized peak ripple magnitude, IEEE Trans. Circuits Syst. 37 (Dec. 1990) 1480–1486.

[9] T. Ciloglu, New initialization methods for discrete coefficient FIR filter design with coefficient scaling and the use of scale factor in the design process, IEEE Trans. Signal Process. 54 (Feb. 2006) 796–800.

[10] A.V. Oppenheim, R.W. Schafer, Discrete-Time Signal Processing, 3rd ed., Pearson, Upper Saddle River, 2010, p. 733.

[11] M.J.D. Powell, Approximation Theory and Methods, Cambridge University Press, Cambridge, 1981, p. 155, pp. 192–193.

[12] E.V. Cheney, Introduction to Approximation Theory, 3rd ed., AMS Chelsea Publishing, Providence, 1982, p. 75, pp. 126–133.

[13] J.C. Mason, D.C. Handscomb, Chebyshev Polynomials, Chapman&Hall/CRC, Boca Raton, 2003, p. 87.

[14] A.F. Timan, Theory of Approximation of Functions of a Real Variable, Dover, New York, 1994, p. 90.

[15] D.M. Kodek, Performance limit of finite word-length FIR digital filters, IEEE Trans. Signal Process. 53 (July 2005) 2462–2469.

[16] D.M. Kodek, LLL algorithm and the optimal finite wordlength FIR design, IEEE Trans. Signal Process. 60 (March 2012) 1493–1498.

Dušan M. Kodek received the B.E.E. degree, the M.E.E. degree, and the Ph.D. degree from the University of Ljubljana, Ljubljana, Slovenia. From 1978 to 1979, he was a Fulbright Visiting Professor with the Department of Electrical Engineering and Computer Science, Princeton University. From 1981 to 1982 he was a manager for laboratory systems in the Physical Acoustics Corporation, Princeton, NJ.

Since 1982 he has been with the Faculty of Electrical Engineering, University of Ljubljana. In 1995 he was a founding dean of the new Faculty of Computer and Information Science, and served as the Dean of the faculty from 1996 to 2001. He is now a Professor and a head of Theoretical Computer Science teaching and conducting research in the computer architecture and digital signal processing areas.

Professor Kodek is the author of three books on computer architecture and is a recipient of three Slovenian awards for research and innovation.