Minimax Design of IIR Digital Filters Using Iterative SOCP

Aimin Jiang, Student Member, IEEE, and Hon Keung Kwan, Senior Member, IEEE

Abstract—In this paper, a novel method for IIR digital filter design using iterative second-order cone programming (SOCP) is proposed under the minimax criterion. The convex relaxation technique is utilized to transform the original nonconvex design problem into an SOCP problem. By solving the relaxed problem, the lower and upper bounds on the optimal value of the original problem can be obtained. In order to reduce the discrepancy between the original and relaxed design problems, an iterative procedure is developed. At each iteration, a linear constraint is further incorporated to guarantee the convergence of the iterative procedure. In practice, the convergence speed can be further improved by introducing a soft threshold variable in this linear constraint. Accordingly, a regularization term is incorporated in the objective function of the design problem at each iteration. The stability of the designed filters can be ensured by a new positive realness based linear constraint. Several examples are presented to demonstrate the effectiveness of the proposed method.

Index Terms—Convex relaxation, infinite impulse response (IIR) digital filters, minimax design, second-order cone programming (SOCP).

I. INTRODUCTION

The design task of IIR digital filters is to approximate a given ideal frequency response by a stable IIR digital filter under some design criterion. If both magnitude and phase (or group delay) responses are concerned, an IIR digital filter design problem is essentially a nonconvex optimization problem due to the presence of the denominator of the transfer function. The nonconvexity property involves two aspects. First, it is hard to transform the approximation error to be minimized into a convex function, which means the globally optimal solution cannot be always obtained. Second, the stability domain cannot be strictly expressed as a convex set if the denominator order is larger than 2. Recently, a number of algorithms [1]–[20] have been proposed to design IIR filters. Some algorithms [1]–[3] design IIR digital filters in an indirect way, that is, an FIR digital filter satisfying the specifications is designed first and then approximated by a reduced-order IIR filter. Generally speaking, it is hard to design IIR filters with accurate cutoff frequencies using this indirect design strategy. Some other algorithms [4]–[8] employ iterative procedures using the Steiglitz-McBride (SM) scheme [27] under different criteria. At each iteration, the denominator of an approximation error is replaced by the one obtained at the previous iteration. Accordingly, the approximation error can be transformed into a convex quadratic function of the current filter coefficients. Then, a linear programming (LP) [6], quadratic programming (QP) [5], [7], [8], or second-order cone programming (SOCP) [4] design problem can be readily obtained and efficiently solved. The obtained filter coefficients are taken into the next iteration until a predetermined stopping condition is satisfied. The argument principle based stability constraint [4] and the positive realness based stability constraint [5]–[8] are employed to guarantee the stability of designed IIR filters. In [9], by introducing an inverse filter corresponding to the denominator of an IIR filter, the numerator and denominator designs can be decoupled into two separate optimization problems, and similar reweighting technique used in the SM design methods can be applied in the iterative procedure to achieve denominator coefficients. Another design strategy [10]–[12] approximates the frequency response of an IIR digital filter by its first-order Taylor series with respect to numerator and/or denominator coefficients. Using the linearized frequency response, an approximation error can be recast in a convex form and minimized during the iterative procedure. In [10], the denominator polynomial is factorized as a product of second-order sections and a first-order section if the denominator order is odd. Then, the stability of designed filters can be guaranteed by a set of linear inequality constraints in terms of these factorized denominator coefficients. The Gauss-Newton (GN) method is employed by [11] to solve the weighted least-squares (WLS) design problem. The Rouché’s theorem based stability constraint is adopted in [11], which is less conservative than the positive realness based stability constraint mentioned earlier. In [12], using the approximate error function, the descent direction for the denominator is first determined, and the best numerator for the updated denominator can then be found. The stability of designed filters is ensured by selecting a sufficiently small step size at each iteration, such that all the updated poles still lie inside the stability domain. A multistage design scheme is proposed in [13] for the WLS IIR filter design, in which the SM, GN, and classical descent methods are used in succession in order to achieve better designs. A new stability constraint in terms of positive realness is utilized in this algorithm. In [14], the minimax design problem is formulated as a generalized eigenvalue problem by using the Remez multiple exchange algorithm. The idea behind the design method [14] is based on a sufficient condition for the optimal rational approximation, which states that an approximation error has a specific number of extreme points over the frequency bands of interest. The Remez exchange algorithm is also employed by [15] to design...
equiripple IIR filters. However, the transfer function of an IIR filter in [15] is in the form of a parallel connection of two allpass filters. Although the effectiveness of these iterative algorithms has been demonstrated by a number of examples in literature, so far the convergences of these iterative procedures have not been strictly guaranteed. Some nonsequential design strategies [16], [17] are also utilized to design IIR digital filters. In [16], the objective function of an unconstrained WLS design problem linearly combines the pure WLS approximation error and some time-domain components, which are used to control the poles’ positions. Using the general purpose optimization procedures, the design problem can be reliably solved. However, the performance of this algorithm depends on the selection of starting point.

Recently, convex optimization has been applied to FIR [21]–[23], allpass [24], [25], and IIR [1], [4]–[8], [10], [13], [17]–[20] digital filter designs. One of the most important advantages of convex optimization is that if a design problem can be equivalently formulated as a convex optimization problem, any local optimal design is also the global optimum. Furthermore, convex optimization problems can be efficiently and reliably solved by interior-point methods [26]. However, in general, an IIR digital filter design problem cannot be directly cast as a convex form. Therefore, convex optimization has to be combined with some other approximation techniques and local search methods. In [17], the denominator of the approximation error in the minimax sense is neglected. By using this simplified approximation error, the design problem formulated in an LP form can then be efficiently solved. However, the obtained filter is essentially not a minimax solution. Some algorithms [18]–[20] utilize the convex relaxation techniques to design IIR filters. In this design strategy, the original nonconvex design problem is first relaxed into a convex optimization problem. Based on the relaxed design problem, iterative procedures can be further applied to reduce the discrepancy between the original and relaxed design problems.

In this paper, we are going to develop a new iterative design procedure in the minimax sense. The most important advantage over other iterative methods described above is that the convergence of the proposed iterative procedure can be guaranteed. A preliminary version of this paper has been presented in [19]. The rest of the paper is organized as follows. The original design problem is first presented in Section II. Then, the convex relaxation technique is introduced to transform the nonconvex problem into a convex form. An iterative procedure is finally developed in Section II. Design examples and conclusions are presented in Section IV and Section V, respectively.

II. MINIMAX DESIGN METHOD

A. Problem Formulation

Let $D(\omega)$ denote a prescribed ideal frequency response over $[0, \pi]$. The transfer function of an IIR digital filter is defined by

$$H(z) = \frac{P(z)}{Q(z)} = \frac{\sum_{n=0}^{N} p_n z^{-n}}{1 + \sum_{m=1}^{M} q_m z^{-m}} = \frac{p^T \varphi_N(z)}{q^T \varphi_M(z)}$$

(1)

where $p = [p_0, p_1, \ldots, p_N]^T$, $q = [q_0, q_1, \ldots, q_M]^T (q_0 = 1)$, $\varphi_N(z) = [1, z^{-1}, \ldots, z^{-N}]^T$, and the superscript $T$ represents the transpose operation of a vector or matrix. In this paper, all the filter coefficients $p_n (n = 0, 1, \ldots, N)$ and $q_m (m = 0, 1, \ldots, M)$ are real values. For convenience in the latter discussion, a weighted complex error is first defined by

$$E(\omega) = W(\omega)\left[\left|D(\omega) - H(e^{j\omega})\right|ight], \forall \omega \in \Omega_I$$

(2)

where $W(\omega)$ is a given nonnegative weighting function, and $\Omega_I$ is the union of frequency bands of interest within $[0, \pi]$.

The design problem of an IIR digital filter in the (weighted) minimax sense can be strictly expressed as

$$\min_{\mathbf{z}} \max_{\omega \in \Omega_I} |E(\omega)|$$

(3)

where $\mathbf{z} = [q^T, p^T]^T (x_0 = 1)$. For clarity, the design problem (3) does not include any explicit stability constraint, which will be incorporated later. By introducing an auxiliary variable $\delta$, the design problem (3) can be formulated as

$$\min_{\mathbf{z}} \delta$$

(4)

s.t. $x_0 = 1$

$$|E(\omega)Q(e^{j\omega})|^2 = \left|\mathbf{c}^T(\omega)\mathbf{x}\right|^2 \leq \delta \left|\mathbf{c}(e^{j\omega})\right|^2$$

(4a)

$$\forall \omega \in \Omega_I$$

where

$$\mathbf{c}_M(\omega) = W(\omega)\left[\Re\{D(\omega)\varphi_M(e^{j\omega})\} - \Im\{D(\omega)\varphi_M(e^{j\omega})\}\right]$$

(5)

$$\mathbf{f}_M(\omega) = \left[\Re\{\varphi_M(e^{j\omega})\} \quad \Im\{\varphi_M(e^{j\omega})\}\right]$$

(6)

In (4b), $\| \cdot \|$ denotes the Euclidean norm of a vector. In (5) and (6), $\Re\{\cdot\}$ and $\Im\{\cdot\}$ are used, respectively, to represent the real and imaginary parts of a complex variable. Note that in the design problem (4), the term $|E(\omega)|$ in (3) has been replaced by its squared value. Since the solution of (3) is also optimal to (4) and vice versa, these two design problems are essentially equivalent to each other.

B. Convex Relaxation

Note that only the magnitude of the denominator is required on the right hand side of the inequality constraint (4b), which is to be used to reformulate the design problem.
It can be verified that [28]

\[
Q(z)Q(z^{-1}) = \left( \sum_{m=0}^{M} q_m z^{-m} \right) \left( \sum_{m=0}^{M} q_m z^{m} \right)
= \sum_{m=-M}^{M} d_m z^{-m}
= R(z)
\]

where the polynomial coefficients of \(R(z)\) are defined by

\[
d_m = d_{-m} = \sum_{i=0}^{M-m} q_d i m, \quad m = 0, 1, \ldots, M.
\]

(7)

It is well known that \(\{d_{im}, m = -M, \ldots, -1, 0, 1, \ldots, M\}\) is an autocorrelation sequence. Some important properties can be directly derived from (8).

1) \(d_0 = \sum_{m=0}^{M} q_m^2 = ||q||^2\).

2) \(d_{-m} = q_0 q_m = q_m\) where \(q_0 = 1\).

3) \(d_{im}^2 = \left| q_m, m = 1, 2, \ldots, M \right| \cdot \left| q_m, m = 1, 2, \ldots, M \right| \cdot \left| q_m, m = 1, 2, \ldots, M \right|^T\).

The first inequality follows from the Cauchy-Schwarz inequality.

By defining \(d = [d_0, d_1, d_2, \ldots, d_M]^T\) and \(s(\omega) = [1, 2 \cos(\omega), \ldots, 2 \cos(M\omega)]^T\), and evaluating (7) on the unit circle, we have

\[
|Q(e^{j\omega})|^2 = \|f^T(\omega)q\|^2
= d^T s(\omega)
= R(e^{j\omega}).
\]

Using (9), the constraint (4b) can be cast as a hyperbolic constraint by replacing \(|Q(e^{j\omega})|^2\) by \(d^T s(\omega)\) on the right hand side of the inequality. It is known that a hyperbolic constraint can be further transformed into a second-order cone (SOC) constraint [29]. On the other hand, the feasible \(q\) and \(d\) should satisfy (9). Then, the design problem (4) can be equivalently formulated as

\[
\begin{align*}
\min_{\delta} & \quad \delta \\
\text{s.t.} & \quad x_0 = 1 \\
& \quad \|e^T(\omega)x\|^2 \leq \delta \cdot d^T s(\omega), \quad \forall \omega \in \Omega_f \\
& \quad \|f^T(\omega)q\|^2 = d^T s(\omega), \quad \forall \omega \in [0, \pi].
\end{align*}
\]

(10)

Note that the hyperbolic constraint (10b) is imposed over \(\Omega_f\), while the quadratic equality constraint (10c) must be satisfied for \(\forall \omega \in [0, \pi]\). Although the trigonometric polynomial coefficients \(d\) are introduced as auxiliary variables in (10), they are tightly related to denominator coefficients \(q\) through (8). In order to establish the equivalence between the design problems (4) and (10), the constraint (8) should be incorporated in (10).

However, the equality constraint (10c) implies that the polynomial \(R(z)\) is nonnegative on the unit circle. Then, based on the theorem of spectral factorization [30], we can find a causal polynomial \(F(z) = \sum_{m=0}^{M} f_m z^{-m}\) with real coefficients such that \(F(z)F(z^{-1}) = R(z)\). Although the spectral factorization is not unique, among all the possible spectral factorizations, there is only one minimum-phase polynomial. In view of the stability requirement of designed IIR filters, then \(F(z)\) should be chosen as the unique minimum-phase polynomial. Consequently, \(Q(z)\) is equivalent to \(F(z)\), and (8) becomes a redundant constraint.

Due to the existence of the quadratic equality constraint (10c), the design problem (10) is still nonconvex and cannot be efficiently solved. However, we can relax it into a convex problem by replacing (10c) by another hyperbolic inequality constraint \(\|f^T(\omega)q\|^2 \leq \delta^* \cdot d^T s(\omega)\). Then the design problem (10) can be transformed as

\[
\begin{align*}
\min_{\delta} & \quad \delta \\
\text{s.t.} & \quad x_0 = 1 \\
& \quad x_m - d_M = 0 \\
& \quad \|e^T(\omega)x\|^2 \leq \delta \cdot d^T s(\omega), \quad \forall \omega \in \Omega_f \\
& \quad \|f^T(\omega)q\|^2 = d^T s(\omega), \quad \forall \omega \in [0, \pi] \\
& \quad x_m \in \mathbb{R}, \quad m = 0, 1, \ldots, M.
\end{align*}
\]

(11)

(11a)

(11b)

(11c)

(11d)

The constraints (11c) and (11d) are imposed, respectively, on \(L\) and \(K\) discrete grid frequency points. Since the equality constraint (10c) is replaced by the SOC constraint (11d), the variables \(q\) and \(d\) in (11) may not satisfy (8) any longer. The resulting difference between \(|Q(e^{j\omega})|^2\) and \(d^T s(\omega)\) can be derived as

\[
\frac{1}{\pi} \int_0^{\pi} \left| Q(e^{j\omega}) \right|^2 - d^T s(\omega) \, d\omega = ||q||^2 - d_0.
\]

(12)

For the sake of the latter discussion, we define an error function \(\lambda(d_0, q)\) for the error obtained in (12) as

\[
\lambda(d_0, q) = ||q||^2 - d_0.
\]

(13)

According to (11d) and (12), we have \(\lambda(d_0, q) \leq 0\). Therefore, by introducing the relaxed SOC constraint (11d), the Property 1 of (8) has been accordingly relaxed as \(||q||^2 \leq d_0\). Although the equality constraint (10c) has been replaced by the relaxed SOC constraint (11d), the nonnegativity of \(R(z)\) on the unit circle is still guaranteed. According to the theorem of spectral factorization, \(\{d_{im}, m = -M, \ldots, -1, 0, 1, \ldots, M\}\) in (11) is still an autocorrelation sequence. Therefore, the Property 3 of (8) can be automatically satisfied. The Property 2 of (8) is ensured by the constraint (11b), which can prefilter out unqualified \(x_M\) and \(d_M\).

Let \(\delta^*\) denote the optimal value (or the minimum squared minimax error) of the original design problem (10), and \(\delta^*_\text{red}\) be the optimal value of the relaxed design problem (11). Note that the feasible domain defined by the relaxed constraint (11d) is larger than that of (10c). Therefore, we always have \(\delta^*_\text{red} \leq \delta^*\), which means a lower bound on the optimal value of (10) can be obtained by solving (11). However, due to the introduction of the relaxed constraint (11d), \(\delta^*_\text{red}\) is not the exact squared minimax error of the IIR filter obtained by (11), which should be equal to \(\delta^*_\text{mm} = \max_{\omega \in \Omega_f} ||E(\omega)||^2\). On the other hand, \(\delta^*_\text{mm}\) serves as an upper bound of \(\delta^*\), i.e., \(\delta^*_\text{red} \leq \delta^* \leq \delta^*_\text{mm}\). Obviously, if the discrepancy between \(\delta^*_\text{red}\) and \(\delta^*_\text{mm}\) can be reduced, the minimum squared minimax error \(\delta^*\) could be located.
C. Iterative Method

In general, there is a discrepancy between \( \delta_{\text{rel}}^* \) and \( \delta_0^* \), and the corresponding \( \mathbf{x} \) and \( \mathbf{d} \) obtained by solving the relaxed problem (11) cannot exactly satisfy the quadratic equality constraint (10c) over the whole frequency band \([0, \pi]\). Therefore, they are not the real solution to the minimax design problem (10). In this subsection, we will develop an iterative procedure, in which a sequence of SOCP problems based on (11) are subsequently solved in order to gradually reduce the discrepancy between \( \mathbf{d}^T \mathbf{s}(\omega) \) and \( \| \mathbf{f}^T(\omega) \mathbf{q} \|^2 \) in (11d). At the \( k \)-th iteration, the filter coefficients \( \mathbf{x} \) and the trigonometric polynomial coefficients \( \mathbf{d} \) are updated by

\[
\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \alpha \Delta \mathbf{x}^{(k)} \quad (14)
\]

\[
\mathbf{d}^{(k)} = \mathbf{d}^{(k-1)} + \alpha \Delta \mathbf{d}^{(k)} \quad (15)
\]

where the step size \( \alpha \) is within \((0, 1)\), \( \mathbf{x}^{(k-1)} \) and \( \mathbf{d}^{(k-1)} \) are obtained at the previous iteration, and the search direction \( \Delta \mathbf{x}^{(k)} = [\mathbf{u}^{(k)}]^T, [\mathbf{u}^{(k)}]^T \) and \( \Delta \mathbf{d}^{(k)} \) will be determined at the current iteration. In \( \Delta \mathbf{x}^{(k)} \), \( \mathbf{u}^{(k)} \) and \( \mathbf{v}^{(k)} \) are used to update denominator and numerator coefficients, respectively. Suppose \( x_0^{(k)} = 1 \) for \( k \geq 0 \). Then, (11a) can be replaced by another linear equality constraint \( \Delta x_0^{(k)} = 0 \). Since the integrand of (12) is always nonpositive, \( |\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)})| \) can be regarded as the total difference between \( \mathbb{L}(Q(\mathbf{e} \Delta \omega)) \) and \( \mathbf{d}^{(k)} \) over \([0, \pi] \) at the \( k \)-th iteration. Based on this observation, the proposed iterative procedure attempts to gradually reduce \(|\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)})| \) as \( k \rightarrow +\infty \). When the discrepancy between \( \| \mathbf{f}^T(\omega) \mathbf{q}^{(k)} \|^2 \) and \( \mathbf{d}^{(k)} \) over \([0, \pi] \) is reduced to 0, the relaxed SOC constraint (11d) will become the equality constraint (10c). Let \( \delta_{\text{rel}} \) denote the optimal value of the relaxed design problem (11) to be solved at the \( k \)-th iteration, and \( \delta_{\text{min}}^{(k)} \) represent the corresponding squared minimax error of the obtained IIR filter. Based on the above analysis, we have

\[
\lim_{k \rightarrow +\infty} \left( \frac{\delta_{\text{rel}}^{(k)}}{\delta_{\text{min}}^{(k)}} - 1 \right) = 0 \quad \text{if} \quad \lim_{k \rightarrow +\infty} \left| \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \right| = 0.
\]

This property implies that a real minimax design can be attained by decreasing \( \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \).

Define the ratio \( \gamma^{(k)} \) as

\[
\gamma^{(k)} = \frac{\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)})}{\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k-1)})} \quad (16)
\]

At the \( k \)-th iteration, we can impose the constraint \( \gamma^{(k)} \leq \gamma < 1 \) on \( \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \), which is equivalent to

\[
\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) = \| \hat{q}^{(k)} \|^2 - \hat{d}_0^{(k)} \geq \gamma \cdot \lambda(\hat{d}_0^{(k-1)}, \hat{q}^{(k-1)}) = \gamma \left( \| \hat{q}^{(k-1)} \|^2 - \hat{d}_0^{(k-1)} \right) \quad (17)
\]

Applying the above inequality recursively, we have

\[
\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \geq \gamma^k \cdot \lambda(\hat{d}_0^{(0)}, \hat{q}^{(0)}) \quad (18)
\]

As \( k \rightarrow +\infty \), the right-hand side of (18) will approach 0. Combined with \( \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \leq 0 \), it can be concluded that

\[
\lim_{k \rightarrow +\infty} \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) = 0.
\]

The major obstacle to incorporate the inequality constraint (17) into the relaxed SOCP design problem (11) is that (17) is nonconvex. Here, the first-order Taylor series approximation is used to linearize \( \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \) around \( \hat{d}_0^{(k-1)} \) and \( \hat{q}^{(k-1)} \). Then, the constraint (17) can be approximated by

\[
\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \approx \lambda(\hat{d}_0^{(k-1)}, \hat{q}^{(k-1)}) \left[ \Delta \hat{d}_0^{(k)} \right]^{\top} \mathbf{u}^{(k)} \nonumber \geq (\gamma - 1) \cdot \lambda(\hat{d}_0^{(k-1)}, \hat{q}^{(k-1)}) \quad (19)
\]

Note that \( \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \) is a convex quadratic function of \( \hat{d}_0^{(k)} \) and \( \hat{q}^{(k)} \). Then, the first-order Taylor series approximation serves as a global under-estimator of \( \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \). Therefore, by imposing (19) on \( \hat{d}_0^{(k)} \) and \( \hat{q}^{(k)} \) (or, equivalently, \( \Delta \hat{d}_0^{(k)} \) and \( \mathbf{u}^{(k)} \)), the inequality \( \lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \geq \gamma \cdot \lambda(\hat{d}_0^{(k-1)}, \hat{q}^{(k-1)}) \) can be definitely ensured. However, since the search direction is restricted in a halfspace defined by (19) instead of the nonconvex set defined by (17), it cannot be guaranteed that the globally optimal solution will be definitely achieved by the iterative procedure which adopts (19) as the search direction at each iteration.

Incorporating (19) into the relaxed problem (11), then at the \( k \)-th iteration the design problem (11) can be reformulated as

\[
\min \delta^{(k)} \quad (20) \\
\text{s.t.} \quad \Delta \hat{d}_0^{(k)} = 0 \quad (20a) \\
\Delta \hat{d}_0^{(k)} - \Delta \hat{d}_0^{(k-1)} = 0 \quad (20b) \\
\Delta \hat{d}_0^{(k)} + 2 \mathbf{q}^{(k-1)} \mathbf{T} \mathbf{u}^{(k)} \nonumber \geq (\gamma - 1) \cdot \lambda(\hat{d}_0^{(k-1)}, \hat{q}^{(k-1)}) \quad (20c) \\
\| \mathbf{e}^T(\omega) \mathbf{e}^{(k-1)} + \mathbf{c}^T(\omega) \Delta \mathbf{x}^{(k)} \|^2 \nonumber \leq \delta^{(k)} \cdot \| \mathbf{d}^{(k-1)} + \Delta \mathbf{d}^{(k)} \|^2 \mathbf{s}(\omega_i) \quad \omega_i \in \Omega, \; i = 1, 2, \ldots, L \quad (20d) \\
\| \mathbf{f}^T(\omega) \mathbf{q}^{(k-1)} + \mathbf{f}^T(\omega_j) \mathbf{u}^{(k)} \|^2 \nonumber \leq \| \mathbf{d}^{(k-1)} + \Delta \mathbf{d}^{(k)} \|^2 \mathbf{s}(\omega_j) \quad \omega_j \in [0, \pi], \; j = 1, 2, \ldots, K \quad (20e)
\]

where \( L \) and \( K \) grid frequency points are used in (20d) and (20e), as in (11c) and (11d). In (20), the decision variables are \( \delta^{(k)} \), \( \Delta \mathbf{x}^{(k)} \) (or \( \mathbf{u}^{(k)} \), \( \mathbf{v}^{(k)} \), and \( \Delta \mathbf{d}^{(k)} \)) After solving (20), the obtained \( \Delta \mathbf{x}^{(k)} \) and \( \Delta \mathbf{d}^{(k)} \) are to be used to update filter coefficients \( \mathbf{x}^{(k-1)} \) and trigonometric polynomial coefficients \( \mathbf{d}^{(k-1)} \) according to (14) and (15).

The iterative procedure will stop when the following condition is satisfied:

\[
\lambda(\hat{d}_0^{(k)}, \hat{q}^{(k)}) \leq \varepsilon \quad (21)
\]
where $\varepsilon$ is a prescribed small positive tolerance. Based on the previous analysis, which has shown that
\[ \lim_{k \to +\infty} \left| \lambda(d_0^{(k)}, q^{(k)}) \right| = 0, \]
the convergence of the iterative procedure can be definitely assured. From (18), it can be further deduced that the iterative procedure will be terminated at the $k$th iteration if
\[ \gamma^k \left| \lambda(d_0^{(0)}, q^{(0)}) \right| \leq \varepsilon. \]
By taking logarithm on both sides of the inequality above, an estimated maximum number $k_{\text{max}}$ of iterations required by the iterative procedure can be obtained by
\[ k_{\text{max}} = \left\lfloor \frac{\ln \varepsilon - \ln \left| \lambda(d_0^{(0)}, q^{(0)}) \right|}{\ln \gamma} \right\rfloor + 1. \tag{22} \]
where $\lfloor x \rfloor$ denotes the largest integer less than or equal to $x$. Since $0 \leq \lambda(d_0^{(k)}, q^{(k)}) - \lambda(d_0^{(k-1)}, q^{(k-1)}) \leq -\lambda(d_0^{(k-1)}, q^{(k-1)})$, then as $k$ is large enough, we can further obtain
\[ \lambda(d_0^{(k)}, q^{(k)}) - \lambda(d_0^{(k-1)}, q^{(k-1)}) = \| u^{(k)} \|^2 + 2d^{(k-1)^T} u^{(k)} - \Delta d_0^{(k)} \approx 0. \tag{23} \]
Note that the constraint (19) indicates that $2d^{(k-1)^T} u^{(k)} - \Delta d_0^{(k)} \geq 0$. Then, it follows from (23) that $\| u^{(k)} \| \approx 0$, which means there is no significant change on $u^{(k)}$ as $k$ is large enough.

III. PRACTICAL CONSIDERATIONS

A. Convergence Speed

As discussed in Section II-C, the convergence of the iterative procedure can be guaranteed if the linear inequality constraint (19) is incorporated. Obviously, a larger $\gamma$ leads to a larger feasible set for the search direction defined by (19). Therefore, $\gamma$ should be chosen as close to 1 as possible in order to achieve a satisfactory design. However, if $\gamma$ is too close to 1, (22) shows that the total number of iterations required by the proposed iterative procedure could be too large. As an attempt to resolve this dilemma, we introduce a new variable $\beta(k) \geq 0$ to replace the term $(\gamma - 1) \cdot \lambda(d_0^{(k-1)}, q^{(k-1)})$ in (19)
\[ -\Delta d_0^{(k)} + 2d^{(k-1)^T} u^{(k)} \geq \beta(k). \tag{24} \]
In (24), $\beta(k)$ serves as a soft threshold such that the feasible set defined by (24) can be automatically converged. Apparently, in order to achieve the fastest convergence speed, we want to maximize $\beta(k)$ (or minimize $-\beta(k)$) at each iteration, while minimizing $\delta(k)$ to reduce the approximation error. A common way to solve this bi-objective optimization problem is to minimize the weighted sum of these two objectives. By introducing a relative weight $\zeta > 0$, the design problem (20) is rewritten as
\[ \begin{align*}
\min_{\delta(k), \beta(k)} & \quad \delta(k) - \zeta \cdot \beta(k) \\
\text{s.t.} & \quad \Delta d_0^{(k)} = 0 \tag{25a} \\
& \quad \Delta u_0^{(k)} = 0 \tag{25b}
\end{align*} \]
\[ \begin{align*}
-\Delta d_0^{(k)} + 2d^{(k-1)^T} u^{(k)} - \beta(k) & \geq 0 \tag{25c} \\
\| u^{(k)} \|^2 + 2d^{(k-1)^T} u^{(k)} - \Delta d_0^{(k)} & \approx 0 \tag{25d} \\
& \quad \left| \begin{array}{c}
\| u^{(k)} \|^2 + 2d^{(k-1)^T} u^{(k)} - \Delta d_0^{(k)} \\
\| u^{(k)} \|^2 + 2d^{(k-1)^T} u^{(k)} - \Delta d_0^{(k)}
\end{array} \right| \leq \delta(k) \cdot \left( d^{(k-1)} + \Delta d^{(k)} \right)^T s(\omega_i) \\
& \quad \omega_i \in \Omega, \quad i = 1, 2, \ldots, L \tag{25e} \\
& \quad \left| u^{(k)} \right|^2 + \left| u^{(k)} \right|^2 \leq \left( d^{(k-1)} + \Delta d^{(k)} \right)^T s(\omega_j) \\
& \quad \omega_j \in [0, \pi], \quad j = 1, 2, \ldots, K. \tag{25f}
\end{align*} \]
The selection of the parameter $\zeta$ is a tradeoff between the convergence speed and the design performance. The convergence speed can be improved by increasing $\zeta$, while the better performance can be attained by decreasing $\zeta$. It seems that the effects of $\zeta$ used in the regularized design problem (25) are similar to those of $\gamma$ used in (20). However, it should be emphasized that given $\gamma$ the ratio $\gamma^{(k)}$ is confined at each iteration by the constraint (19), and, hence, the convergence speed cannot be further increased. In contrast, the restriction on $\gamma^{(k)}$ has been removed in (24) by introducing the soft threshold $\beta(k)$. In general, the modified design procedure can achieve faster convergence speed, which has been verified by many simulations we tried so far. In practice, as $\zeta$ is small enough, decreasing $\zeta$ contributes less to the performance improvement, and the convergence speed of the iterative procedure could be too slow for practical applications. If at each iteration, the following condition is still satisfied for some $\gamma < 1$
\[ \beta(k) \geq (\gamma - 1) \cdot \lambda(d_0^{(k-1)}, q^{(k-1)}) \tag{26} \]
the convergence of the iterative procedure (25) can also be strictly guaranteed. However, it should be noticed that (26) is only a sufficient condition for the convergence of the iterative procedure, which implies that even without (26), the iterative procedure could still converge to the final solution when $\zeta$ is appropriately selected. The effects of $\zeta$ on the design results will be illustrated in Example 1 to be presented in Section IV-A.

B. Stability Constraint

A sufficient condition for the stability of designed filters in terms of positive realness was introduced in [13], which can be stated as: If $Q^{(k-1)}(z)$ is a Schur polynomial, i.e., all roots of $Q^{(k-1)}(z)$ lie inside the unit circle, and the transfer function $G^{(k)}(z) = 1 + u^{(k)}(z)/Q^{(k-1)}(z)$ is strictly positive real (SPR), i.e.,
\[ \text{Re} \left\{ G^{(k)}(e^{j\omega}) \right\} > 0, \quad \forall \omega \in [0, \pi] \tag{27} \]
where $u^{(k)}(z) = u^{(k)^T} \varphi_M(z) \varphi_M^T(z) e^{j\theta_0^{(k)}} = 0$, then the weighted sum of $Q^{(k-1)}(z)$ and $u^{(k)}(z)$, i.e., $Q^{(k)}(z) = Q^{(k-1)}(z) + \alpha u^{(k)}(z)$ for $\alpha \in [0, 1]$, is a Schur polynomial. According to this condition, a stability domain with an interior point $q^{(k-1)}$ can be defined as $D_{\alpha} = \{ u^{(k)} | G^{(k)}(z) \text{ is SPR} \}$. The condition
that \(G^{(k)}(z)\) is SPR is equivalent to requiring that [see (28), shown at the bottom of the page] is real and positive on the unit circle. Since the denominator of (28) is positive on the unit circle, it follows that the symmetric numerator polynomial of (28) must be positive on the unit circle, which is further cast in [13] as a linear matrix inequality (LMI) constraint independent of frequency \(\omega\). It has been proved [13] that this stability constraint defines a larger feasible domain than the Rouché’s theorem based stability constraint [11].

Since SOCP problems cannot cope with LMI constraints, we express the stability constraint \(G^{(k)}(e^{j\omega}) + G^{(k)}(e^{-j\omega}) \geq 0\) as the following linear inequality constraints:

\[
\begin{align*}
\Re \left\{ Q^{(k-1)}(e^{j\omega}) \phi_M^T(e^{j\omega}) \right\} \cdot \mathbf{w}^{(k)} &\geq \mu - \left| Q^{(k-1)}(e^{j\omega}) \right|^2 \omega_j \in [0, \pi], \ j = 1, 2, \ldots, K \quad (29)
\end{align*}
\]

where \(\mu\) is a specified small positive number. If all the poles of designed IIR filters are required to lie inside a prescribed circle of radius \(\rho < 1\) for robust stability, \(\phi_M(e^{j\omega})\) and \(Q^{(k-1)}(e^{j\omega})\) in (29) should be replaced by \(\phi_M(\rho e^{j\omega})\) and \(Q^{(k-1)}(\rho e^{j\omega})\), respectively. In general, the parameter \(\mu\) can be selected within \([10^{-3}, 10^{-6}]\). Simulation results show that the design results are not very sensitive to the selection of \(\mu\).

C. Selection of Initial IIR Digital Filter

For iterative design methods, the selection of the initial filter is a critical step to find a satisfactory solution. Without any prior knowledge of IIR filters, initial filters can be chosen as optimal FIR filters designed under the same specifications as suggested in [11]. Some other algorithms utilize more complicated multi-stage initialization strategy [13].

In our design, the initial filters are obtained by solving the relaxed SOCP problem (11). For stability, the constraint (29) should be incorporated in (11). The initial denominator can be simply assumed as \(q_m^{(1)} = 1\) and \(q_m^{(1)} = 0\) for \(m = 1, 2, \ldots, M\). In this situation, the stability constraint (29) is equivalent to the positive reality based stability constraint proposed in [17]. Although only the lower and upper bounds on the optimal value \(\delta^*\) of the original nonconvex problem (10) can be obtained, we find that the corresponding \(\mathbf{x}\) and \(\mathbf{d}\) can always lead to satisfactory solutions for all the design problems that have been tried so far. Some other initial guesses can also be utilized as the initial filters of the iterative procedure. It should be noted that \(\mathbf{q}^{(0)}\) and \(\mathbf{d}^{(0)}\) should satisfy \(d_M^{(0)} = d_M^{(0)}\) and \(\|Q^{(0)}(e^{j\omega})\|^2 \leq d_M^{(0)} T \mathbf{s}(\omega)\) for \(\omega \in [0, \pi]\).

Finally, the major steps of the proposed design algorithm are summarized as follows.

Step 1: Given the ideal frequency response \(D(\omega)\), filter orders \(N\) and \(M\), and weighting function \(W(\omega)\), set \(k = 0\) and solve the relaxed design problem (11). The obtained filter and trigonometric polynomial coefficients are chosen as the initial guess \(\mathbf{z}^{(0)}\) and \(\Delta \mathbf{d}^{(0)}\).

Step 2: Set \(k = k + 1\), and solve for \(\Delta \mathbf{x}^{(k)}\) and \(\Delta \mathbf{d}^{(k)}\) the SOCP problem (25) with stability constraints (29).

Step 3: Update the coefficients \(\mathbf{z}^{(k)}\) and \(\mathbf{d}^{(k)}\) by (14) and (15). If the stopping condition (21) is satisfied, terminate the iterative procedure; otherwise, go to Step 2 and continue.

Some remarks about the proposed design method are made below:

1. In practice, after obtaining the final filter, some local search methods can be further applied to refine the design result. In our postprocessing, the denominator coefficients are kept fixed and the numerator coefficients can be updated by solving the following SOCP problem:

\[
\begin{align*}
\min \quad & \delta \\
\text{s.t.} \quad & |B(\omega_i)p - b(\omega_i)| \leq \delta \\
& \omega_i \in \Omega_l, \ i = 1, 2, \ldots, L \quad (30a)
\end{align*}
\]

where

\[
\begin{align*}
B(\omega) =& W(\omega) \left[ \begin{array}{c}
\text{Re} \left\{ \mathbf{z}_M(\omega) \right\} \\
\text{Im} \left\{ \mathbf{z}_M(\omega) \right\}
\end{array} \right] \\
b(\omega) =& W(\omega) \left[ \begin{array}{c}
\text{Re} \left\{ D(\omega) \right\} \\
\text{Im} \left\{ D(\omega) \right\}
\end{array} \right].
\end{align*}
\]

For a given \(\mathbf{q}\), the numerator obtained by solving (30) is optimal.

2. According to the previous analysis, it is clear that the parameter \(\zeta\) used in (25) should be appropriately selected. Through a large number of simulations, we find that generally \(\zeta\) can be chosen within \([10^{-6}, 10]\). In practice, \(\zeta\) can also be varied during the iterative procedure. Based on our simulation observation, we find that both \(|\lambda(d_M^{(k)}, \mathbf{q}^{(k)})|\) and the approximation error can be dramatically reduced at the first several iterations even if \(\zeta \ll 1\). As \(k\) increases, the convergence speed becomes slow. Moreover, as \(k\) is large enough, it is hard to further reduce the approximation error. This observation suggests that instead of a fixed \(\zeta\), we can employ a variable \(\zeta\) in (25) during the iterative procedure. At the beginning of the iterative procedure, \(\zeta\) can be chosen as a small value such that the feasible set defined by (24) can be as large as possible. As \(k\) increases, \(\zeta\) can be gradually increased. Then, the convergence speed can be improved. Example 4 will be presented in Section IV-D to demonstrate the effectiveness of the utilization of a variable \(\zeta\).

\[
G^{(k)}(z) + G^{(k)}(z^{-1}) = \frac{2Q^{(k-1)}(z)Q^{(k-1)}(z^{-1}) + u^{(k)}(z)Q^{(k-1)}(z^{-1}) + Q^{(k-1)}(z)u^{(k)}(z^{-1})}{Q^{(k-1)}(z)Q^{(k-1)}(z^{-1})} \quad (28)
\]
In this section, four examples are presented to demonstrate the effectiveness of the proposed method. We use the SeDuMi [31] in MATLAB environment to solve the SOCP problems (11) and (25). Besides the peak and $L_2$ errors of magnitude (MAG) and group delay (GD) over $\Omega_f$, we also employ the minimax error defined by

$$e_{\text{MM}} = \max_{\omega \in \Omega_f} |E(\omega)|$$

to evaluate the design results. Without explicit declaration, the weighting function is always set to $W(\omega) = 1$ for $\forall \omega \in \Omega_f$ and $W(\omega) = 0$ otherwise. Similarly, the admissible maximum pole radius $\rho$ is always set equal to 1, unless it is explicitly specified. The parameter $K$ is always equal to 101. Let $S$ be the set of equally-spaced grid points over $[0, \pi]$, i.e., $S = \{\omega_j | \omega_j = (j - 1) \times \pi / (K - 1), j = 1, 2, \ldots, K\}$. Then the constraints (11c), (20d), (25e), and (30a) are imposed on a set of grid points taken from $S$, i.e., $\{\omega_i | \omega_i \in S \cap \Omega_f\}$. Generally speaking, a larger $K$ can lead to a more accurate design. However, in practice, this effect is almost negligible when $K$ is large enough, e.g., 100 or more. With a larger $K$, the proposed iterative procedure needs more computation time to converge to the final solution. It should be mentioned that the total number of iterations is normally not changed. The extra computation time is spent to construct the extra constraints and solve the SOCP problem of a larger size at each iteration. Our simulation experience shows that when $K$ is chosen between 100 and 500, the computation time is acceptable. In all the examples, the step size $\alpha_i$ in (14) and the parameter $\varepsilon$ in (21) are set to 0.5 and $10^{-3}$, respectively. The positive number $\mu$ used in the stability constraint (29) is always chosen as $10^{-3}$.

### A. Example 1

The first example is to design a lowpass IIR filter whose specifications are the same as those adopted in [11]. The ideal frequency responses are defined as

$$D(\omega) = \begin{cases} e^{\omega J_2(\omega)}, & 0 \leq \omega \leq 0.4\pi \leq 0.5\pi \leq \omega < \pi. \\ 0, & \end{cases}$$

Filter orders are chosen as $N = 15$ and $M = 4$. In this design, the parameter $\zeta$ used in (25) is set as 0.001. After 24 iterations, the iterative procedure converges to the final solution. The maximum pole radius of the obtained filter is 0.8598. The numerator and denominator coefficients are given, respectively, by

\begin{align*}
&[1.0796e-002, 1.7887e-002, -1.7854e-002, 4.5345e-002, 3.4012e-002, 2.2196e-001, 3.7787e-001, 3.7109e-001, 2.2094e-001, 7.5230e-002, 1, -4.5908e-001, 8.9299e-001, -2.5445e-001, 8.1335e-002];
\end{align*}

The magnitude and group delay responses are shown in Fig. 1. The magnitude of the weighted complex error $E(\omega)$ is plotted in Fig. 2. For comparison, we also employ the iterative algorithm proposed in [10] to design an IIR filter under the same set of specifications. Instead of an $M$th-order polynomial used in (1), the denominator utilized by [10] is expressed as a product of second-order factors and a first-order factor if $M$ is odd, i.e.,

$$Q(z) = (1 + b_0 z^{-1}) \prod_{i=1}^{I} (1 + b_{i,1} z^{-1} + b_{i,2} z^{-2})$$

where $I = (M - 1)/2$ if $M$ is odd or $M/2$ if $M$ is even. In so doing, the first-order Taylor series approximation can be directly applied on the frequency response $H(e^{j\omega})$ with respect to the numerator coefficients $p_n$ and the factorized denominator coefficients $b_0$, $b_{1,1}$, and $b_{2,2}$, and subsequently the design problem at each iteration can be formulated as an SOCP. The advantage of adopting the factorized denominator is that the prescribed maximum polynomial...
radius constraint can be cast as a set of linear inequality constraints in terms of \( b_0, b_{b1}, \) and \( b_{b2} \), which are sufficient and necessary. At the beginning of the iterative procedure, all the poles are simply placed at the origin. The initial numerator is obtained by solving (30) with the initial denominator. The maximum pole radius of the IIR filter designed by [10] is 0.8590. All the error measurements of both designs are summarized in Table I. It can be observed that the proposed method can achieve slightly better performance in \( \epsilon_{\text{MDM}} \) than the SOCP method [10].

In order to illustrate the effects of \( \zeta \) on final design results, we repeat the experiment with different \( \zeta \) within \([5 \times 10^{-4}, 5 \times 10^{-3}]\) under the same set of specifications. Fig. 3 shows the variation of the minimax error \( \epsilon_{\text{MDM}} \) versus the parameter \( \zeta \). We can find that, generally speaking, the design performance can be improved by decreasing \( \zeta \). In all the designs, the iterative procedure converges to the final solution within 28 iterations. However, as \( \zeta \) is small enough (in this example, \( \zeta \leq 10^{-5} \)), it is difficult to further improve the design performance. Moreover, when \( \zeta \) is too small (in this example \( \zeta \leq 10^{-4} \)), the iterative procedure could converge in a slow speed. Fig. 3 suggests a way to find a suitable \( \zeta \): First of all, we can choose a large value for \( \zeta \) (e.g., 1). Then, we gradually reduce \( \zeta \) until the improvement of the design performance is negligible or the iterative procedure cannot converge within a prescribed maximum number (e.g., 50) of iterations.

**B. Example 2**

The second example is to design a highpass IIR filter [8]. The filter orders are chosen as \( N = M = 14 \), and the ideal frequency response is defined by

\[
D(\omega) = \begin{cases} 
2^{-\frac{j\omega}{12}}, & 0.525 \leq \omega < \pi \\
0, & 0 \leq \omega \leq 0.475\pi.
\end{cases}
\]

Originally, the maximum pole radius is set as \( \rho = 1 \). However, the design results show that there is a magnitude overshoot within the transition band. Thereby, we reduce \( \rho \) from 1 to 0.96. The parameter \( \zeta \) is set equal to 0.3. After 30 iterations, the iterative procedure converges to the final solution. The maximum pole radius of the obtained IIR filter is 0.9559. The numerator and denominator coefficients are given, respectively, by \(-9.1215e-003, 1.7383e-002, 5.5492e-003, -3.6005e-003, -9.0162e-003, 6.9193e-003, 9.9406e-003, -1.5344e-002, 1.3480e-002, 4.9121e-002, 5.3284e-002, -1.856e-001, 3.5661e-001, -2.5639e-001, 1.8654e-001, 1.7636e-001, 1.2253e+000, 7.5411e-001, 4.5514e-001, 3.1573e-001, 2.5391e-001, 1.4183e-001, -4.7181e-003, -8.9064e-002, -9.1586e-002, -7.0589e-002, -6.7160e-002, -5.2190e-002, -3.5435e-002 \).

The design results are shown in Fig. 4. The magnitude of the weighted complex error is given in Fig. 5. The variation of discrepancy between \( \delta_{\text{MM}}^{(k)} \) and \( \delta_{\text{rel}}^{(k)} \) in logarithmic scale, i.e., \( 10 \times \log_{10}(\delta_{\text{MM}}^{(k)} / \delta_{\text{rel}}^{(k)}) \), versus the iteration number \( k \) is presented in Fig. 6. The changes of \( \delta_{\text{MM}}^{(k)} \) and \( \delta_{\text{rel}}^{(k)} \) in logarithmic scale at various iterations are also plotted in Fig. 6. It can be seen that at the beginning of the iterative procedure (in this example \( k \leq 5 \)), \( \delta_{\text{MM}}^{(k)} \) and \( \delta_{\text{rel}}^{(k)} \) decrease fast. Then, the iterative procedure reaches a steady stage until the stopping condition is satisfied. In contrast, the optimal value \( \delta_{\text{rel}}^{(k)} \) of the design problem (25) is first increased, and then gradually reduced. Actually, in all the designs we tried so far, \( \delta_{\text{MM}}^{(k)} \) and \( \delta_{\text{rel}}^{(k)} \) change at each iteration in a similar way. For comparison, we also utilize the SM method [6] to design an IIR filter under the same set of specifications except the maximum pole radius which is still chosen as \( \rho = 1 \). The initial denominator for the SM method [6] is simply chosen as \( q^{(0)} = [1, 0, \ldots, 0]^T \), and the initial numerator is chosen as the optimal FIR filter of order \( N = 14 \), which can be obtained by solving (30) with the denominator \( q^{(0)} \). The maximum pole radius of the obtained filter is 0.9427. All the error measurements for both designs are listed in Table II for comparison. Apparently, the proposed
method can achieve better performance than the SM method [6].

C. Example 3

The third example is to design a two-band IIR digital filter [16] with the desired frequency response given by

$$D(\omega) = \begin{cases} e^{-j14.2\omega}, & 0 \leq \omega \leq 0.46\pi \\ 0.5e^{-j30\omega}, & 0.5\pi \leq \omega < \pi. \end{cases}$$

The maximum pole radius is set as 0.95. A group of IIR filters are designed by the proposed method, each of them totally having 31 filter coefficients, i.e., $N + M + 1 = 31$. The value of denominator order $M$ changes from 0 to 15. Table III shows the minimax error of each design. It can be observed that the best design is attained when $M = 6$ and $N = 24$. The corresponding $\zeta$ used in this design is 0.005. The maximum pole radius of the obtained filter in the best design is 0.9486, and all the numerator and denominator coefficients are listed below: $-2.6325e-003$, $1.2557e-002$, $3.2471e-003$, $4.5004e-003$, $7.3894e-003$, $4.2937e-003$, $-8.8385e-003$, $-6.6589e-003$.

---

Table II

<table>
<thead>
<tr>
<th>Method</th>
<th>$\phi_{SM}$</th>
<th>Passband MAG (Peak/L_2 in dB)</th>
<th>Stopband MAG (Peak/L_2 in dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>4.302e-2</td>
<td>-27.46/-37.09</td>
<td>3.82/3.08e-1</td>
</tr>
<tr>
<td>SM[6]</td>
<td>5.288e-2</td>
<td>-25.73/-33.65</td>
<td>3.94/3.53e-1</td>
</tr>
</tbody>
</table>
1.5448e-002, 1.2607e-002, -3.3186e -002, -3.3382e -002, 1.3066e-001, 4.3430e-001, 6.8400e-001, 7.0396e-001, 4.9731e-001, 1.9775e-001, -7.8242e-002, 2.2166e-001, -3.0361e -001, 2.5119e-001, -1.3502e -001, 4.0037e-002, and 1.2759e-001, 1.1538e+000, 1.1822e-001, 2.0907e-001, 8.2517e-004, -1.7973e -002. The magnitude and group delay responses within the frequency bands of interest are shown in Fig. 7. The magnitude of the weighted complex error is plotted in Fig. 8. The minimax errors of all the designs are summarized in Table III. The design algorithm in [16] is originally proposed under the weighted integral of the squared error (WISE) criterion, which linearly combines the WLS criterion and time-domain components to ensure the stability of obtained filters. In order to utilize the WISE method to design IIR filters in the minimax sense, an iterative procedure is introduced. At each iteration, the weighting function is multiplied by the envelope of $|E(k)(\omega)|$, where $E(k)(\omega)$ denotes the weighted frequency response error of the IIR filter obtained at the current iteration, such that the minimax error can be accordingly minimized at the next iteration by solving the WLS design problem. For comparison, the minimax errors for IIR filters designed by the WISE method are also listed in Table III. It can be seen that the proposed method achieve much better performance than the WISE method [16].

D. Example 4

The last example is to design a full-band differentiator [17]. The ideal frequency response is given by

$$D(\omega) = \frac{\omega}{\pi} e^{j0.5\pi\tau_s(-\tau_s+0.5\omega)} , \ 0 \leq \omega \leq \pi$$

where $\tau_s$ assumes integer values. In the phase of $D(\omega)$ defined above, a half of sample delay is added to eliminate the discontinuity of the desired phase response [17]. The filter order is chosen as $N = M = 17$. In this example, we adopt a variable $\zeta(k)$ in (25). At the $k$th iteration, $\zeta^{(k)}$ is set as $\zeta^{(k)} = 0.01k$ for $k \geq 1$ and used in (25) to determine

<table>
<thead>
<tr>
<th>$M$</th>
<th>Proposed</th>
<th>WISE[16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6.49e-2</td>
<td>2.605e-1</td>
</tr>
<tr>
<td>1</td>
<td>6.668e-2</td>
<td>2.625e-1</td>
</tr>
<tr>
<td>2</td>
<td>1.133e-1</td>
<td>1.133e-1</td>
</tr>
<tr>
<td>3</td>
<td>1.155e-2</td>
<td>1.141e-1</td>
</tr>
<tr>
<td>4</td>
<td>1.073e-2</td>
<td>1.211e-1</td>
</tr>
<tr>
<td>5</td>
<td>1.077e-2</td>
<td>1.236e-1</td>
</tr>
<tr>
<td>6</td>
<td>1.054e-2</td>
<td>1.180e-1</td>
</tr>
<tr>
<td>7</td>
<td>1.122e-2</td>
<td>1.182e-1</td>
</tr>
</tbody>
</table>

Authorized licensed use limited to: UNIVERSITY OF WINDSOR. Downloaded on December 2, 2009 at 14:56 from IEEE Xplore. Restrictions apply.
the search direction $\Delta x^{(k)}$ and $\Delta q^{(k)}$. Naturally, there are some other ways to select $\zeta^{(k)}$ during the iterative procedure. We change $\tau_n$ from 8 to 17. The best design can be attained as $\tau_n = 15$. After 19 iterations, the iterative procedure converges. The maximum pole radius of the obtained filter is 0.9635. The numerator and denominator coefficients are given, respectively, by $-3.0503e-003, -2.7413e-003, -3.8702e-004, -1.8694e-006, -2.0119e-004, 1.2524e-004, -2.3072e-004, 3.4058e-004, -3.7793e-004, 5.8881e-004, -9.4826e-004, 1.7485e-003, -3.1625e-003, 7.9066e-003, -2.8664e-002, 3.5891e-001, 1.8208e-002, -3.5417e-001, and 1, 1.0531e + 000, 6.6768e-002, -1.0020e-002, 3.7019e-003, -1.7434e-003, 7.5298e-004, -5.1406e-004, 3.2175e-004, -3.5136e-004, 2.4046e-005, -1.6413e-004, 1.0602e-005, -1.7735e-004, -1.7353e-004, -6.6872e-004, -1.1515e-003, and -8.5683e-004.

The designed characteristics and the approximation errors of magnitude and group delay responses are given in Fig. 9. It can be observed that near the origin, the group delay (or phase response) of the designed filter has a large error. However, it can be neglected since the magnitude on the frequencies near the origin is almost equal to zero. This can be verified by the magnitude of the complex error, i.e., $|E(\omega)|$, which is shown in Fig. 10. Therefore, the group delay error within $[0, 0.01\pi]$ is neglected when computing all the error measurements of group delay listed in Table IV. The best design result for the LP method [17] can be obtained as $\tau_n = 14$. The corresponding filter coefficients have been reported in [17]. The maximum pole radius of the obtained filter is 0.9821. The error measurements are also summarized in Table IV. The proposed method can achieve better design compared with the LP method [17].

![Fig. 9. Characteristics and errors of designed filter in Example 4.](image)

![Fig. 10. Magnitude of weighted complex error of designed IIR filter in Example 4.](image)

<table>
<thead>
<tr>
<th>TABLE IV</th>
<th>MEASUREMENTS OF DESIGN RESULTS IN EXAMPLE 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>$\tau_n$</td>
</tr>
<tr>
<td>Proposed</td>
<td>14</td>
</tr>
<tr>
<td>LP [17]</td>
<td>15</td>
</tr>
</tbody>
</table>

V. Conclusion

In this paper, an iterative method for designing IIR digital filter in the minimax sense has been presented. Due to the non-convexity of the original design problem, the convex relaxation technique is first applied to obtain the lower and upper bounds on the optimal value of the original design problem. In order to reduce the discrepancy between the original and relaxed problems, an iterative procedure is then employed. At each iteration, the linear constraint (19) is incorporated to guarantee the con-
vergence of the iterative procedure. In practice, to increase the convergence speed, this linear constraint can be further modified by replacing the right hand side of (19) by a soft threshold $\beta(k)$. Accordingly, a regularization term is introduced in the objective function of (25). For stability, a new positive realness based constraint is incorporated in the iterative procedure.

Compared with other iterative design methods, the obvious advantage of the proposed method is that the convergence can be guaranteed, if the regularization coefficient $\zeta$ in (24) is appropriately selected. In order to find a suitable $\zeta$, we can initially choose a large $\zeta$ for the design, and then gradually reduce $\zeta$ until the design performance cannot be further improved. Although a globally optimal design cannot be definitely obtained by the proposed method, through a large number of simulations, we find that a satisfactory design can almost always be obtained, using the strategy described above to choose $\zeta$.

REFERENCES


