

Part VI: Design of digital filters and filter banks by optimization: Applications

- The purpose of this part is to give a rough idea on how to use linear and nonlinear optimization for synthesizing DSP algorithms.
- There are sets of transparencies of two talks as well as one long article.
- What to read for the examination: Why the two-step procedure described in the two talks as well as in the article are very useful in the cases it can be applied.
- Please do not read all the details!!
- Note that the first set of transparencies concentrates on the use of the Dutta-Vidyasagar algorithm that has been implemented in FORTRAN. The file `fminimax.m` in the MATLAB OPTIMIZATION TOOLBOX can be used equally well for the same purpose. Please do not hesitate to contact the lecturer if you like to use this file.

PRACTICAL EXAMPLES OF OPTIMIZATION IN DIGITAL SIGNAL PROCESSING APPLICATIONS

Tapio Saramäki

Signal Processing Laboratory

Tampere University of Technology

P. O. Box 553, FIN-33101 Tampere, Finland

E-mail: ts@cs.tut.fi

- Needs for Constrained Nonlinear Optimization
- Dutta-Vidyasagar Algorithm (S. R. K. Dutta and M. Vidyasagar, "New algorithms for constrained minimax optimization," *Mathematical Programming*, vol. 13, pp. 140–155, 1977)
- How to Use This Algorithm for Solving Two Example Problems

NEEDS FOR CONSTRAINED NONLINEAR OPTIMIZATION

- When designing digital signal processing algorithms, there exist several optimization problems where one response of the system is desired to be optimized in some sense (usually in the minimax or least-mean-square sense) subject to the given constraints.
- The constraints may include among others:
 - 1) One or more responses depending on the same adjustable parameters are restricted to stay within the given limits in the time or frequency domains.
 - 2) There exist relations between the adjustable parameters.
- When designing direct-form linear-phase FIR filters linear programming can be used in many cases.

- For other types of filters and filter banks, lazy people may use genetic algorithms or simulated annealing if they happen to have powerful enough computers as well as enough time to wait for the solution that is not necessarily the real optimum.
- This talk concentrates on the use of the Dutta-Vidyasagar algorithm.
- Two example problems are included illustrating the use of this algorithm.

DUTTA-VIDYASAGAR ALGORITHM

- This algorithm has turned out to be very powerful in solving various kinds of constrained optimization problems in digital signal processing.
- It can be applied to the following optimization problem: Find \mathbf{x} containing the adjustable parameters to minimize

$$\epsilon = \max_{1 \leq i \leq I} \{f_i(\mathbf{x})\}$$

subject to

$$g_j(\mathbf{x}) \leq 0 \quad \text{for } j = 1, 2, \dots, J$$

and

$$h_l(\mathbf{x}) = 0 \quad \text{for } l = 1, 2, \dots, L.$$

- The main idea in the algorithm is to gradually find ϕ and \mathbf{x} to minimize the following function:

$$\begin{aligned}
 P(\mathbf{x}, \phi) = & \sum_{i|f_i(\mathbf{x}) > \phi} [f_i(\mathbf{x}) - \phi]^2 \\
 & + \sum_{j|g_j(\mathbf{x}) > 0} w_j [g_j(\mathbf{x})]^2 + \sum_{l=0}^L v_l [h_l(\mathbf{x})]^2.
 \end{aligned}
 \tag{A}$$

- In equation (A), the first summation contains only those $f_i(\mathbf{x})$'s for $i = 1, 2, \dots, I$ which are larger than ϕ . Similarly, the second summation contains only those $g_j(\mathbf{x})$'s for $j = 1, 2, \dots, J$ which are larger zero.
- The w_j 's and v_l 's are the weights given by the user.
- If ϕ is very large, then \mathbf{x} can be found to make $P(\mathbf{x}, \phi)$ zero or practically zero. On the other hand, if ϕ is too small, then $P(\mathbf{x}, \phi)$ cannot be made zero.
- To key idea is to find the minimum of ϕ for which \mathbf{x} can be found such that $P(\mathbf{x}, \phi)$ becomes zero or practically zero.
- In this case, $\epsilon \approx \phi$.

STEPS FOR IMPLEMENTING THE DUTTA-VIDYASAGAR ALGORITHM

Step 1: Set $B_L = 0$, $B_U = 10^4$, $\phi_1 = B_U$, and $k = 1$.

Step 2: Find $\hat{\mathbf{x}}_k$ to minimize $P(\mathbf{x}, \phi_k)$.

Step 3: Evaluate

$$M^M = \phi_k + \sqrt{P(\hat{\mathbf{x}}_k, \phi_k)/n},$$

where n is the number of the $f_i(\hat{\mathbf{x}}_k)$'s satisfying $f_i(\hat{\mathbf{x}}_k) > \phi_k$ and

$$M^T = \phi_k + \frac{P(\hat{\mathbf{x}}_k, \phi_k)}{\sum_{i|f_i(\hat{\mathbf{x}}_k) > \phi_k} [f_i(\hat{\mathbf{x}}_k) - \phi_k]}.$$

Step 4: If $M^T \leq B_U$, then set $\phi_{k+1} = M^T$. Otherwise, set $\phi_{k+1} = M^M$. Also set $\phi_0 = \phi_{k+1} - \phi_k$.

Step 5: Set $B_L = M^M$ and $S = P(\hat{\mathbf{x}}_k, \phi_k)$.

Step 6: Set $k = k + 1$.

Step 7: Find $\hat{\mathbf{x}}_k$ to minimize $P(\mathbf{x}, \phi_k)$.

Step 8: If $(B_U - B_L)/B_U \leq \epsilon_1$ or $\phi_0/\phi_k \leq \epsilon_1$, then stop. Otherwise, go to the next step.

Step 9: If $P(\hat{\mathbf{x}}_k, \phi_k) > \epsilon_2$, then go to Step 3. Otherwise, if $S \leq \epsilon_3$, then stop. If none is true, then set $B_U = \phi_k$, $S = 0$, $\phi_k = B_L$, and go to Step 7.

COMMENTS

- In the above algorithm, $\epsilon_1 = \epsilon_2 = \epsilon_3 = 10^{-8}$ can be used.
- Also other selections can be used and are worth trying.
- A very crucial issue to arrive at least at a local optimum is to perform optimization at Steps 2 and 7 effectively. One alternative is to use the Fletcher-Powell algorithm (R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," *Comput. J.*, vol. 6, pp. 163–168, 1963).
- When applying the Fletcher-Powell algorithm the partial derivatives of the objective function with respect to the unknowns are needed.
- Another very crucial issue is to find good starting point values for elements of the adjustable vector \mathbf{x} .
- The effectiveness of the above algorithm lies in the fact that at Steps 2 and 7 it uses the least-mean-square optimization when updating ϕ .
- In this case, the objective function is well-behaved.

OPTIMIZATION EXAMPLE I

- Consider a lowpass filter transfer function

$$H(z) = \frac{1}{2}[A(z) + B(z)],$$

where

$$A(z) = \frac{-r_1 + z^{-1}}{1 - r_1 z^{-1}} \prod_{k=2}^N \frac{(r_k)^2 - 2r_k \cos \theta_k z^{-1} + z^{-2}}{1 - 2r_k \cos \theta_k z^{-1} + (r_k)^2 z^{-2}}$$

and

$$B(z) = \prod_{k=1}^M \frac{(R_k)^2 - 2R_k \cos \Theta_k z^{-1} + z^{-2}}{1 - 2R_k \cos \Theta_k z^{-1} + (R_k)^2 z^{-2}}$$

are stable allpass filters of orders $2N - 1$ and $2M$, respectively.

- The poles of $A(z)$ are located at $z = r_1$ and $z = r_k \exp(\pm j\theta_k)$ for $k = 2, 3, \dots, N$, whereas the poles of $B(z)$ are located at $z = R_k \exp(\pm j\Theta_k)$ for $k = 1, 2, \dots, M$.
- For a lowpass filter design, it is required that the orders $2N - 1$ and $2M$ differ by one.
- The overall filter order is

$$L = 2N - 1 + 2M.$$

- The amplitude response of our filter is expressible as

$$|H(\Phi, e^{j\omega})| = |\cos[(\phi_A(\omega) - \phi_B(\omega))/2]|$$

and the phase response in the passband

$$\arg H(\Phi, e^{j\omega}) = [\phi_A(\omega) + \phi_B(\omega)]/2,$$

where

$$\Phi = [r_1, r_2, \dots, r_N, \theta_2, \theta_3, \dots, \theta_N, \\ R_1, R_2, \dots, R_M, \Theta_1, \Theta_2, \dots, \Theta_M]$$

denotes the adjustable parameter vector, whereas

$$\phi_A(\omega) = \phi^{(1)}(r_1, \omega) + \sum_{k=2}^N \phi^{(2)}(r_k, \theta_k, \omega)$$

and

$$\phi_B(\omega) = \sum_{k=1}^M \phi^{(2)}(R_k, \Theta_k, \omega).$$

- Here,

$$\phi^{(1)}(r, \omega) = -\omega - 2 \tan^{-1} \left[\frac{r \sin(\omega)}{1 - r \cos(\omega)} \right]$$

and

$$\phi^{(2)}(r, \theta, \omega) = -2\omega - 2 \tan^{-1} \left[\frac{r \sin(\omega - \theta)}{1 - r \cos(\omega - \theta)} \right] \\ - 2 \tan^{-1} \left[\frac{r \sin(\omega + \theta)}{1 - r \cos(\omega + \theta)} \right].$$

OPTIMIZATION PROBLEMS

- We state the following two approximation problems:
- *Optimization Problem I.* Given ω_p , ω_s , δ_p , δ_s as well as the filter order L , find Φ and ψ , the slope of a linear phase response, to minimize

$$\Delta = \max_{0 \leq \omega \leq \omega_p} |\arg H(\Phi, e^{j\omega}) - \psi\omega| \quad (1a)$$

subject to

$$1 - \delta_p \leq |H(\Phi, e^{j\omega})| \leq 1 \quad \text{for } \omega \in [0, \omega_p] \quad (1b)$$

and

$$|H(\Phi, e^{j\omega})| \leq \delta_s \quad \text{for } \omega \in [\omega_s, \pi]. \quad (1c)$$

- *Optimization Problem II.* Given ω_p , ω_s , δ_p , δ_s as well as the filter order L , find Φ and ψ to minimize Δ as given by Eq. (1a) subject to the conditions of Eqs. (1b) and (1c) and

$$\frac{d|H(\Phi, e^{j\omega})|}{d\omega} \leq 0 \quad \text{for } \omega \in (\omega_p, \omega_s). \quad (1d)$$

- The main difference between the above problems is that for Problem II the amplitude response is forced to be monotonously decreasing in the transition band.

MODIFIED ALGORITHM FOR PROBLEM I

- We discretize the passband and the stopband regions into the frequency points $\omega_i \in [0, \omega_p]$, $i = 1, 2, \dots, N_p$ and $\omega_i \in [\omega_s, \pi]$, $i = N_p+1, \dots, N_p+N_s$.
- The resulting discrete minimax problem that can be solved using the Dutta-Vidyasagar algorithm is to find Φ and ψ to minimize

$$\epsilon = \max_{1 \leq i \leq N_p} \{f_i(\Phi, \psi)\}$$

subject to

$$g_i(\Phi, \psi) \leq 0, \quad i = 1, 2, \dots, N_p + N_s,$$

where

$$f_i(\Phi, \psi) = |\arg H(\Phi, e^{j\omega_i}) - \psi\omega_i|, \quad i = 1, 2, \dots, N_p$$

and

$$g_i(\Phi, \psi) = \begin{cases} ||H(\Phi, e^{j\omega_i})| \\ -(1 - \frac{\delta_p}{2})| - \frac{\delta_p}{2}, & i = 1, \dots, N_p \\ |H(\Phi, e^{j\omega_i})| - \delta_s, & i = N_p + 1, \dots, \\ & N_p + N_s. \end{cases}$$

MODIFIED ALGORITHM FOR PROBLEM II

- We discretize the passband, the transition band, and the stopband regions into the frequency points $\omega_i \in [0, \omega_p]$, $i = 1, 2, \dots, N_p$, $\omega_i \in (\omega_p, \omega_s)$, $i = N_p + 1, \dots, N_p + N_t$, and $\omega_i \in [\omega_s, \pi]$, $i = N_p + N_t + 1, \dots, N_p + N_t + N_s$.
- The resulting discrete minimax problem that can be solved using the Dutta-Vidyasagar algorithm is to find Φ and ψ to minimize

$$\epsilon = \max_{1 \leq i \leq N_p} \{f_i(\Phi, \psi)\}$$

subject to

$$g_i(\Phi, \psi) \leq 0, \quad i = 1, 2, \dots, N_p + N_t + N_s,$$

where

$$f_i(\Phi, \psi) = |\arg H(\Phi, e^{j\omega_i}) - \psi\omega_i|, \quad i = 1, 2, \dots, N_p$$

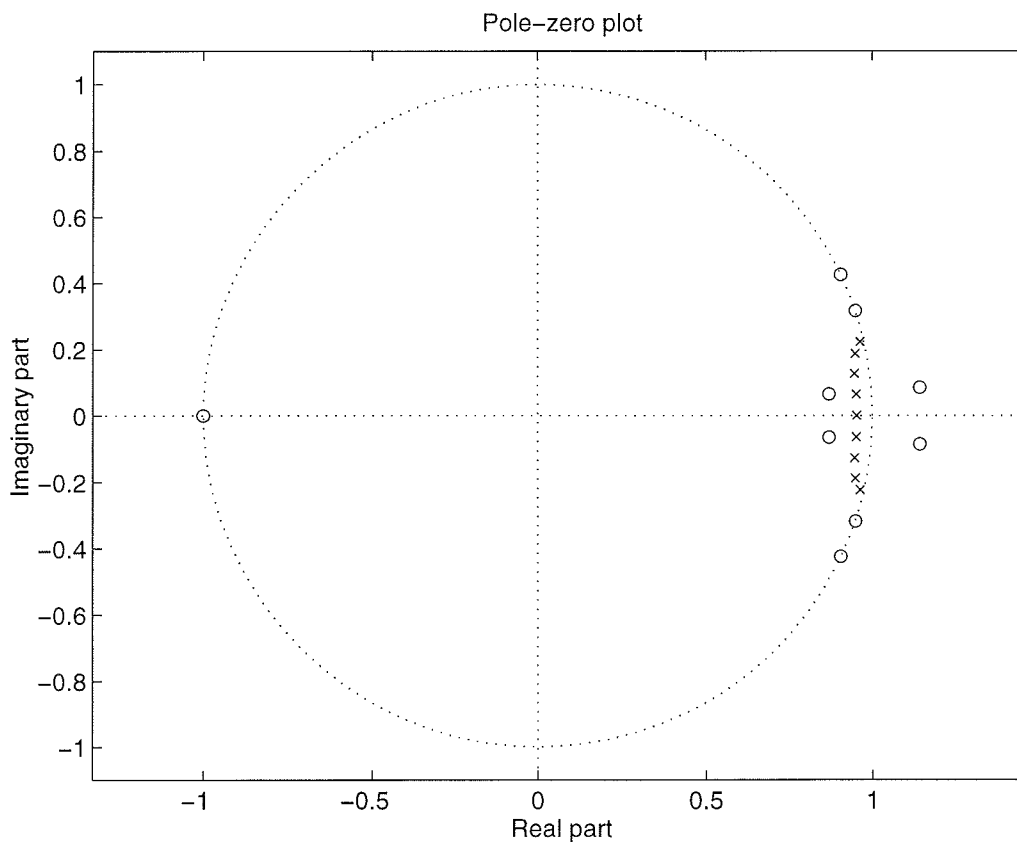
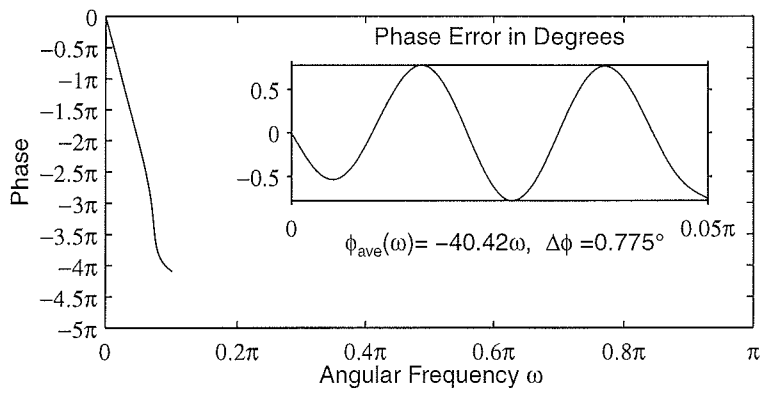
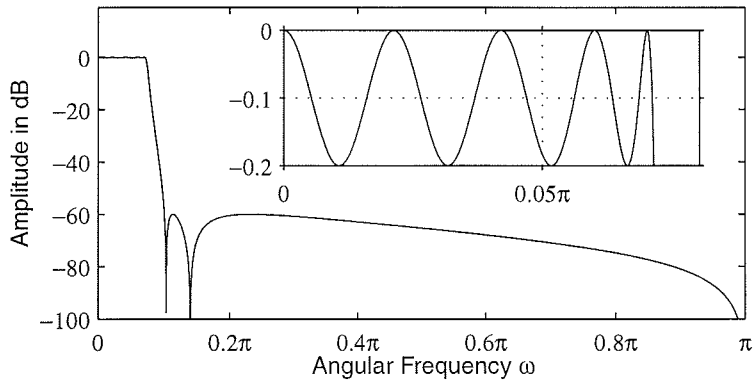
and

$$g_i(\Phi, \psi) = \begin{cases} |H(\Phi, e^{j\omega_i})| \\ -(1 - \frac{\delta_p}{2})| - \frac{\delta_p}{2}, & i = 1, \dots, N_p \\ |H(\Phi, e^{j\omega_i})| \\ -|H(\Phi, e^{j\omega_{i-1}})|, & i = N_p + 1, \dots, N_p + N_t \\ |H(\Phi, e^{j\omega_i})| - \delta_s, & i = N_p + N_t + 1, \dots, \\ & N_p + N_t + N_s. \end{cases}$$

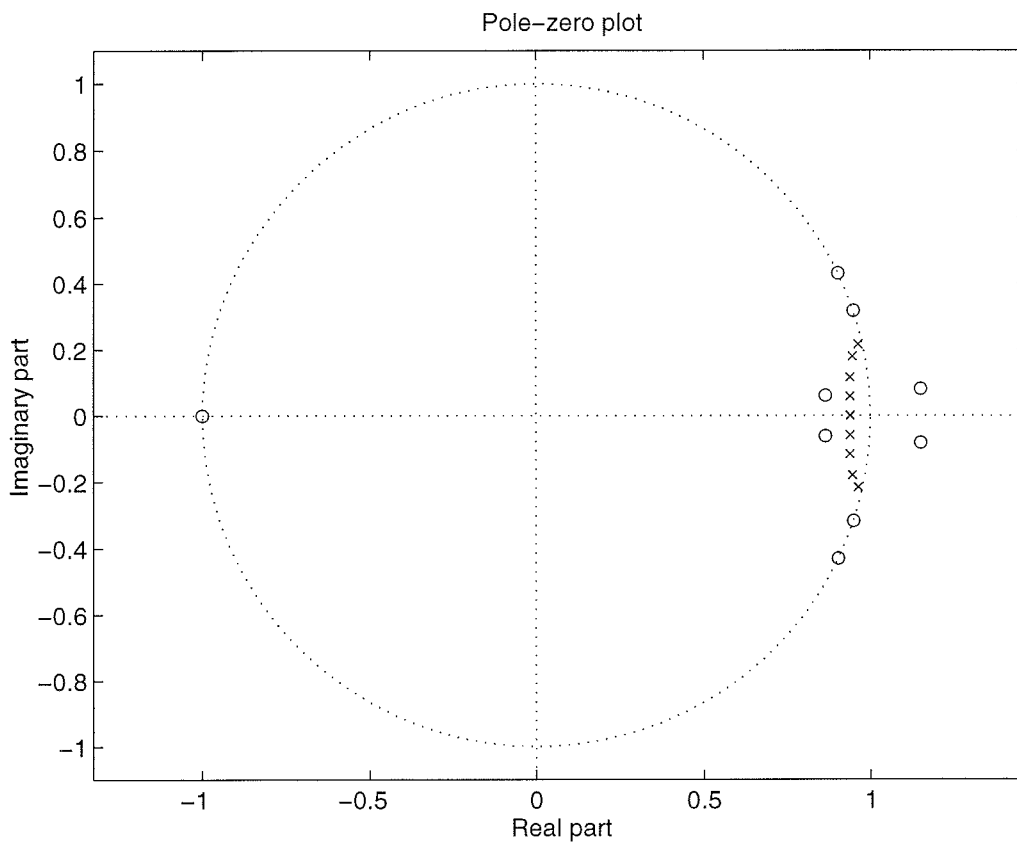
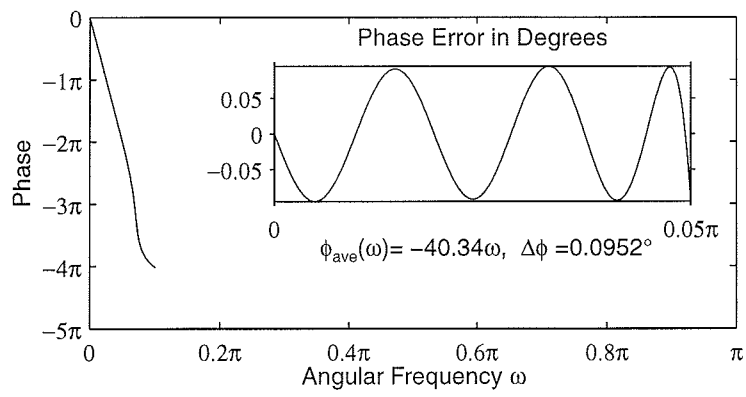
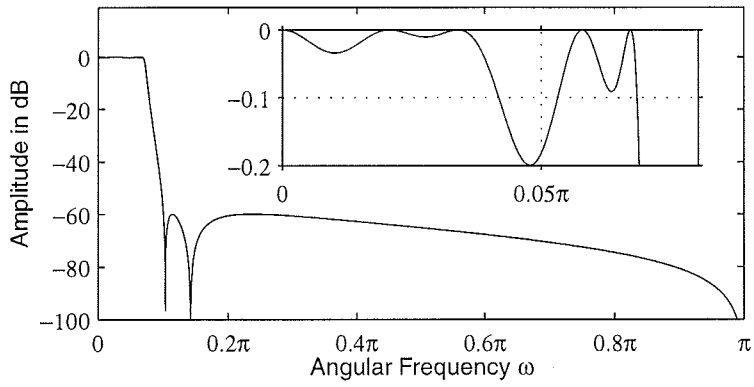
EXAMPLE SPECIFICATIONS

- $\omega_p = 0.05\pi$, $\omega_s = 0.1\pi$, $\delta_p = 0.0228$ (0.2 dB pass-band variation), and $\delta_s = 10^{-3}$ (60 dB stopband attenuation).
- The minimum order of the elliptic filter meeting these criteria is 5.
- Satisfactory phase performances are achieved by increasing the filter order from five to nine.
- In the following, there are three transparencies illustrating the characteristics of the optimized filters.
- The first one is the initial filter for Problem I. This has been generated using a simple algorithm described in the enclosed article.
- The second and third transparencies are for the optimized filters for Problems I and II, respectively.
- As seen from these three transparencies, the maximum phase errors for these three filters are 0.775, 0.0952, and 0.419 degrees, respectively.

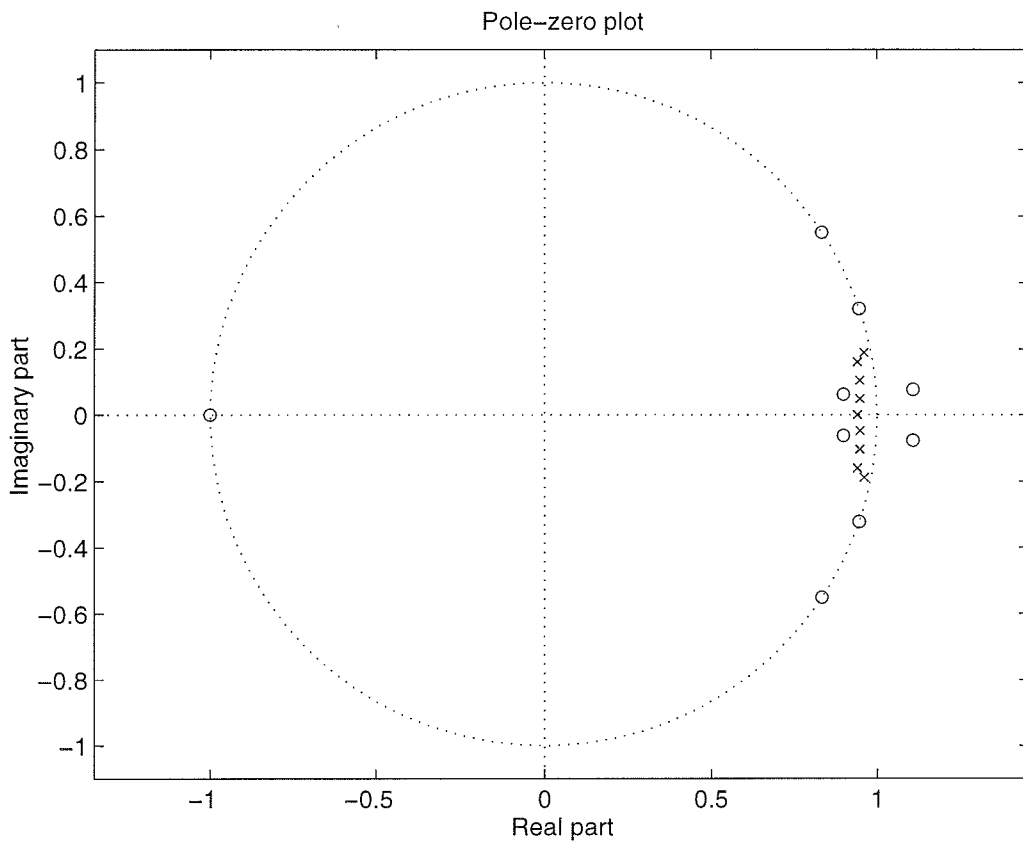
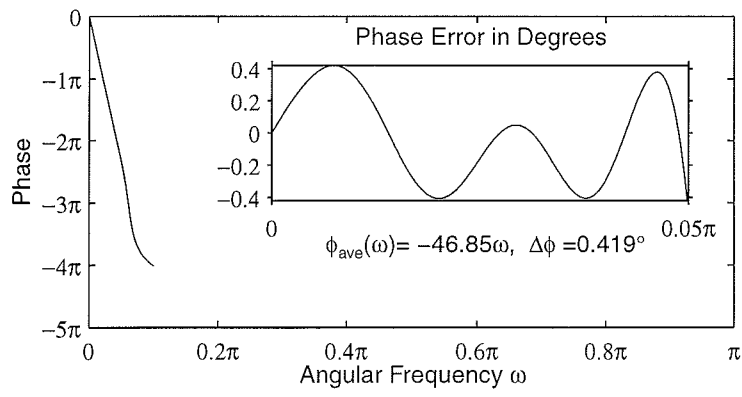
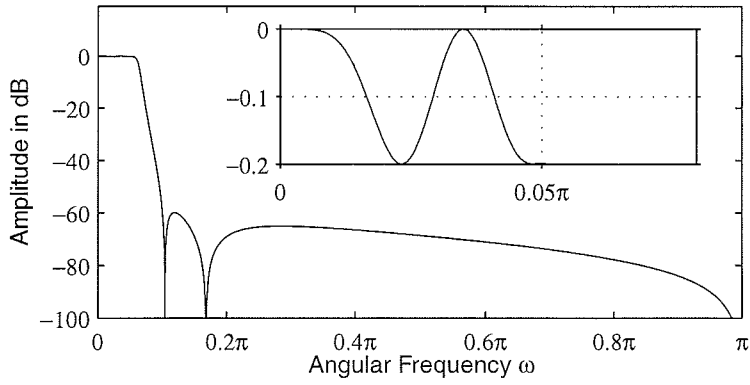
INITIAL FILTER FOR PROBLEM I



OPTIMIZED FILTER FOR PROBLEM I



OPTIMIZED FILTER FOR PROBLEM II



OPTIMIZATION EXAMPLE II

- Consider a minimum-phase FIR filter

$$E_0(\Phi, z) = \sum_{n=0}^M e_0(n) z^{-n},$$

where M is odd and

$$\Phi = [e_0(0), e_0(1), \dots, e_0(M)]$$

denotes the adjustable parameter vector.

- The corresponding maximum-phase filter is then

$$E_1(\Phi, z) = \sum_{n=0}^M e_0(M - n) z^{-n}.$$

- The common amplitude response of these filters is given by

$$|E_0(\Phi, e^{j\omega})| = \sqrt{S(\Phi, \omega)},$$

where

$$S(\Phi, \omega) = A(\Phi, \omega)^2 + B(\Phi, \omega)^2$$

with

$$A(\Phi, \omega) = \sum_{n=0}^M e_0(n) \cos(n\omega)$$

and

$$B(\Phi, \omega) = - \sum_{n=0}^M e_0(n) \sin(n\omega).$$

- We state the following optimization problem: Given M and $\omega_s > \pi/2$, find the coefficients $e_0(n)$ to minimize

$$\int_{\omega_s}^{\pi} |E_0(\Phi, e^{j\omega})|^2 d\omega = \int_{\omega_s}^{\pi} S(\Phi, \omega) \omega$$

subject to the condition that the impulse-response coefficients of the linear-phase FIR filter transfer function

$$T(\Phi, z) = E_0(\Phi, z)E_1(\Phi, z) = \sum_{n=0}^{2M} t(n)z^{-n}$$

satisfies

$$t(M) = 1/2$$

$$t(M + 2r) = 0 \quad \text{for } r = \pm 1, \pm 2, \dots, (M - 1)/2.$$

- It can be shown that the above constraint can be met by requiring that at $(M+1)/2$ distinct points in the region $[0, \pi/2]$ it is satisfied

$$S(\Phi, \omega) + S(\Phi, \pi - \omega) = 1.$$

- Also, the above integral can be minimized by sampling $S(\Phi, \omega)$ along a dense grid of frequencies on $[\omega_s, \pi]$ and by finding Φ to minimize the sum of these samples.

MODIFIED ALGORITHM FOR OUR PROBLEM

- We discretize the stopband region into the frequency points $\omega_i \in [\omega_s, \pi]$, $i = 1, 2, \dots, K$ and the region $[0, \pi/2]$ into the points $\hat{\omega}_l \in [0, \pi/2]$, $l = 1, 2, \dots, L = (M + 1)/2$.
- The resulting discrete minimax problem is to find Φ

$$\epsilon = f_1(\Phi)$$

subject to

$$h_l(\Phi) = 0, \quad l = 1, 2, \dots, L,$$

where

$$f_1(\Phi) = \sum_{i=1}^K S(\Phi, \omega_i)$$

and

$$h_l(\Phi) = S(\Phi, \hat{\omega}_l) + S(\Phi, \pi - \hat{\omega}_l) - 1, \quad l = 1, 2, \dots, L.$$

USE OF OUR RESULT

- By selecting

$$H_0(z) = E_0(\Phi, z),$$

$$H_1(z) = E_1(\Phi, -z),$$

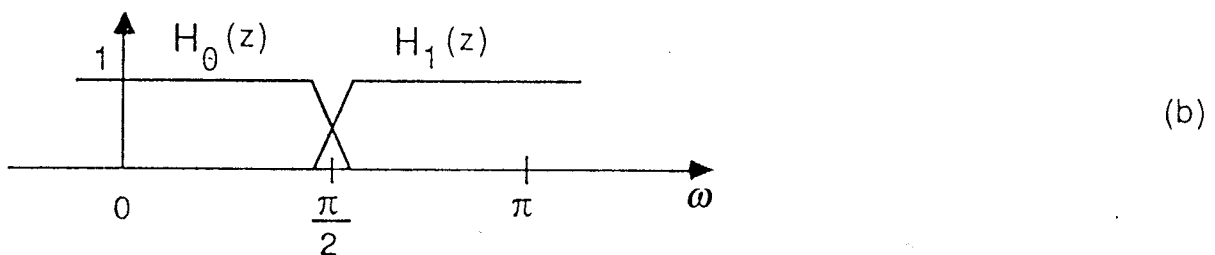
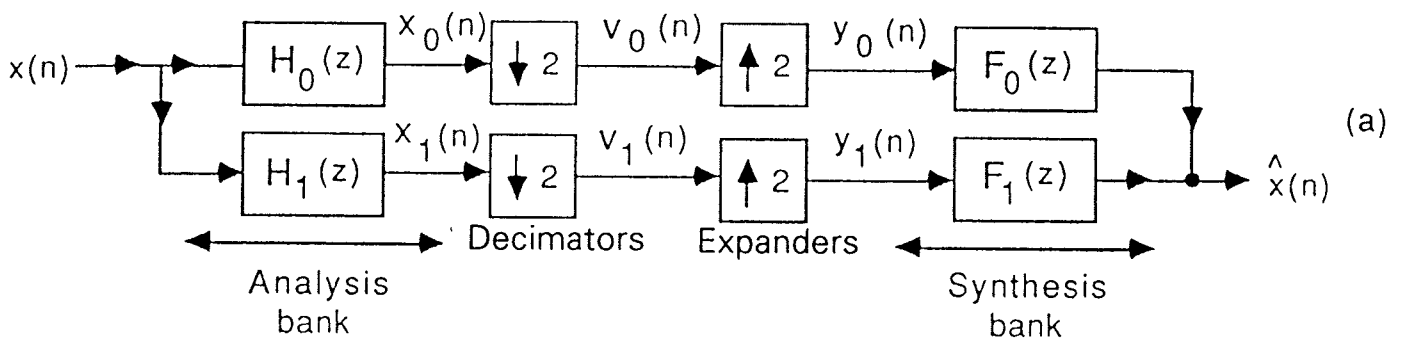
$$F_0(z) = 2H_1(-z),$$

and

$$F_1(z) = -2H_0(-z),$$

we can generate a perfect-reconstruction QMF filter bank shown below.

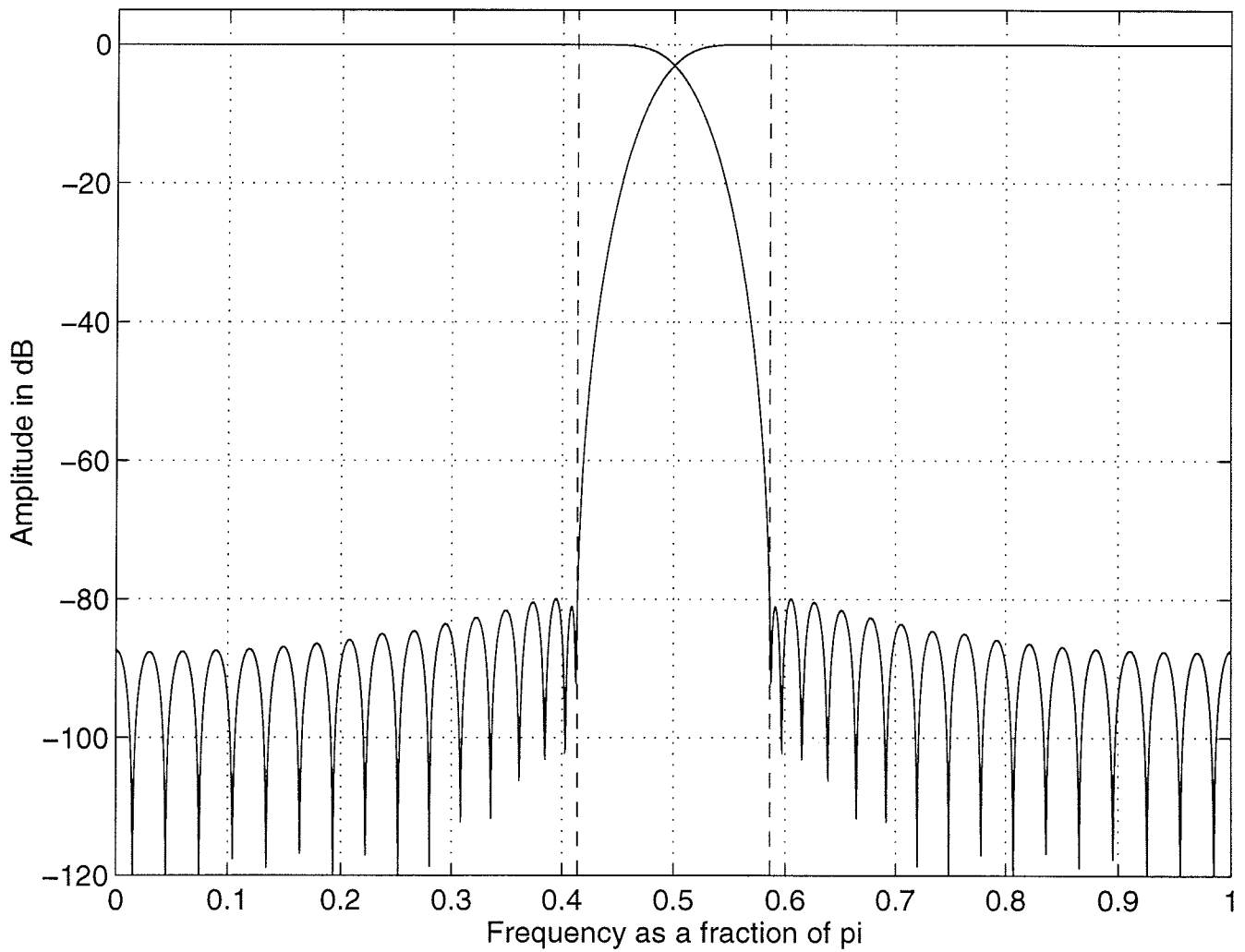
- This means that $\hat{x}(n) = x(n - M)$, that is, the output is a delayed version of the input.

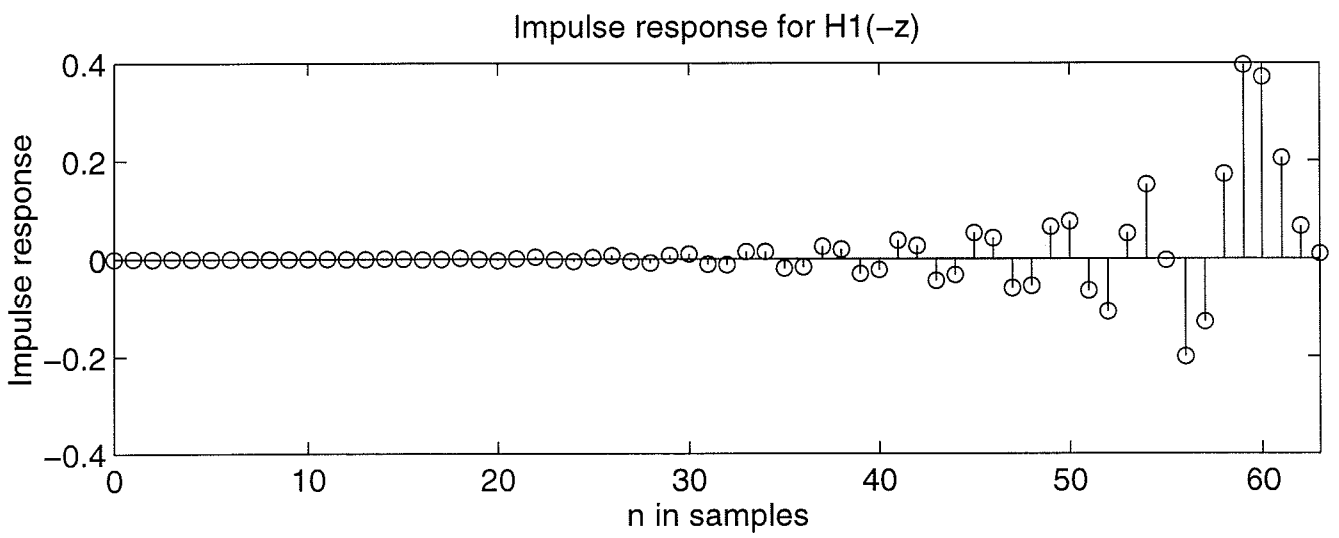
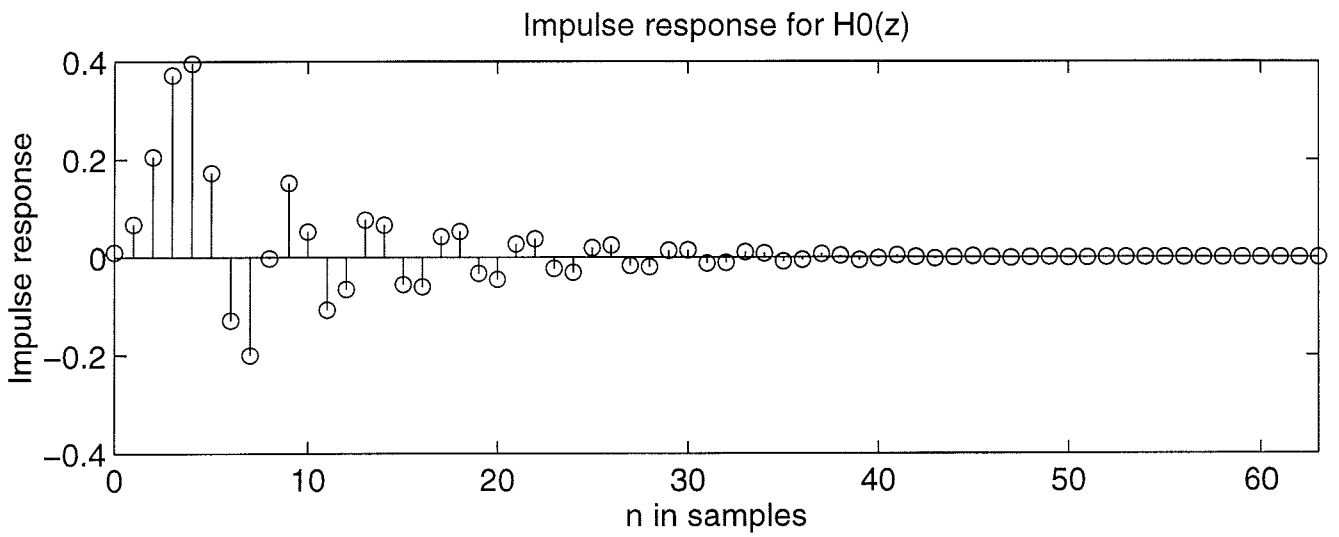


EXAMPLE SPECIFICATIONS

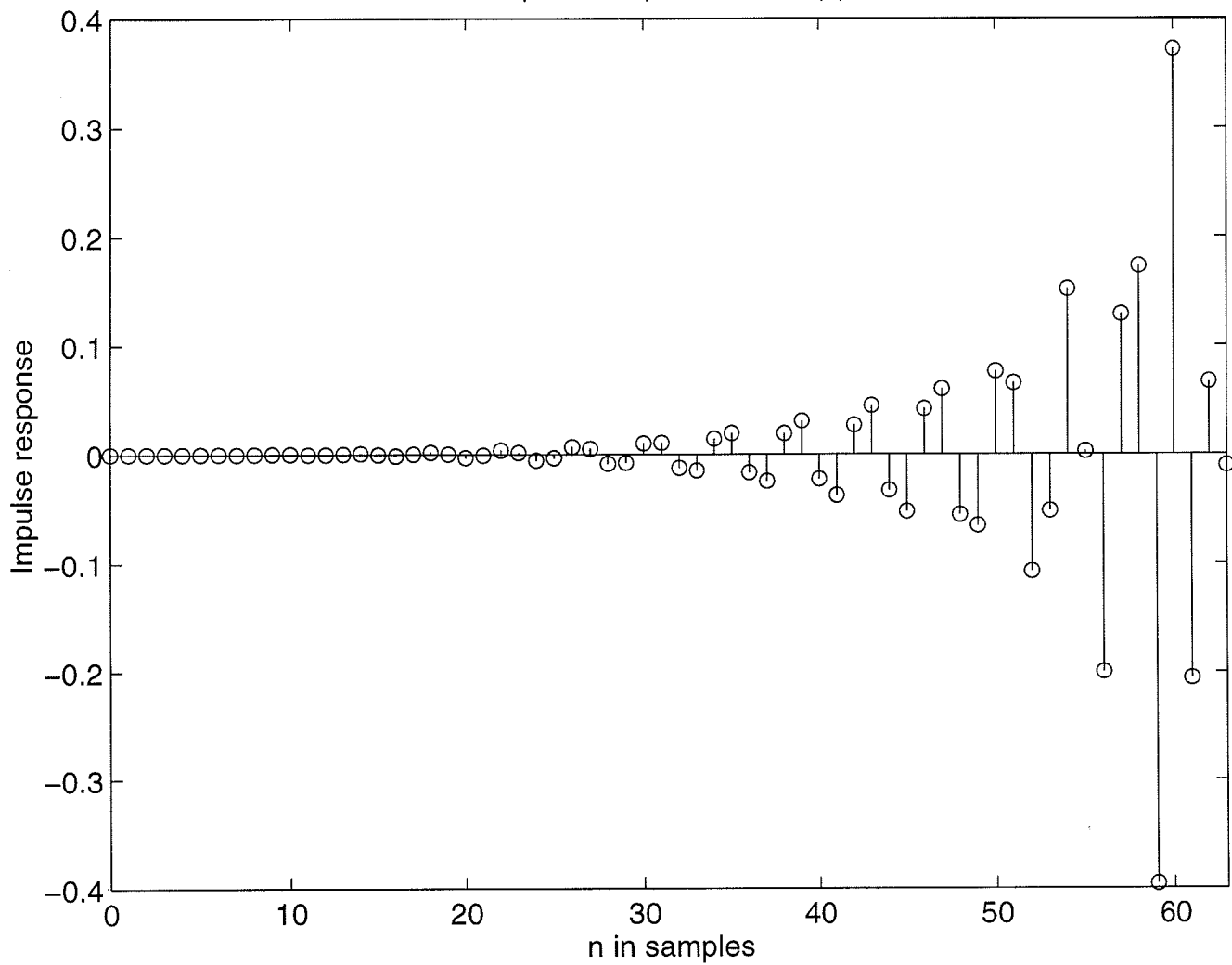
- $M = 63$ and $\omega_s = 0.586\pi$.
- The following nine transparencies illustrate the performance of the resulting filter bank.

$H_0(z)$ is the lowpass filter and $H_1(z)$ is the highpass filter

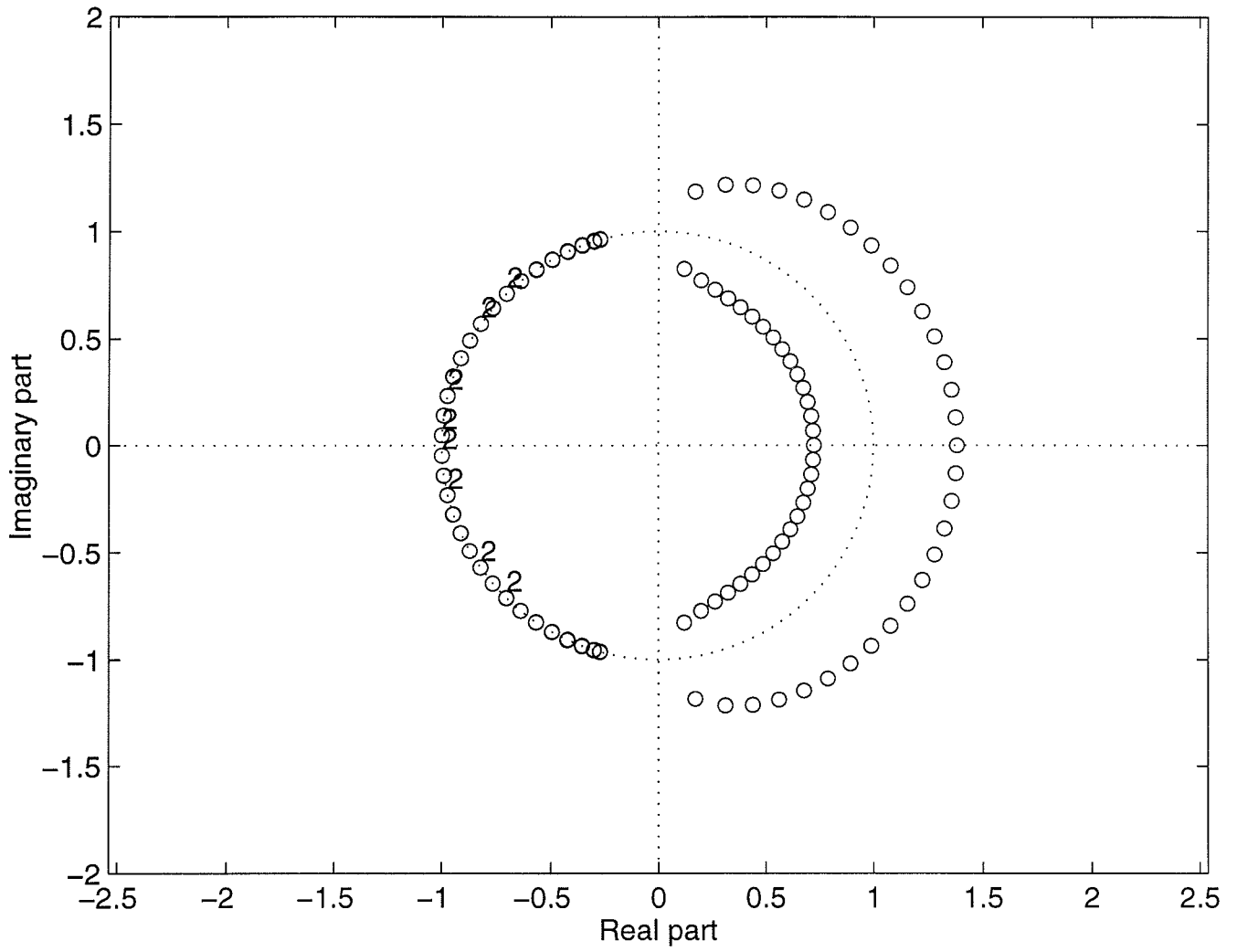




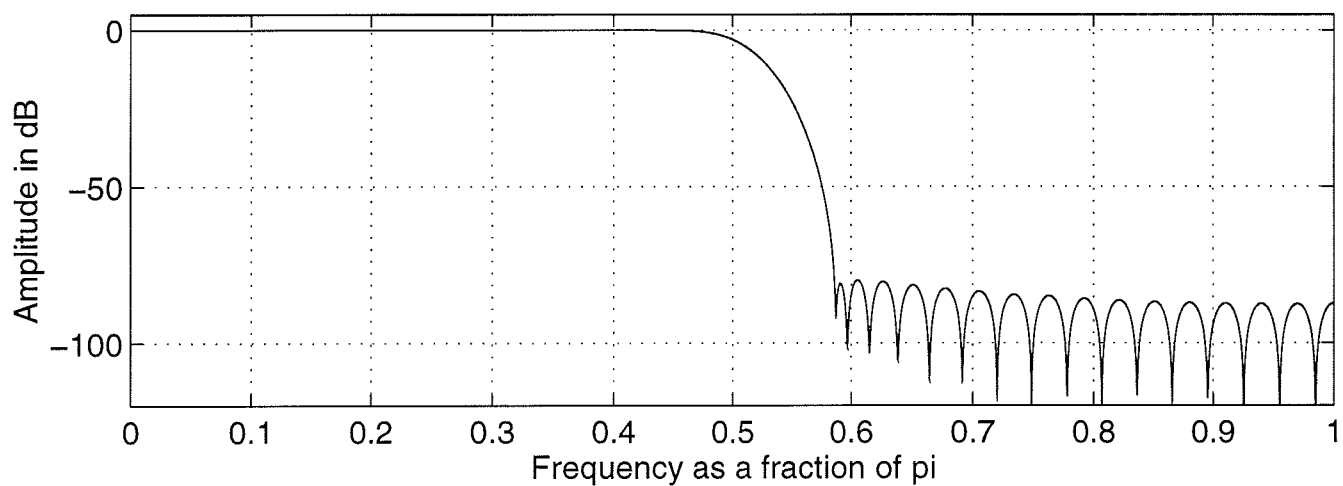
Impulse response for $H_1(z)$



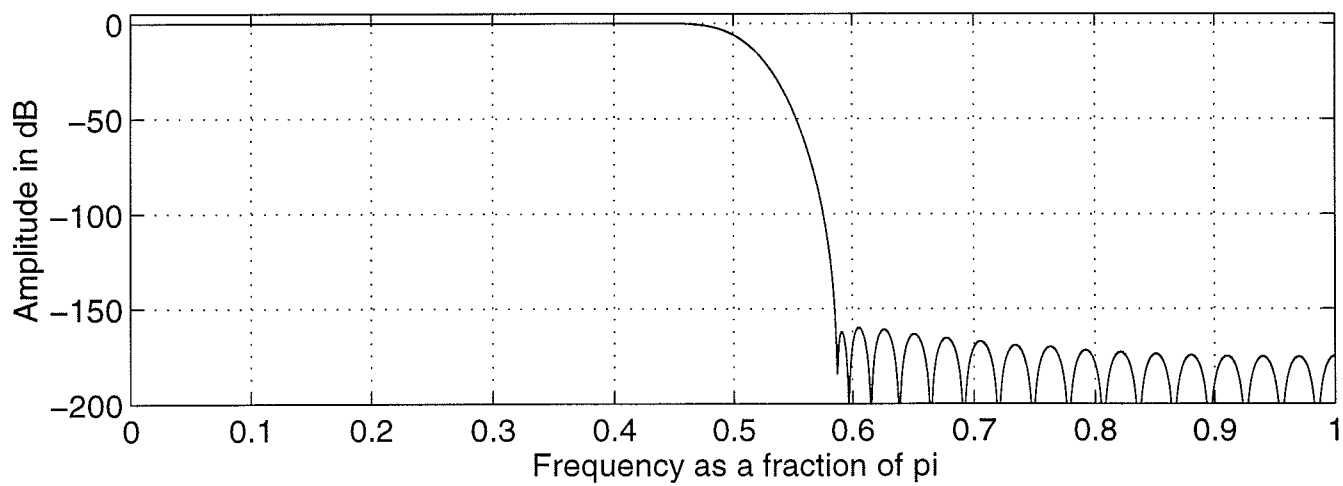
Zeros for $H_0(z)H_1(-z)$



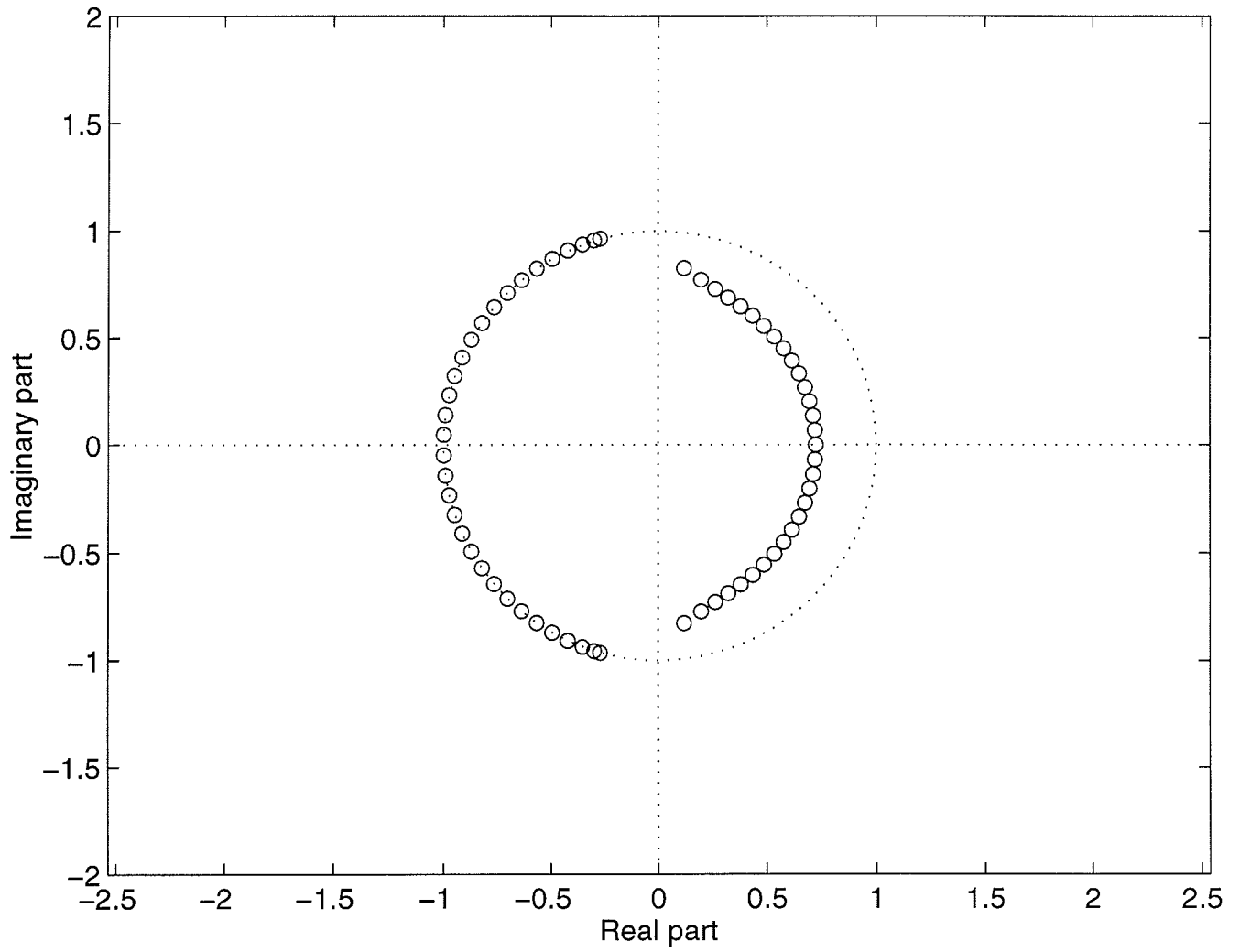
Amplitude responses for $H_0(z)$ and $H_1(-z)$



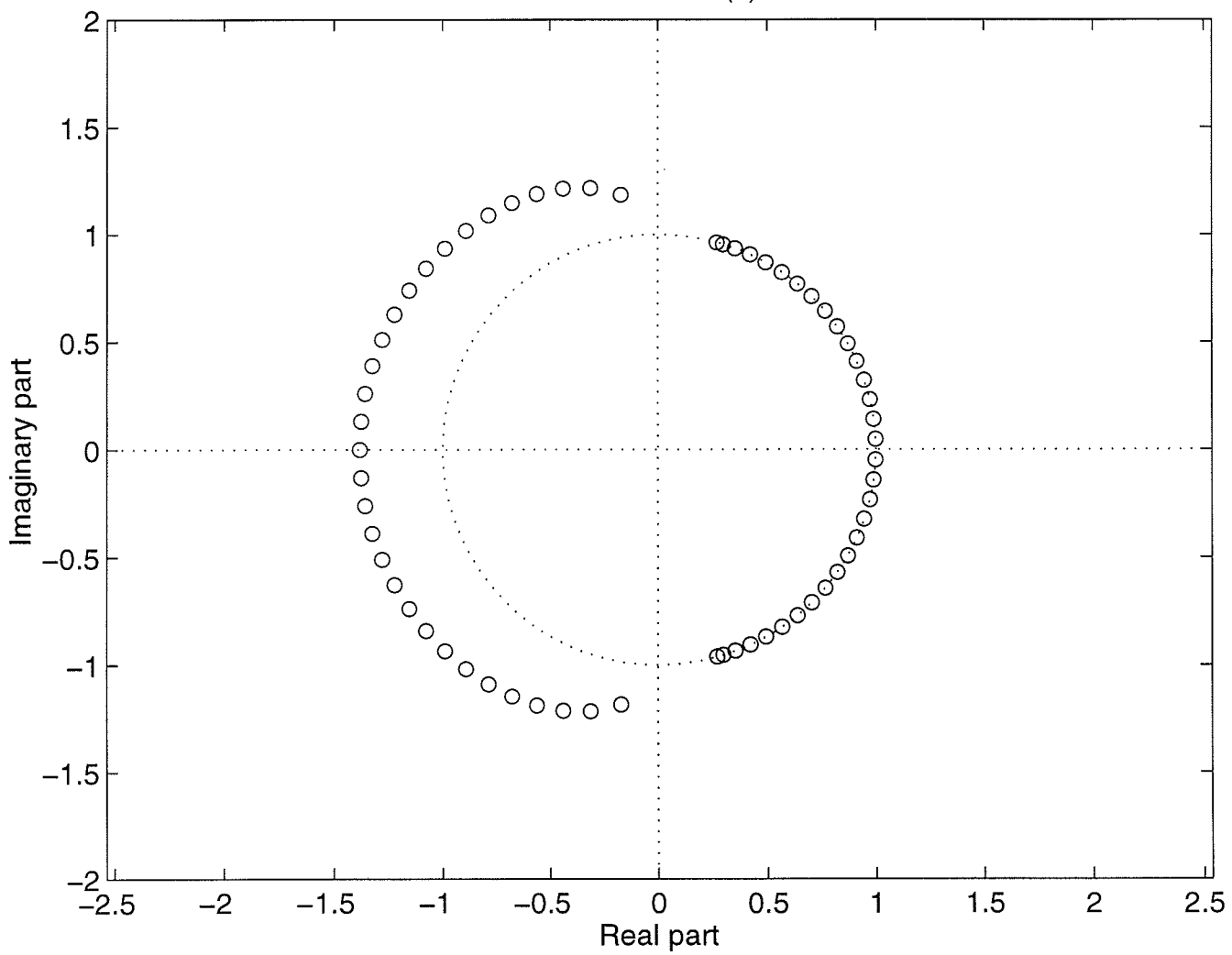
Amplitude response for $H_0(z)H_1(-z)$



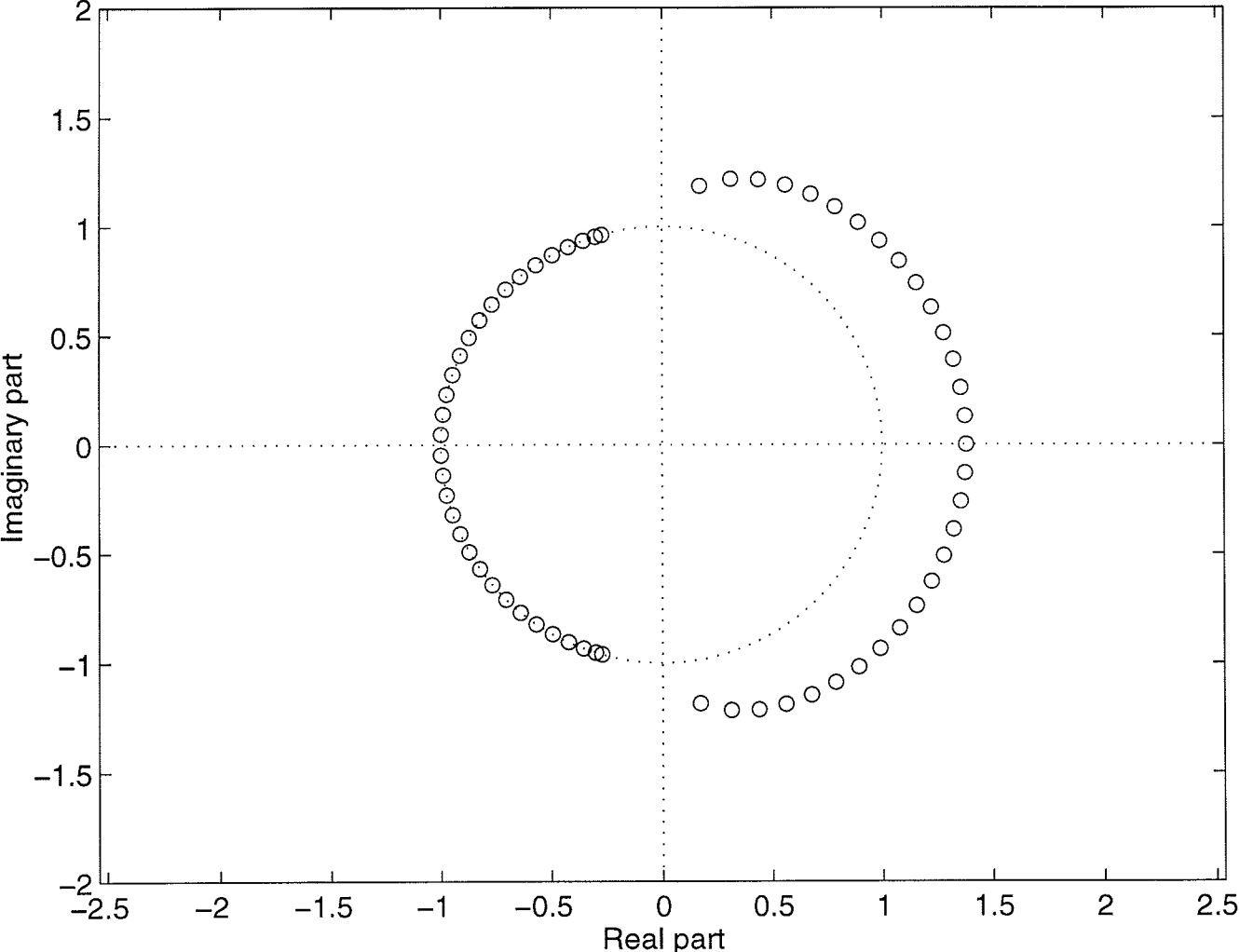
Zeros for $H_0(z)$



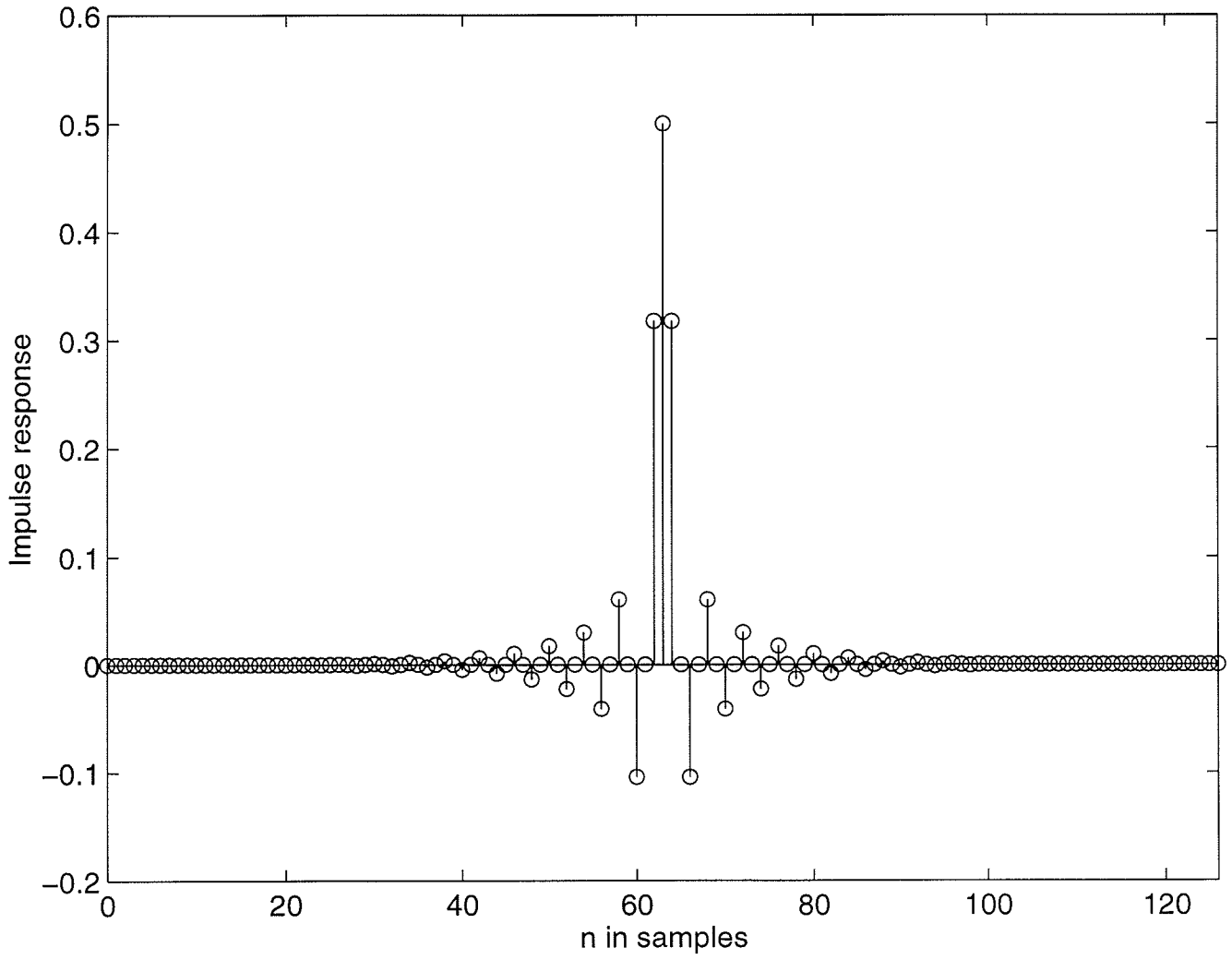
Zeros for H1(z)



Zeros for $H1(-z)$



Impulse response for $H_0(z)H_1(-z)$



DESIGN OF DIGITAL FILTERS AND FILTER BANKS BY OPTIMIZATION: APPLICATIONS

Tapio Saramäki and Juha Yli-Kaakinen

Signal Processing Laboratory,
Tampere University of Technology,
Finland

e-mail: ts@cs.tut.fi



Tampere University of Technology
Signal Processing Laboratory

ABSTRACT

This paper emphasizes the usefulness and the flexibility of optimization for finding optimized digital signal processing algorithms for various constrained and unconstrained optimization problems. This is illustrated by optimizing algorithms in six different practical applications:

- Optimizing nearly perfect-reconstruction filter banks subject to the given allowable errors,
- minimizing the phase distortion of recursive filters subject to the given amplitude criteria,
- optimizing the amplitude response of pipelined recursive filters,
- optimizing the modified Farrow structure with an adjustable fractional delay,
- finding the optimum discrete values for coefficient representations for various classes of lattice wave digital filters, and
- finding the multiplierless coefficient representations for the linear-phase finite impulse response filters.

WHY THERE IS A NEED TO USE OPTIMIZATION?

Among others, there exist following three reasons:

1) Thanks to dramatic advances in VLSI circuit technology and signal processors, more complicated DSP algorithms can be implemented faster and faster.

- In order to generate effective DSP products, old algorithms have to be reoptimized or new ones should be generated subject to the implementation constraints.

2) All the subalgorithms in the overall DSP product should be of the same quality:

- In the case of lossy coding, nearly perfect-reconstruction filter banks are more beneficial: lower overall delay and shorter filters.

3) There are various problems where one response is desired to optimized subject to the given criteria for other responses:

- A typical example is to design recursive filters such that the phase is made as linear as possible subject to the given amplitude criteria.

TWO-STEP PROCEDURE

It has turned out that the following procedure is very efficient:

1) Find in a systematic simple manner a suboptimum solution.

2) Improve this solution using a general-purpose nonlinear optimization procedure:

- Dutta-Vidyasagar algorithm
- Sequential quadratic programming

Desired Form: Find the adjustable parameters included in the vector Φ to minimize

$$\rho(\Phi) = \max_{1 \leq i \leq I} f_i(\Phi) \quad (1)$$

subject to constraints

$$g_l(\Phi) \leq 0 \quad \text{for } l = 1, 2, \dots, L \quad (2)$$

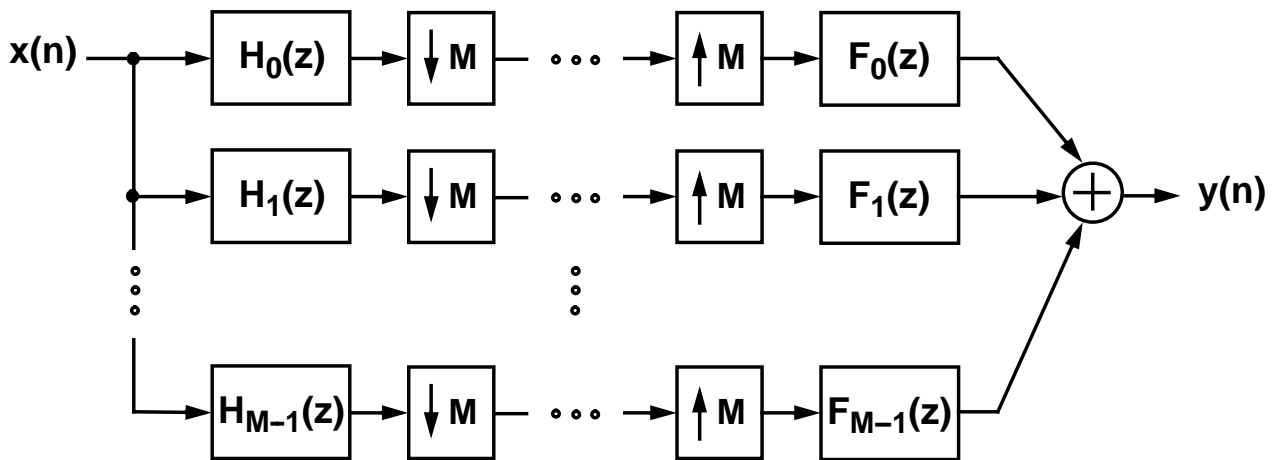
and

$$h_m(\Phi) = 0 \quad \text{for } m = 1, 2, \dots, M. \quad (3)$$

COMMENTS

- The proposed two-step procedure is very efficient when a good start-up solution being rather close to the optimum solution can found.
- For each problem under consideration the way of generating this initial solution is very different.
- A good understanding of the problem at hand is needed.
- If a good enough start-up solution cannot be found or there are several local optima, then simulated annealing or genetic algorithms can be used.

NEARLY PERFECT-RECONSTRUCTION COSINE-MODULATED FILTER BANKS



Linear-phase prototype filter:

$$H_p(z) = \sum_{n=0}^N h_p(n) z^{-n}, \quad (4)$$

Filters in the bank:

$$h_k(n) = 2h_p(n) \cos \left[(2k + 1) \frac{\pi}{2M} \left(n - \frac{N}{2} \right) + (-1)^k \frac{\pi}{4} \right] \quad (5)$$

$$f_k(n) = 2h_p(n) \cos \left[(2k + 1) \frac{\pi}{2M} \left(n - \frac{N}{2} \right) - (-1)^k \frac{\pi}{4} \right]. \quad (6)$$

INPUT-OUTPUT RELATION

$$Y(z) = T_0(z)X(z) + \sum_{l=1}^{M-1} T_l(z)X(ze^{-j2\pi l/M}), \quad (7a)$$

where

$$T_0(z) = \frac{1}{M} \sum_{k=0}^{M-1} F_k(z)H_k(z) \quad (7b)$$

and for $l = 1, 2, \dots, M - 1$

$$T_l(z) = \frac{1}{M} \sum_{k=0}^{M-1} F_k(z)H_k(ze^{-j2\pi l/M}). \quad (7c)$$

Here, $T_0(z)$ is the reconstruction transfer function and the remaining ones are aliased transfer functions. It is desired that $T_0(z) = z^{-N}$ and the remaining transfer functions are zero.

STATEMENT OF THE PROBLEMS

Problem I: Given ρ , M , and N , find the coefficients of $H_p(z)$ to minimize

$$E_2 = \int_{\omega_s}^{\pi} |H_p(e^{j\omega})|^2 d\omega, \quad (8a)$$

where

$$\omega_s = (1 + \rho)\pi/(2M) \quad (8b)$$

subject to

$$1 - \delta_1 \leq |T_0(e^{j\omega})| \leq 1 + \delta_1 \quad \text{for } \omega \in [0, \pi] \quad (8c)$$

and for $l = 1, 2, \dots, M - 1$

$$|T_l(e^{j\omega})| \leq \delta_2 \quad \text{for } \omega \in [0, \pi]. \quad (8d)$$

Problem II: Given ρ , M , and N , find the coefficients of $H_p(z)$ to minimize

$$E_\infty = \max_{\omega \in [\omega_s, \pi]} |H_p(e^{j\omega})| \quad (9)$$

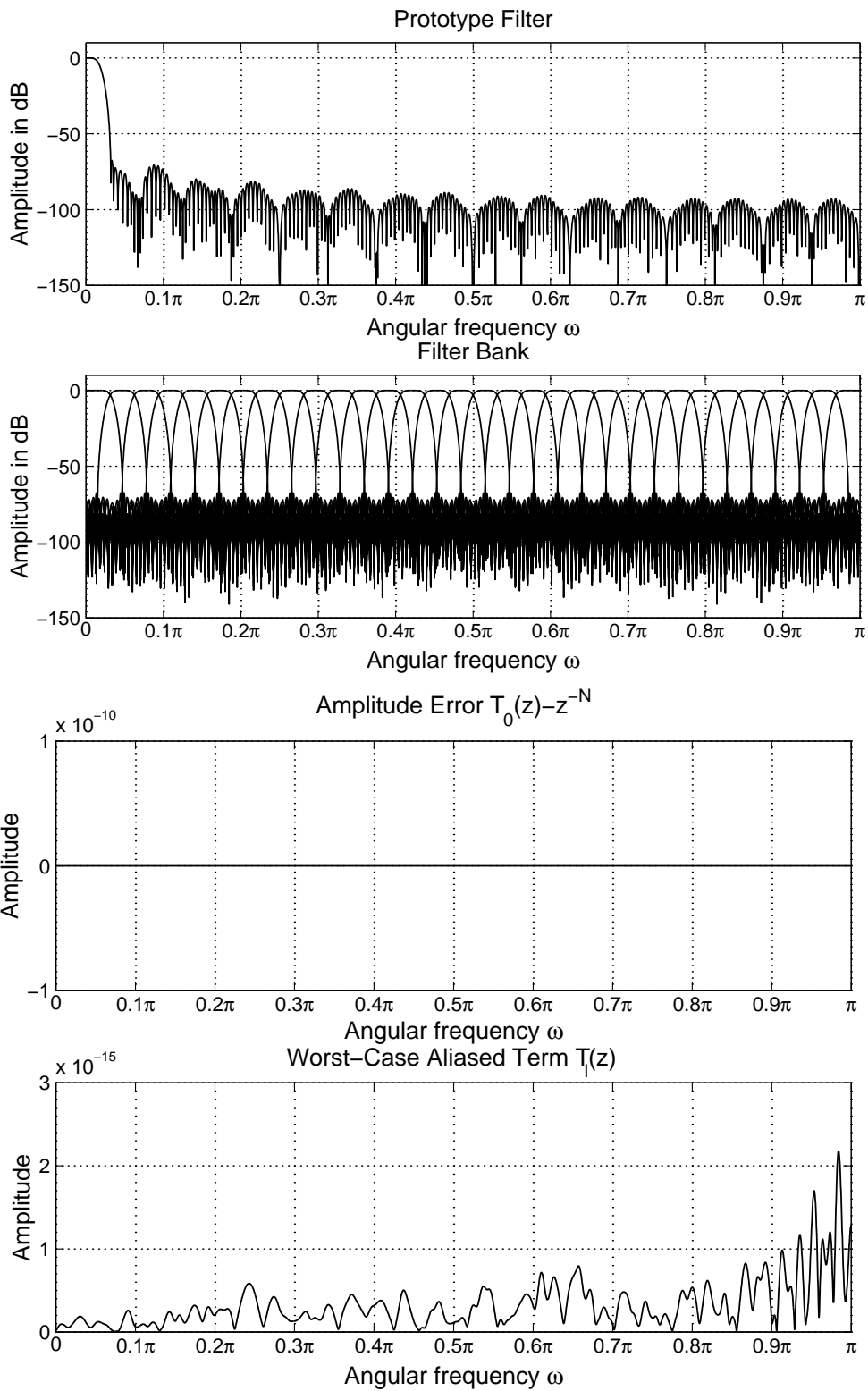
subject to the conditions of Eqs.(8c) and (8d).

COMPARISONS BETWEEN FILTER BANKS WITH $M = 32$ and $\rho = 1$

Boldface numbers indicate that these parameters have been fixed in the optimization.

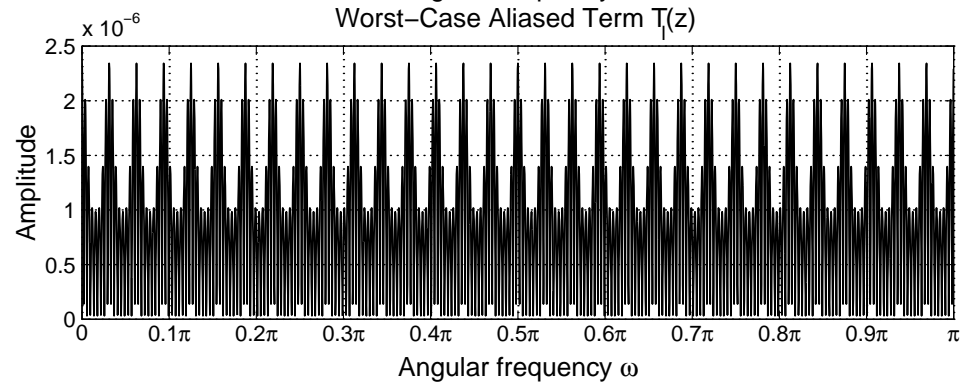
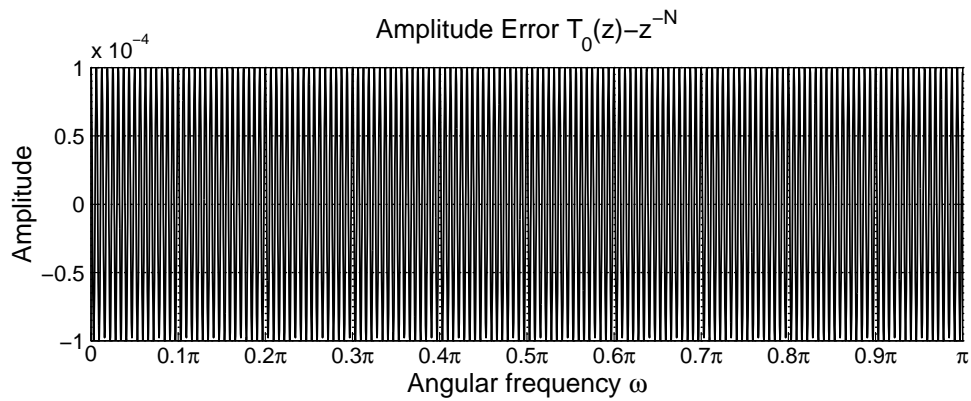
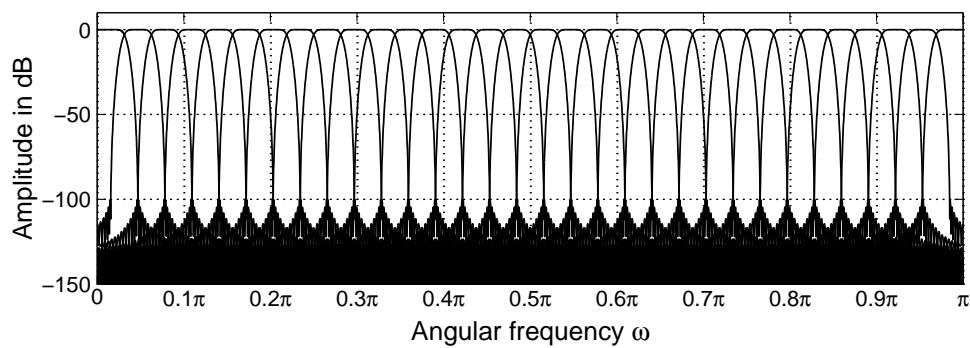
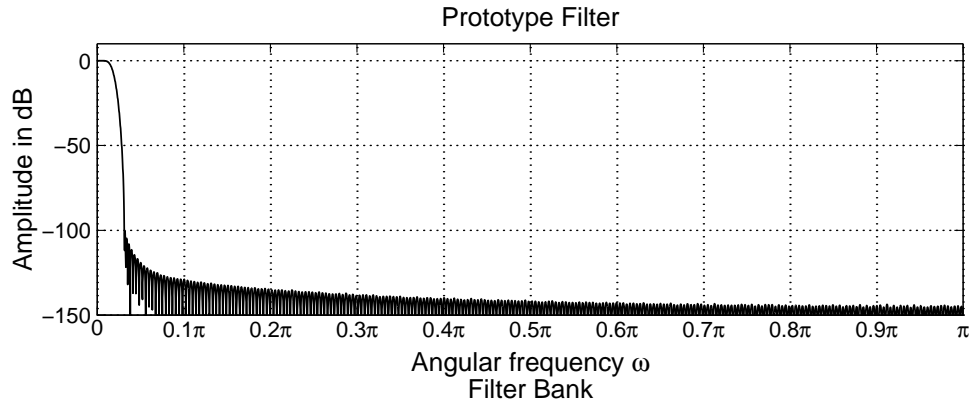
Criterion	K	N	δ_1	δ_2	E_∞	E_2
Least Squared	8	511	0	0 –∞ dB	$1.2 \cdot 10^{-3}$ –58 dB	$7.4 \cdot 10^{-9}$
Minimax	8	511	0	0 –∞ dB	$2.3 \cdot 10^{-4}$ –73 dB	$7.5 \cdot 10^{-8}$
Least Squared	8	511	10^{-4}	$2.3 \cdot 10^{-6}$ –113 dB	$1.0 \cdot 10^{-5}$ –100 dB	$5.6 \cdot 10^{-13}$
Minimax	8	511	10^{-4}	$1.1 \cdot 10^{-5}$ –99 dB	$5.1 \cdot 10^{-6}$ –106 dB	$3.8 \cdot 10^{-11}$
Least Squared	8	511	0	$9.1 \cdot 10^{-5}$ –81 dB	$4.5 \cdot 10^{-4}$ –67 dB	$5.4 \cdot 10^{-10}$
Least Squared	8	511	10^{-2}	$5.3 \cdot 10^{-7}$ –126 dB	$2.4 \cdot 10^{-6}$ –112 dB	$4.5 \cdot 10^{-14}$
Least Squared	6	383	10^{-3}	0.00001 –100 dB	$1.7 \cdot 10^{-4}$ –75 dB	$8.8 \cdot 10^{-10}$
Least Squared	5	319	10^{-2}	0.0001 –80 dB	$8.4 \cdot 10^{-4}$ –62 dB	$2.7 \cdot 10^{-9}$

PERFECT-RECONSTRUCTION FILTER BANK with $N = 511$



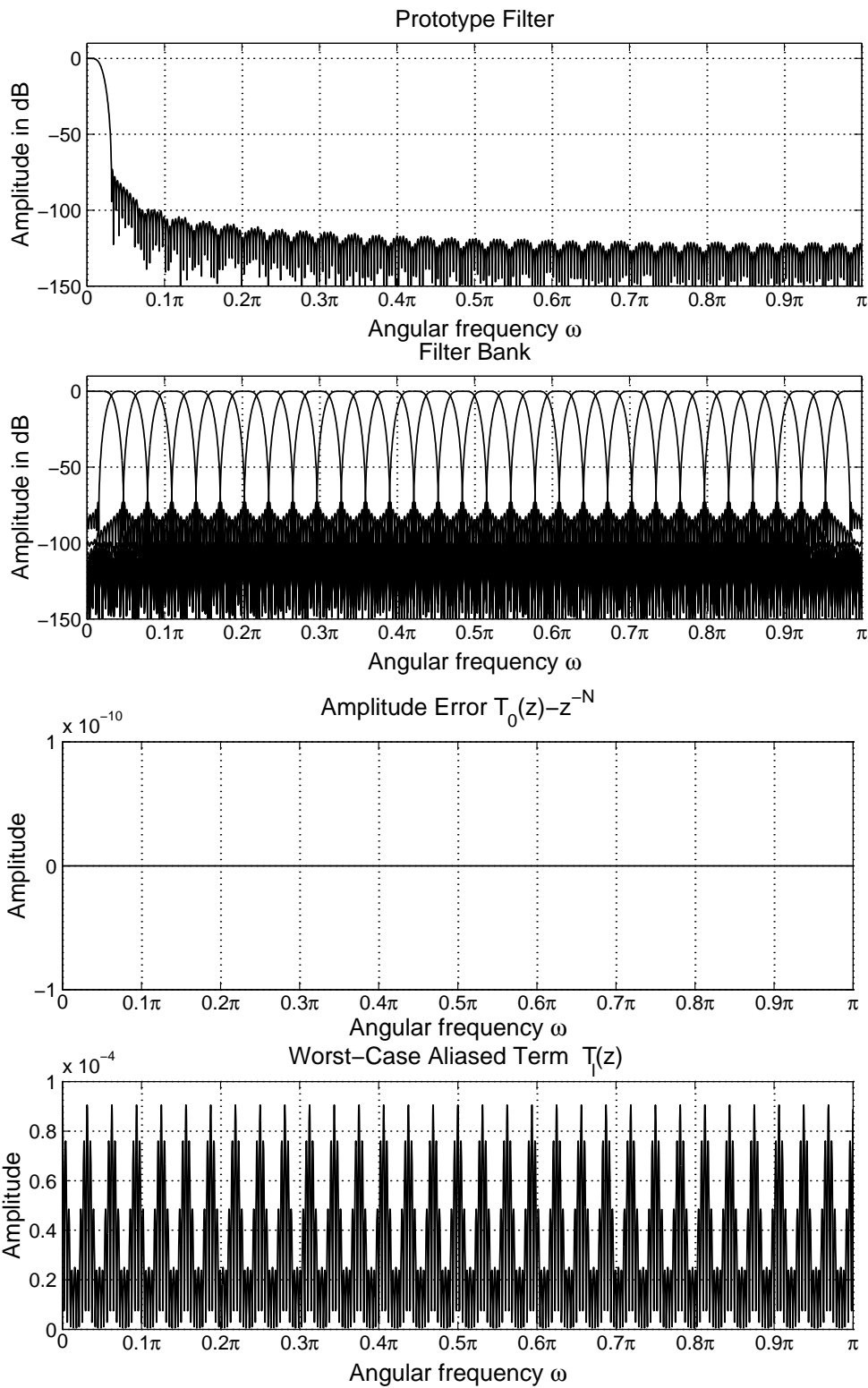
NEARLY PR FILTER BANK with

$$N = 511 \text{ for } \delta_1 = 0.0001$$



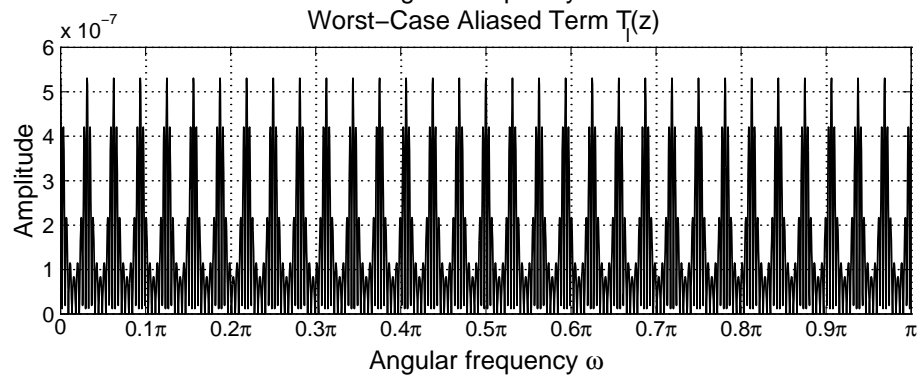
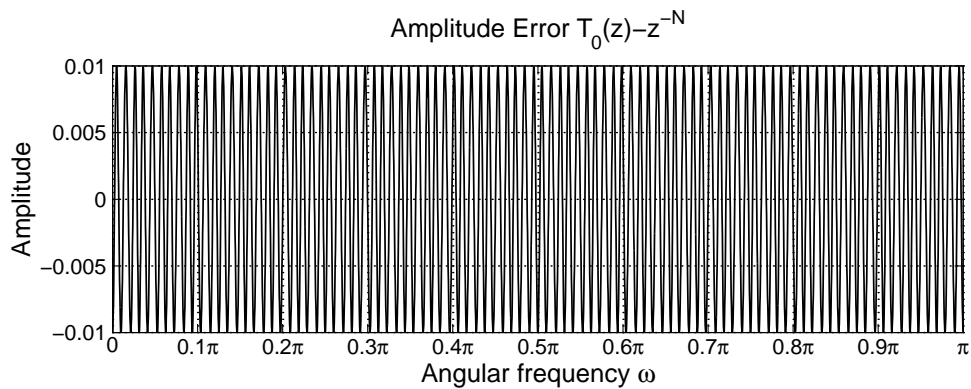
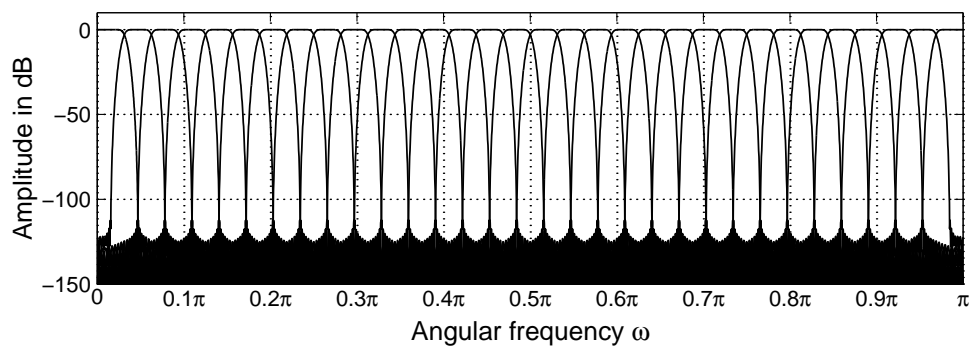
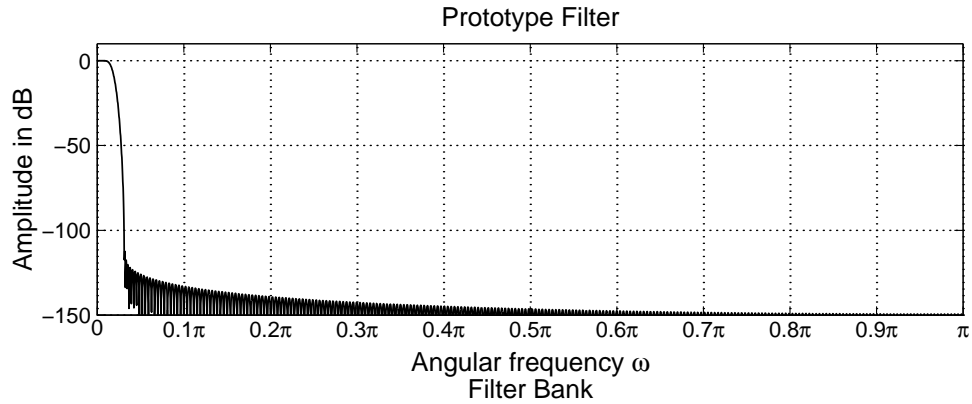
NEARLY PR FILTER BANK with

$$N = 511 \text{ for } \delta_1 = 0$$



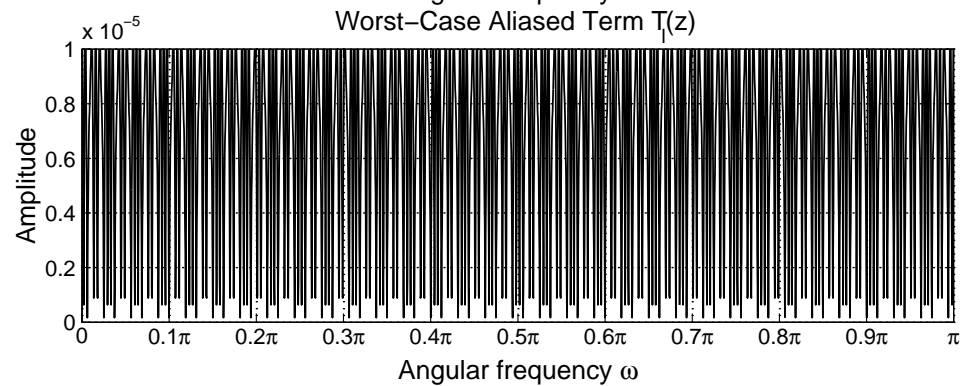
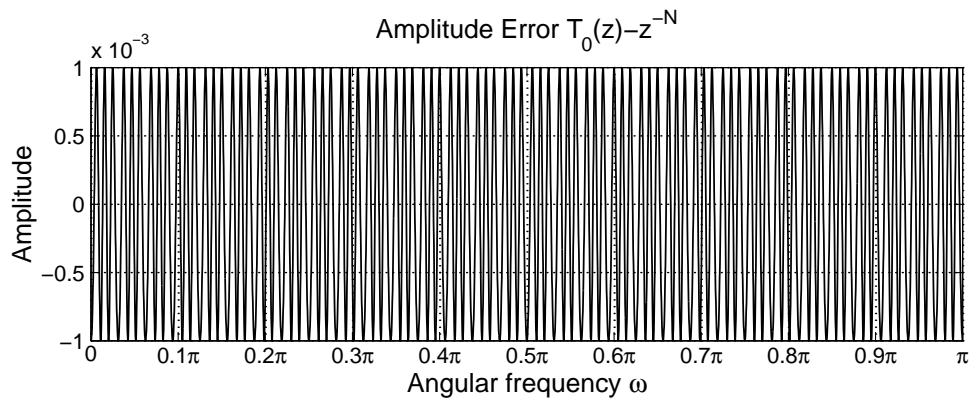
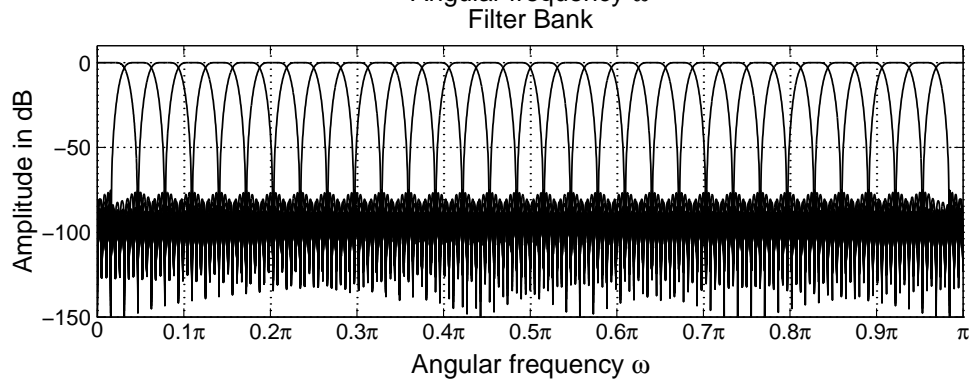
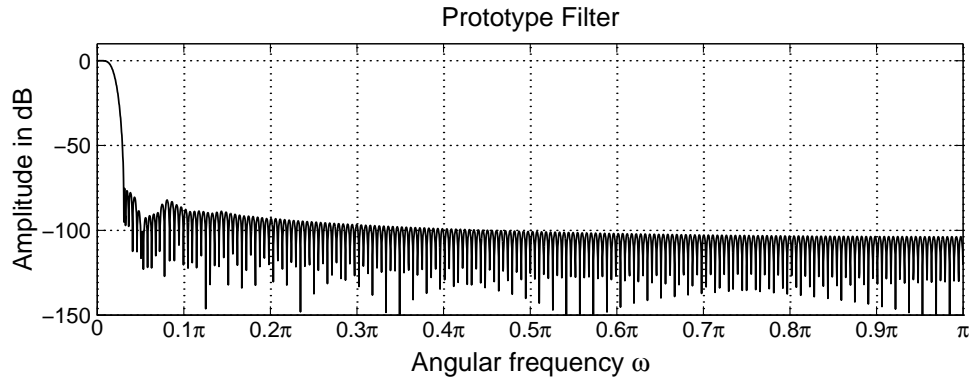
NEARLY PR FILTER BANK with

$$N = 511 \text{ for } \delta_1 = 0.01$$



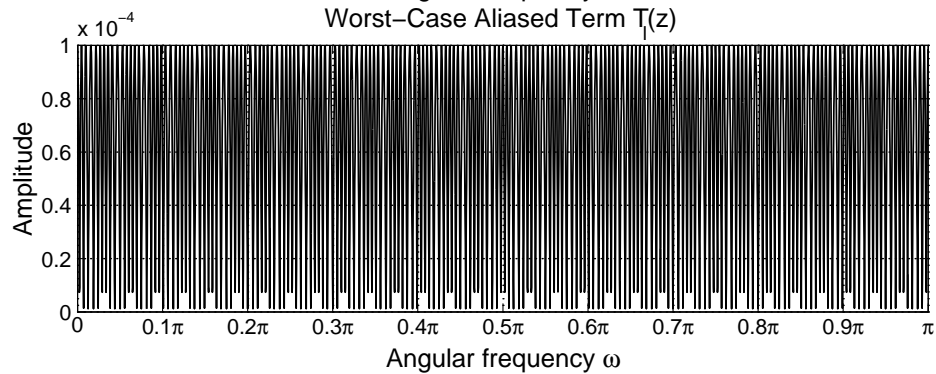
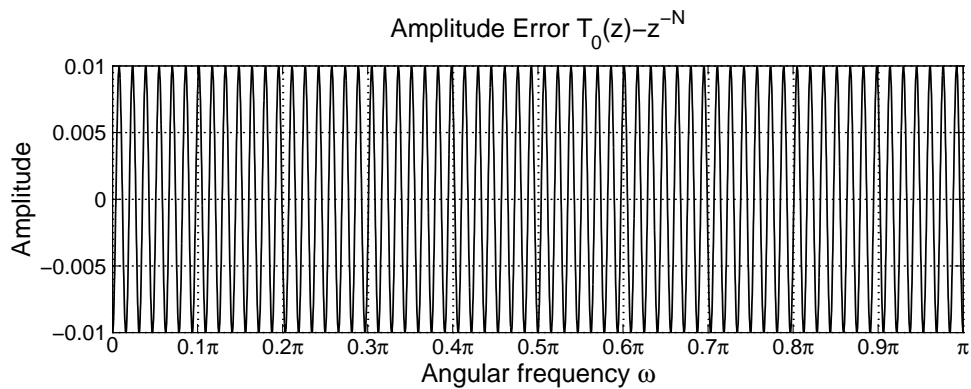
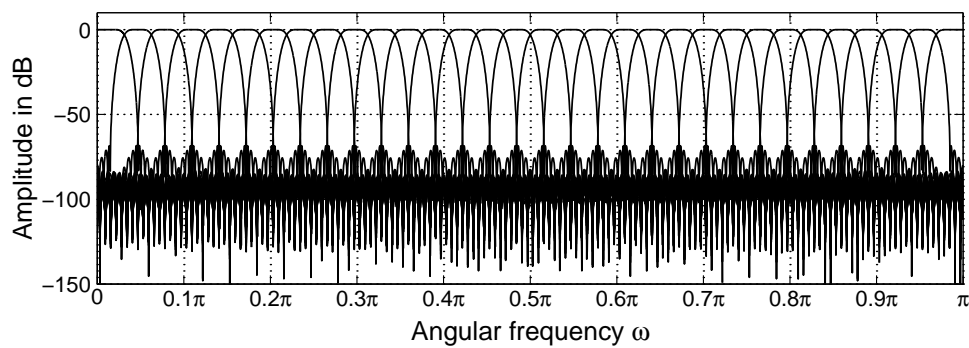
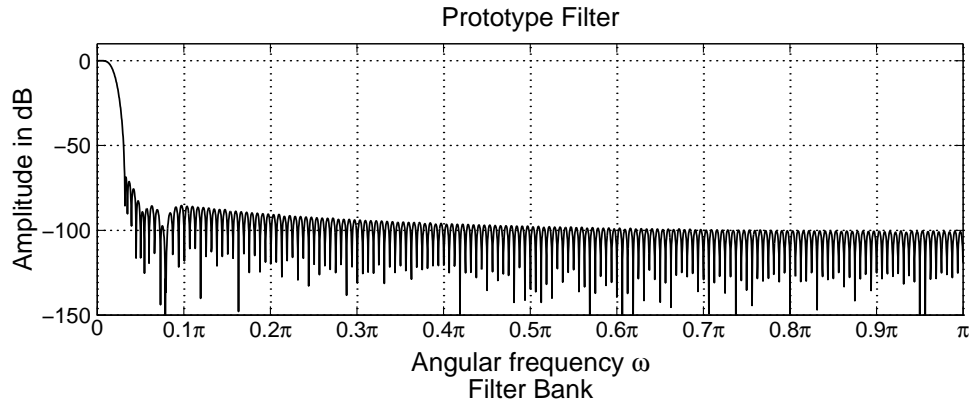
NEARLY PR BANK with $N = 383$ for

$$\delta_1 = 0.001 \text{ and } \delta_2 = 0.00001$$



NEARLY PR BANK with $N = 319$ for

$$\delta_1 = 0.01 \text{ and } \delta_2 = 0.0001$$



DESIGN OF APPROXIMATELY LINEAR PHASE RECURSIVE DIGITAL FILTERS

- It is shown how to minimize the maximum deviation of the passband phase of a recursive digital filter subject to the given amplitude criteria.
- The filters under consideration are conventional cascade-form filters and lattice wave digital (LWD) filters (parallel connections of two all-pass filters).
- There exist very efficient schemes for designing initial filters in the lowpass case.
- Before stating the problems, we denote the overall transfer function by $H(\Phi, z)$, where Φ is the adjustable parameter vector.
- The unwrapped phase response of the filter is denoted by $\arg H(\Phi, e^{j\omega})$.

STATEMENT OF THE PROBLEMS

Approximation Problem I: Given ω_p , ω_s , δ_p , and δ_s , as well as the filter order N , find Φ and ψ , the slope of a linear phase response, to minimize

$$\Delta = \max_{0 \leq \omega \leq \omega_p} |\arg H(\Phi, e^{j\omega}) - \psi\omega| \quad (10a)$$

subject to

$$1 - \delta_p \leq |H(\Phi, e^{j\omega})| \leq 1 \quad \text{for } \omega \in [0, \omega_p], \quad (10b)$$

$$|H(\Phi, e^{j\omega})| \leq \delta_s \quad \text{for } \omega \in [\omega_s, \pi], \quad (10c)$$

and

$$|H(\Phi, e^{j\omega})| \leq 1 \quad \text{for } \omega \in (\omega_p, \omega_s). \quad (10d)$$

Approximation Problem II: Given ω_p , ω_s , δ_p , and δ_s , as well as the filter order N , find Φ and ψ to minimize Δ as given by Eq. (10a) subject to the conditions of Eqs. (10b) and (10c) and

$$\frac{d|H(\Phi, e^{j\omega})|}{d\omega} \leq 0 \quad \text{for } \omega \in (\omega_p, \omega_s). \quad (10e)$$

EXAMPLE: $\omega_p = 0.05\pi$, $\omega_s = 0.1\pi$,
 $\delta_p = 0.0228$ (0.2-dB passband ripple), and
 $\delta_s = 10^{-3}$ (60-dB attenuation)

Elliptic Filter: The minimum order is **five**.

Cascade-Form Filter for Problem I: For the filter of order **seven** the maximum deviation from $\phi_{\text{ave}}(\omega) = -47.06\omega$ is 0.29 degrees.

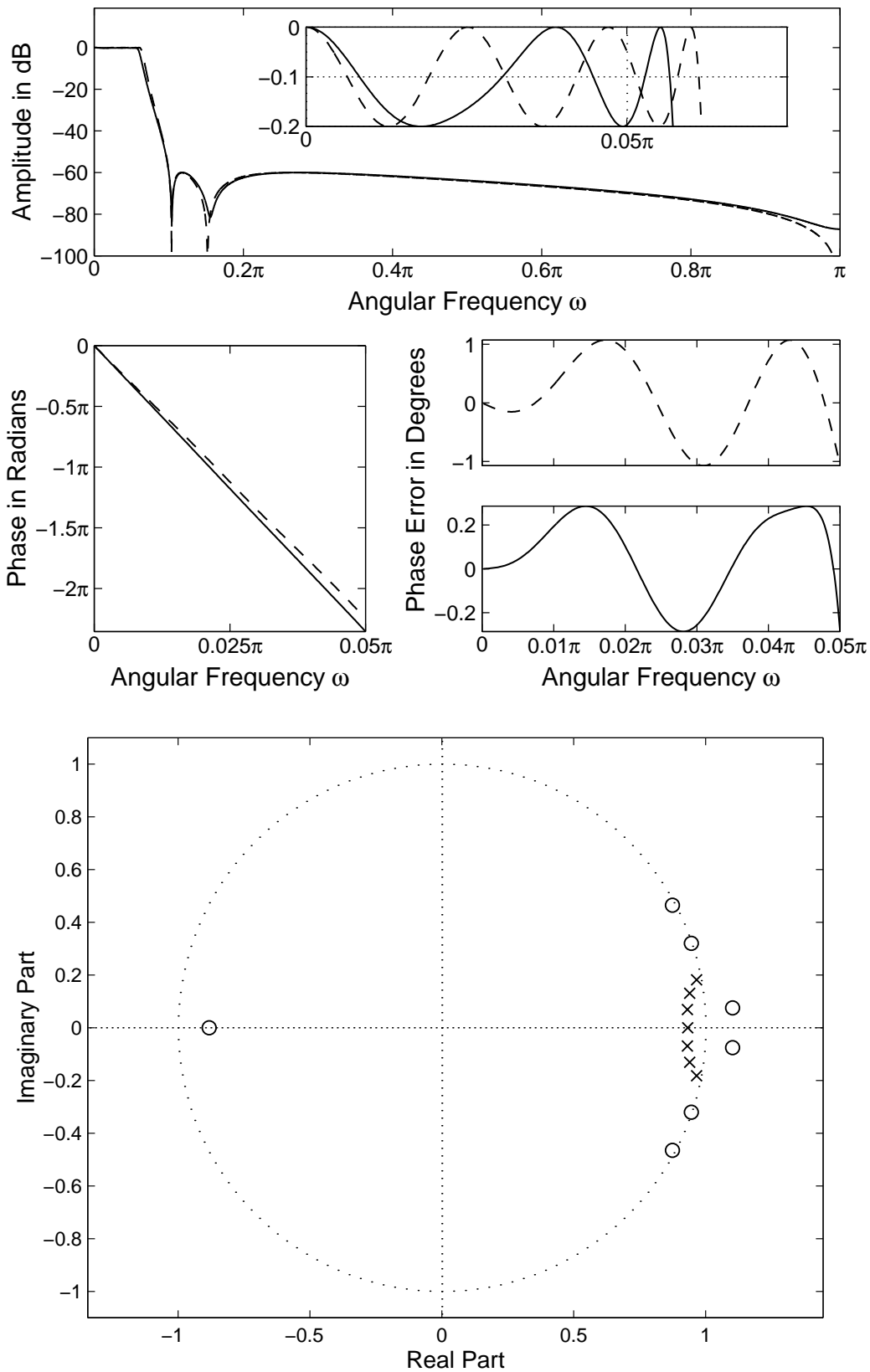
Cascade-Form Filter for Problem II: For the filter of order **seven** the maximum deviation from $\phi_{\text{ave}}(\omega) = -47.56\omega$ is 0.50 degrees.

Lattice Wave Digital Filter for Problem I: For the filter of order **nine** the maximum deviation from $\phi_{\text{ave}}(\omega) = -40.38\omega$ is 0.094 degrees.

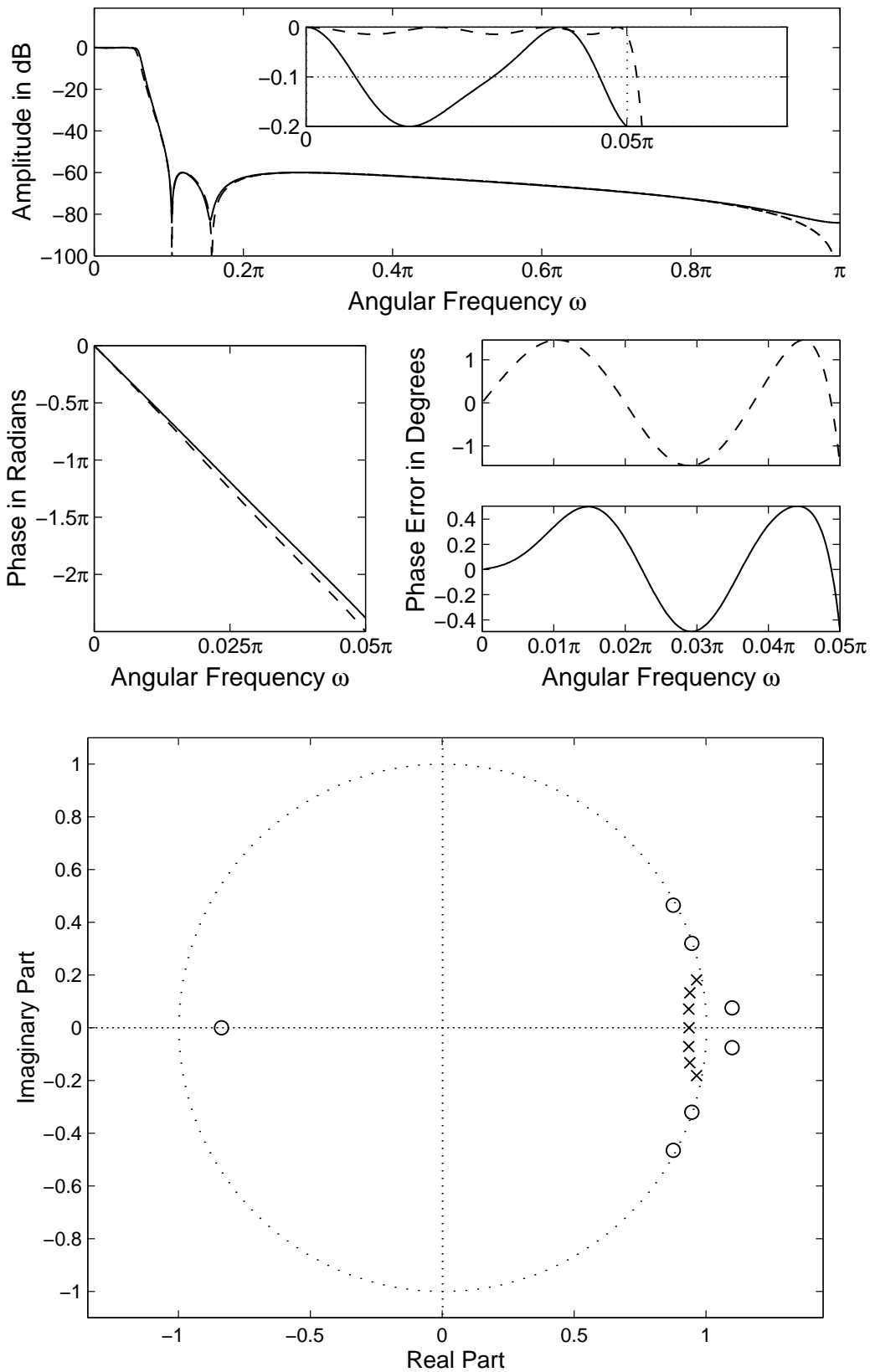
Lattice Wave Digital Filter for Problem II: For the filter of order **nine** the maximum deviation from $\phi_{\text{ave}}(\omega) = -42.92\omega$ is 0.27 degrees.

Linear-Phase FIR Filter: Minimum order is 107 and delay is 53.5 samples.

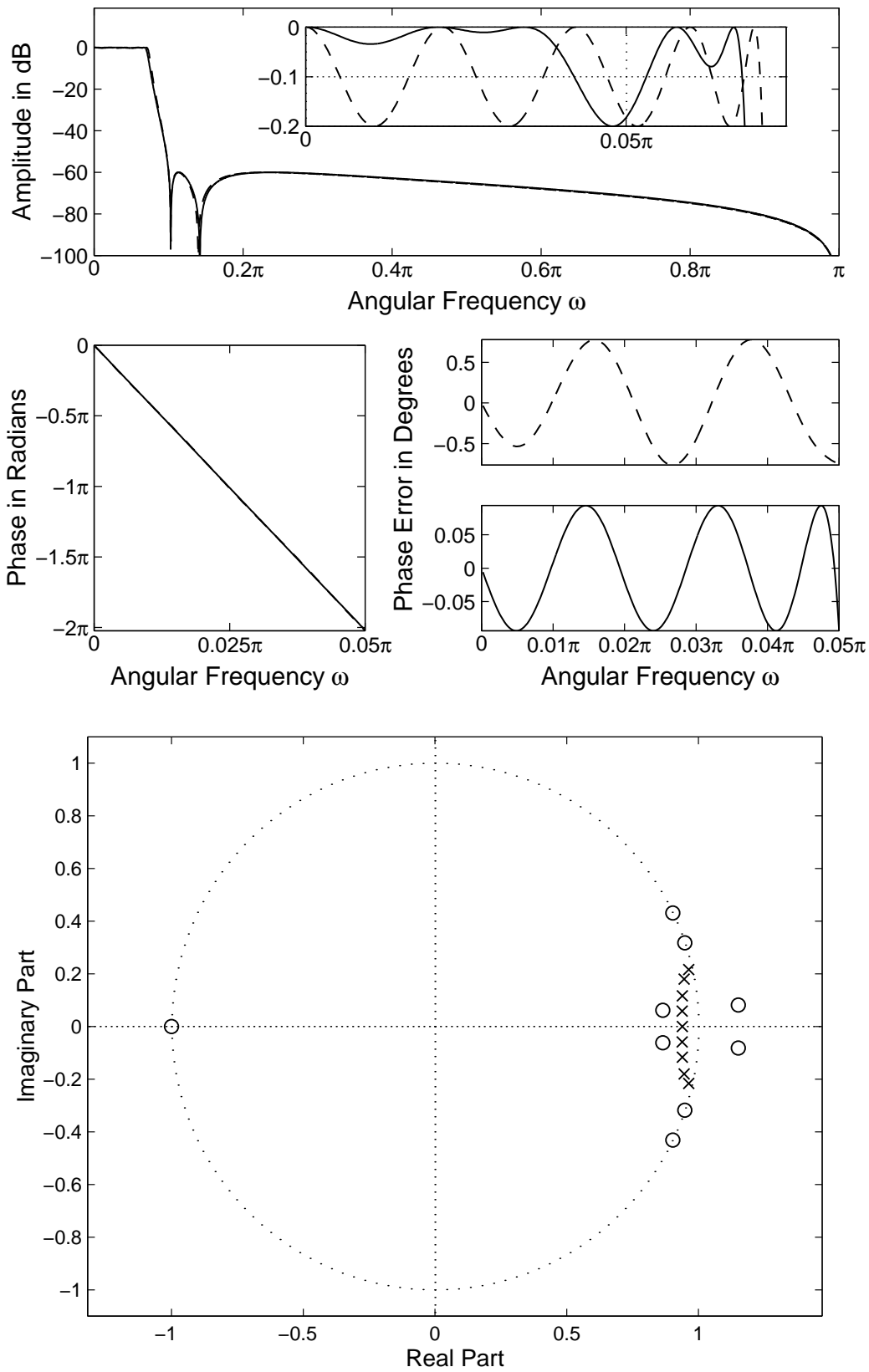
Cascade-Form Filter for Problem 1



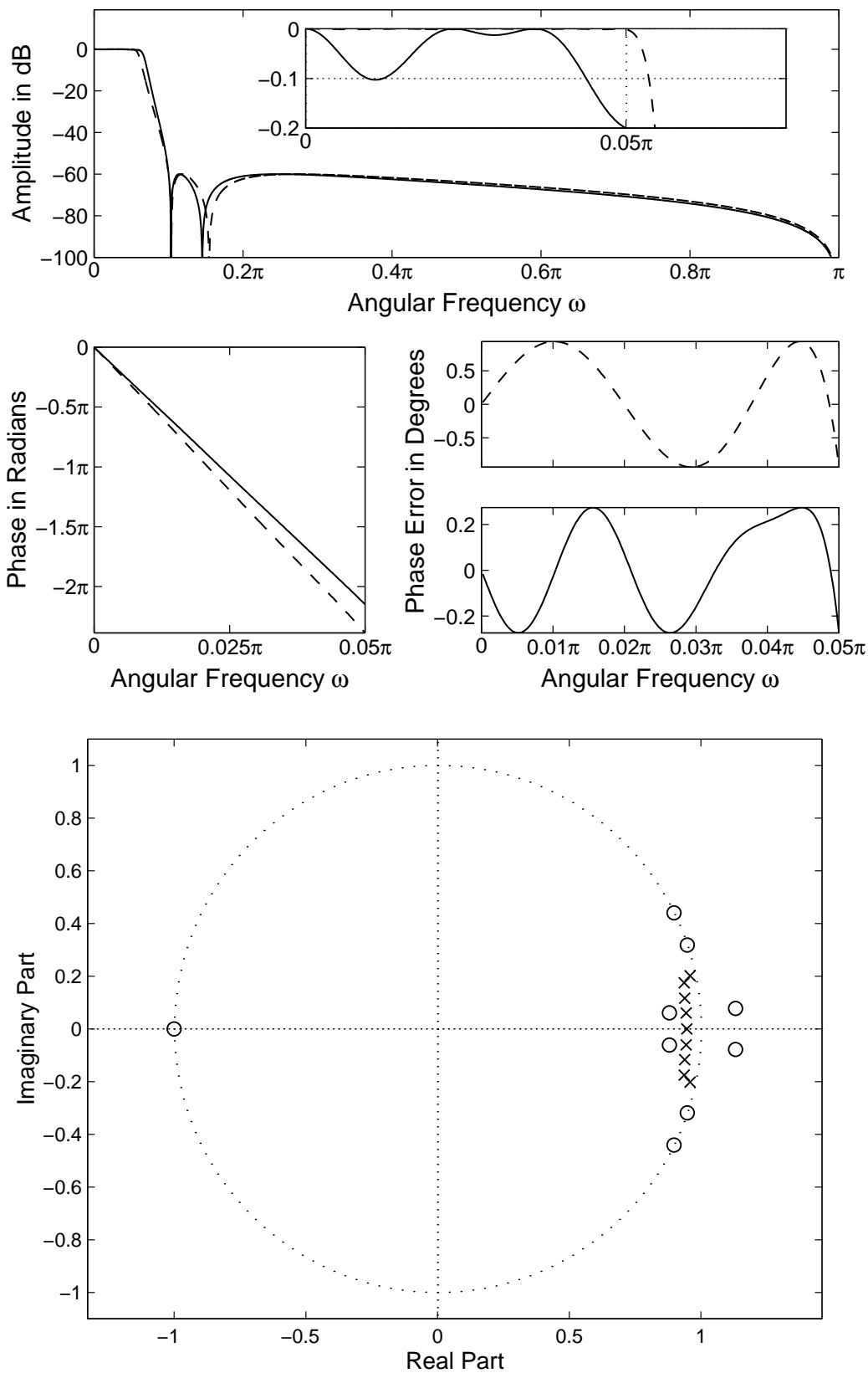
Cascade-Form Filter for Problem II



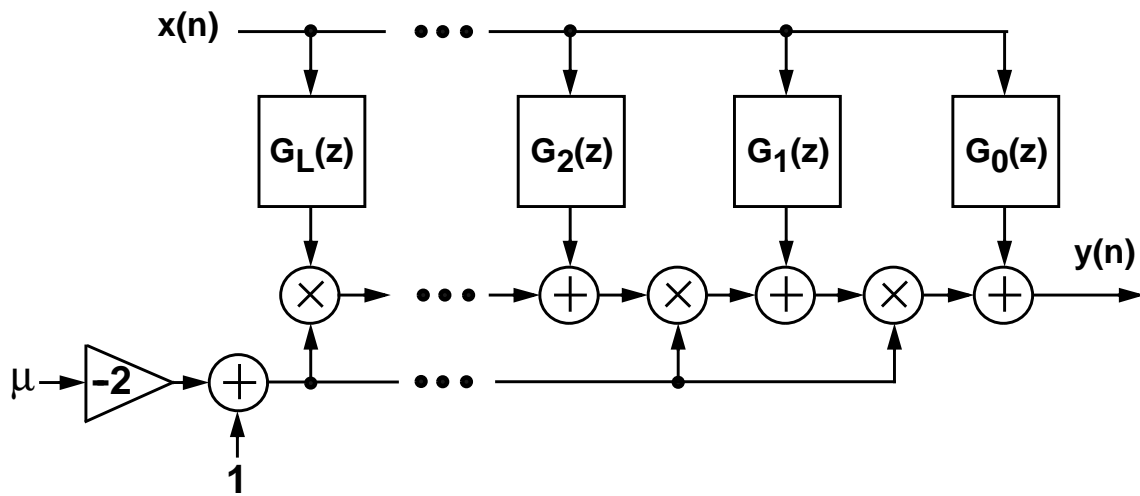
Lattice Wave Digital Filter for Problem I



Lattice Wave Digital Filter for Problem II



MODIFIED FARROW STRUCTURE WITH ADJUSTABLE FRACTIONAL DELAY



Fixed linear-phase filters for $l = 0, 1, \dots, L$:

$$G_l(z) = \sum_{n=0}^{N-1} g_l(n) z^{-n} \quad (11a)$$

where N is an even integer and

$$g_l(n) = \begin{cases} g_l(N-1-n) & \text{for } l \text{ even} \\ -g_l(N-1-n) & \text{for } l \text{ odd.} \end{cases} \quad (11b)$$

Delay: $N/2 - 1 + \mu$, where the fractional delay $0 \leq \mu < 1$ is directly the adjustable parameter of the structure.

TRANSFER FUNCTION

The overall transfer function is given by

$$H(\Phi, z, \mu) = \sum_{n=0}^{N-1} h(n, \Phi, \mu) z^{-n}, \quad (12a)$$

where

$$h(n, \Phi, \mu) = \sum_{l=0}^L g_l(n) (1 - 2\mu)^l \quad (12b)$$

and Φ is the adjustable parameter vector

$$\Phi = [g_0(0), g_0(1), \dots, g_0(N/2 - 1), g_1(0), g_1(1), \dots, \\ g_1(N/2 - 1), \dots, g_L(0), g_L(1), \dots, g_L(N/2 - 1)]. \quad (12c)$$

AMPLITUDE AND PHASE DELAY RESPONSES

The frequency, amplitude, and phase delay responses of the proposed Farrow structure are given by

$$H(\Phi, e^{j\omega}, \mu) = \sum_{n=0}^{N-1} h(n, \Phi, \mu) e^{-j\omega n}, \quad (13a)$$

$$|H(\Phi, e^{j\omega}, \mu)| = \left| \sum_{n=0}^{N-1} h(n, \Phi, \mu) e^{-j\omega n} \right|, \quad (13b)$$

and

$$\tau_p(\Phi, \omega, \mu) = -\arg(H(\Phi, e^{j\omega}, \mu))/\omega, \quad (13c)$$

respectively.

STATEMENT OF THE PROBLEM

Optimization Problem: Given L , N , Ω_p , and ϵ , find the adjustable parameter vector Φ to minimize

$$\delta_p = \max_{0 \leq \mu < 1} \left[\max_{\omega \in \Omega_p} |\tau_p(\Phi, \omega, \mu) - (N/2 - 1 + \mu)| \right] \quad (14a)$$

subject to

$$\delta_a = \max_{0 \leq \mu < 1} \left[\max_{\omega \in \Omega_p} \left| |H(\Phi, e^{j\omega}, \mu)| - 1 \right| \right] \leq \epsilon. \quad (14b)$$

EXAMPLES

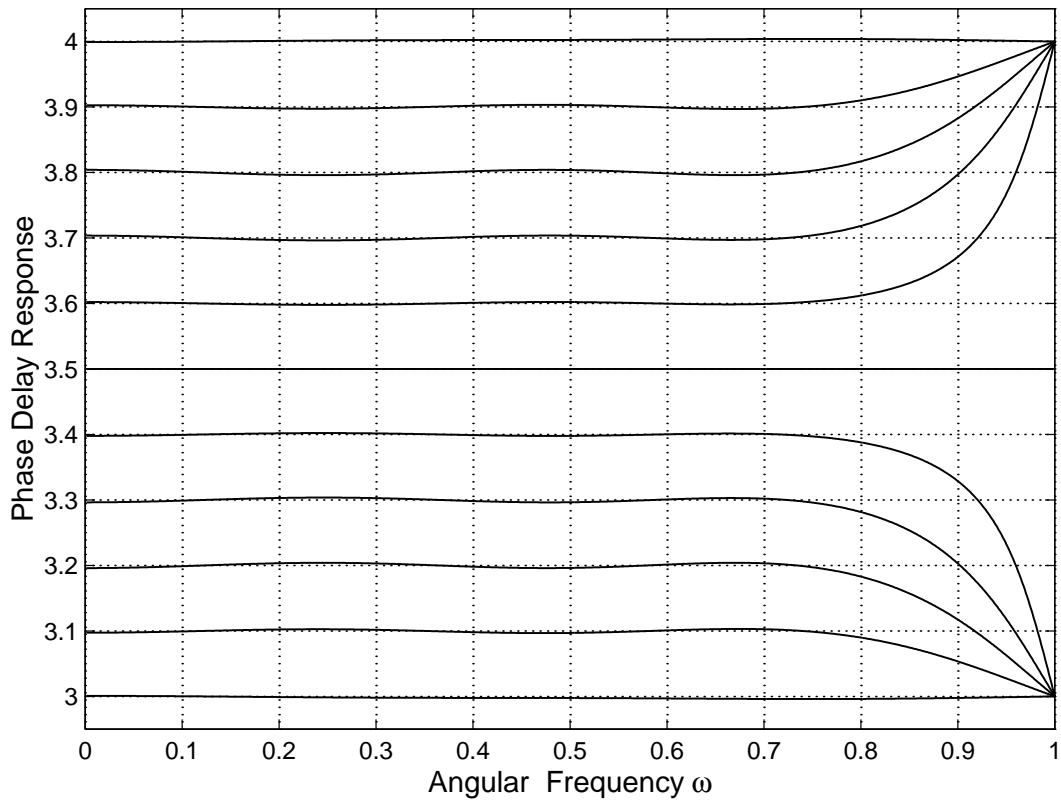
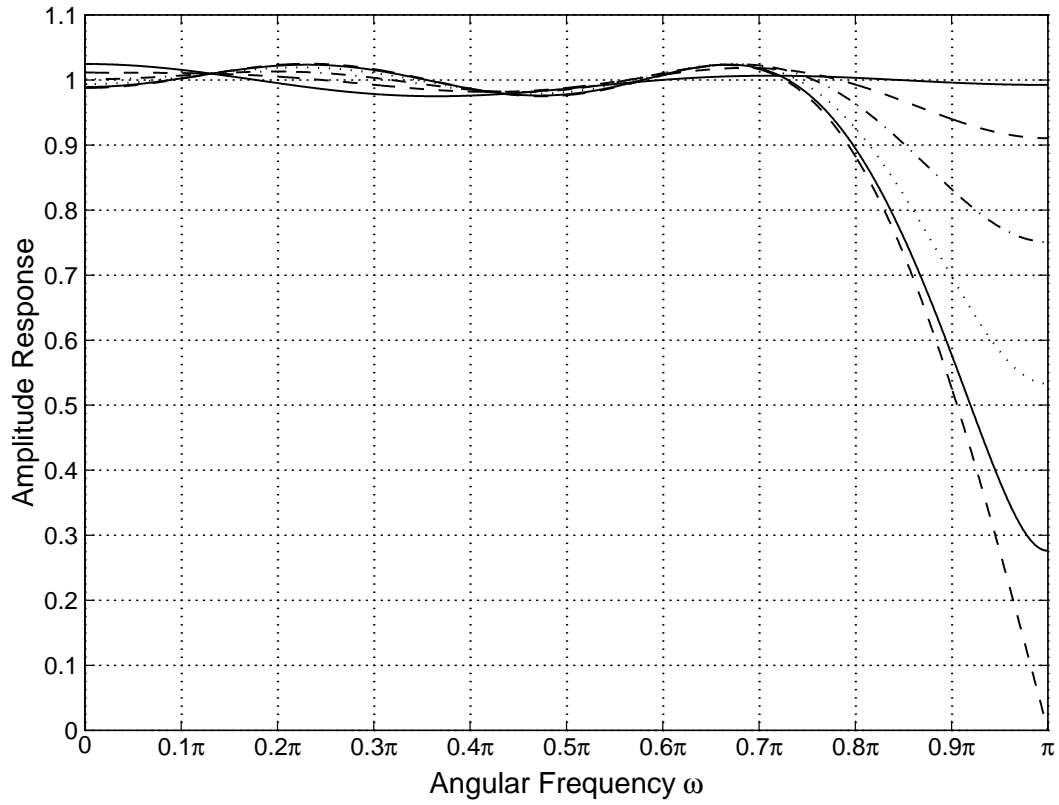
Example 1: $\Omega_p = [0, 0.75\pi]$, $\epsilon = 0.025$, and $\delta_p \leq 0.01$.

$\delta_p = 0.00402$ is achieved by $N = 8$ and $L = 3$.

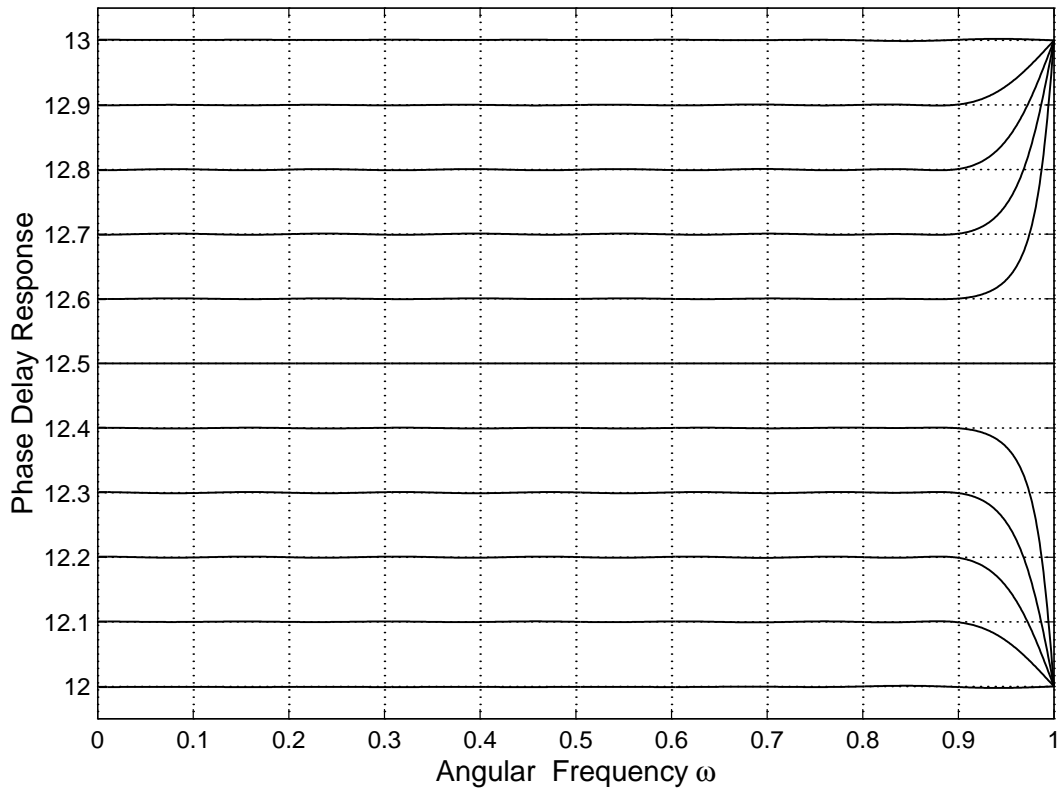
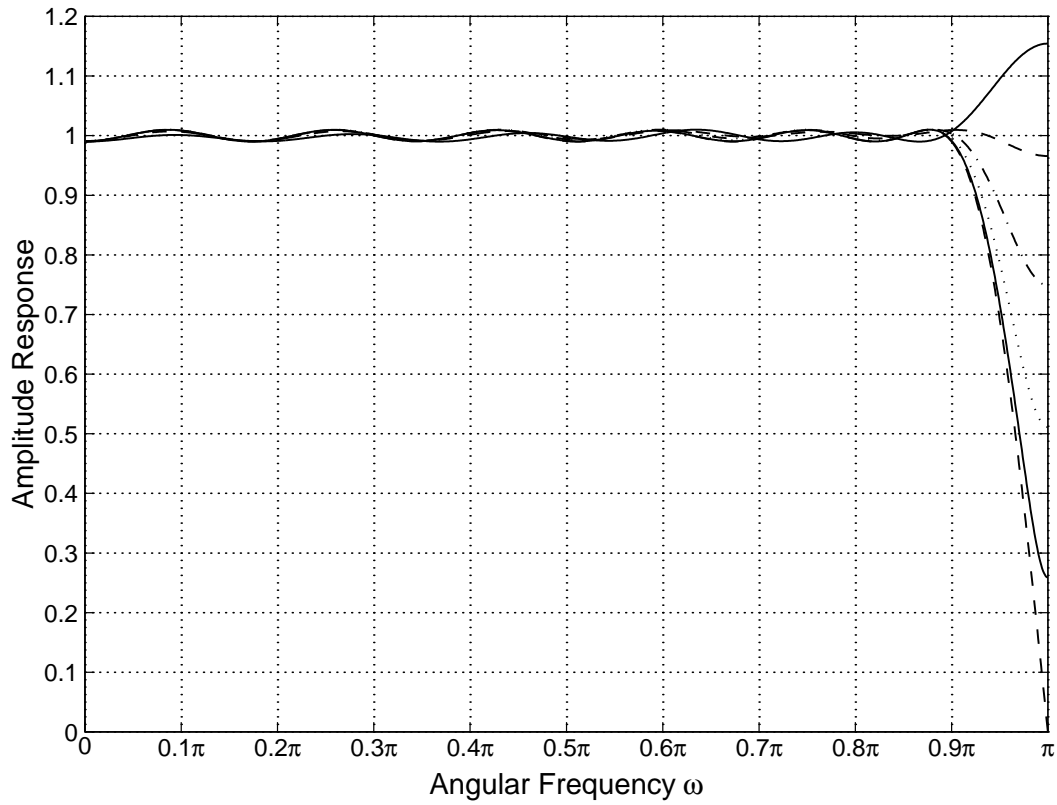
Example 2: $\Omega_p = [0, 0.9\pi]$, $\epsilon = 0.01$, and $\delta_p \leq 0.001$.

The criteria are met by $N = 26$ and $L = 4$.

RESPONSES FOR EXAMPLE 1



RESPONSES FOR EXAMPLE 2



DESIGN OF LATTICE WAVE DIGITAL FILTERS WITH SHORT COEFFICIENT WORDLENGTH

- It is shown how the coefficients of the various classes of lattice wave digital (LWD) filters can be conveniently quantized.
- The filters under consideration are
 - the conventional LWD filters,
 - cascades of low-order LWD filters providing a very low sensitivity and roundoff noise, and
 - LWD filters with an approximately linear phase in the passband.
- There exist very efficient schemes for designing initial filters in the lowpass case.
- Before stating the problems, we denote the overall transfer function as $H(\Phi, z)$, where Φ is the adjustable parameter vector.

Overall Transfer Function

The most general form of the transfer function is given by

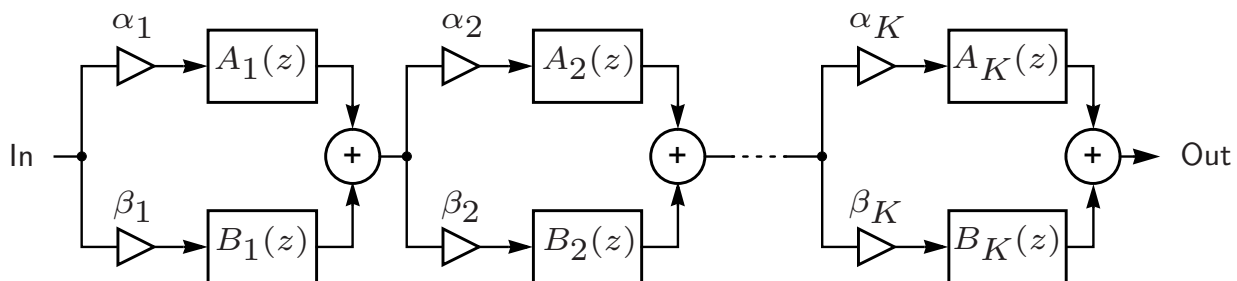
$$H(\Phi, z) = \prod_{k=1}^K H_k(\Phi, z), \quad (15a)$$

where

$$H_k(\Phi, z) = \alpha_k A_k(z) + \beta_k B_k(z). \quad (15b)$$

Here, $A_k(z)$'s and $B_k(z)$'s are stable all-pass filters of orders M_k and N_k , respectively.

For conventional and approximately linear-phase LWD filters $K = 1$.



Statement of the Problems

In VLSI applications it is desirable to express the coefficient values in the form

$$\sum_{r=1}^R a_r 2^{-P_r}, \quad (16)$$

where each of the a_r 's is either 1 or -1 and the P_r 's are positive integers in the increasing order.

The target is to find all the coefficient values included in Φ , in such a way that:

1. R , the number of powers of two, is made as small as possible and
2. P_R , the number of fractional bits, is made as small as possible.

Problem I: Find K , the number of subfilters, the M_k 's and N_k 's, as well as the adjustable parameter vector Φ in such a way that:

1. $H(\Phi, z)$ meets the criteria given by

$$1 - \delta_p \leq |H(\Phi, e^{j\omega})| \leq 1 \quad \text{for } \omega \in [0, \omega_p] \quad (17a)$$

$$|H(\Phi, e^{j\omega})| \leq \delta_s \quad \text{for } \omega \in [\omega_s, \pi]. \quad (17b)$$

2. The coefficients included in Φ are quantized to achieve the above-mentioned target for their representations.

Problem II: Find Φ as well as τ , the slope of the linear-phase response, in such a way that:

1. $H(\Phi, z)$ meets the criteria given by Eq. (17) and

$$|\arg[H(\Phi, e^{j\omega_i})] - \tau\omega| \leq \Delta \quad \text{for } i = 1, 2, \dots, L_p,$$

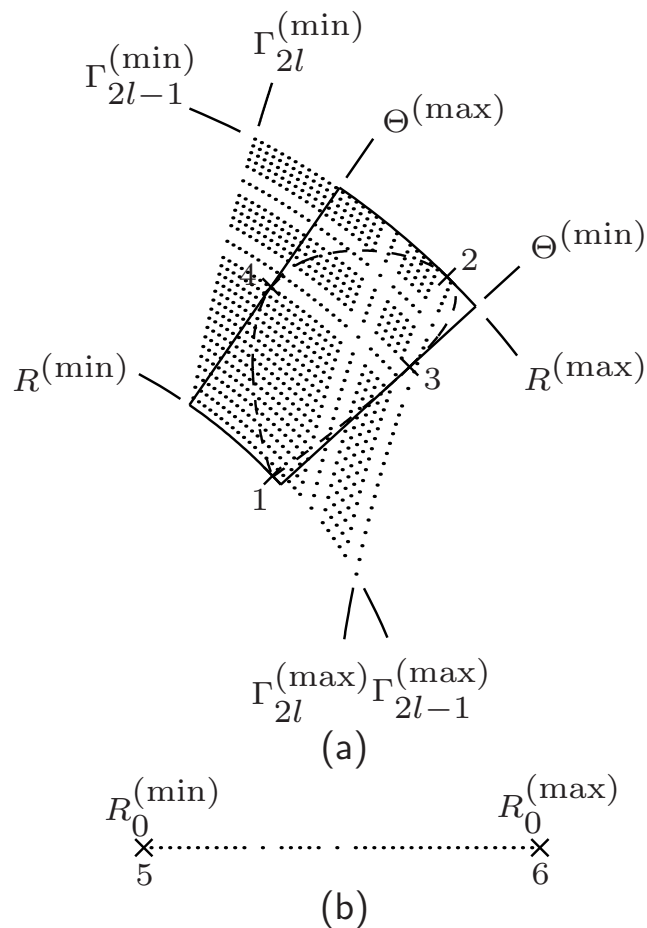
where $\arg[H(\Phi, e^{j\omega})]$ denotes the unwrapped phase response of the filter and Δ is the maximum allowable phase error from the linear-phase response.

2. The coefficients are quantized to achieve the above-mentioned target for their representations.

Quantization Algorithm

The coefficient optimization is performed in two stages:

1. A nonlinear optimization algorithm is used for determining a parameter space of the infinite-precision coefficients including the feasible space where the filter meets the criteria.



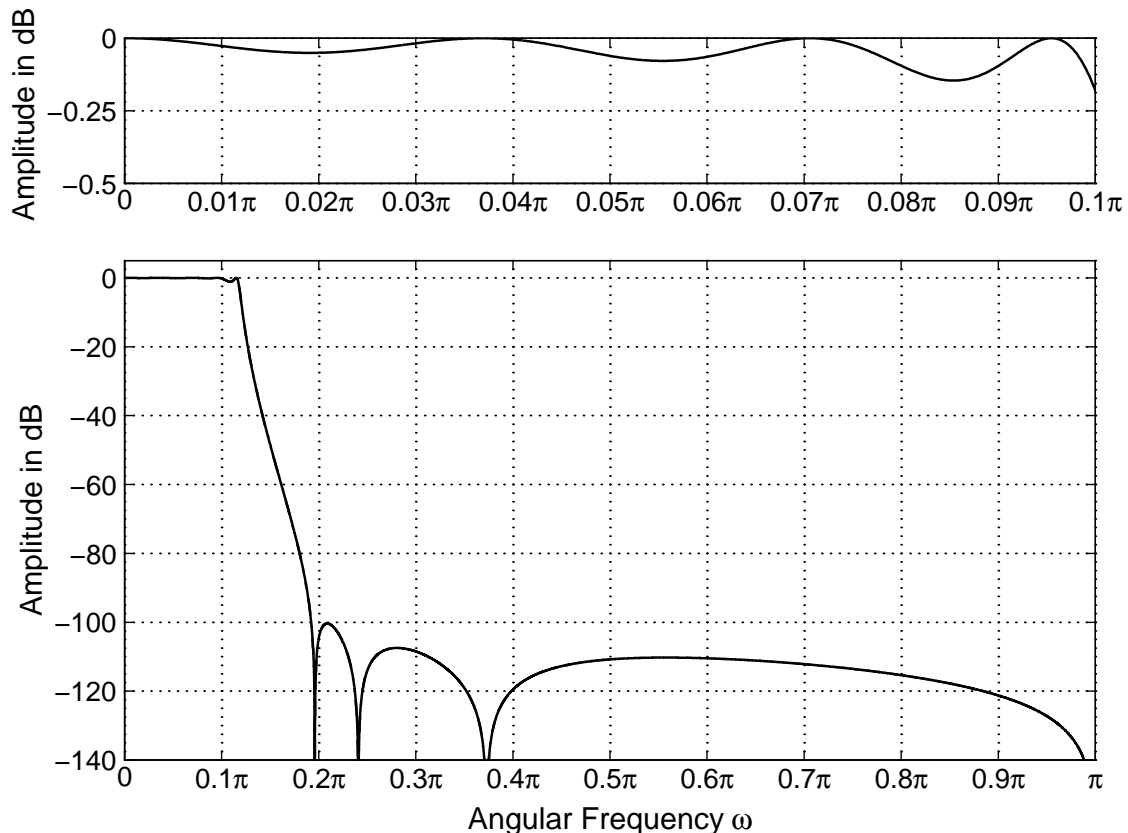
2. The filter parameters in this space are searched in such a manner that the resulting filter meets the given criteria with the simplest coefficient representation forms.

The algorithm guarantees that the optimum finite-wordlength solution can be found for both the fixed-point and the multiplierless coefficient representations.

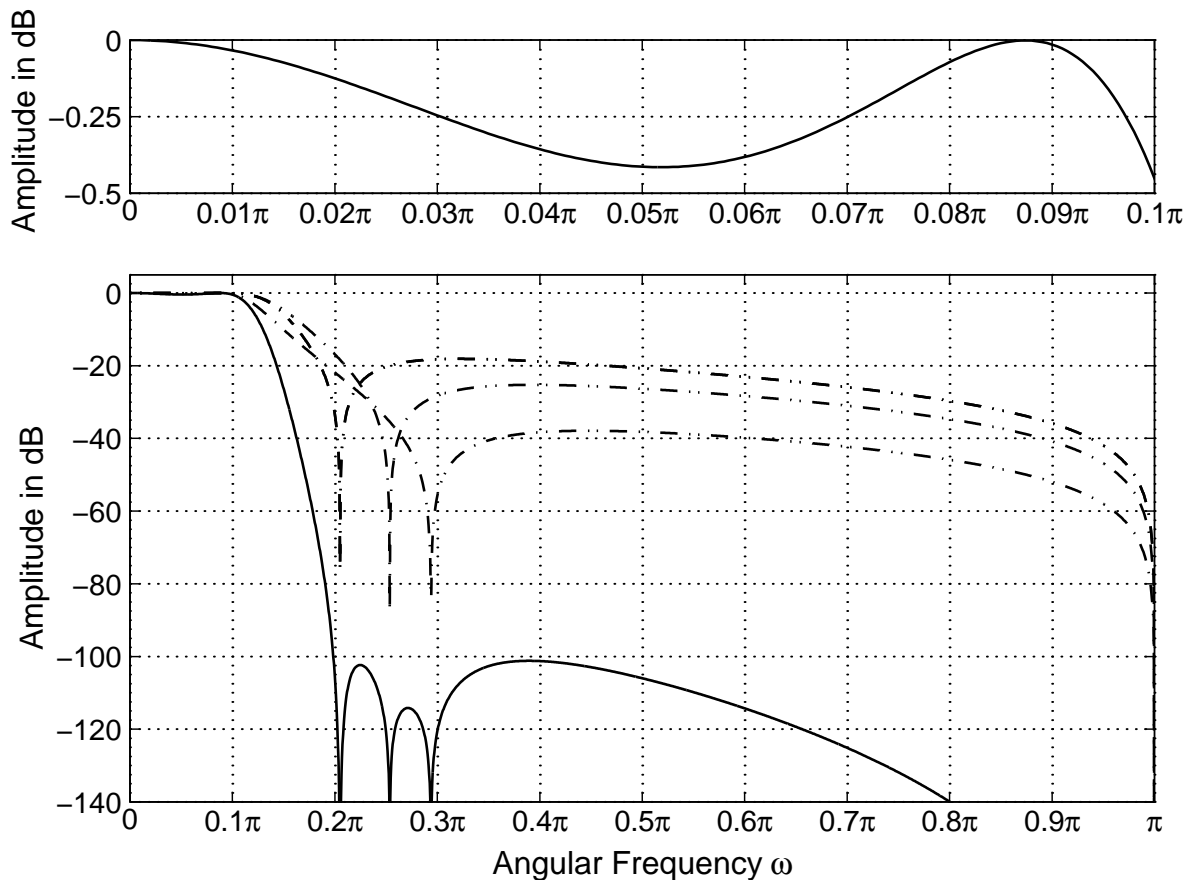
Numerical Examples

Filter specifications: $\delta_p = 0.0559$ (0.5-dB passband variation), $\delta_s = 10^{-5}$ (100-dB stopband attenuation), $\omega_p = 0.1\pi$, and $\omega_s = 0.2\pi$.

Ninth-order direct LWD filter is required to meet the criteria ($M_1 = 5$ and $N_1 = 4$).



Alternatively, cascade of **four 3rd-order** LWD filters is needed to satisfy the amplitude specifications.



For the cascade of four LWD filters, only **5** fractional bits are needed for coefficient implementation compared to **9** bits required by the direct LWD filter.

The number of adders required to implement all the coefficients are **12** and **6**, for the direct and cascade implementations, respectively.

The price paid for this is a slight increase in the overall filter order (from nine to twelve).

Optimized Finite-Precision Adaptor Coefficient Values for the Cascade of Four LWD Filters

$A(z)$	$B(z)$
$\gamma_0^{(1,2)} = 2^{-1} + 2^{-3}$	$\hat{\gamma}_1^{(1,2)} = -1 + 2^{-2} - 2^{-5}$ $\hat{\gamma}_2^{(1,2)} = 1 - 2^{-3} + 2^{-5}$
$\gamma_0^{(3)} = 2^{-1} + 2^{-3} + 2^{-5}$	$\hat{\gamma}_1^{(3)} = -1 + 2^{-2}$ $\hat{\gamma}_2^{(3)} = 1 - 2^{-3} + 2^{-5}$
$\gamma_0^{(4)} = 1 - 2^{-2} + 2^{-5}$	$\hat{\gamma}_1^{(4)} = -1 + 2^{-2} - 2^{-4}$ $\hat{\gamma}_2^{(4)} = 1 - 2^{-4}$

Approximately Linear-Phase LWD Filter

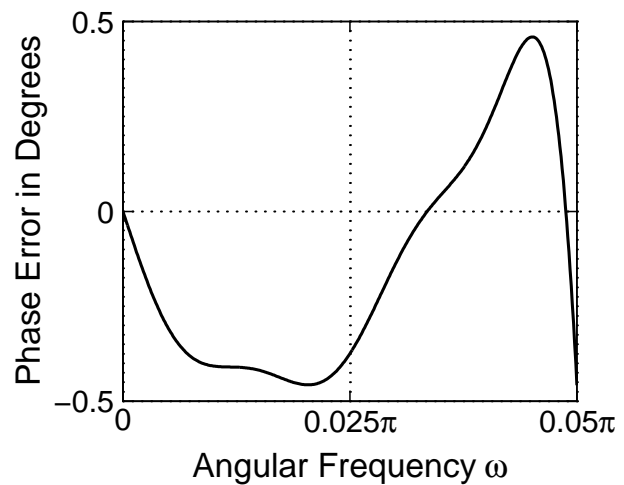
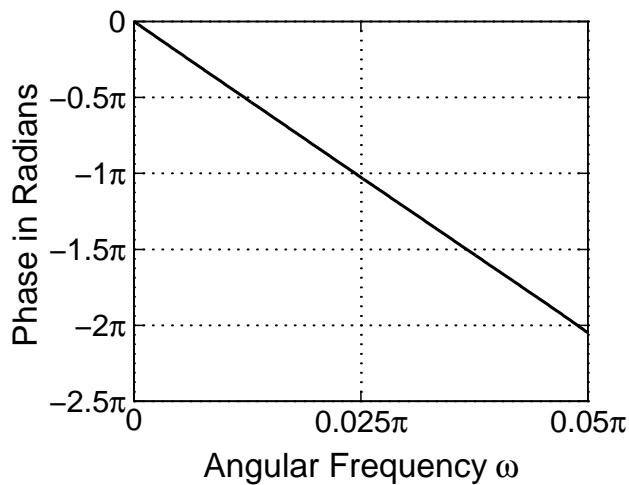
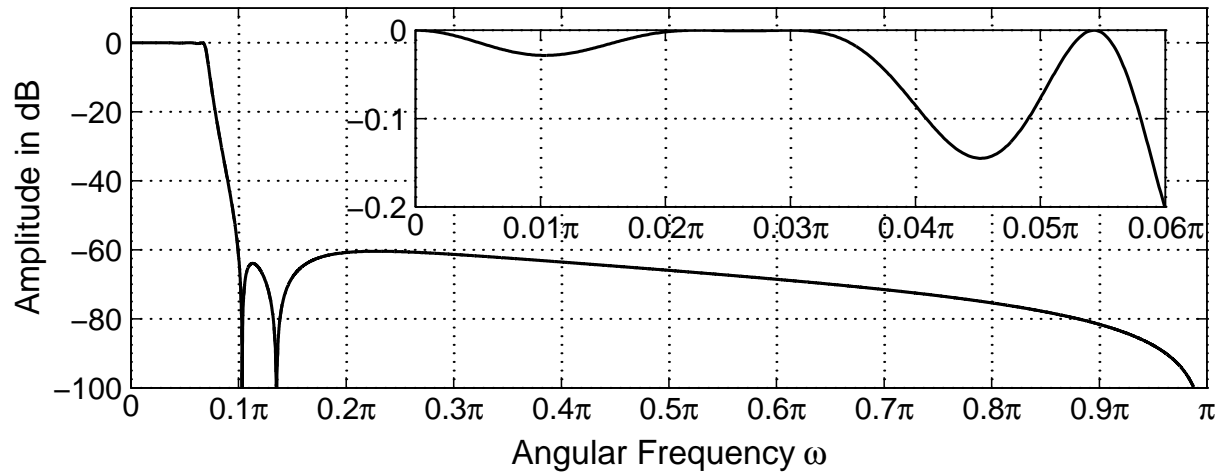
Filter specifications: $\delta_p = 0.0228$ (0.2-dB passband variation), $\delta_s = 10^{-3}$ (60-dB stopband attenuation), $\omega_p = 0.05\pi$, and $\omega_s = 0.1\pi$.

The minimum order of an elliptic filter to meet the amplitude specifications in **five**. An excellent phase performance is obtained by increasing the filter order to **nine**.

For the optimal infinite-precision filter the phase error is 0.09399 degrees.

To allow some tolerance for the quantization, the maximum allowable phase error is increased to 0.5 degrees.

Amplitude and Phase Responses for the Quantized Filter



For the optimized filter, only 10 adders with 11 fractional bits are required to implement all the adaptor coefficients.

The phase error for the optimized filter is 0.45855 degrees.

A SYSTEMATIC ALGORITHM FOR THE DESIGN OF MULTIPLIERLESS FIR FILTERS

In this application, we show how the coefficients of the linear-phase FIR filters can be conveniently quantized using optimization techniques.

The zero-phase frequency response of a linear-phase N th-order FIR filter can be expressed as

$$H(\omega) = \sum_{n=0}^M h(n) \text{Trig}(\omega, n), \quad (18)$$

where the $h(n)$'s are the filter coefficients and $\text{Trig}(\omega, n)$ is an appropriate trigonometric function depending on whether N is odd or even and whether the impulse response is symmetrical or antisymmetrical. Here, $M = N/2$ if N is even, and $M = (N + 1)/2$ if N is odd.

Desired Coefficient Representation Form

The general form for expressing the FIR filter coefficient values as a sums of signed-powers-of-two (SPT) terms is given by

$$h(n) = \sum_{k=1}^{W_n+1} a_{k,n} 2^{-P_{k,n}} \quad \text{for } n = 1, 2, \dots, M, \quad (19)$$

where $a_{k,n} \in \{-1, 1\}$ and $P_{k,n} \in \{1, 2, \dots, L\}$ for $k = 1, 2, \dots, W_n + 1$. In this representation form, each coefficient $h(n)$ has W_n adders and the maximum allowable wordlength is L bits.

Statement of the Problem

When finding the optimized simple discrete-value representation forms for the coefficients of FIR filters, it is a common practise to accomplish the optimization in such a manner that the scaled response meets the given amplitude criteria.

In this case, the criteria for the filter can be expressed as

$$1 - \delta_p \leq H(\omega)/\beta \leq 1 + \delta_p \quad \text{for } \omega \in [0, \omega_p] \quad (20a)$$

$$-\delta_s \leq H(\omega)/\beta \leq \delta_s \quad \text{for } \omega \in [\omega_s, \pi], \quad (20b)$$

where

$$\beta = \frac{1}{2} [\max H(\omega) + \min H(\omega)] \quad \text{for } \omega \in [0, \omega_p] \quad (20c)$$

is the average passband gain.

These criteria are preferred to be used when the filter coefficients are desired to be quantized on a highly nonuniform discrete grid as in the case of the power-of-two coefficients.

Optimization Problem

Given ω_p , ω_s , δ_p , and δ_s , as well as L , the number of fractional bits, and the maximum allowed number of SPT terms per coefficient, find the filter coefficients $h(n)$ for $n = 1, 2, \dots, M$ as well as β to minimize implementation cost, in such a manner that:

1. The magnitude criteria, as given by Eq. (20), are met and
2. the normalized peak ripple (NPR), as given by

$$\delta_{\text{NPR}} = \max\{\Delta_p/W, \Delta_s\}, \quad (21)$$

is minimized.

Here, Δ_p and Δ_s are the passband and stopband ripples of the finite-precision filter scaled by $1/\beta$ and $W = \delta_p/\delta_s$.

Coefficient Optimization

The coefficient optimization is performed in two stages:

1. For each of the filter coefficient $h(n)$ for $n = 0, 1, \dots, M - 1$ the largest and smallest values of the coefficient are determined in such a manner that the given amplitude criteria are met subject to $h(M) = 1$.

This restriction, simplifying the overall procedure, can be stated without loss of generality since the scaling constant β , as defined by Eq. (20), can be used for achieving the desired passband amplitude level.

These problems can be solved conveniently by using linear programming.

2. It has been experimentally observed that the parameter space defined above forms the feasible space where the filter specifications are satisfied. After finding this larger space, all what is needed is to check whether in this space there exists a combination of the discrete coefficient values with which the overall criteria are met.

Optimization of Infinite-Precision Coefficients

The goal is achieved by solving $2M$ problems of the following form. Find the filter coefficients $h(n)$ for $n = 0, 1, \dots, M - 1$ as well as β to minimize ψ subject to the conditions

$$\begin{aligned} & \sum_{n=0}^{M-1} h(n) \text{Trig}(\omega_i, n) - \beta(\delta_p + 1) \leq -\text{Trig}(\omega_i, M), \\ - & \sum_{n=0}^{M-1} h(n) \text{Trig}(\omega_i, n) - \beta(\delta_p - 1) \leq -\text{Trig}(\omega_i, M), \end{aligned}$$

for $\omega_i \in [0, \omega_p]$ and

$$\begin{aligned} & \sum_{n=0}^{M-1} h(n) \text{Trig}(\omega_i, n) - \beta\delta_s \leq -\text{Trig}(\omega_i, M), \\ - & \sum_{n=0}^{M-1} h(n) \text{Trig}(\omega_i, n) - \beta\delta_s \leq -\text{Trig}(\omega_i, M), \end{aligned}$$

for $\omega_i \in [\omega_s, \pi]$.

Here ψ is $-h(n)$ or $h(n)$ where $h(n)$ is one among the filter coefficients $h(n)$ for $n = 0, 1, \dots, M - 1$.

Optimization of Finite-Precision Coefficients

In the above procedure, $h(M)$ was fixed to be unity. The search for a proper combinations of discrete values can be conveniently accomplished by using a scaling constant $\alpha \equiv h(M)$.

For this constant, all the existing values between $1/3$ and $2/3$ are selected from the look-up table containing all the possible power-of-two numbers for a given wordlength and a given maximum number of SPT terms per coefficient.

Then for each value of $\alpha = h(M)$, the largest and smallest values of the infinite-precision coefficients are scaled in the look-up table as

$$\hat{h}(n)^{(\min)} = \alpha h(n)^{(\min)} \quad \text{for } n = 0, 1, \dots, M \quad (23a)$$

$$\hat{h}(n)^{(\max)} = \alpha h(n)^{(\max)} \quad \text{for } n = 0, 1, \dots, M \quad (23b)$$

and the magnitude response is evaluated for each combination of the power-of-two numbers in the ranges $[\hat{h}(n)^{(\min)}, \hat{h}(n)^{(\max)}]$ for $n = 0, 1, \dots, M$ to check whether the filter meets the amplitude criteria.

Numerical Examples

Example 1: $N = 37$, $\delta_p = \delta_s = 10^{-3}$, $\omega_p = 0.3\pi$, and $\omega_s = 0.5\pi$.

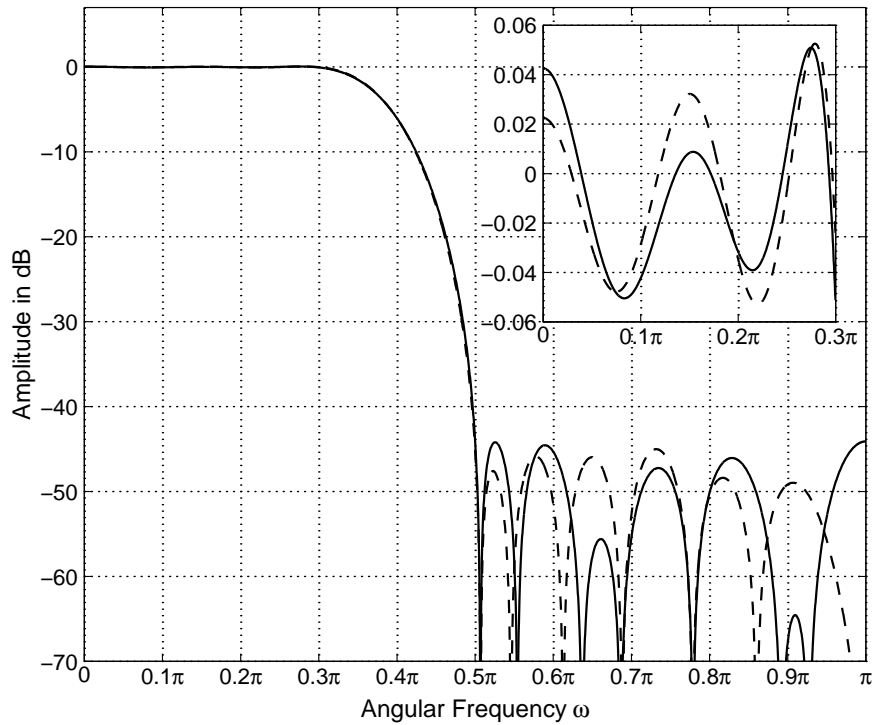
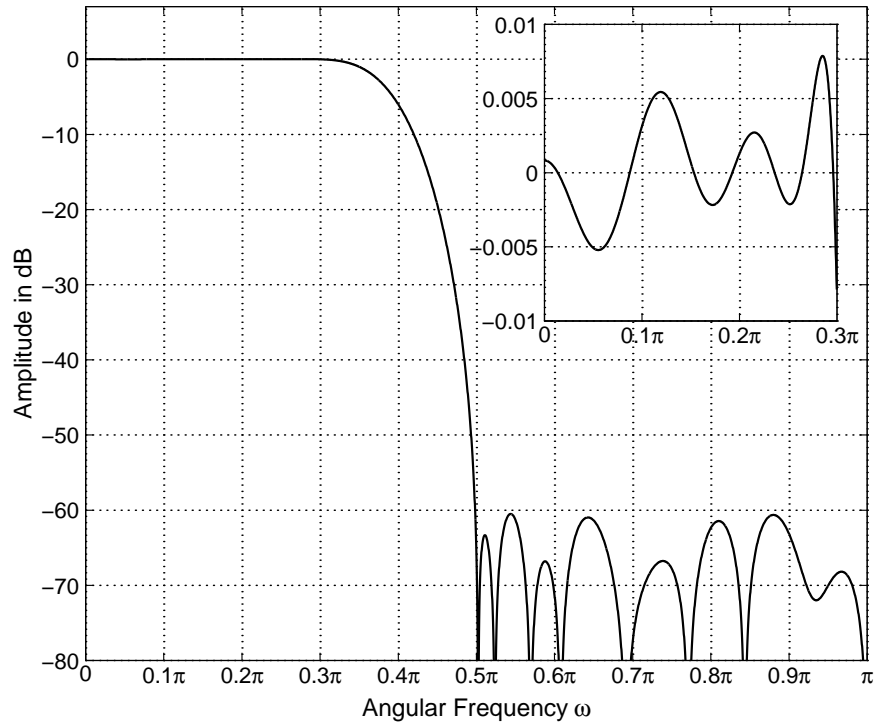
Method	δ_{NPR} (dB)	No. Powers of Two	No. Adders
Lim and Parker	-62.08	43	-
Chen and Willson	-60.87	40	-
Proposed	-60.48	34	48

Example 2: $N = 24$, $\delta_p = \delta_s = 0.005$, $\omega_p = 0.3\pi$, and $\omega_s = 0.5\pi$.

Method	δ_{NPR} (dB)	No. Powers of Two	No. Adders
Samueli	-42.17	24	35
Li <i>et al.</i>	-43.33	24	-
Chen and Willson	-43.97	24	33
Proposed	-44.09	21	30

If redundancies within the coefficients are utilized only 26 adders are required to meet the specifications.

Responses for Examples 1 and 2



Optimized Finite-Precision Coefficient Values for the FIR Filter in Example 1

$$\begin{aligned}h(0) &= h(37) = -2^{-11} \\h(1) &= h(36) = 0 \\h(2) &= h(35) = +2^{-9} - 2^{-12} \\h(3) &= h(34) = +2^{-9} \\h(4) &= h(33) = -2^{-9} - 2^{-11} \\h(5) &= h(32) = -2^{-7} + 2^{-9} - 2^{-11} \\h(6) &= h(31) = 0 \\h(7) &= h(30) = +2^{-6} - 2^{-8} \\h(8) &= h(29) = +2^{-7} + 2^{-9} \\h(9) &= h(28) = -2^{-6} + 2^{-8} - 2^{-10} \\h(10) &= h(27) = -2^{-5} + 2^{-8} + 2^{-12} \\h(11) &= h(26) = 0 \\h(12) &= h(25) = +2^{-4} - 2^{-6} - 2^{-9} \\h(13) &= h(24) = +2^{-5} + 2^{-8} + 2^{-10} \\h(14) &= h(23) = -2^{-4} + 2^{-6} - 2^{-10} \\h(15) &= h(22) = -2^{-3} + 2^{-6} + 2^{-8} \\h(16) &= h(21) = 0 \\h(17) &= h(20) = +2^{-2} + 2^{-6} \\h(18) &= h(19) = +2^{-1}\end{aligned}$$

Effective Implementation Exploiting Coefficient Symmetry for the Multiplierless FIR Filter in Example 1

