Constrained Least Square Design of FIR Filters Without Specified Transition Bands*

Ivan W. Selesnick, Markus Lang and C. Sidney Burrus
Department of Electrical and Computer Engineering - MS 366
Rice University, Houston, TX 77251-1892, USA
selesi@ece.rice.edu, lang@ece.rice.edu, csb@ece.rice.edu

September 27, 1995

Abstract

This paper puts forth the notion that explicitly specified transition bands have been introduced in the filter design literature in part as an indirect approach for dealing with discontinuities in the desired frequency response. We suggest that the use of explicitly specified transition bands is sometimes inappropriate because, to satisfy a meaningful optimality criterion, their use implicitly assumes a possibly unrealistic assumption on the class of input signals.

This paper also presents an algorithm for the design of peak constrained lowpass FIR filters according to an integral square error criterion that does not require the use of specified transition bands. This rapidly converging, robust, simple multiple exchange algorithm uses Lagrange multipliers and the Kuhn-Tucker conditions on each iteration. The algorithm will design linear and minimum phase FIR filters and gives the best L_2 filter and a continuum of Chebyshev filters as special cases.

It is distinct from many other filter design methods because it does not exclude from the integral square error a region around the cut-off frequency, and yet, it overcomes Gibbs' phenomenon without resorting to windowing or 'smoothing out' the discontinuity of the ideal lowpass filter.

EDICS: SP 2.1.1 FIR Systems

Contact Author:
Ivan W. Selesnick
Phone: (713) 527-8101 x3508
FAX: (713) 524-5237

`Submitted to: IEEE Transactions on Signal Processing - June 1994`
`Revised - April 1995`
`Revised - Sept 1995`

*This research has been supported by DARPA under an NDSEG fellowship, by the Alexander von Humboldt Foundation, by NSF grant MIP-9316588, and by BNR.

List of Figures

1 Introduction

We consider the definition of optimality for digital filter design and suggest that a constrained least squared error criterion with no specified transition band is a useful complement to existing approximation criteria for filter design.

Consider, for example, a basic lowpass filtering scenario in which a signal of interest whose spectrum occupies the frequency range $(0, \omega_o)$ is embedded in an additive noise signal whose spectrum occupies the entire frequency range $(0, \pi)$. In this case, without further assumptions, no transition band naturally arises from the problem of removing the noise from the signal of interest: No part of the passband is more or less critical than any other part of the passband. Similarly, no part of the stopband is more or less critical than any other part of the stopband. In many practical cases there is *no* separation of the passband and stopband by a transition (or 'don't care') band between them. Indeed, spectra of the desired and undesired signals often overlap.

An important exception to this is the design of filters used to select one out of two or more signals which have been designed to occupy well separated frequency bands. In this case, and other cases in which the signals to be filtered have no energy in a transition band, the use of transition bands in filter design is well motivated - the transition band constitutes a non-critical part of the frequency response. But even in these cases, there is in the "guard bands" usually some noise or other undesirable signals that one wants to remove. It is for this reason that transition region anomalies in the frequency response are undesirable. Indeed, when large peaks occur in the 'don't care' transition band of certain multiband Chebyshev filter designs [24], engineers decide they *do* care and alter the specifications or employ modified algorithms [32, 33] to eliminate the peaks. In many cases, a transition band is introduced to reduce or remove the oscillations in the frequency response near the band edges caused by the Gibbs' phenomenon, not because transition bands naturally arises from the physics of the problem.

For the meaningful design of filters it is necessary to choose an error criterion that does not implicitly require unrealistic assumptions on the signals, such as the existence of a band separating desired signals and noise. Although these statements are informal, below we draw upon the results of Weisburn, Parks, and Shenoy [38] to give a mathematical justification for the inclusion of the transition region in the measure of approximation error.

This paper (an earlier version of which is [28]) is organized as follows: In section 2 we establish notation and discuss ways in which Gibbs' phenomenon has previously been treated. We also discuss the constrained L_2 approach of Adams [1, 2] to filter design and the results of Weisburn, Parks, and Shenoy [38]. In section 3 we take into account the results of Weisburn et al. to modify the approach of Adams. The resulting design method adopts the view of Adams and is, at the same time, motivated by the results of Weisburn et al. The rapidly converging, robust, multiple exchange algorithm algorithm presented in section 4 produces peak constrained least square lowpass FIR filters. The result is a versatile design algorithm that will design linear and minimum phase FIR filters and which gives the best L_2 filter and a continuum of Chebyshev filters as special cases. Section 5 gives some interpretations and extensions of this method. A simple Matlab program that illustrates the algorithm for odd length lowpass filter design is given in the appendix.

2 Preliminaries

The frequency response $H(\omega)$ of an FIR filter is given by the discrete-time Fourier transform of its impulse response $h(n)$:

$$H(\omega) = \sum_{n=0}^{N-1} h(n) e^{-j\omega n}. \tag{1}$$

If $h(n) = h(N-1-n)$, then $H(\omega)$ has linear phase and can be written as

$$H(\omega) = A(\omega) e^{-jM\omega} \tag{2}$$

where $A(\omega)$ is the real-valued amplitude and $M = (N-1)/2$ for length-N filters [22]. For simplicity, symmetric odd length filters will be discussed in this paper, in which case $A(\omega)$ can be written as

$$A(\omega) = \frac{1}{\sqrt{2}} a(0) + \sum_{n=1}^{M} a(n) \cos n\omega \tag{3}$$

where the impulse response coefficients $h(n)$ are related to the the cosine coefficients $a(n)$ by

$$h(n) = \begin{cases} \frac{1}{2} a(M-n) & \text{for } 0 \le n \le M-1 \\ \frac{1}{\sqrt{2}} a(0) & \text{for } n = M \\ \frac{1}{2} a(n-M) & \text{for } M+1 \le n \le N-1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Let $D(\omega)$ denote the desired amplitude. For example, see figure 2 in which the desired amplitude of an ideal lowpass filter is shown. Approximating this discontinuous function by the cosine polynomial $A(\omega)$ given in equation (3) is the most basic filter design problem. Many of the various methods in the filter design literature can be distinguished by the ways in which they treat this discontinuity. Indeed, controlling the behavior of $A(\omega)$ in the region around the discontinuity has strongly influenced the development of filter design methods.

Two primary measures of approximation error are used in filter design. Let $E(\omega) = D(\omega) - A(\omega)$. (i) The weighted integral square error (or 'L_2 error') is given by

$$||E(\omega)||_2 = \left(\frac{1}{\pi} \int_0^{\pi} W(\omega)(A(\omega) - D(\omega))^2 \, d\omega \right)^{\frac{1}{2}}. \tag{5}$$

(ii) The weighted Chebyshev error is given by

$$||E(\omega)||_\infty = \max_{\omega \in [0,\pi]} |W(\omega)(A(\omega) - D(\omega))|. \tag{6}$$

In both cases, $W(\omega)$ is a nonnegative error weighting function. When $W(\omega)$ is set to unity over $[0, \pi]$, the approximation measures above are called the unweighted (or uniformly weighted) integral square error and the unweighted (or uniformly weighted) Chebyshev error. The simplest method to design optimal FIR filters minimizes $||A(\omega) - D(\omega)||_2$ and the resulting filter we call the best L_2 filter. As is well known, if the error weighting function is set to unity over $[0, \pi]$, then the best L_2 filter is obtained by truncating the Fourier series of $D(\omega)$ (the rectangular window method). Hence, for simple $D(\omega)$, a closed form expression for the filter is easily found. But $W(\omega) = 1$ is not generally used in practice because best L_2 filters with this error weighting possess large peak errors near the band edges. Moreover, the peak value of these 'overshoots' does not diminish with increasing filter length. To overcome this behavior, known as Gibbs' phenomenon, three main approaches have been employed: (i) the use of non-rectangular windows, (ii) the use of transition functions to continuously connect adjacent bands, and (iii) the use of zero-weighted transition bands placed between adjacent bands.

The use of these approaches has spawned a variety of filter design procedures having the following two desirable properties:

1. The procedure produces filters that do not suffer from Gibbs' phenomenon.

2. The procedure can be implemented using a computationally efficient numerical algorithm.

For example, variable order spline transition functions can be used with the integral square error approximation measure to obtain expressions for filters having good response behavior around the discontinuity [8]. Alternatively, when a zero-weighted transition band is used, the best weighted L_2 filter can be found by solving a system of linear equations [37]. In this case Gibbs' phenomenon is also eliminated: the peak error diminishes as the filter length increases. The use of zero-weighted transition bands also permits the meaningful use of the Chebyshev norm, an error measure for which the Parks-McClellan (PM) program produces optimal linear phase filters [22]. In fact, in order to design lowpass filters by minimizing the Chebyshev norm, either a transition function or a zero-weighted transition band *must* be specified: For the discontinuous lowpass response given in figure 2, the filter minimizing the unweighted Chebyshev error has a Chebyshev error of one half and is not unique.

Unfortunately, each of the three enumerated methods has its shortcoming:

1. **Windows:** Although the multiplication of the Fourier coefficients of $D(\omega)$ with a non-rectangular window is very simple, the method is generally considered sub-optimal because it is difficult to use it to minimize meaningful error measures (but see [12]) and because error weighting in the sense of (5) is not achieved.

2. **Transition Functions:** Although modifying the desired amplitude $D(\omega)$ (so that it is no longer discontinuous) yields approximations that do not suffer from Gibbs' phenomenon, the method does not directly approximate the original discontinuous desired lowpass amplitude function.

3. **Zero-Weighted Transition Bands:** By a zero-weighted transition band, we mean a region placed between two adjacent bands where $W(\omega)$ is taken to be 0. Because $A(\omega) - D(\omega)$ is not weighted there, zero-weighted transition bands are sometimes called 'don't care' regions. The use of zero-weighted transition bands make the approximation problem easier. But if they are used, then unless the input signals have no energy in the transition band, the optimality of the best Chebyshev and L_2 filters in the operator norm sense [38] is problematic.

Therefore, although the use of these approaches makes easier the approximation of the discontinuous desired amplitude, they are all rather indirect methods for dealing with the discontinuity.

2.1 Adams' Error Criterion

An early comparison of error criteria was made by Tufts and Francis [36]. More recently Adams [1, 2] described perhaps the most meaningful error criterion for filter design to date and suggested an iterative algorithm to design the corresponding best linear phase filters.

As Adams has noted, L_2 filter design is based on the assumption that the size of the peak errors can be ignored. Likewise, filter design according to the Chebyshev norm assumes the L_2 measure of approximation error is irrelevant. In practice, however, both of these criteria are important, a point Adams elaborates in [1]. Furthermore, Adams finds that the peak error of a best L_2 filter can be significantly reduced with only a slight increase in the L_2 error. Similarly, the L_2 error of an equiripple filter can be reduced with only a slight increase in the Chebyshev error. In Adams' terminology, both equiripple filters and best L_2 filters are *inefficient*.

Consider lowpass filter design. To obtain filters having a better trade-off between these two criteria, Adams uses zero-weighted transition bands and proposes that $||A(\omega) - D(\omega)||_2$ be minimized subject to a constraint on the Chebyshev error. This is formulated as a quadratic program [1] as follows:

$$\min_{\mathbf{a}\in\mathbf{R}^{M+1}} ||A(\omega) - D(\omega)||_2^2 \tag{7}$$

such that

$$L(\omega) \leq A(\omega) \leq U(\omega) \quad \text{for all } \omega \in [0, \omega_p] \cup [\omega_s, \pi]. \tag{8}$$

In (7), Adams weights the integral square error by the weight function:

$$W(\omega) = \begin{cases} W_p & \text{for all } \omega \in [0, \omega_p] \\ 0 & \text{for all } \omega \in [\omega_p, \omega_s] \\ W_s & \text{for all } \omega \in [\omega_s, \pi]. \end{cases} \tag{9}$$

In (8), the upper and lower bound functions $L(\omega)$ and $U(\omega)$ are given by

$$L(\omega) = \begin{cases} 1 - \delta_p & \text{for all } \omega \in [0, \omega_p] \\ -\delta_s & \text{for all } \omega \in [\omega_s, \pi] \end{cases} \tag{10}$$

and by

$$U(\omega) = \begin{cases} 1 + \delta_p & \text{for all } \omega \in [0, \omega_p] \\ \delta_s & \text{for all } \omega \in [\omega_s, \pi] \end{cases} \tag{11}$$

where δ_p and δ_s are the maximum allowed deviations from 1 and 0 in the passband and stopband.

This constrained L_2 approach allows the user to control the trade-off between the L_2 and Chebyshev errors, and produces best L_2 and Chebyshev filters as special cases. Of course, for a fixed filter length and a fixed δ_p and δ_s (each less than 0.5), it is not possible to obtain an arbitrarily narrow transition band. Therefore, if the band edges ω_p and ω_s are taken to be too close together, then the quadratic program (7,8) has no solution. Similarly, for a fixed ω_p and ω_s, if δ_p and δ_s are taken too small, then there is again no solution. In the terminology of quadratic programming [10], the feasible region is empty.

Although the algorithm in [1] tests for optimality upon termination by checking the non-negativity of Lagrange multipliers, during the iterations it does not enforce this non-negativity. For this reason, it may converge to a non-optimal filter. In [17, 18] an algorithm for the design of nonlinear phase FIR filters according to the same error measure is proposed. It inspects the signs of Lagrange multipliers on each iteration so that, if the algorithm converges, then the filter to which it converges is guaranteed to be optimal. However, the algorithm in [18] is also not guaranteed to converge. But it was found that for lowpass filter design, whenever there exists a filter satisfying the constraints specified by the user, the algorithm converges in practice. To develop exchange algorithms for solving (7,8) that are guaranteed to converge to optimal filters, it is necessary to modify the algorithms of [1, 17, 18]. In [2], Adams et al. describe in detail appropriate modifications based on [11] that guarantee convergence. In [35], Sullivan and Adams extend the algorithm of [2] to the design of nonlinear phase FIR filters with constraints on the group delay.

2.2 Approximation by Operator Norms

Weisburn, Parks, and Shenoy [38] present a rigorous motivation for the use of the Chebyshev and L_2 error measures and discuss the use of zero-weighted transition bands (see also [30, 31]). Using the theory of operator norms, they show that best Chebyshev filters minimize a worst case error signal energy, while best L_2 filters minimize a worst case pointwise error in the time domain. However, to show this optimality in the operator norm sense when a zero-weighted transition band is used, a hypothetical ideal prefilter is placed at the input, the frequency response of which is zero on the don't care region and 1 everywhere else [38].

Weisburn et al. begin by defining $E(\omega) = A(\omega) - D(\omega)$ as the frequency response of an error filter, E. They view the error filter as an operator, and use an operator norm as a measure of approximation. The filter design problem then becomes one of finding A to minimize $||A - D||$ where the norm is an operator norm. To define the norm of an operator, it is necessary to define a norm on the class of input signals, which we will denote $||\cdot||_i$, and a norm on the class of output signals, $||\cdot||_o$. Note that the norms used for the input and output signals do not have to be the same. The operator norm of E is then defined as

$$||E|| = \sup_{x \in U} \frac{||Ex||_o}{||x||_i} \tag{12}$$

where x represents the input signal of the error filter E and $y = Ex$ represents the corresponding output signal. The supremum is over the space of input signals U. This ratio indicates the amount by which the filter E 'magnifies' or 'attenuates' the input signal with respect to the chosen input and output norms.

It turns out that when the class of input signals is taken to be the space of all finite energy sequences and when the input and output signal norms are both taken to be the l_2 norm, ($||x||_i = \sqrt{\sum_n x(n)^2}$ and $||\cdot||_i = ||\cdot||_o$), then the operator norm is equal to the unweighted Chebyshev norm, $||E(\omega)||_\infty$, defined in equation (6). Therefore, the best Chebyshev filter minimizes a worst case output *signal energy* over a set of bounded energy input signals [38].

On the other hand, if the input signal space and norm is kept the same, but the output signal norm is taken to be $||y||_o = |y(n)|$ for some fixed index n, then the operator norm is equal to the unweighted integral square error, $||E(\omega)||_2$, defined in equation (5). Note that it is independent of the index n. Consequently, the best L_2 filter minimizes a worst case *pointwise* error in the time domain over a set of bounded energy input signals [38].

Suppose $D(\omega)$ is the usual discontinuous ideal lowpass frequency response. Further, suppose an error weighting function is used which is zero in a specified transition band. If the signals in the input class have no energy in the transition band, then the filters obtained by minimizing the weighted Chebyshev and L_2 norms are optimal in the operator norm sense [38]. However, if the input signals do have energy in the specified transition bands, then the way in which the filters are optimal is unclear in general.

Suppose on the other hand, that a transition function is used to modify $D(\omega)$ so that it has a smooth transition between the passband and stopband. Then the best filters obtained according to any approximation measure do not correspond to the original desired discontinuous frequency response $D(\omega)$.

3 A New Criterion for Lowpass Filter Design

In light of the preceding discussion, we suggest that the use of explicitly specified transition bands in FIR filter design began in part with the desire to reduce peak errors near the band edges and that these 'don't care' regions are, in some cases, used because they facilitate the approximation problem, not because they naturally arise from the filter problem under consideration.

Furthermore, because the minimization of the Chebyshev norm for lowpass filter design requires the use of a transition band, we propose for some problems that the L_2 norm be the primary optimization measure and that the peak errors be controlled by a constraint.[1] In the next section we present an algorithm for the design of peak constrained lowpass FIR filters according to an integral square error criterion that does not require the use of specified transition bands.

Let us define $E(\omega) = A(\omega) - D(\omega)$, where $D(\omega)$ is the ideal discontinuous lowpass amplitude. It will be convenient to make the meaning of the term 'peak errors' more precise. By peak errors of a lowpass filter, we mean the values of $|E(\omega)|$ at the *local minima and maxima* of $A(\omega)$. With this definition, the peak errors do not include values of $E(\omega)$ at the edges of a transition band (see figure 4).

Let ω_o be the cut-off frequency (the frequency at which the ideal lowpass amplitude is discontinuous). The design problem we propose uses the L_2 weight function and lower and upper bound functions given by:

1. $W(\omega) = 1$ for all $\omega \in [0, \pi]$
2. $U(\omega) = 1 + \delta_p$, $L(\omega) = 1 - \delta_p$, for all $\omega \in [0, \omega_o]$
3. $U(\omega) = \delta_p$, $L(\omega) = -\delta_s$, for all $\omega \in (\omega_o, \pi]$

[1]There are cases where the Chebyshev error should be truly minimized (e.g. narrow band interference at an unknown frequency) but simply reducing it or constraining it is generally sufficient.

We propose the following problem formulation:

new criterion Minimize the unweighted integral square error (5) subject to the constraint that the local minima and maxima of $A(\omega)$ lie within the specified lower and upper bound functions, $L(\omega)$ and $U(\omega)$.

The filters produced by the algorithm described below produces lowpass linear phase filters according to this criterion. They have frequency response amplitudes that are very similar in appearance to those obtained using the approach described by Adams. However, there are three main differences:

1. There is no region around the discontinuity that is excluded from the integral square error approximation measure. The algorithm minimizes the square error over the entire frequency range from 0 to π subject to the peak constraints.

2. The transition region is *implicitly* defined by the constrained L_2 minimization procedure. It is not dictated by the specification of transition band edges. Indeed, band edges are not specified.

3. There is no minimum achievable peak error size. That is, filters having arbitrarily small peak errors can be obtained. The problem of infeasibility encountered in the quadratic program formulation does not arise.

Because the approach taken here weights the L_2 error over the entire interval $[0, \pi]$, there are no natural band edges to use in (8) of the quadratic program formulation. For this reason, the approach to filter design described here can *not* be described by a quadratic program. It should be noted that the new criterion also applies to multiband filter design, but for clarity, only the lowpass case is examined in this paper. The extension to arbitrary responses is not developed here but it is expected that some extensions to nonpiece-wise constant responses are possible.

The use of this criterion for lowpass filter design is suitable for situations where the maximum peak error size is to be controlled and where there is no reason to assume that the input signals to be filtered have no energy in a transition band. If the peak errors are unimportant, then the sinc function obtained by minimizing an unweighted unconstrained L_2 is more suitable. When it is known that input signals do not have energy in a transition band (or very little), then the use of a specified transition band is well motivated. In this case, the transition band constitutes a non-critical part of the response. Examples of this are when well separated signals are to be filtered.

For such applications, the PM algorithm, the algorithm of Adams', or the linear programming algorithms of [33] are better suited.

The approach we propose in this paper and the algorithm we describe are intended to complement the existing methods by providing an approach with few assumptions for a basic lowpass filtering problem.

4 A New Algorithm for Lowpass Filter Design

The algorithm described below produces lowpass linear phase FIR filters according to the new criterion described in the previous section. It is a rapidly converging, robust, simple multiple exchange algorithm that uses Lagrange multipliers and the Kuhn-Tucker conditions on each iteration [10, 34]. Although we have not proven its convergence, the algorithm converges in practice when used for lowpass filter design. A Matlab program that implements this algorithm is especially simple and is given in the appendix. For multiband filter design, the algorithm must be modified to obtain robust convergence [29].

The algorithm will design linear and minimum phase lowpass FIR filters and gives the best L_2 filter and a continuum of Chebyshev filters as special cases. The algorithm can be modified to allow different L_2 error weighting in different bands and to allow other types of constraints. We have designed lowpass filters of lengths over 3,000 and have used loose and tight constraints that differed in the passband and stopband by factors as much as 10^6.

4.1 The Equality Constrained Minimization Problem

The amplitude $A(\omega)$ of the filter minimizing the L_2 error subject to peak constraints will touch the lower and upper bound functions at certain extremal frequencies of $A(\omega)$. (By extremal frequencies of $A(\omega)$ we mean local minima and maxima of $A(\omega)$). If these frequencies were known in advance, then the filter could be found by minimizing $||E||_2^2$ subject to equality constraints at these frequencies. The procedure below determines the appropriate set of frequencies by solving a sequence of equality constrained quadratic minimization problems. The solution to each minimization problem is found by solving a linear system of equations.

This iterative algorithm is based on those of [1, 18]. The constraints are on the values of $A(\omega_i)$ for the frequency points ω_i in a constraint set. On each iteration, the constraint set is updated so that at convergence, the only frequency points at which equality constraints are imposed are those where $A(\omega)$ touches the constraint. The equality constrained problem is solved with Lagrange multipliers. The algorithm below associates an inequality constrained problem with each equality constrained one. According to the Kuhn-Tucker conditions, the solution to the *equality* constrained problem solves the corresponding *inequality* constrained problem if all the Lagrange multipliers are non-negative (where the signs of the multipliers are defined appropriately). If on some iteration a multiplier is negative, then the solution to the equality constrained problem does not solve the corresponding inequality constrained one. For this reason, before the constraint set is updated in the algorithm described below, constraints corresponding to negative multipliers (when they appear) are sequentially dropped from the constraint set. In this way, an inequality constrained problem is solved on each iteration, albeit over a possibly smaller constraint set. It turns out that in the special case of lowpass filter design considered here, this simple iterative technique converges in practice.

Let the constraint set S be a set of frequencies $S = \{\omega_1, \dots, \omega_r\}$ with $\omega_i \in [0, \pi]$. Let S be partitioned into two sets, S_l and S_u, where S_l is the set of frequencies where we wish to impose the equality constraint

$$A(\omega) = L(\omega), \tag{13}$$

while S_u is the set of frequencies where we wish to impose the equality constraint

$$A(\omega) = U(\omega). \tag{14}$$

Let us have $S_l = \{\omega_1, \dots, \omega_q\}$ and $S_u = \{\omega_{q+1}, \dots, \omega_r\}$. To minimize $||E(\omega)||_2$ subject to these constraints we form the Lagrangian [10, 20, 34]:

$$\mathcal{L} = ||E(\omega)||_2^2 - \sum_{i=1}^{q} \mu_i [A(\omega_i) - L(\omega_i)] + \sum_{i=q+1}^{r} \mu_i [A(\omega_i) - U(\omega_i)]. \tag{15}$$

Necessary conditions for $||E(\omega)||_2$ to be minimized subject to the constraints above, are obtained by setting the derivative of $\mathcal{L}$ with respect to a_k and μ_i to zero. This yields the following equations:

$$\frac{\partial ||E(\omega)||_2^2}{\partial a_k} - \sum_{i=1}^{q} \mu_i \frac{\partial A(\omega_i)}{\partial a_k} + \sum_{i=q+1}^{r} \mu_i \frac{\partial A(\omega_i)}{\partial a_k} = 0 \tag{16}$$

for $0 \le k \le M$,

$$A(\omega_i) = L(\omega_i) \tag{17}$$

for $1 \le i \le q$, and

$$A(\omega_i) = U(\omega_i) \tag{18}$$

for $q + 1 \le i \le r$. According to the Kuhn-Tucker conditions, when the Lagrange multipliers $\mu_1, \dots, \mu_r$ are all nonnegative, then the solution to equations (16,17,18) minimizes $||E(\omega)||_2$ subject to the inequality constraints

$$A(\omega_i) \ge L(\omega_i) \tag{19}$$

for $1 \le i \le q$, and

$$A(\omega_i) \le U(\omega_i) \tag{20}$$

for $q+1 \le i \le r$.

Recalling equation (3), and letting $W(\omega) = 1$, equations (16,17,18) can be written as

$$\mathbf{a} + \mathbf{G}^t \boldsymbol{\mu} = \mathbf{c} \tag{21}$$

and

$$\mathbf{G}\mathbf{a} = \mathbf{d} \tag{22}$$

where $\mathbf{a}$ is the length $M+1$ vector of unknown filter parameters $\mathbf{a} = (a_0, \dots, a_M)^t$, and $\boldsymbol{\mu}$ is the vector of Lagrange multipliers, one for each frequency in the constraint set: $\boldsymbol{\mu} = (\mu_1, \dots, \mu_r)^t$. $\mathbf{G}$ is the r by $M+1$ matrix of cosine terms that calculates the amplitude response $A(\omega)$ of the filter $\mathbf{a}$ at the frequencies in the constraint set S. The elements of $\mathbf{G}$ are given by

$$G_{i,0} = \frac{-1}{\sqrt{2}} \tag{23}$$

$$G_{i,k} = -\cos k\omega_i \tag{24}$$

for $1 \le i \le q$, $1 \le k \le M$ and

$$G_{i,0} = \frac{1}{\sqrt{2}} \tag{25}$$

$$G_{i,k} = \cos k\omega_i \tag{26}$$

for $q+1 \le i \le r$, $1 \le k \le M$. $\mathbf{c}$ is the vector of coefficients for the unconstrained optimal L_2 filter given by

$$c_0 = \frac{\sqrt{2}}{\pi} \int_0^\pi D(\omega)\, d\omega \tag{27}$$

$$c_k = \frac{2}{\pi} \int_0^\pi D(\omega) \cos k\omega \, d\omega. \tag{28}$$

Notice that the elements of $\mathbf{c}$ do not depend upon the constraints imposed on $A(\omega)$, in fact, they are the Fourier coefficients. Indeed, if there are no constraints imposed upon $A(\omega)$, then $a(n)$ equals $c(n)$, the best L_2 filter. The term $\mathbf{d}$ is the vector of values the amplitude response $A(\omega)$ is made to interpolate and is given by

$$d_i = -L(\omega_i) \tag{29}$$

for $1 \le i \le q$, and

$$d_i = U(\omega_i). \tag{30}$$

for $q+1 \leq i \leq r$. It is necessary to introduce minus signs for the lower bound constraints so that at the solution to the inequality constrained minimization problem, all of the Lagrange multipliers will have the same sign. Equations (21,22) can be combined into one matrix equation:

$$\begin{bmatrix} \mathbf{I}_{M+1} & \mathbf{G}^t \\ \mathbf{G} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}. \tag{31}$$

In equation (31), $\mathbf{I}_{M+1}$ is the $(M+1)$ by $(M+1)$ identity matrix. Similar expressions can be derived for the even length filter and the odd symmetric filters [22].

It is easy to verify that

$$\begin{aligned} \boldsymbol{\mu} &= (\mathbf{G}\mathbf{G}^t)^{-1}(\mathbf{G}\mathbf{c} - \mathbf{d}) && (32) \\ \mathbf{a} &= \mathbf{c} - \mathbf{G}^t\boldsymbol{\mu} && (33) \end{aligned}$$

is the solution to equation (31). Therefore, if the number of constraints (r) is small compared to the number of filter coefficients $(M+1)$, then the system (31) is computationally very inexpensive to solve. It requires the solution to an $r \times r$ system of linear equations. This is attributed to the use of $W(\omega) = 1$ over $[0, \pi]$. It is interesting to note that on each iteration the cosine coefficients $\mathbf{a}$ are obtained by *adding* to the best L_2 (Fourier) coefficients $\mathbf{c}$, a correction term. This is in contrast to the window method, in which the best L_2 coefficients are *multiplied* by a window.

For non-uniform weighting functions, the identity matrix in (31) becomes a full symmetric matrix [7, 8] and equations (16,17,18) become

$$\begin{bmatrix} \mathbf{R} & \mathbf{G}^t \\ \mathbf{G} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \tag{34}$$

where the elements of the vector $\mathbf{c}$ are given by

$$\begin{aligned} c_0 &= \frac{\sqrt{2}}{\pi}\int_0^\pi W(\omega)D(\omega)\,d\omega && (35) \\ c_k &= \frac{2}{\pi}\int_0^\pi W(\omega)D(\omega)\cos k\omega\,d\omega && (36) \end{aligned}$$

and the elements of the matrix $\mathbf{R}$ are given by

$$\begin{aligned} R_{0,0} &= \frac{1}{\pi}\int_0^\pi W(\omega)\,d\omega && (37) \\ R_{0,k} = R_{k,0} &= \frac{\sqrt{2}}{\pi}\int_0^\pi W(\omega)\cos k\omega\,d\omega && (38) \\ R_{k,l} = R_{l,k} &= \frac{2}{\pi}\int_0^\pi W(\omega)\cos k\omega\cos l\omega\,d\omega && (39) \end{aligned}$$

and where $\mathbf{G}$ and $\mathbf{d}$ are the same as above. It is easy to verify that

$$\begin{aligned} \boldsymbol{\mu} &= (\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^t)^{-1}(\mathbf{G}\mathbf{R}^{-1}\mathbf{c} - \mathbf{d}) && (40) \\ \mathbf{a} &= \mathbf{R}^{-1}(\mathbf{c} - \mathbf{G}^t\boldsymbol{\mu}) && (41) \end{aligned}$$

is the solution to equation (34).

4.2 The Exchange Iterations

The equality constrained optimization procedure described above is performed at each step of an iterative algorithm. At each iteration, the constraint set frequencies are updated, in much the same way as are the reference set frequencies of the Remez algorithm.

The algorithm begins with an empty constraint set so that the first filter designed is the best unconstrained L_2 filter. Then constraints are iteratively imposed upon $A(\omega)$ at selected frequencies until the best constrained L_2 filter is obtained. The constraint set is updated (i) by locating the local maxima of $A(\omega)$ that exceed the upper constraint function $U(\omega)$ and (ii) by locating the local minima of $A(\omega)$ that fall below the lower constraint function $L(\omega)$. Note that the 'induced' band edges of the passband and stopband are not extremal frequencies of $A(\omega)$. Unlike the program of Adams and the Parks-McClellan program, the band edges are not included in the constraint set (in the case of the PM program, the reference set). The mechanism that yields a sharp transition between the passband and stopband is the inclusion of the transition region in the integral square error.

The algorithm can be summarized in the following steps.

1. **Initialization:** Initialize the constraint set to the empty set: $S = \emptyset$.

2. **Minimization with Equality Constraints:** Calculate the Lagrange multipliers associated with the filter that minimizes $||E(\omega)||_2$ subject to the equality constraints $A(\omega_i) = L(\omega_i)$ for $\omega_i \in S_l$, and $A(\omega_i) = U(\omega_i)$ for $\omega_i \in S_u$. (Solve equation (32)).

3. **Kuhn-Tucker Conditions:** If there is a constraint set frequency ω_i for which the Lagrange multiplier μ_i is negative, then remove from the constraint set the frequency corresponding to the most negative multiplier and go back to step 2. Otherwise, calculate the new cosine coefficients using equation (33) and go on to step 4.

4. **Multiple Exchange of Constraint Set:** Set the constraint set S equal to $S_l \cup S_u$ where S_l is the set of frequency points ω_i in $[0, \pi]$ satisfying both $A'(\omega_i) = 0$ and $A(\omega_i) \leq L(\omega_i)$, and where S_u is the set of frequency points ω_i in $[0, \pi]$ satisfying both $A'(\omega_i) = 0$ and $A(\omega_i) \geq U(\omega_i)$.

5. **Check for Convergence:** If $A(\omega) \geq L(\omega) - \epsilon$ for all frequency points in S_l and if $A(\omega) \leq U(\omega) + \epsilon$ for all frequency points in S_u then convergence has been achieved. Otherwise, go back to step 2.

According to the Kuhn-Tucker conditions, because $\boldsymbol{\mu} \geq \mathbf{0}$ in ensured for each set of computed cosine coefficients $\mathbf{a}$, each filter minimizes the L_2 error subject to the inequality constraints (19,20) over a set of frequencies. At convergence, the constraint set frequencies are exactly those extrema of $A(\omega)$ where $A(\omega)$ touches the lower and upper bound function. It should be noted that negative multipliers generally appear only during the early iterations of the algorithm. ϵ in step 4 is a small number (like 10^{-6}) indicating the numerical accuracy desired. In appendix A, some issues concerning the convergence of the algorithm and the optimality of the filters it produces are discussed.

A flowchart is shown in figure 1. The Matlab program below implements this algorithm with $W(\omega) = 1$. For the sake of space and clarity, it uses a grid of frequency values. However, it is much preferable to refine the location of the extremal frequencies by Newton's method; otherwise a rather dense grid is sometimes required for convergence. The use of Newton's method is easily incorporated.

The computational complexity is $O(M^3)$ per iteration, however, the computation required for each iteration depends upon the size of the constraint set for that iteration. Some of the efficient computational techniques that have been used to improve the implementation of the Parks-McClellan program can also be applied to the algorithm described here [3, 4, 5, 6, 9, 32]. These techniques are used to (1) increase the speed of execution, (2) reduce memory requirements, and (3) improve numerical accuracy of the result.

Example 1: We let $D(\omega)$ be the usual ideal lowpass filter with a cut-off frequency $\omega_o = 0.3\pi$ shown in figure 2. We use 31 cosine coefficients ($M = 30$ and the filter length is 61), $U(\omega) = D(\omega) + \delta$, and $L(\omega) = D(\omega) - \delta$ with $\delta = 0.02$. We use the L_2 weighting function $W(\omega) = 1$. In 4 iterations, the above described algorithm converges to the frequency response amplitude shown in figure 4. In the figure, the circular marks indicate the 14 constraint set frequency points upon convergence. Compared to the best unconstrained L_2 filter shown in figure 3, the constrained filter in figure 4 has a considerably smaller peak error near the band edge. This is achieved with a small increase in the transition width and the L_2 error. The L_2 error and the peak error associated with the best unconstrained L_2 filter are $||E||_2^2 = 0.003375$ and 0.09369 respectively. For the constrained L_2 filter, they are $||E||_2^2 = 0.003858$ and 0.02 respectively. The resulting 'induced' band edges of the constrained L_2 filter are $\omega_{ip} = 0.2728\pi$ and $\omega_{is} = 0.3270\pi$.

The convergence of the algorithm is illustrated in figure 5. The amplitude $A_0(\omega)$ shown in figure 5(a) is the best unconstrained L_2 filter (given by the sinc function). The first set of constraints formed are taken to be the extrema of $A_0(\omega)$ that violate the upper and lower bound constraints. When equation (32) is used with this constraint set to calculate $\boldsymbol{\mu}$ it turns out that two of the Lagrange multipliers are negative. (The two constraints corresponding to these two negative Lagrange multipliers are the local maximum and the local minimum nearest $\omega = 0$.) In this example, after the constraint set frequency corresponding to the more negative of these two negative multipliers is removed from the constraint set and equation (32) is used again, there is still one negative multiplier. After the corresponding constraint is removed from the constraint set and equation (32) is used a third time to calculate $\boldsymbol{\mu}$, it is found that all the multipliers are positive. Now equation (33) is used to compute the filter coefficients $\mathbf{a}$, the new amplitude $A_1(\omega)$ is shown in figure 5(b). The circular marks in this figure indicate the constraint set used to obtain $A_1(\omega)$. $A_1(\omega)$ interpolates $L(\omega)$ and $U(\omega)$ at the constraint points because equations (32,33) were derived from equations (16,17,18) and because equations (17,18) are interpolation equations. As above, the extrema of $A_1(\omega)$ that violate the upper and lower bound constraints are used to form a new constraint set. Equation (32) is used with this constraint set to compute a new set of multipliers. It is found that all the new multipliers are positive and so no constraints are removed from this constraints set. Equation (33) is then used and the amplitude $A_2(\omega)$ is shown in figure 5(c). $A_3(\omega)$ is obtained similarly, also without the appearance of negative multipliers.

It should be noted that this example and the following examples were generated using a version of the program that uses a grid to approximately locate the extrema of $A(\omega)$ and Newton's method to refine these frequencies.

5 Interpretations and Extensions

There are several observations and interpretations of this algorithm that may be helpful in understanding it in relation to other approaches and in modifying it for other applications.

The constraint set at each step in the iteration contains the candidates for the final extremal frequencies that touch the constraint. Satisfying these constraints forces $A(\omega)$ to interpolate $L(\omega)$ and $U(\omega)$ at the frequencies in the constraint set. This is quite similar to the behavior of the Remez algorithm used in the Parks-McClellan program. The process of the algorithm has three major differences to the PM program:

1. The number of reference set frequencies in the PM program is fixed and does not change throughout the algorithm, while here, the number of constraint set frequencies does change and is generally smaller than the number used in the PM program. Also, the alternation of the signs of $E(\omega)$ that holds for a Parks-McClellan equiripple filter and is enforced at each step of the Remez algorithm does not necessarily hold here.

2. The size of the Chebyshev error changes (increases) on each step of the PM program, while here it is prescribed. Here, the L_2 error generally increases and the induced transition band generally widens.

3. In the PM program, the number of reference set frequencies is $M+2$, this is one greater than the number of unknown cosine coefficients. In our algorithm, unless the constraints are tight, the number of frequencies is less than the number of unknowns, and those degrees of freedom not used to satisfy interpolation constraints are used to minimize the integral square error at each step.

Nevertheless, the overall behavior is similar to that of the PM program. As in the Remez algorithm used in the PM program, on each iteration (1) an optimization problem is solved over a finite set of frequencies and (2) the set of frequencies is updated. Indeed, the interpolation step of the Remez algorithm can be interpreted as an optimization problem: The filter found at each iteration of the Remez algorithm minimizes the Chebyshev error over the updated reference set [23].

Often it is desirable to include equality constraints on the value and derivatives of $A(\omega)$ at prescribed frequency points. An application is given, for example, in [21], where flatness at $\omega = 0$ is achieved by imposing appropriate equality constraints. The inclusion of these and other linear equality constraints on the cosine coefficients in the approach described in this paper is straight forward. It requires only the use of extra Lagrange multipliers, the signs of which are unimportant. For example, magnitude squared design of minimum phase filters can be accomplished by taking the lower bound function $L(\omega)$ to be 0 in the stopband and by spectrally factoring the resulting nonnegative frequency response amplitude. A polynomial root finding algorithm and the computation of spectral factors is discussed in [19].

5.1 Chebyshev Solutions

Observe that if, for a fixed number of filter coefficients, the constraint on the weighted Chebyshev error $||E(\omega)||_\infty$ in [1, 18] is chosen too small, (or equivalently, δ_p and δ_s in (7,8) are chosen too small) then no filter satisfies the constraint. In this case the algorithms of [1, 18] can not converge. Although this problem can be avoided by computing the minimum value of $||E(\omega)||_\infty$ with the Parks-McClellan program, it is interesting to note that there is no minimum δ_p and δ_s below which the approach described in this paper fails to converge. If δ_p and δ_s are taken to be small, then the transition region between the passband and stopband simply becomes wider. Therefore, for a fixed ω_o, by decreasing δ_p and δ_s, a continuum of Parks-McClellan equiripple filters are obtained.

Example 2: We use the same desired lowpass amplitude as in example 1, we keep $M = 30$, but we use $\delta = 0.004$. We use the unity L_2 weighting function. The resulting frequency response shown in figure 6 is obtained in 6 iterations and the size of the constraint set upon convergence is 30. Here, the peak error is significantly reduced with a corresponding increase in the transition width and L_2 error. The L_2 error associated with this filter is $||E||_2^2 = 0.004780$. The resulting 'induced' band edges of this constrained L_2 filter are $\omega_{ip} = 0.2576\pi$ and $\omega_{is} = 0.3421\pi$.

Note that although this filter was not designed using the Parks-McClellan program, it corresponds to a best Chebyshev filter for appropriately chosen band edges because then the alternation theorem will be satisfied. For this example, this algorithm takes about the same number of iterations as does the Parks-McClellan program when the PM program is executed without a superior initialization of reference frequencies such as those given in [6, 9, 32]. This comparison is made to evaluate the convergence in the number of iterations, not to compare execution times.

It is interesting that the filter in example 2 is both (i) a best peak constrained L_2 filter, and (ii) a best Chebyshev filter for an appropriate transition band. There are other algorithms for the design of equiripple filters with specified peak errors [13, 14, 15, 16, 26, 27], but the peak constrained L_2 approach used here gives a way to design a subset of such filters which incorporates the L_2 error.

5.2 Trade-off Curves

In [1] Adams provides a curve illustrating the trade-off between the weighted L_2 error and the Chebyshev error for the filters produced by his approach. It is quite convincing that the filters on the endpoints of this curve do not provide the most desirable trade-off. The same is true here.

By decreasing the peak error δ to 0, a curve illustrating the trade-off between the unweighted integral square error and δ can be obtained. Figure 7 shows the curve for length 61 filters designed using the approach of this paper, where the cut-off frequency is $\omega_o = 0.3\pi$ and $\delta = \delta_p = \delta_s$. δ is varied to obtain the curve in the figure. The circle at the right end of the figure indicates the best unconstrained L_2 filter. As δ is decreased from that point, the L_2 error increases as illustrated. The circular mark at $\delta \approx 0.0086$ indicates the point where δ first becomes small enough to produce a Parks-McClellan equiripple filter. In this example, all points on the curve to the left of this point represent Parks-McClellan equiripple filters. Because there is no smallest value for δ, the curve approaches a point on the $\delta = 0$ axis. We conjecture that the point on the $\delta = 0$ axis this curve approaches represents a maximally flat (digital Butterworth) filter [13].

We should note that even though all points on the curve for $\delta < 0.0086$ in figure 7 represent equiripple filters, this is not true in general. For example, suppose δ_s is decreased towards zero and δ_p is kept constant. Then it is sometimes the case that the set of δ_s, for which the points on the curve represent equiripple filters, is a union of disjoint intervals.

5.3 Specified Band Edges

The approach taken in this paper does not preclude the specification of a transition band edge in the design of a lowpass filter. Let $\omega_p < \omega_o$ be a specified passband edge. This subsection describes how to append the constraint

$$L(\omega_p) \leq A(\omega_p) \leq U(\omega_p) \tag{42}$$

to the formulation of section 3. By doing so, the frequency response amplitude $A(\omega)$ will be guaranteed to lie between the lower and upper bound functions $L(\omega)$ and $U(\omega)$ for all ω in the passband $[0, \omega_p]$. This is because the only way in which $A(\omega)$ can violate the lower and upper bound constraints in the passband is by violating one of them at ω_p or by violating one of them at a local extremal. Because the algorithm described in section 4.2 ensures that $L(\omega)$ and $U(\omega)$ are not violated at the local extremals of $A(\omega)$, it is sufficient to append the single constraint (42). The appropriate modification to the algorithm of section 4.2 is described below and a program is available from the authors or electronically (see below).

Appending the constraint (42) to the existing constraints requires modifying only the way in which the constraint set is updated. There are two issues that must be addressed.

First notice that the constraint (42) may be satisfied by the filter produced by the basic algorithm of section 4.2 - the constraint is met with no additional effort. This is the case when the induced passband edge ω_{ip} is closer to ω_o than is the specified passband edge ω_p. When this is not the case, it is necessary to simply append ω_p to the constraint set. To detect, during the iterative algorithm, exactly when it is necessary to include ω_p in the constraint set, the following decision rule is used: Let ω_a be the passband extremum of $A(\omega)$ closest to ω_o. If $\omega_a < \omega_p < \omega_o$ and $A(\omega_p) < L(\omega_p)$, then include ω_p in the constraint set, otherwise leave the constraint set unchanged.

The second issue that must be addressed is the occurrence of an over-constrained problem on some iteration. After appending ω_p to the constraint set, the number of constraints may outnumber the number of cosine coefficients by one. When this occurs, equations (21,22) are in general over-determined and can not be solved. Consequently, a frequency must be removed from the constraint set before the algorithm can proceed. This occurs only when the constraints are relatively tight, in which case the algorithm described here reproduces a modified Parks-McClellan algorithm [26]. Likewise, the rule for deciding which frequency to remove is similar to the rule used in [26] and is described below. Note that in this situation, the constraint set necessarily contains 0 and π. Step 4 of the algorithm described in section 4.2 becomes:

4 **Multiple Exchange of Constraint Set:** Let S_l be the set of frequency points ω_i in $[0, \pi]$ satisfying both $A'(\omega_i) = 0$ and $A(\omega_i) \leq L(\omega_i)$. Let S_u be the set of frequency points ω_i in $[0, \pi]$ satisfying both $A'(\omega_i) = 0$ and $A(\omega_i) \geq U(\omega_i)$.

Let ω_a be the passband extremum of $A(\omega)$ closest to ω_o. If $\omega_a < \omega_p < \omega_o$ and $A(\omega_p) < L(\omega_p)$, then let $S_l = S_l \cup \omega_p$.

If $\omega = 0$ is a local maximum of $A(\omega)$, let $E_0 = A(0) - U(0)$, otherwise set $E_0 = L(0) - A(0)$. If $\omega = \pi$ is a local maximum of $A(\omega)$, let $E_\pi = A(\pi) - U(\pi)$, otherwise set $E_\pi = L(\pi) - A(\pi)$.

If $|S_l| + |S_u| = M + 2$ and $E_0 < E_\pi$, then remove 0 from S_l or S_u, whichever contains 0.

If $|S_l| + |S_u| = M + 2$ and $E_0 \geq E_\pi$, then remove π from S_l or S_u, whichever contains π.

Set the constraint set S equal to $S_l \cup S_u$.

A stopband edge ω_s can be specified instead of a passband edge in exactly the same way. If both the passband and the stopband are to be specified simultaneously, then the problem can be posed as a quadratic program as discussed above, and Adams' approach should be used. When both band edges are specified, it is possible that no solution exists because the transition band can not be arbitrarily sharp. Note that here a distinction is being made between the cut-off frequency ω_o and the band edges ω_p and ω_s ($\omega_p \leq \omega_o \leq \omega_s$).

Example 3: We use the same desired lowpass frequency response as in the previous two examples with $\delta = 0.020$ and $M = 30$, but we require that the passband edge be located at 0.285π. We use the L_2 weighting function $W(\omega) = 1$. In 7 iterations the algorithm converges to the frequency response shown in figure 8. The size of the constraint set upon convergence is 28. The L_2 error associated with this filter is $||E||_2^2 = 0.006893$. The resulting 'induced' stopband edge of this constrained L_2 filter is $\omega_{is} = 0.3376\pi$.

5.4 The Use of L_2 Weighting

Although the algorithm described in section 4.2 was introduced with a uniform L_2 weighting function, the algorithm can also be used with an L_2 weighting function that equals zero in a specified transition band, if so desired. As discussed above, when it is known that the signals in the input class have no (or little) energy in a transition region, the use of a zero-weighted (or lightly-weighted) transition band is well motivated. The only modification required is the substitution of equation (34) for equation (31). The method for performing the multiple exchange of the constraint set remains unchanged. In this case, the L_2 weighting function possesses band edges, but the upper and lower bound functions are not enforced at these two frequencies. As in section 4.2, the upper and lower bound functions are used to constrain the frequency response amplitude only at its local extrema. This variation of the algorithm combines the approach of Adams' with the approach suggested in section 3 in which a uniform L_2 weighting is used. Like the uniformly L_2 weighted approach, this variation avoids the infeasibility problems associated with the quadratic program approach.

5.5 Remarks on Comparisons

As mentioned above, when the present algorithm yields an equiripple filter, it gives the same result as the PM algorithm when the PM algorithm is used with the appropriate specifications. Also, the frequency responses of the filters produced by the algorithm described in this paper are similar to those obtained by the approach of Adams. It should be noted that although the responses are similar, they are not exactly the same in general. This is because the quadratic program formulation Adams gives is not equivalent to the formulation given in section 3. The differences lie in the weighting of the error and the way in which the constraints are imposed. Note that when the algorithm presented here and that of Adams each give an equiripple PM filter with the same lower and upper bound functions and the same transition band ('induced' in the case of the present algorithm) then certainly the two filters are identical.

The approaches of [13, 14, 15, 16] should also be mentioned. While they employ implicitly defined transition bands and provide direct control of the peak errors, those algorithms (i) do not incorporate the L_2 error into the design procedure and (ii) provide only approximate control of the location of the cut-off frequency. (Although for lowpass filter design limitation (ii) can be overcome [26, 27]). In addition, it should also be noted that while the program of [33] is very flexible, it does not incorporate the L_2 error.

5.6 Multiband Filters

When used for the design of multiband filters, the simple algorithm we have described for lowpass filter design does not, in general, converge. We have have found as in [2] that it is necessary to use a single point update procedure for some iterations to obtain robust convergence [29]. Thus, by maintaining the approach of using a sequence of equality constrained L_2 minimizations, best peak constrained L_2 multiband filters can be readily designed without the use of 'don't care' regions. A program for the multiband case using the criterion described in this paper is also available from the authors or electronically. Adams et al. address algorithm issues concerning the design of multiband filters via the quadratic program formulation in detail in [2].

6 Conclusion

We have considered the design of optimal filters and have discussed the implicit assumptions associated with the use of explicitly specified transition bands in the frequency domain design of FIR filters. We have also put forth the notion that explicitly specified transition bands have been introduced in the filter design literature in part as an indirect approach for dealing with Gibbs' phenomenon occuring at the discontinuities in the desired frequency response. Moreover, the results of Weisburn, Parks, and Shenoy suggest that if a 'don't care' region is used for filter design, then unless the input signals have no energy in the 'don't care' region, the optimality of the best Chebyshev and L_2 filters in the operator norm sense is problematic. Because the minimization of the Chebyshev norm requires the use of a specified transition band, this suggests that the Chebyshev criterion is better suited as a constraint rather than the primary optimization criterion. This is also consistent with the motivation Adams gives to support the constrained L_2 approach described in [1].

This paper (i) proposes that the unweighted (or weighted) integral square error be minimized such that the peak errors lie within the specified tolerances, and (ii) describes a simple multiple exchange algorithm for lowpass filter design according to this design formulation that is robust and efficient. Because the proposed approach does not ignore the L_2 error around the band edge, it does not implicitly assume that signals in the input class have no frequency content there. In addition, the constraints imposed upon the peak errors can be made arbitrarily small. For a fixed cut-off frequency, it also gives the best L_2 filter and a continuum of Chebyshev filters as special cases. With the weighting function $W(\omega) = 1$, the optimal filter coefficients are obtained by making a simple *additive* correction to the Fourier series coefficients.

The approach taken in this paper is distinct from many other filter design methods because it does not exclude from the integral square error a region around the cut-off frequency, and yet, it overcomes Gibbs' phenomenon without resorting to windowing or 'smoothing out' the discontinuity of the ideal lowpass filter. The algorithm is also appealing because it can be implemented with an especially simple Matlab program. Versions of this and other programs (multiband, etc.) can be found on the World Wide Web at URL *http://www-dsp.rice.edu*.

7 Acknowledgment

We wish to express appreciation to Professor Hans W. Schüßler for his interest, careful review of the manuscript, observations, and valuable suggestions. We also wish to thank the anonymous reviewers. Responding to their comments helped us improve the paper.

Appendices

A On Optimality and Convergence

To discuss the optimization problem discussed above, define the feasible set Q:

$$Q = \{\mathbf{a} \in \mathbb{R}^{M+1} : \ L(\omega_i) \leq A(\omega_i) \leq U(\omega_i) \text{ for the local extremal frequencies } \omega_i \text{ of } A(\omega)\}. \quad (43)$$

The problem is to minimize $||E||_2^2$ over the set Q. Since Q is closed and $||A - D||_2^2$ is a convex function of $\mathbf{a}$, a minimizer exists. Usually uniqueness of a minimizer is established by ascertaining the convexity of the feasible set. However, the set Q is not convex as is easily explained: Note that a maximally flat frequency response will always be feasible (it has extremal frequencies only at 0 and π, where $A(\omega)$ equals 1 and 0 respectively). If a filter is obtained by averaging a feasible equiripple (PM) filter and maximally flat (Herrmann) filter, then the frequency response amplitude of the filter will have local extrema around the cut-off frequency which will generally violate the upper and lower bound functions $L(\omega)$ and $U(\omega)$. Therefore Q is not convex. Because Q lacks convexity, it is necessary to use a different perspective to discuss optimality, as follows.

Although the minimization of $||E||_2^2$ over Q is not a quadratic program, it is closely related to one. Suppose the L_2 weighting function is set to unity throughout the following discussion. If the algorithm converges, then the filter obtained by the algorithm is optimal in the sense that it is the solution to a meaningful quadratic program: Given a filter produced by the algorithm, define the two 'induced' band edges as follows: Define the induced passband edge ω_{ip} to be the highest frequency less than ω_o at which $A(\omega)$ equals $L(\omega)$. Similarly, define the induced stopband edge ω_{is} to be the lowest frequency greater than ω_o at which $A(\omega)$ equals $U(\omega)$. See figure 9. Second, label the two extremals of $A(\omega)$ that are adjacent to ω_{ip} and ω_{is}: Let ω_a be the frequency at which $A(\omega)$ achieves its first local maximum to the left of ω_o. Similarly, let ω_b be the frequency at which $A(\omega)$ achieves its first local minimum to the right of ω_o. With these definitions of ω_a, ω_b, ω_{ip} and ω_{is}, it can be said that, if ω_p and ω_s are two frequencies satisfying $\omega_a \leq \omega_p \leq \omega_{ip}$, and $\omega_{is} \leq \omega_s \leq \omega_b$, then the filter obtained by the algorithm described above solves the quadratic program (7,8), where the weight function $W(\omega) = 1$ for all $\omega \in [0, \pi]$ is used, not the weight function of expression (9).

The intervals $[\omega_a, \omega_{ip}]$ and $[\omega_{is}, \omega_b]$ give insight into the properties of the solution. If $\omega_s \in [\omega_{is}, \omega_b]$ but ω_p is taken to be a frequency less than ω_a, then the solution to the quadratic program (7,8) is a filter having peak errors in the passband that exceed δ_p. On the other hand, if ω_p is taken to be a frequency greater than ω_{ip}, then the L_2 error of the resulting filter must be greater (because it is subject to additional constraints). Similar statements are true for ω_s. Therefore, these values have the following two properties:

1. Filters obtained by solving the quadratic program (7,8, but with $W(\omega) = 1 \, \forall \, \omega$) with narrower transition widths ($\omega_s - \omega_p < \omega_{is} - \omega_{ip}$) necessarily have a greater L_2 error (assuming same the peak constraints).

2. Filters obtained by solving the quadratic program (7,8, but with $W(\omega) = 1 \, \forall \, \omega$) with wider transition widths ($\omega_s - \omega_p > \omega_b - \omega_a$) have peak errors around ω_o that exceed δ_p and/or δ_s.

Although we have not proven that the filters obtained by the algorithm are global minimizers of $||E||_2^2$ over Q, the quadratic program analysis given here is highly suggestive.

Regarding the convergence of the algorithm, the non-convexity of Q has little bearing. Non-convexity of the feasible set is a problem for algorithms that proceed by updating one feasible solution to obtain another feasible solution by moving within the feasible set. The algorithm described above begins with the best *unconstrained* minimizer, and each filter produced during the course of the algorithm is infeasible. This sequence of filters approaches the feasible set. The algorithm terminates exactly when feasibility is achieved. The progress of this kind of algorithm is affected less by the lack of feasible set convexity.

The quadratic program analysis above also suggests that a unique minimizer exists (the QP has a unique minimizer), although we do not give a proof of this here. Note that because the algorithm always begins with the best unconstrained L_2 filter, there is no ambiguity about the initial filter used in the iterative procedure above. However, as mentioned above, we have not proven the convergence of the algorithm presented in this paper, but have found it to converge reliably in practice for lowpass filter design. Indeed, even though it can be posed as a sequence of similar quadratic programs, the convergence of this algorithm is not supported by the theory of quadratic programming. That it does in fact converge in practice is due to properties of the particular problem of lowpass filter design. When used for bandpass filter design, for example, this algorithm does not converge in general. For multiband filter design we have found as in [2] that it is necessary to modify the update procedure for some iterations to obtain robust convergence. It should be noted, however, that because the constrained L_2 approach of Adams can be formulated as a quadratic program, the existence of a unique optimal solution to the problem posed by him is assured, and the algorithm of [2] is guaranteed to converge to it .

B A Matlab Program

The Matlab program `cl2lp` below implements the algorithm described in section 4.2 of the paper. The program `local_max` computes the indexes of a vector corresponding to its local maxima by comparing each vector element with the two adjacent elements. This version of `cl2lp` does not use Newton's method to refine the position of the extrema of $A(\omega)$, so a possibly dense grid is required in order to obtain convergence ($L \geq 2^{\lceil \log_2 10m \rceil}$). A version using Newton's method (and other programs) can be obtained from the authors or electronically on the World Wide Web.

```
function h = cl2lp(m,wo,up,lo,L)
% Constrained L2 Low Pass FIR filter design
% Author: Ivan Selesnick, Rice University, 1994
% See: Constrained Least Square Design of FIR
% Filters Without Specified Transition Bands
% by I.W.Selesnick, M.Lang, C.S.Burrus
%   h  : 2*m+1 filter coefficients
%   m  : degree of cosine polynomial
%   wo : cut-off frequency in (0,pi)
%   up : [upper bound in passband, stopband]
%   lo : [lower bound in passband, stopband]
%   L  : grid size
% example
%   up = [1.02, 0.02]; lo = [0.98, -0.02];
%   h = cl2lp(30,0.3*pi,up,lo,2^11);

r = sqrt(2);              w = [0:L]'*pi/L;
Z = zeros(2*L-1-2*m,1);   q = round(wo*L/pi);
u = [up(1)*ones(q,1); up(2)*ones(L+1-q,1)];
l = [lo(1)*ones(q,1); lo(2)*ones(L+1-q,1)];
c = 2*[wo/r; [sin(wo*[1:m])./[1:m]]']/pi;
a = c;       % best L2 cosine coefficients
mu = [];     % Lagrange multipliers
SN = 1e-6;   % Small Number
while 1
  % ----- calculate A -------------------------
  A = fft([a(1)*r;a(2:m+1);Z;a(m+1:-1:2)]);
  A = real(A(1:L+1))/2;
  % ----- find extremals ----------------------
  kmax = local_max(A);    kmin = local_max(-A);
  kmax = kmax( A(kmax) > u(kmax)-10*SN );
  kmin = kmin( A(kmin) < l(kmin)+10*SN );
  % ----- check stopping criterion ------------
  Eup = A(kmax)-u(kmax); Elo = l(kmin)-A(kmin);
  E = max([Eup; Elo; 0]); if E < SN, break, end
  % ----- calculate new multipliers -----------
  n1 = length(kmax);         n2 = length(kmin);
  O  = [ones(n1,m+1); -ones(n2,m+1)];
```

```
  G  = O .* cos(w([kmax;kmin])*[0:m]);
  G(:,1) = G(:,1)/r;
  d  = [u(kmax); -l(kmin)];
  mu = (G*G')\(G*c-d);
  % ----- remove negative multiplier ----------
  [min_mu,K] = min(mu);
  while min_mu < 0
    G(K,:) = []; d(K) = [];
    mu = (G*G')\(G*c-d);
    [min_mu,K] = min(mu);
  end
  % ----- determine new coefficients ----------
  a = c-G'*mu;
end
h = [a(m+1:-1:2); a(1)*r; a(2:m+1)]/2;
```

```
function k = local_max(x)
% finds location of local maxima
s = size(x); x = [x(:)].'; N = length(x);
b1 = x(1:N-1)<x(2:N); b2 = x(1:N-1)>x(2:N);
k = find(b1(1:N-2)&b2(2:N-1))+1;
if x(1)>x(2), k = [k, 1]; end
if x(N)>x(N-1), k = [k, N]; end
k = sort(k); if s(2) == 1, k = k'; end
```

References

[1] J. W. Adams. FIR digital filters with least squares stop bands subject to peak-gain constraints. *IEEE Trans. on Circuits and Systems*, 39(4):376–388, April 1991.

[2] J. W. Adams, J. L. Sullivan, R. Hashemi, C. Ghadimi, J. Franklin, and B. Tucker. New approaches to constrained optimization of digital filters. In *Proc. IEEE Int. Symp. Circuits and Systems*, pages 80–83, May 1993.

[3] J. W. Adams and A. N. Willson, Jr. On the fast design of high-order FIR digital filters. *IEEE Trans. on Circuits and Systems*, 32(9):958–960, September 1985.

[4] A. Antoniou. Accelerated procedure for the design of equiripple nonrecursive digital filters. *IEE Proceedings, Part G*, 129(1):1–10, February 1982.

[5] A. Antoniou. New improved method for the design of weighted-Chebyshev, nonrecursive, digital filters. *IEEE Trans. on Circuits and Systems*, 30(10):740–750, October 1983.

[6] F. Bonzanigo. Some improvements to the design programs for equiripple FIR filters. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 274–277, 1982.

[7] C. S. Burrus. Multiband least squares FIR filter design. *IEEE Trans. on Acoust., Speech, Signal Proc.*, 43(2):412–421, February 1995.

[8] C. S. Burrus, A. W. Soewito, and R. A. Gopinath. Least squared error FIR filter design with transition bands. *IEEE Trans. on Acoust., Speech, Signal Proc.*, 40(6):1327–1340, June 1992.

[9] S. Ebert and U. Heute. Accelerated design of linear or minimum phase FIR filters with a Chebyshev magnitude response. *IEE Proceedings, Part G*, 130(6):267–270, December 1983.

[10] R. Fletcher. *Practical Methods of Optimization.* John-Wiley and Sons, 1987.

[11] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33, 1983.

[12] R. A. Gopinath. Some thoughts on least-square error optimal windows. In *Proc. IEEE Int. Symp. Circuits and Systems*, volume 2, pages 413–416, May-June 1994.

[13] O. Herrmann. Design of nonrecursive filters with linear phase. *Electron. Lett.*, 6(11):328–329, May 28 1970. Also in [25].

[14] O. Herrmann and H. W. Schuessler. On the design of nonrecursive digital filters. *IEEE Arden House Workshop*, January 1970.

[15] E. Hofstetter, A. Oppenheim, and J. Siegel. A new technique for the design of nonrecursive digital filters. In *Proc. of the Fifth Annual Princeton Conference on Information Sciences and Systems*, pages 64–72, October 1971. Also in [25].

[16] E. Hofstetter, A. Oppenheim, and J. Siegel. On optimum nonrecursive digital filters. In *Proc. of the Ninth Allerton Conference on Circuits and System Theory*, pages 789–798, 1971. Also in [25].

[17] M. Lang and J. Bamberger. Nonlinear phase FIR filter design with minimum LS error and additional constraints. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 57–60, April 1993.

[18] M. Lang and J. Bamberger. Nonlinear phase FIR filter design according to the l_2 norm with constraints for the complex error. *Signal Processing*, 36(1):31–40, March 1994. Reprinted in July issue to correct typesetting errors.

[19] M. Lang and B.-C. Frenzel. Polynomial root finding. *IEEE Signal Processing Letters*, 1(10):141–143, October 1994.

[20] D. G. Luenberger. *Introduction to Linear and Nonlinear Programming.* Addison-Wesley, 1984.

[21] G. W. Medlin, J. W. Adams, and C. T. Leondes. Lagrange multiplier approach to the design of FIR filters for multirate applications. *IEEE Trans. on Circuits and Systems*, 35(10):1210–1219, October 1988.

[22] T. W. Parks and C. S. Burrus. *Digital Filter Design.* John Wiley and Sons, 1987.

[23] M. J. D. Powell. *Approximation Theory and Methods.* Cambridge University Press, 1981.

[24] L. Rabiner, J. Kaiser, and R. Schafer. Some considerations in the design of multiband finite-impulse-response digital filters. *IEEE Trans. on Acoust., Speech, Signal Proc.*, 22(6):462–472, December 1974.

[25] L. R. Rabiner and C. M. Rader, editors. *Digital Signal Processing.* IEEE Press, 1972.

[26] I. W. Selesnick and C. S. Burrus. Exchange algorithms that complement the Parks-McClellan algorithm for linear phase FIR filter design. *IEEE Trans. on Circuits and Systems II.* To appear.

[27] I. W. Selesnick and C. S. Burrus. Some exchange algorithms complementing the Parks-Mcclellan program for filter design. In *International Conference on Digital Signal Processing, Limassol, Cyprus*, pages 804–809, June 1995.

[28] I. W. Selesnick, M. Lang, and C. S. Burrus. Constrained least square design of FIR filters without specified transition bands. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 2, pages 1260–1263, May 1995.

[29] I. W. Selesnick, M. Lang, and C. S. Burrus. A simple algorithm for constrained least square design of multiband FIR filters without specified transition bands. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Submitted 1995.

[30] R. G. Shenoy, D. Burnside, and T. W. Parks. Fourier analysis of linear periodic systems and multirate filter design. In *Paper Summaries for the 1992 IEEE DSP Workshop*, page 2.4.1, 1992.

[31] R. G. Shenoy, D. Burnside, and T. W. Parks. Linear periodic systems and multirate filter design. *IEEE Trans. on Acoust., Speech, Signal Proc.*, 42(9):2242–2256, September 1994.

[32] D. J. Shpak and A. Antoniou. A generalized Remez method for the design of FIR digital filters. *IEEE Trans. on Circuits and Systems*, 37(2):161–174, February 1990.

[33] K. Steiglitz, T. W. Parks, and J. F. Kaiser. METEOR: A constraint-based FIR filter design program. *IEEE Trans. on Signal Processing*, 40(8):1901–1909, August 1992.

[34] G. Strang. *Introduction to Applied Mathematics.* Wellesley-Cambridge Press, 1986.

[35] J. L. Sullivan and J. W. Adams. A new nonlinear optimization algorithm for asymmetric FIR digital filters. In *Proc. IEEE Int. Symp. Circuits and Systems*, volume 2, pages 541–544, May-June 1994.

[36] D. W. Tufts and J. T. Francis. Designing digital low-pass filters - comparison of some methods and criteria. *IEEE Trans. Audio Electroacoust.*, 18(4):487–494, December 1970.

[37] D. W. Tufts, D. W. Rorabacher, and W. E. Mosier. Designing simple effective digital filters. *IEEE Trans. Audio Electroacoust.*, 18(2):142–158, June 1970.

[38] B. A. Weisburn, T. W. Parks, and R. G. Shenoy. Error criteria for filter design. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 3, pages 565–568, April 1994.

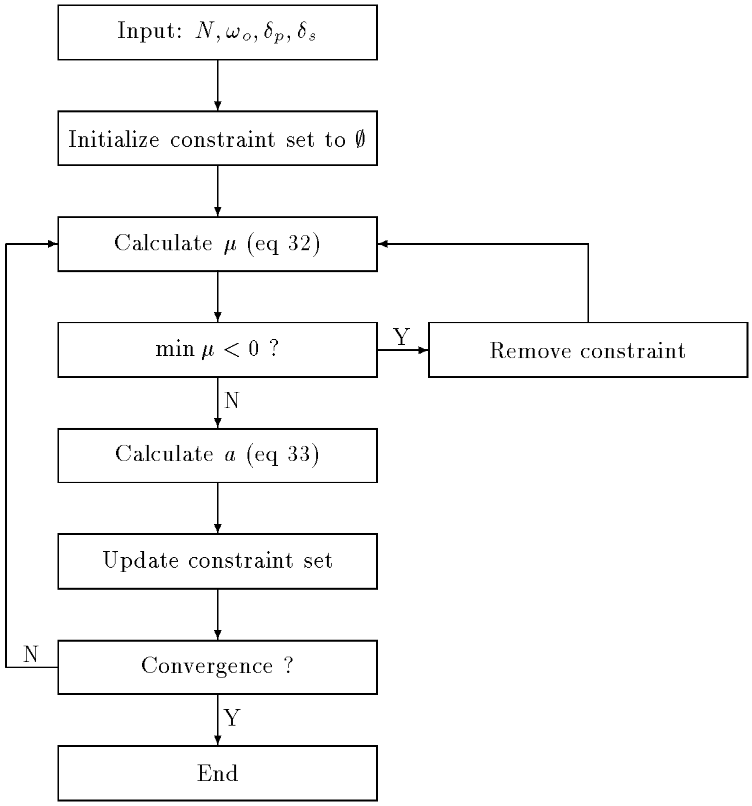


Figure 1: Flowchart for the exchange algorithm for the constrained least square design of lowpass filters.

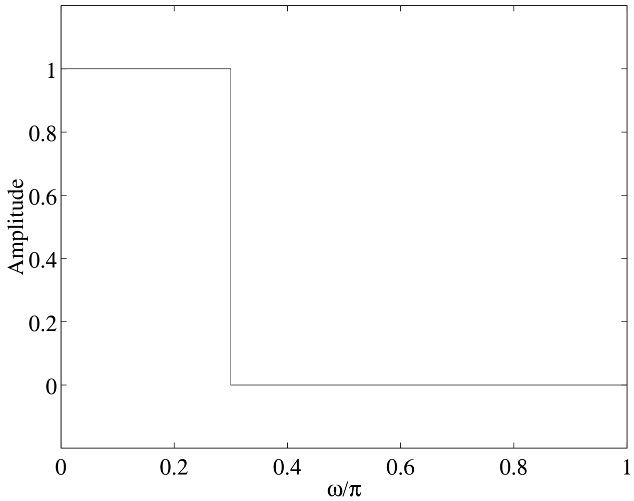


Figure 2: The desired amplitude of an ideal lowpass filter.

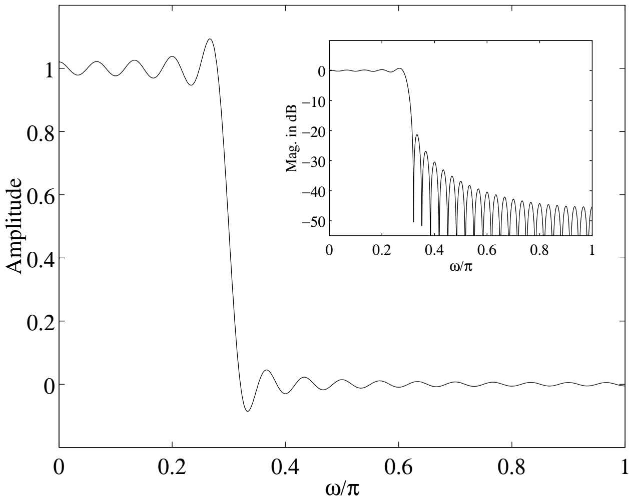


Figure 3: Best unconstrained L_2 filter ($N = 61$, $\omega_o = 0.3\pi$), given by the sinc function. $||E||_2^2 = 0.003375$, peak error $= 0.09369$.

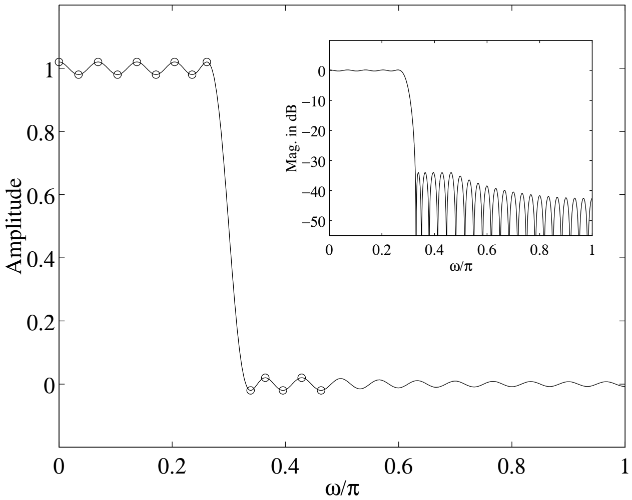
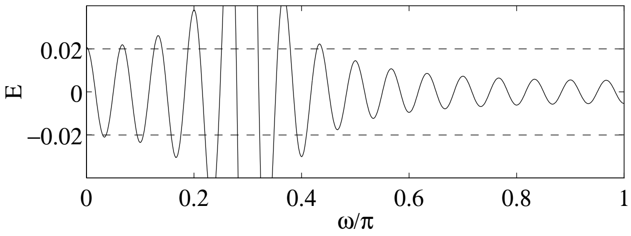
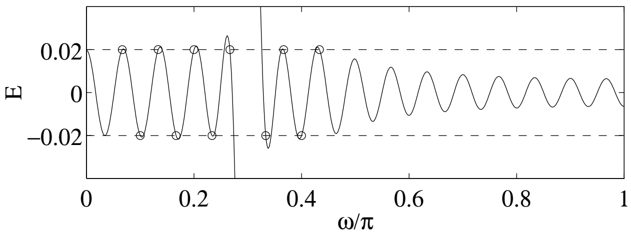


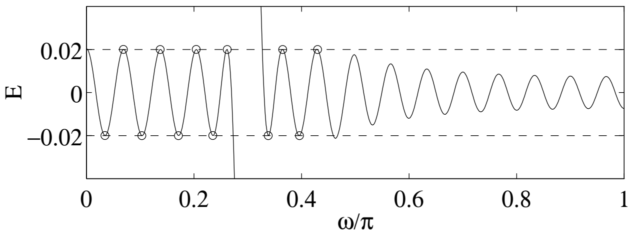
Figure 4: Best constrained L_2 filter ($N = 61$, $\omega_o = 0.3\pi$) with $\delta_p = \delta_s = 0.02$. $||E||_2^2 = 0.003858$. The 'induced' band edges are $\omega_{ip} = 0.2728\pi$ and $\omega_{is} = 0.3270\pi$.



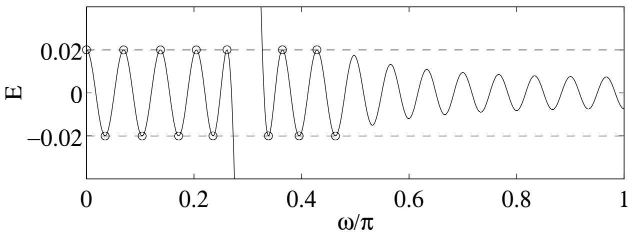
(a) Iteration 0. Maximum constraint violation = 0.07369.



(b) Iteration 1. Maximum constraint violation = 0.006481.



(c) Iteration 2. Maximum constraint violation = 0.001308.



(d) Iteration 3. Maximum constraint violation = 0.000003541.

Figure 5: Illustration of the convergence behavior of the algorithm.

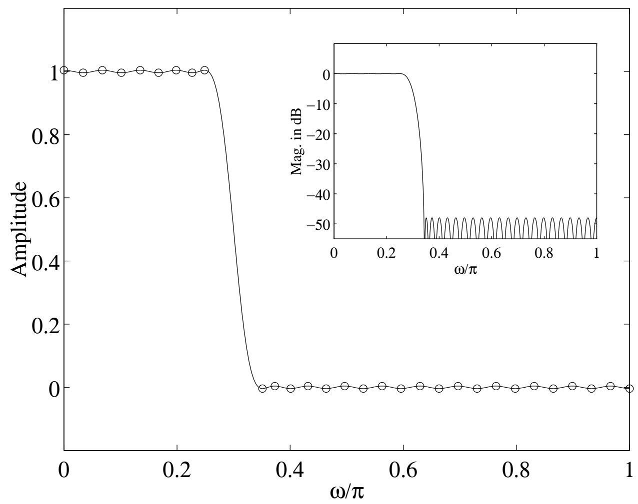


Figure 6: Best constrained L_2 filter ($N = 61$, $\omega_o = 0.3\pi$) with $\delta_p = \delta_s = 0.004$. $||E||_2^2 = 0.004780$. The 'induced' band edges are $\omega_{ip} = 0.2576\pi$ and $\omega_{is} = 0.3421\pi$.

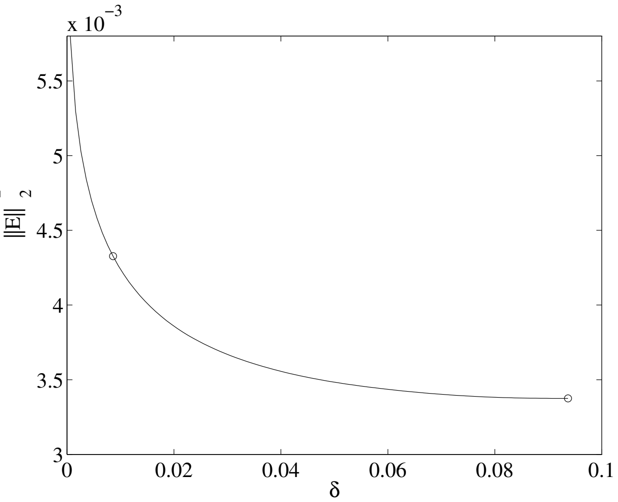


Figure 7: The trade-off curve for integral square error and peak size error: $||E||_2^2$ versus δ_s. This curve illustrates the trade-off for filters of length 61 designed with $\omega_o = 0.3\pi$, and $\delta = \delta_p = \delta_s$.

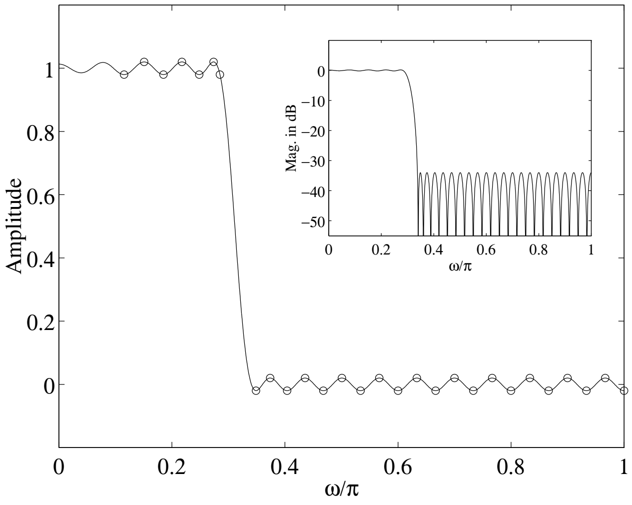
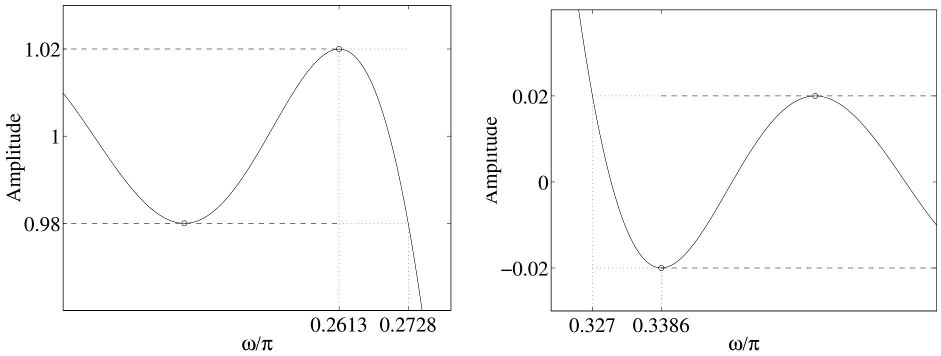


Figure 8: Best constrained L_2 filter ($N = 61$, $\omega_o = 0.3\pi$) with $\delta_p = \delta_s = 0.02$ and a specified passband edge at 0.285π. $||E||_2^2 = 0.006893$. The 'induced' stopband edge is $\omega_{is} = 0.3376\pi$.



(a) passband detail

(b) stopband detail

Figure 9: Passband and stopband details of the amplitude shown in figure 4. $\omega_a = 0.2613\pi$, $\omega_{ip} = 0.2728\pi$, $\omega_{is} = 0.3270\pi$, $\omega_b = 0.3386\pi$.