Optimal Design of Magnitude Responses of Rational Infinite Impulse Response Filters

Charlotte Yuk-Fan Ho\textsuperscript{1}, Student Member, IEEE, Bingo Wing-Kuen Ling\textsuperscript{2}, Yan-Qun Liu\textsuperscript{3}, Peter Kwong-Shun Tam\textsuperscript{4}, Member, IEEE, and Kok-Lay Teo\textsuperscript{5}, Senior Member, IEEE

Abstract—This paper considers a design of magnitude responses of optimal rational infinite impulse response (IIR) filters. The design problem is formulated as an optimization problem in which a total weighted absolute error in the passband and stopband of the filters (the error function reflects a ripple square magnitude) is minimized subject to the specification on this weighted absolute error function defined in the corresponding passband and stopband, as well as the stability condition. Since the cost function is nonsmooth and nonconvex, while the constraints are continuous, this kind of optimization problem is a nonsmooth nonconvex continuous functional constrained problem. To address this issue, our previous proposed constraint transcription method is applied to transform the continuous functional constraints to equality constraints. Then the nonsmooth problem is approximated by a sequence of smooth problems and solved via a hybrid global optimization method. The solutions obtained from these smooth problems converge to the global optimal solution of the original optimization problem. Hence, small transition bandwidth filters can be obtained.

Index Terms—Rational IIR filters, constraint transcription method, hybrid global optimization method.

I. INTRODUCTION

Although it is more difficult for rational IIR filters to have linear phase frequency responses when compared to that for finite impulse response (FIR) filters, costs for implementing the rational IIR filters are usually lower than that for the FIR filters at given passband and stopband specifications. Hence, rational IIR filters are preferred in many industrial and engineering applications in which phase responses are not very important \cite{1}-\cite{3}. In particular, in a sigma delta modulator, it consists of a discrete-time filter. Since a sigma delta modulator is operated in an oversampling manner, a narrow band filter is required. As a result, a rational IIR filter is preferred because the cost for employing an FIR filter is too high. Due to the quantization process, the phase information is seriously corrupted by the quantizer. Hence, the phase information cannot be exploited and it is not very important for the design of a sigma delta modulator.

One of the most common methods for designing rational IIR filters is via eigenfilter approaches \cite{4}-\cite{8}, in which optimal solutions can be found by computing the eigenvalues of the error matrices. Another method is via a WISE approach \cite{19}. An optimal solution can be found by computing a gradient of the corresponding cost function. A model matching approach \cite{17} was also proposed. This method is to model rational IIR filters as FIR filters and then minimized the difference between a norm of these two classes of filters. However, since all these methods \cite{4}-\cite{8}, \cite{17}, \cite{19} are based on formulating their design problems as unconstrained optimization problems, the stability, as well as the size of the ripple magnitudes in passbands and stopbands of the filters, are not guaranteed. Moreover, they required phase information for the desired filter responses. In some applications, such as the applications in sigma delta modulators \cite{1}, phase responses are not very important. Imposing extra phase information on desired filter responses may cause degradation on filter performances.

In order to tackle parts of these issues, rational IIR filter design problems are formulated as constrained optimization problems subject to various constraints. These optimization problems are solved via the Gauss-Newton method \cite{18}. However, this method replied on smooth cost functions and is easy to trap at local minima because these optimization problems are not convex. In order to avoid computing the gradients of cost functions, these design problems are formulated as constrained iterative design problems \cite{9}-\cite{16}. Filter coefficients are designed based on initialized denominator coefficients and the iteration of the design process until the denominator coefficients converged. Since these approaches required an initialization of denominator coefficients, the global optimal solutions, as well as the convergence of the iterative process, are not guaranteed.

There were some other methods proposed for designing rational IIR filters, such as via half band filters \cite{20}. However, this approach is not applied if filters are not halfband ones. Another method based on controlling frequency response of filters continuously was proposed \cite{21}. As it is a kind of adaptive filter design techniques, the filters are time-varying.

If only the magnitude response of rational IIR filters is designed, then we can formulate the design problems as optimization problems. The cost of the corresponding optimization problems can be defined as the total weighted absolute error in the passbands and stopbands of the filters, in which the error function reflects the ripple square magnitudes, subject to constraints based on the specification on this weighted absolute error function in the corresponding passbands and stopbands, as well as to a stability condition of the filters. However, this kind of optimization problem is difficult to solve because it involves a nonsmooth nonconvex
cost and continuous functional constraints.

To solve the optimization problems with continuous functional constraints, one may sample these continuous functional constraints and convert to finite discrete constraints [14], [22]. However, it is not guaranteed that solutions obtained satisfy the original continuous functional constraints. Although the difference between the exact upper bounds of discretized constraint functions and that of the corresponding continuous functional constraint functions decrease as the number of grid points increases, the computational complexity increases. To find the global optimal solution of nonconvex problems, one may apply the bridging method [23]. However, this method is applied only for one-dimensional optimization problems.

In this paper, a magnitude design of rational IIR filters is formulated as a nonsmooth nonconvex optimization problem with continuous functional constraints. Our previous proposed constraint transcription method [24] is applied to transform these continuous functional constraints to equality constraints. The global optimal solution can be obtained via the hybrid global optimization method [25]. The obtained numerical experiments show that very small transition bandwidth filters can be obtained.

The outline of this paper is as follows. The problem formulation is presented in Section II. The numerical experiments are shown in Section III. Finally, a conclusion is summarized in Section IV.

II. PROBLEM FORMULATION

Consider a general rational IIR filter with frequency response

\[
H(\omega) = \frac{e^{-j\omega D} \sum_{m=0}^{M} b_m e^{-j\omega m \eta}}{1 + \sum_{n=1}^{N} a_n e^{-j\omega n \eta}}, \tag{1}
\]

where \(j = \sqrt{-1}\), \(D\) relates to the delay of the filter, \(M\) and \(N\) are, respectively, the number of non-zero roots of the polynomials of \(e^{-j\omega}\) in the numerator and denominator, \(b_m\) for \(m = 0, 1, \ldots, M\) and \(a_n\) for \(n = 1, 2, \ldots, N\) are, respectively, the filter coefficients in the numerator and denominator. Solving the optimal filter design problem is equivalent to determine the values of \(a_n\) for \(n = 1, 2, \ldots, N\) and \(b_m\) for \(m = 0, 1, \ldots, M\). It is worth noting that \(D \in \mathbb{N}\) is not important for the magnitude design problem, where \(\mathbb{N}\) denotes the set of all real numbers.

Here, we only consider filters with real coefficients, which are the most usual cases in most applications [1]-[3]. So \(a_n, b_m \in \mathbb{N}\) for \(n = 1, 2, \ldots, N\) and \(m = 0, 1, \ldots, M\). It is worth noting that both the causal and noncausal filters can be designed via the following approach. That means, \(M\) can be greater than, equal or less than \(N\). \(D\) can be positive, zero or negative numbers, and not necessary to be an integer.

Let the desired magnitude response of \(H(\omega)\) be \(\bar{H}(\omega)\),

where \(\bar{H}(\omega) \geq 0 \quad \forall \omega \in [-\pi, \pi]\). We want to achieve

\[
\left| e^{-j\omega D} \sum_{m=0}^{M} b_m e^{-j\omega m \eta} \right|^2 \leq \left(\bar{H}(\omega)\right)^2, \tag{2}
\]

where \(\| \) denotes the modulus of the corresponding complex function. For example, there are many ways to formulate an error function. For example, consider a second order rational IIR filter with \(D = 0\), \(b_0 = 2.816335701763035 \times 10^{-3}\), \(b_1 = 1.877557134508662 \times 10^{-3}\) and \(b_2 = 2.816335701763063 \times 10^{-3}\). The plot of \(E(\omega)\) against \((a_1, a_2)\) at \(\omega = 0\) is shown in Figure 1. It can be seen from the figure that \(E(0)\) is not differentiable along the line \(a_1 + a_2 + 1 = 0\). Besides, since this error function consists of taking the modulus operators inside the square operator, there does not exist any method for solving this kind of nonsmooth problem. Although there are some alternative methods to define the error function so that the error function is smooth, for example,

\[
E(\omega) \equiv \left[ e^{-j\omega D} \sum_{m=0}^{M} b_m e^{-j\omega m \eta} \right]^2 - \bar{H}(\omega)^2 + \sum_{n=1}^{N} \left| a_n e^{-j\omega n \eta}\right|^2. \tag{3}
\]

In this case, \(E(\omega)\) is differentiable with respect to the filter coefficients. Let the filter coefficients in the numerator and denominator be, respectively,

\[x_n = [b_0, b_1, \ldots, b_d]^T\] \tag{5}

and

\[x_d = [a_1, a_2, \ldots, a_n]^T,\] \tag{6}

where the superscript \(^T\) denotes the transpose. Define

\[\eta_n(\omega) = [1, e^{-j\omega}, \ldots, e^{-j\omega D}]^T,\] \tag{7}

and

\[\eta_d(\omega) = [e^{-j\omega}, e^{-j2\omega}, \ldots, e^{-jN\omega}]^T,\] \tag{8}

then

\[
E(\omega) = \left[\eta_n(\omega)^T x_n - \bar{H}(\omega)^2 \right] + \left(\eta_d(\omega)^T x_d \right)^2. \tag{9}
\]

Denote the passband and stopband of the filter be, respectively, \(B_p\) and \(B_s\). In order to design a rational IIR filter having good frequency selectivity, total ripple energy in both the passband and stopband of the filter should be minimized. Hence, we define a cost function as follows:

\[
J(x_n, x_d) = \int_{\omega \in [B_p \cup B_s]} W(\omega) E(\omega) d\omega, \tag{10}
\]
where \( W(\omega) > 0 \) \( \forall \omega \in B_p \cup B_s \) is a weighting function. This cost function can represent the total weighted absolute ripple square magnitude in the passband and stopband of the filter because \( |E(\omega)| \) represents the absolute ripple square magnitude. It is worth noting that \( |E(\omega)| \) is still a nonsmooth function. However, since the modulus operator is taken outside a smooth function, this kind of optimization problem can be solved via the constraint transcription method [24] and will be discussed below.

Although the cost function can be used to minimize the total weighted absolute ripple square magnitude in the passband and stopband of the filter, there may have a very serious overshoot. Hence, a specification based on the weighted absolute ripple square magnitude is defined as follows:

\[
\tilde{W}(\omega)|E(\omega)| \leq \tilde{\delta}(\omega) \quad \forall \omega \in B_p \cup B_s, \tag{11a}
\]

where \( \tilde{W}(\omega) > 0 \) \( \forall \omega \in B_p \cup B_s \) is a weighting function and \( \tilde{\delta}(\omega) > 0 \) \( \forall \omega \in B_p \cup B_s \) relates to the allowable weighted absolute ripple square magnitude in both the passband and stopband of the filter. This constraint is equivalent to:

\[
\tilde{W}(\omega)|E(\omega)| \leq \tilde{\delta}(\omega) \quad \forall \omega \in B_p \cup B_s, \tag{11b}
\]

and

\[
-\tilde{\delta}(\omega) \leq \tilde{W}(\omega)|E(\omega)| \quad \forall \omega \in B_p \cup B_s. \tag{11c}
\]

In order to guarantee that the designed filter is stable, we need to satisfy the following condition:

\[
\Re\left(\{i(\Omega(\omega))\} \times x_j\right) < 0 \quad \forall \omega \in [-\pi, \pi]. \tag{12}
\]

Hence, the rational IIR filter design problem can be formulated as the following optimization problem:

**Problem (P)**

\[
\min_{\{x_j\}} \tilde{J}(x_j) = \int_{B_p \cup B_s} W(\omega)|E(\omega)|d\omega, \tag{13a}
\]

subject to \( \tilde{g}_l(x_j, \omega) = \tilde{W}(\omega)|E(\omega)| - \tilde{\delta}(\omega) = 0 \quad \forall \omega \in B_p \cup B_s \), \( \tilde{g}_l(x_j, \omega) = -\tilde{W}(\omega)|E(\omega)| - \tilde{\delta}(\omega) = 0 \quad \forall \omega \in B_p \cup B_s \), \( \tilde{g}_l(x_j, \omega) = \Re\left(\{i(\Omega(\omega))\} \times x_j\right) = 0 \quad \forall \omega \in [-\pi, \pi]. \tag{13b}
\]

It is worth noting that problem \( \tilde{P} \) consists of a nonsmooth nonconvex cost and continuous functional constraints. This kind of optimization problem is difficult to solve. In order to tackle this issue, our proposed constraint transcription method [24] is applied to convert these continuous functional constraints to equality constraints and discussed as follows. Since

\[
\max\{\tilde{g}_l(x_j, \omega)\} \omega = \begin{cases} 0 & \tilde{g}_l(x_j, \omega) \leq 0, \\
\text{positive value} & \tilde{g}_l(x_j, \omega) > 0 \end{cases}, \tag{14}
\]

by defining

\[
\tilde{g}_l(x_j, \omega) = \int_{B_p \cup B_s} \max\{\tilde{g}_l(x_j, \omega)\} d\omega, \tag{15}
\]

then we have:

\[
\tilde{g}_l(x_j, \omega) = \begin{cases} 0 & \forall \omega \in B_p \cup B_s, \tilde{g}_l(x_j, \omega) \leq 0, \\
\text{positive value} & \exists \omega \in B_p \cup B_s, \tilde{g}_l(x_j, \omega) > 0 \end{cases}. \tag{16}
\]

Hence, the satisfaction of the constraint defined by

\[
\forall \omega \in B_p \cup B_s \quad \tilde{g}_l(x_j, \omega) \leq 0 \quad \text{is equivalent to the equality constraint defined by} \quad \tilde{g}_l(x_j, \omega) = 0. \quad \text{Since}
\]

\[
(\max\{\tilde{g}_l(x_j, \omega)\} \omega)^2 = \begin{cases} 0 & \tilde{g}_l(x_j, \omega) \leq 0, \\
\text{positive value} & \tilde{g}_l(x_j, \omega) > 0 \end{cases}, \tag{17}
\]

\[
\forall \omega \in \max\{\tilde{g}_l(x_j, \omega)\} \omega = 0 \quad \text{when} \quad \tilde{g}_l(x_j, \omega) = 0, \quad \text{so} \quad \forall \omega \in \max\{\tilde{g}_l(x_j, \omega)\} \omega = 0. \quad \text{Moreover, since}
\]

\[
2\tilde{g}_l(x_j, \omega) = \max\{\tilde{g}_l(x_j, \omega)\} \omega = \tilde{g}_l(x_j, \omega) \quad \text{when} \quad \tilde{g}_l(x_j, \omega) = 0 \quad \text{and}
\]

\[
2\max\{\tilde{g}_l(x_j, \omega)\} \omega = \max\{\tilde{g}_l(x_j, \omega)\} \omega = \tilde{g}_l(x_j, \omega) \quad \text{by definition}
\]

\[
\tilde{g}_l(x_j, \omega) = \int_{B_p \cup B_s} \max\{\tilde{g}_l(x_j, \omega)\} d\omega, \tag{18}
\]

we have

\[
\forall \omega \in \max\{\tilde{g}_l(x_j, \omega)\} \omega = 0 \quad \text{when} \quad \tilde{g}_l(x_j, \omega) = 0 \quad \text{and}
\]

\[
2\max\{\tilde{g}_l(x_j, \omega)\} \omega = \max\{\tilde{g}_l(x_j, \omega)\} \omega = \tilde{g}_l(x_j, \omega) \quad \text{by definition}
\]

\[
\forall \omega \in \max\{\tilde{g}_l(x_j, \omega)\} \omega = 0. \tag{19}
\]

As a result, we have

\[
\forall \omega \in \max\{\tilde{g}_l(x_j, \omega)\} \omega = 2 \int_{B_p \cup B_s} \max\{\tilde{g}_l(x_j, \omega)\} d\omega, \tag{20}
\]

Similarly, by defining

\[
\tilde{g}_l(x_j, \omega) = \int_{B_p \cup B_s} \max\{\tilde{g}_l(x_j, \omega)\} d\omega \quad \text{and}
\]

\[
\tilde{g}_l(x_j, \omega) = \int_{[-\pi, \pi]} \max\{\tilde{g}_l(x_j, \omega)\} d\omega, \tag{21}
\]

we have

\[
\forall \omega \in \max\{\tilde{g}_l(x_j, \omega)\} \omega = 2 \int_{B_p \cup B_s} \max\{\tilde{g}_l(x_j, \omega)\} d\omega \quad \text{when}
\]

\[
\forall \omega \in \max\{\tilde{g}_l(x_j, \omega)\} \omega = 2 \int_{[-\pi, \pi]} \max\{\tilde{g}_l(x_j, \omega)\} d\omega. \tag{22}
\]

As \( \tilde{g}_l(x_j, \omega) \), \( \tilde{g}_l(x_j, \omega) \), and \( \tilde{g}_l(x_j, \omega) \) are continuously differentiable with respect to \( (x_j, \omega) \) and \( x_j \), respectively, the optimization problem \( \tilde{P} \) is equivalent to the following optimization problem, denoted as problem \( P \):

**Problem (P)**

\[
\min_{\{x_j\}} \tilde{J}(x_j) = \int_{B_p \cup B_s} W(\omega)|E(\omega)|d\omega, \tag{23a}
\]

subject to \( \tilde{g}_l(x_j, \omega) = 0 \), \( \tilde{g}_l(x_j, \omega) = 0 \), \( \tilde{g}_l(x_j, \omega) = 0. \tag{23b}
\]

However, problem \( P \) is still a nonsmooth nonconvex problem, where the nonsmooth function appears in the cost. Thus, standard optimization software packages, such as Matlab Optimization toolbox, in theory, cannot be applied directly. To overcome this difficulty, the nonsmooth absolute function \( |E(\omega)| \) \( \forall \omega \in B_p \cup B_s \) is handled in the following manner. \( \forall \omega \in B_p \cup B_s \) and \( \varepsilon > 0 \), consider the following function:
\[
E_\epsilon(\omega) = \begin{cases} 
\frac{|E(\omega)|}{\epsilon}, & |E(\omega)| \geq \frac{\epsilon}{2}, \\
\left(\frac{|E(\omega)|^2}{\epsilon} + \frac{\epsilon}{4}\right), & |E(\omega)| < \frac{\epsilon}{2}.
\end{cases}
\] (26)

Clearly, the function \(E_\epsilon(\omega)\) possesses the following properties:

i) \(\forall \omega \in B_\rho \cup B_{\epsilon}, E_\epsilon(\omega)\) is continuously differentiable with respect to \((x_n, x_j)\).

ii) \(\forall (x_n, x_j)\) and \(\forall \omega \in B_\rho \cup B_{\epsilon}, E_\epsilon(\omega) \geq |E(\omega)|.\)

iii) \(\forall (x_n, x_j)\) and \(\forall \omega \in B_\rho \cup B_{\epsilon}, E_\epsilon(\omega) - |E(\omega)| \leq \frac{\epsilon}{2}.\)

iv) \(\forall (x_n, x_j), (x_n^*, x_j^*)\) minimizes \(|E(\omega)|\) if and only if it minimizes \(E_\epsilon(\omega)\).

By virtue of these properties, \(E_\epsilon(\omega)\) is an ideal approximation of the nonsmooth function \(|E(\omega)|\). By replacing \(E_\epsilon(\omega)\) for \(|E(\omega)|\) in the cost function (25a), we obtain

\[
J_\epsilon(x_n, x_j) = \int_{\beta_n \cup \beta_j} W(\omega)E_\epsilon(\omega)d\omega,
\]
(27)

where the function \(J_\epsilon(x_n, x_j)\) is now continuously differentiable with respect to \((x_n, x_j)\) \(\forall \epsilon > 0\). Hence, we can approximate the nonsmooth optimization problem \(P\) by a smooth optimization problem, where the cost function (27) is to be minimized subject to the equality constraints defined in (25b), (25c) and (25d). Let this optimization problem be referred to as problem \(Q_\epsilon\) as follows:

**Problem (\(Q_\epsilon\))**

\[
\min_{(x_n, x_j)} J_\epsilon(x_n, x_j) = \int_{\beta_n \cup \beta_j} W(\omega)E_\epsilon(\omega)d\omega,
\]
(28a)

subject to \(g_1(x_n, x_j) = 0,\)

\(g_2(x_n, x_j) = 0,\)

\(g_3(x_n, x_j) = 0.\)

(28b) \(\) \(\) \(\)

(28c) \(\) \(\) \(\)

(28d)

\(\forall \epsilon > 0\), let \((x_n^*, x_j^*)\) be an optimal solution to the approximate problem \(Q_\epsilon\). Furthermore, let \((x_n^*, x_j^*)\) be an optimal solution to the original problem \(P\). Then, there are two questions to be answered. First, how much does \(J_\epsilon(x_n, x_j)\) differ from \(\tilde{J}(x_n^*, x_j^*)\)? Second, what is the relationship between \(\{x_n^*, x_j^*\}\) and \(\{x_n^*, x_j^*\}\)? To address the first question, we have the following theorem:

**Theorem 1**

Let \((x_n^*, x_j^*)\) and \((x_n^*, x_j^*)\) be, respectively, optimal solutions to problems \(Q_\epsilon\) and \(P\). Then

\[
0 \leq J_\epsilon(x_n, x_j) - \tilde{J}(x_n^*, x_j^*) \leq \frac{\epsilon}{4} \int_{\beta_n \cup \beta_j} W(\omega)d\omega.
\]
(29)

**Proof**

By virtue of property (ii) of the function \(E_\epsilon(\omega)\), we have

\[
J_\epsilon(x_n^*, x_j^*) = \tilde{J}(x_n^*, x_j^*) \leq \min_{(x_n, x_j)} \tilde{J}(x_n, x_j) = \tilde{J}(x_n^*, x_j^*).
\]

Hence,

\[
J_\epsilon(x_n^*, x_j^*) - \tilde{J}(x_n^*, x_j^*) \geq 0. \tag{30}
\]

Next, from property (iii) of the function \(E_\epsilon(\omega)\), we have

\[
0 \leq J_\epsilon(x_n^*, x_j^*) - \tilde{J}(x_n^*, x_j^*) \leq \frac{\epsilon}{4} \int_{\beta_n \cup \beta_j} W(\omega)d\omega. \tag{31}
\]

But

\[
J_\epsilon(x_n^*, x_j^*) \leq \tilde{J}(x_n^*, x_j^*), \tag{32}
\]

so we have

\[
J_\epsilon(x_n^*, x_j^*) \leq \frac{\epsilon}{4} \int_{\beta_n \cup \beta_j} W(\omega)d\omega. \tag{33}
\]

Hence, this completes the proof.

To address the second question, we have the following theorem:

**Theorem 2**

Let \(\{x_n^*, x_j^*\}\) be a sequence of optimal solutions to the corresponding sequence of approximate problems \(\{Q_\epsilon\}\). Then an accumulation point exists and it is an optimal solution to the original problem \(P\).

**Proof**

Since \(J_\epsilon(x_n, x_j)\) is continuous with respect to both \((x_n, x_j)\) and \(\epsilon\), \(\{x_n^*, x_j^*\}\) is a convergent sequence and there exists an accumulation point \((x_n, x_j)\) and a subsequence of the sequence \(\{x_n^*, x_j^*\}\), which is again denoted by the original sequence, such that \(\|x_n^*, x_j^* - (x_n, x_j)\| \rightarrow 0\) as \(\epsilon \rightarrow 0\), where \(\|\cdot\|\) denotes the Euclidean norm. By Theorem 1, as

\[
0 \leq J_\epsilon(x_n, x_j) - \tilde{J}(x_n^*, x_j^*) \leq \frac{\epsilon}{4} \int_{\beta_n \cup \beta_j} W(\omega)d\omega,
\]
(28d)

we have

\[
J_\epsilon(x_n, x_j) \rightarrow \tilde{J}(x_n^*, x_j^*) \quad \text{as} \quad \epsilon \rightarrow 0. \]

Hence, this completes the proof.

Based on these two theorems, problem \(\tilde{P}\) can be solved via solving a sequence of approximate problems \(\{Q_\epsilon\}\) by an iterative technique stated in [24] with decreasing value of \(\epsilon\) and the algorithm is summarized as follows:

**Algorithm 1**

Step 0: Initialize \(\epsilon_i > 0\) and \(k = 1\).

Step 1: Solve problem \(Q_{\epsilon_i}\) by hybrid global optimization method discussed in [25]. Denote the solution by \((x_n^*, x_j^*)\).

Step 2: Set \(\epsilon_{i+1} = \frac{\epsilon_i}{L}\), where \(L > 1\) is a prespecified number.

Step 3: If \(\|x_n^*, x_j^* - (x_n^*, x_j^*)\| \leq \beta\), where \(\beta > 0\) is a prescribed small number depending on the accuracy desired, then stop. Otherwise, set \(k = k + 1\) and go to Step 1.

In Algorithm 1, we can see that \(\epsilon_k \rightarrow 0\) as \(k \rightarrow +\infty\) because \(L > 1\). Hence, according to Theorem 2, we can see that the solution obtained \(\{x_n^*, x_j^*\}\) converges to the global optimal...
solution of problem $P$.

There are three parameters in the Algorithm 1, namely, $\varepsilon_1$, $L$ and $\beta$. $\varepsilon_1$ determines how close the approximate problem $Q_{\varepsilon_1}$ and the original problem $P$. The smaller the value of $\varepsilon_1$, the more close will be the problem $Q_{\varepsilon_1}$ to problem $P$, and hence less number of iterations of Algorithm 1 is required. However, the cost function becomes less smooth. $L$ also determines the number of iterations required. Similarly, the larger the value of $L$, the less number of iterations is required, but the cost function becomes less smooth even for small values of $k$. Practically, we find that if $\varepsilon_1 \approx 10^{-3}$ and $L \approx 10$, then the number of iterations required and the cost function will be, respectively, small and smooth enough for most optimization problems [24]. $\beta$ controls the acceptable precision of the obtained solution. The smaller the value of $\beta$, the more accurate is the solution. However, the cost function becomes less smooth. Hence less number of iterations of Algorithm 1 is required.

The hybrid global optimization method [25] guarantees the global optimal solution, the rate of convergence of the algorithm depends on the initial choice of $(x_{*a}^l, x_{*d}^l)$. In order to have a fast rate of convergence, $(x_{*a}^l, x_{*d}^l)$ should be selected as close to the global optimal solution. For the rational IIR filter design problems, the solutions obtained using the elliptic filter design method may be a good choice of this initial guess because the solution obtained by the elliptic filter design method is a suboptimal solution.

### III. Numerical Experiments

In this paper, a unit DC gain highpass halfband filter, that is $H(\omega)=\begin{cases} 1 & \omega \in B_{r(i)} \\ 0 & \omega \in B_{s(i)} \end{cases}$, where $B_{r(i)}=\left[\pi, \frac{\pi}{2}-\Delta\right] \cup \left[\frac{\pi}{2}+\Delta, \pi\right]$ and $B_{s(i)}=\left[-\frac{\pi}{2}+\Delta, \frac{\pi}{2}-\Delta\right]$, in which $2\Delta$ denotes the transition bandwidth of the filter, is designed for the illustration of the effectiveness of the proposed method. Halfband filters with unit DC gain are selected for illustration because they are found in many engineering applications, such as in wavelet applications. For other filters with different DC gains, such as lowpass filters, bandpass filters, band reject filters, notch filters, highpass filters with other passbands and stopbands, the design method can be applied directly.

To evaluate the effectiveness of the proposed method, our result is compared with the one obtained using the iterative approach [9] and that using an elliptic filter. These two design methods are chosen for comparisons because that using the iterative approach [9] would be of great value for the readers working in this field, while that using an elliptic filter because the design objectives are the same. For the iterative design approach, it was reported in [9] that the magnitude response of the filter in the passband and stopband is approximately bounded by, respectively, 0.1406dB and $-27.8974$dB, if the filter order is 14. The corresponding magnitude response is shown in Figure 2a, the zoom in the passband, stopband and the transition band are shown in, respectively, Figure 3a, 4a and 5a. It can be seen from the figure that the transition bandwidth is 0.2756. To compare this result with that using an elliptic filter, we use the Matlab function “ellip” to implement the filter and set the filter order, as well as the passband and stopband specifications same as that reported in [9]. The corresponding magnitude response is shown in Figure 2b, while the zoom in the passband, stopband and the transition
band are shown in, respectively, Figure 3b, 4b and 5b. It can be seen from the figure that the transition bandwidth of the filter is $1.382 \times 10^{-3}$. For our design, we set both $W(\omega) = 1$ and $\tilde{W}(\omega) = 1$ for simplicity reason. In fact, other positive weighting functions can be applied directly. For the parameters in the algorithm, we choose $\varepsilon_1 = 10^{-3}$, $L = 10$, $\beta = 10^{-6}$ and $(\mathbf{x}_{1,n}, \mathbf{x}_{2,n})$ as the elliptic filter coefficients as discussed in Section II. After running three iterations, the optimization algorithm terminates because the stopping criterion satisfies. The magnitude response of the filter is shown in Figure 2c, while the zoom in the passband, stopband and the transition band are shown in, respectively, Figure 3c, 4c and 5c. The phase responses and the pole-zero plots of these designed filters are shown in, respectively, Figure 6 and Figure 7, while the filter coefficients are listed in Table 1. It can be checked that the transition bandwidth of our designed filter is $6.258 \times 10^{-4}$, which is $0.2271\%$ of that using the iterative approach and $45.2822\%$ of that using an elliptic filter. Our result performs much better than that using an iterative design approach [9] because this design approach requires a desired phase response and this information is necessary and cannot be removed from the design procedure. By an extra imposing a desired phase response on the design, the magnitude response will be trade-off. Our result also performs better than that using an elliptic filter because the one obtained using an elliptic filter is a local optimal solution, while our result is a global optimal solution.

It is worth noting that our proposed design method can be applied to a strong specification if a solution exists. Since there is a tradeoff between a filter length and a reduction on the passband and stopband ripple magnitudes, there does not exist any design that gives a filter with very short filter length but very large reduction on the passband and stopband ripple magnitudes. If there exists a stable filter such that it satisfies the specifications on the passband and stopband ripple magnitudes at a relatively short filter length, then a global optimal solution for the optimization problem exists. Since our proposed design method guarantees to obtain the global optimal solution, our proposed design method works properly under a strong specification if a solution exists.

In order to test the rate of convergence of the algorithm, four different initial guesses of $(\mathbf{x}_{1,n}, \mathbf{x}_{2,n})$ are used. These four initial guesses give the same global optimal solution. The design time for choosing $(\mathbf{x}_{1,n}, \mathbf{x}_{2,n})$ as the elliptic filter coefficients is 2 seconds, that as the Chebyshev Type I filter coefficients is 10 minutes, that as the Chebyshev Type II filter coefficients is 15 minutes, and that as the one obtained using the iterative approach [9] is 1.5 hours, where all numerical experiments are running using a PC with Pentium 1.2GHz CPU and 256M bytes DDRAM. From these results, we can conclude that the required design time will be shorter if the initial guess is closer to the global optimal solution.

IV. CONCLUSION

The main contribution of this paper is to formulate an optimum rational IIR filter design problem as a nonsmooth nonconvex optimization problem subject to continuous functional constraints. Our previous proposed constraint transcription method is applied to transform the continuous functional constraints to equality constraints. A hybrid global optimization method is applied to find the global optimal solution. According to our numerical experiments, small transition bandwidth filters are obtained.

ACKNOWLEDGEMENT

The work obtained in this paper was supported by a research grant (project number G-YD26) from The Hong Kong Polytechnic University, the Centre for Multimedia Signal Processing, The Hong Kong Polytechnic University, the CRGC grant (project number PolyU 5105/01E) from the Research Grants Council of Hong Kong, as well as a research grant from Australian Research Council.

REFERENCES


Figure 1: Plot of $E(0)$ against different denominator coefficients. It can be seen that $E(0)$ is not differentiable with respect to the denominator coefficients.

Figure 2. Magnitude responses of various filters. (a) Filter designed via the iterative approach [9]. (b) Filter designed via an elliptic filter. (c) Filter designed via our proposed approach. All the passband and stopband ripple magnitudes are the same.

Figure 3. Zoom of the magnitude responses in the passband. (a) Filter designed via the iterative approach [9]. (b) Filter designed via an elliptic filter. (c) Filter designed via our proposed approach. All the passband and stopband ripple magnitudes are the same.

Figure 4. Zoom of the magnitude responses in the stopband. (a) Filter designed via the iterative approach [9]. (b) Filter designed via an elliptic filter. (c) Filter designed via our proposed approach. All the stopband ripple magnitudes are the same.

Figure 5. Zoom of the magnitude responses in the transition band. (a) Filter designed via the iterative approach [9]. The transition bandwidth of the filter is 0.2756. (b) Filter designed via an elliptic filter. The transition bandwidth of the filter is $1.382\times 10^{-3}$. (c) Filter designed via our proposed approach. The transition bandwidth of the filter is $6.258\times 10^{-4}$. 
Figure 6. Phase responses of various filters. (a) Filter designed via the iterative approach [9]. (b) Filter designed via an elliptic filter. (c) Filter designed via our proposed approach. The phase response of the filter designed using the iterative approach [9] is approximately linear, while those of via an elliptic filter and our design method are nonlinear.

Figure 7. Pole-zero plots of various filters. (a) Filter designed via the iterative approach [9]. (b) Filter designed via an elliptic filter. (c) Filter designed via our proposed approach. The poles and zeros of the filter designed using the iterative approach [9] are spread over a wide region in the complex plane, while those of the filters designed via an elliptic filter and our design method are located in a small region in the complex plane.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00216452</td>
<td>1</td>
</tr>
<tr>
<td>0.01422263</td>
<td>0.85183505</td>
</tr>
<tr>
<td>0.01070862</td>
<td>1.40277206</td>
</tr>
<tr>
<td>-0.00173896</td>
<td>1.15151797</td>
</tr>
<tr>
<td>-0.00056474</td>
<td>0.95948743</td>
</tr>
<tr>
<td>0.01648599</td>
<td>0.81499579</td>
</tr>
<tr>
<td>0.01450372</td>
<td>0.68685173</td>
</tr>
<tr>
<td>-0.01424925</td>
<td>0.52355787</td>
</tr>
<tr>
<td>-0.01101010</td>
<td>0.35584783</td>
</tr>
<tr>
<td>0.04991283</td>
<td>0.23561367</td>
</tr>
<tr>
<td>0.04637753</td>
<td>0.15595331</td>
</tr>
<tr>
<td>-0.19765241</td>
<td>0.08260605</td>
</tr>
<tr>
<td>0.33139538</td>
<td>0.01749834</td>
</tr>
<tr>
<td>-0.25569216</td>
<td>-0.01347975</td>
</tr>
<tr>
<td>0.13691517</td>
<td>-0.01109415</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Numerator coefficients of an elliptic filter</th>
<th>Denominator coefficients of an elliptic filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.14216384559483</td>
<td>1</td>
</tr>
<tr>
<td>-0.33386656595210</td>
<td>0.12993638934053</td>
</tr>
<tr>
<td>1.17827982115402</td>
<td>5.51295641219684</td>
</tr>
<tr>
<td>-2.02956611583674</td>
<td>0.67443232223880</td>
</tr>
<tr>
<td>3.90213338343642</td>
<td>12.68806334048509</td>
</tr>
</tbody>
</table>

Table 1. Filter coefficients of various filters.