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
(*selesi@rice.edu*)

ABSTRACT

We consider the design of digital filters and discuss the inclusion of explicitly specified transition bands in the frequency domain design of FIR filters. We put forth the notion that explicitly specified transition bands have been introduced in the filter design literature as an indirect and often inadequate approach for dealing with discontinuities in the desired frequency response.

We also present a rapidly converging, robust, simple algorithm for the design of optimal peak constrained least square lowpass FIR filters that does not require the use of transition bands. This versatile 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.

1. INTRODUCTION

We consider the definition of optimality for digital filter design and conclude that a constrained least squared error criterion with no transition band is often the best approximation measure for many physical filtering problems. This comes from noticing that transition bands are usually introduced to reduce the Gibbs effect for least squares approximation or to permit the use of Chebyshev approximation.

The basic lowpass filter, for example, is usually designed to separate a desired signal from an undesired signal or noise, the spectrums of which exist in bands of frequencies designated the passband and stopband respectively. In most practical cases there is *no* separation of the passband and stopband to give a transition (or "don't care") band between them. Indeed, spectra of the desired and undesired signals often overlap and one is hard pressed to specify a point that separates the pass and stop bands and one certainly cannot give a band to separate them. In most cases, a transition band is introduced to reduce or remove the oscillations in the frequency response near the band edges caused by the Gibbs effect, not because transition bands naturally arises from the physics of the problem. And when large peaks occur in the "don't care" transition band of certain Chebyshev filter designs, engineers decide they *do* care and alter the specifications to eliminate the peaks.

For the meaningful design of filters it is necessary to choose an error criterion carefully. Moreover, the error criterion should not implicitly require unrealistic assumptions on the signals, such as the existence of a band separating desired signals and noise.

1.1. Error Criteria

In the following discussion we draw upon [8] in which Weisburn, Parks and Shenoy present a rigorous motivation for the use of the Chebyshev and L_2 error measures and discuss the use of zero-weighted transition bands. They show that best Chebyshev filters minimize the *energy* of the worst case error signal, while best L_2 filters minimize the *pointwise value* of the the worst case error signal. However, the use of zero-weighted transition bands implicitly requires assumptions on the class of input signals if the corresponding filters are to possess an optimality property.

If a zero-weighted transition band is used, then the best Chebyshev and L_2 filters are optimal in the meaningful way described in [8] *only* if the signals in the input class have no frequency content in the transition band. This assumption on the class of input signals is sometimes difficult to justify.

In light of the preceding discussion, we suggest that the use of explicitly specified transition bands in FIR filter design began with the desire to reduce peak errors near the band edges and that these "don't care" regions are often a somewhat artificial contrivance used to make possible the design of attractive filters. We find that the use of explicitly specified transition bands is sometimes inappropriate and undesirable because, to satisfy a meaningful optimality criterion, their use requires unrealistic assumptions on the class of input signals.

1.2. A New Approach to FIR Filter Design

We present a rapidly converging, robust, simple, multiple exchange algorithm for the design of optimal peak constrained least square lowpass FIR filters that does not require the use of transition bands. The algorithm uses Lagrange multipliers and the Kuhn-Tucker (KT) conditions, as suggested by Adams [1] and further developed in [2, 4], to guarantee optimality upon convergence. This design algorithm will design linear and minimum phase lowpass FIR filters. It gives the best L_2 filter and a *continuum* of Chebyshev filters as special cases and allows arbitrary error weighting. However, with the error weighting $W(f) = 1$, the optimal filter is obtained by making a simple additive correction to the Fourier series coefficients.

This work has been supported by DARPA under an NDSEG fellowship, BNR, NSF and Humboldt.

The algorithm can be modified to allow different error weighting in different bands, to allow other types of constraints, and to achieve complex approximation. We have designed lowpass filters of lengths over 3,000 and have used loose and tight constraints that differed in the pass and stop bands by factors as much as 1,000,000. Although we have not proven its convergence for lowpass filter design, the algorithm never failed to converge to the optimal solution. We feel this new approach could be a useful method for many FIR filter designs.

2. THE FILTER DESIGN METHOD

We begin by reviewing strategies employed for the reduction of large peak errors in FIR filter design. After discussing the constrained L_2 approach and viewpoint of Adams [1], we introduce our approach to filter design. With it, we adopt the valuable insight of Adams and at the same time we do not use a zero-weighted transition band, a transition function, or a window. To achieve this, we formulate the constrained L_2 approach differently than does Adams. We then describe an algorithm that designs FIR filters according to this new formulation and give examples of its efficacy.

2.1. Preliminaries

Define the error function $E(f) = A(f) - D(f)$ where $A(f)$ and $D(f)$ are the realized and desired real-valued frequency response amplitudes of a linear phase FIR filter. The L_2 measure is given by $||E||_2^2 = \int_0^{0.5} W(f)^2 E(f)^2 df$.

The simplest method to design optimal FIR filters minimizes $||E||_2$ and the resulting filter we call the best L_2 filter. As is well known, if $W(f) = 1$, then the best L_2 filter is obtained by truncating the Fourier series of $D(f)$ (the rectangular window method). But this is not done in practice because the resulting filters possess large peak errors near the band edges. To overcome this behavior, known as Gibbs phenomenon, two approaches have been employed: (i) non-rectangular windowing and (ii) the introduction of explicit transition bands.

Although windows are simple to use, they are generally considered sub-optimal because it is difficult to use them to minimize meaningful error measures and because error weighting is not allowed.

By a transition band, we mean a region placed between two bands where either $W(f)$ is taken to be 0 or a function is used to (continuously) connect the two bands. Because $E(f)$ is not weighted there, zero-weighted transition bands are sometimes called "don't care" regions.

2.2. Adams' Error Criterion

In [1] Adams described perhaps the most meaningful error criterion to date and suggested an algorithm to design the corresponding best filters. Using zero-weighted transition bands Adams asks that $||E||_2$ be minimized subject to a constraint on the Chebyshev error, $||E||_\infty$. He provides an excellent motivation for this approach, an approach which yields best L_2 and Chebyshev filters as special cases. Adams notes that it is possible to reduce the Chebyshev error of a best L_2 filter with only a slight increase in the L_2 error.

2.3. The New Problem Formulation

Our problem formulation is similar to that of [1, 2, 7, 4]: the error measure we minimize is $||E||_2$, but with $W(f) = 1$ over $[0, 0.5]$, and we impose a constraint on the maximum value of $|E(f)|$, but we impose this constraint only where $\frac{\partial E(f)}{\partial f} = 0$. The associated minimization problem is:

$$\min ||E||_2$$

$$\text{subject to } |E(f)| \leq T(f), \ \forall\, f \text{ for which } E'(f) = 0.$$

When $D(f)$ is discontinuous this constraint is *different* than a constraint on $||E||_\infty$. This constraint addresses directly the size of the "overshoot" near the band edge, for it literally constrains the peaks of $|A(f)|$ at its local maxima. It is straight forward to use arbitrary weighting functions and to impose more general constraints.

2.4. The Algorithm

The algorithm we use solves a succession of *equality* constrained square error minimization problems where the constraints are on $A(f)$ for the frequency points in a constraint set. The constraint set is updated so that at convergence the only frequency points at which constraints are imposed are those where $A(f)$ touches the constraint. The equality constrained problem is solved with Lagrange multipliers [6]. According to the KT conditions, the *equality* constrained problem solves the corresponding *inequality* constrained problem if all the multipliers are non-negative.

Let the constraint set be $\{f_1, \ldots, f_L\}$. If f_i is a candidate local maximum (minimum) of $A(f)$, then it is necessary to impose the constraint $A(f_i) \leq D(f_i) + T(f_i)$ $(A(f_i) \geq D(f_i) - T(f_i))$. Together, these can be written as $s_i A(f_i) \leq s_i D(f_i) + T(f_i)$ where s_i is 1 and -1 respectively.

To minimize $||E||_2$ subject to $s_i A(f_i) = s_i D(f_i) + T(f_i)$, the use of Lagrange multipliers yields the equations

$$\frac{\partial ||E||_2^2}{\partial a_k} + \frac{1}{2} \sum_{i=1}^{L} \mu_i s_i \frac{\partial A(f_i)}{\partial a_k} = 0 \qquad (1)$$

$$E(f_i) = s_i T(f_i). \qquad (2)$$

For an even symmetric odd length filter, $A(f)$ can be written as $\frac{a_0}{\sqrt{2}} + \sum_{k=1}^{m} a_k \cos 2\pi k f$ [5] so eqs (1) and (2) become

$$\begin{bmatrix} \mathbf{I}_{m+1} & \mathbf{G}^t \\ \mathbf{G} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mu \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \qquad (3)$$

where the unknowns are the coefficients, $\mathbf{a} = (a_0, \ldots, a_m)^t$, and the multipliers, $\mu = (\mu_1, \ldots, \mu_L)^t$. In eq (3),

$$\mathbf{c}_0 = 2\sqrt{2} \int_0^{0.5} D(f) df \qquad \mathbf{c}_k = 2 \int_0^{0.5} D(f) \cos 2\pi k f df \qquad (4)$$

$$\mathbf{d}_i = s_i D(f_i) + T(f_i) \qquad (5)$$

$$\mathbf{G}_{i,0} = \frac{s_i}{\sqrt{2}} \qquad \mathbf{G}_{i,k} = s_i \cos 2\pi k f_i. \qquad (6)$$

Note that $\mathbf{c}$ are the Fourier series coefficients and that

$$\mu = (\mathbf{G}\mathbf{G}^t)^{-1}(\mathbf{G}\mathbf{c} - \mathbf{d}) \qquad \mathbf{a} = \mathbf{c} - \mathbf{G}^t\mu \qquad (7)$$

is the solution to eq (3). Therefore, if the number of constraints (L) is small compared to the number of filter coefficients ($m+1$), then (3) is computationally simple to solve. On each iteration the cosine coefficients, $\mathbf{a}$, are obtained by *adding* the correction $\mathbf{G}^t\mu$ to the best L_2 (Fourier) coefficients, $\mathbf{c}$. This is in contrast to the window method, in which the best L_2 coefficients are *multiplied* by a window.

The algorithm begins with the best unconstrained L_2 filter. Then constraints are iteratively imposed upon $A(f)$ at selected frequencies until the best constrained L_2 filter is obtained. The algorithm can be summarized as follows:

1. Initialize the constraint set to the empty set.
2. Use eq (7) to calculate the multipliers and the filter that minimizes $||E||_2$ subject to the equality constraints $s_i A(f_i) = s_i D(f_i) + T(f_i)$ for all f_i in the constraint set.
3. If there is a constraint set frequency f_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, go on to step 4.
4. Set the constraint set equal to the set of frequency points satisfying both (i) $E'(f) = 0$ and (ii) $|E(f)| \geq T(f)$. If $A(f_i)$ is a local maximum (minimum), then set $s_i = 1$ ($s_i = -1$). If $|E(f)| \leq T(f) + \epsilon$ for all frequency points in the new constraint set, then convergence has been achieved. Otherwise, go back to step 2.

As in [2, 4, 7], according to the KT conditions, optimality is guaranteed upon convergence because $\mu \geq \mathbf{0}$.

The Matlab program below implements this algorithm. For the sake of space and clarity, it uses a grid of frequency values. It is much preferable, however, to refine the location of the extremal frequencies by Newton's method; otherwise a rather dense grid is sometimes required for convergence. Some of the computational techniques [3] used to improve the Parks-McClellan program can also be used here.

Example 1: We let $D(f)$ be the ideal low-pass filter with a band edge at 0.15 and $T(f) = 0.025$. We use 31 cosine coefficients (the filter length is 61). In 3 iterations, the proposed algorithm converges to the response in fig 1. The circular marks indicate the constraint set frequency points. Compared to the best unconstrained L_2 filter in fig 2, the constrained filter 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.

3. CHEBYSHEV SOLUTIONS

The proposed algorithm gives as special cases a *continuum* of best Chebyshev filters. First, observe that if, for a fixed filter length, the constraint on $||E||_\infty$ in [1, 2, 4] is chosen too small, then no filter satisfies the constraint and the algorithms of [1, 2, 4] can not converge. However, there is *no* minimum $T(f)$ below which the proposed approach fails to converge. If $T(f)$ is taken to be small, then the transition between the bands simply becomes wider, as we find in example 2.

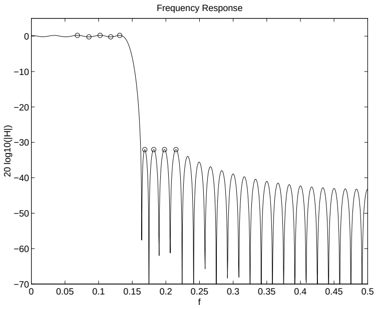

Figure 1: The frequency response for example 1.

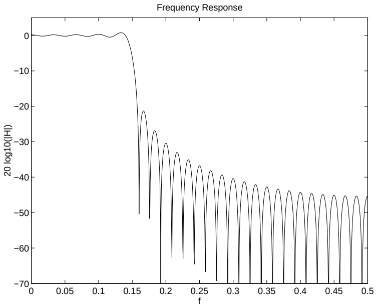

Figure 2: The frequency response of the best unconstrained L_2 filter.

Example 2: We use the same desired response with $m = 30$, but take $T(f) = (0.025)^2$. The resulting response in fig 3 is obtained in 6 iterations. The peak error is significantly reduced with a corresponding increase in transition width and L_2 error.

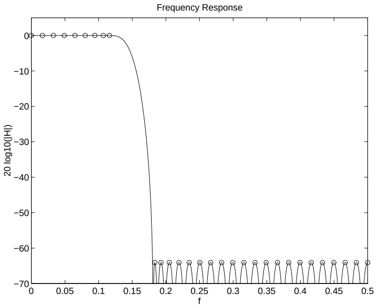

Figure 3: The frequency response for example 2.

Although this filter was not designed by the Parks-McClellan program, it is a best Chebyshev filter for an appropriate transition band. By varying the constraint, a continuum of best Chebyshev filters is obtained. This is in contrast to the approach in [1, 2, 4] where a transition band is specified so that only one value of the tolerance gives rise to a Chebyshev solution.

4. MULTIBAND FILTERS

For the design of multiband filters, the algorithm described above must be modified to obtain robust convergence. Although it works for some multiband specifications, when the tolerance $T(f)$ is taken to be relatively small, we have found that it is often necessary to use a single point update procedure for some iterations. We are currently developing algorithms for the multiband case.

5. CONCLUSION

We have discussed the inclusion of explicitly specified transition bands in the design of optimal FIR filters. We have put forth the notion that explicitly specified transition bands have been introduced in the filter design literature as an indirect approach for dealing with discontinuities in the desired frequency response. However, we have found that by imposing appropriate constraints, Gibbs phenomenon can be eliminated without the use of explicitly specified transition bands. Furthermore, (i) the elimination of Gibbs phenomenon does not depend on smoothing the desired (discontinuous) frequency response, and (ii) the proposed approach does not ignore the L_2 error around the band edge, and thereby does not implicitly assume that signals in the input class have no frequency content there.

The algorithm we have described for our new design formulation is robust and efficient. It also gives the best L_2 filter and a continuum of Chebyshev filters as special cases. In addition, the constraints imposed upon the "overshoot" can be made arbitrarily small. The proposed algorithm allows arbitrary error weighting, however, with the weighting function $W(f) = 1$, the optimal filter coefficients are obtained by making a simple *additive* correction to the Fourier series coefficients.

6. 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. of 1993 ISCAS*, 1993.

[3] 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.

[4] M. Lang and J. Bamberger. Nonlinear phase FIR filter design according to the l_2 norm with constraints for the complex error. *EURASIP Signal Processing*, 36(1), January 1994.

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

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

[7] J. L. Sullivan and J. W. Adams. A new nonlinear optimization algorithm for asymmetric fir digital filters. In *Proc. of 1994 ISCAS*, 1994.

[8] B. A. Weisburn, T. W. Parks, and R. G. Shenoy. Error criteria for filter design. In *Proc. of 1994 ICASSP*, 1994.

7. A MATLAB PROGRAM

```
function h = cl2lp(m,wo,up,lo,L)
% Constrained L2 Low Pass FIR filter design
% Author: Ivan Selesnick, Rice University, 1994
% Please retain this header and cite:
% 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 = [1.025, 0.025]; lo = [0.975, -0.025];
%   h = cl2lp(30,0.3*pi,up,lo,2^10);

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-7;   % Small Number
while 1
  % ----- calculate H -------------------------
  H = fft([a(1)*r;a(2:m+1);Z;a(m+1:-1:2)]);
  H = real(H(1:L+1))/2;
  % ----- find extremals ----------------------
  kmax = local_max(H);    kmin = local_max(-H);
  kmax = kmax( H(kmax) > u(kmax)-SN );
  kmin = kmin( H(kmin) < l(kmin)+SN );
  % ----- check stopping criterion ------------
  Eup = H(kmax)-u(kmax); Elo = l(kmin)-H(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 x = local_max(c)
% finds location of local maxima
s = size(c); c = [c(:)].'; N = length(c);
b1 = c(1:N-1)<c(2:N); b2 = c(1:N-1)>c(2:N);
x = find(b1(1:N-2)&b2(2:N-1))+1;
if c(1)>c(2), x = [x, 1]; end
if c(N)>c(N-1), x = [x, N]; end
x = sort(x); if s(2) == 1, x = x'; end
```