Enhanced Steiglitz-McBride Procedure for Minimax IIR Digital Filters

Wu-Sheng Lu
Dept. of Electrical and Computer Engineering
University of Victoria
Victoria, BC, Canada
Email: wslu@ece.uvic.ca

Takao Hinamoto
Hiroshima University
Higashi-Hiroshima, 739-8527, Japan
Email: hinamoto@ieee.org

***Abstract*—This paper presents an enhanced Steiglitz-McBride (SM) procedure for the design of stable minimax IIR digital filters. It is well known that minimax design of IIR filters is typically initiated with a nonconvex formulation, followed by a procedure to relax the original design problem to a sequence of convex sub-problems to be solved iteratively. We proposed an enhanced SM procedure that leads to improved convex relaxation relative to the conventional SM techniques, hence to improved designs. In addition, we make an observation that the state-of-the-art convex stability constraint based on strictly positive realness is equivalent to that deduced from an enhanced version of Rouché theorem from complex analysis. Design examples are presented to evaluate the new design algorithm.**

I. Introduction

The class of IIR digital filters is well known for achieving high selectivity with considerably better computational efficiency relative to FIR counterparts [1]-[3] at the cost of non-constant group delay, stability issues, and nonconvex design formulation. In the past decades a great deal of efforts has been made to develop new design techniques for IIR digital filters with improved performance, see for example [4]-[15], [19] and the references therein. Among other things, design methods based on the Steiglitz-McBride (SM) procedure [6] are found effective, especially when it is appropriately combined with other design techniques [11], [15]. Typically design of an IIR filter is initiated with a nonconvex formulation, followed by a procedure where the original problem is relaxed to a sequence of convex sub-problems to be solved iteratively. It is in this part of the design the SM procedure plays an effective role to convexify the original formulation. Essentially the SM procedure is a zeroth-order technique because, at certain point of the design process, it holds the denominator fixed to that obtained in preceding iteration. In this paper, we propose a first-order SM procedure in a minimax design setting and show that it results in convex sub-problems that resemble the original formulation more closely. As such, the design is expected to gain faster convergence to a likely improved design. In dealing with the stability issue for IIR filters, a successful technique [7] is based on the well-known Rouché theorem [14]. Linear matrix inequality type of convex stability conditions based on strictly positive realness (SPR) [11] are deduced and shown to be superior to that based on Rouch theorem. Linear constraints approximating those from SPR are derived in [15]. In this paper, we make an observation that there is an enhanced version of Rouché's theorem from complex analysis [16], which appears to have been overlooked in the past, that can be used to deduce convex stability constraints that are equivalent to those obtained using SPR. Design examples are presented to evaluate the new design algorithm.

II. Problem Formulation

Consider designing an IIR filter of order (n, m) with transfer function

$$H(z) = \frac{b(z)}{a(z)} = \frac{b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}}{1 + a_1 z^{-1} + \cdots + a_n z^{-n}} \tag{1}$$

such that $H(e^{j\omega})$ best approximates a desired frequency response $H_d(\omega)$ in minimax sense over a frequency region of interest Ω subject to stability. The design problem can be formulated as finding parameter vectors $\boldsymbol{a} = [a_1 \ \cdots \ a_n]^T$ and $\boldsymbol{b} = [b_0 \ \cdots \ b_m]^T$ by solving the constrained problem

$$\underset{\boldsymbol{a},\boldsymbol{b}}{\text{minimize}} \quad \max_{\omega\in\Omega} W(\omega)|H(e^{j\omega}) - H_d(\omega)| \tag{2a}$$

$$\text{subject to:} \quad H(z) \text{ stable} \tag{2b}$$

where $W(\omega) \geq 0$ is a known weighting function over Ω.

III. An Enhanced SM Procedure for Problem (2)

By introducing an upper bound β for $W(\omega)|H(e^{j\omega}) - H_d(\omega)|$ over Ω and treating β as an additional design variable, problem (2) becomes

$$\underset{\boldsymbol{a},\boldsymbol{b},\beta}{\text{minimize}} \ \beta \tag{3a}$$

$$\text{subject to: } W(\omega)|H(e^{j\omega}) - H_d(\omega)| \leq \beta \text{ for } \omega \in \Omega \tag{3b}$$

$$H(z) \ \text{ stable} \tag{3c}$$

Now we write the constraint in (3b) as

$$W(\omega)|b(e^{j\omega}) - H_d(\omega)a(e^{j\omega})| \leq \beta |a(e^{j\omega})| \tag{4}$$

In the $(k+1)$th iteration, we update the kth design variables $\{\boldsymbol{a}_k, \boldsymbol{b}_k, \beta_k\}$ to $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k + \boldsymbol{\delta}_a$, $\boldsymbol{b}_{k+1} = \boldsymbol{b}_k + \boldsymbol{\delta}_b$, and $\beta_{k+1} = \beta_k + \delta_\beta$ where $\{\boldsymbol{\delta}_a, \boldsymbol{\delta}_b, \delta_\beta\}$ are updates to be determined with $\boldsymbol{\delta}_a = [\delta_1^{(a)} \ \cdots \ \delta_n^{(a)}]^T$ and $\boldsymbol{\delta}_b = [\delta_0^{(b)} \ \cdots \ \delta_m^{(b)}]^T$.

If we denote the denominator and numerator polynomials associated with $\boldsymbol{a}_{k+1}$ and $\boldsymbol{b}_{k+1}$ by $a_{k+1}(e^{j\omega})$ and $b_{k+1}(e^{j\omega})$, respectively, iteration (4) in $(k+1)$th iteration becomes

$$W(\omega)|b_{k+1}(e^{j\omega}) - H_d(\omega)a_{k+1}(e^{j\omega})| \leq \beta_{k+1}|a_{k+1}(e^{j\omega})| \tag{5}$$

Note that constraint (5) is not convex because of the presence of term $|a_{k+1}(e^{j\omega})|$ on its right-hand side. If we replace

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

the right-hand side of (5) with its first-order approximation, namely,

$$\beta_{k+1}|a_{k+1}(e^{j\omega})| \approx |a_k(e^{j\omega})|\beta_{k+1} + \beta_k \nabla_a^T |a_k(e^{j\omega})| \boldsymbol{\delta}_a$$

where $\nabla_a |a_k(e^{j\omega})|$ denotes the gradient of $|a(e^{j\omega})|$ with respect to parameter vector $\boldsymbol{a}$ evaluated at $\boldsymbol{a} = \boldsymbol{a}_k$, then a convex relaxation of (5) is obtained as

$$\begin{aligned} &W(\omega)|b_{k+1}(e^{j\omega}) - H_d(\omega)a_{k+1}(e^{j\omega})| \\ &\le\ |a_k(e^{j\omega})|\beta_{k+1} + \beta_k \nabla_a^T |a_k(e^{j\omega})| \boldsymbol{\delta}_a \end{aligned} \tag{6}$$

Note that if the conventional SM procedure were applied to (5), the right-hand side of its convex relaxation would be $|a_k(e^{j\omega})|\beta_{k+1}$ which is the first term on the right-hand side of (6). Since the second term on the right side of (5) involves gradient $\nabla_a |a_k(e^{j\omega})|$, (6) can be regarded as a *first-order* SM procedure while the conventional SM is regarded as a *zeroth-order* procedure.

If we denote $H_d(\omega) = e^{-jk\omega} A_d(\omega)$ with desired passband group delay κ,

$$\begin{aligned} \boldsymbol{c}(\omega) &= [\cos(\kappa+1)\omega \cdots \cos(\kappa+n)\omega]^T \\ \boldsymbol{s}(\omega) &= [\sin(\kappa+1)\omega \cdots \sin(\kappa+n)\omega]^T \\ \hat{\boldsymbol{c}}(\omega) &= [1\ \ \cos\omega \cdots \cos m\omega]^T \\ \hat{\boldsymbol{s}}(\omega) &= [0\ \ \sin\omega \cdots \sin m\omega]^T \end{aligned}$$

(6) can be explicitly expressed in terms of updates $\{\boldsymbol{\delta}_a, \boldsymbol{\delta}_b, \delta_\beta\}$ as

$$W(\omega)\left\| \boldsymbol{M}(\omega)\begin{bmatrix}\boldsymbol{\delta}_a \\ \boldsymbol{\delta}_b\end{bmatrix} + \boldsymbol{c}_k(\omega)\right\|_2 \le \boldsymbol{p}_k(\omega)^T \begin{bmatrix}\delta_\beta \\ \boldsymbol{\delta}_a\end{bmatrix} + q_k(\omega) \tag{7}$$

where

$$\begin{aligned} \boldsymbol{M}(\omega) &= \begin{bmatrix} -A_d(\omega)\boldsymbol{c}^T(\omega) & \hat{\boldsymbol{c}}^T(\omega) \\ -A_d(\omega)\boldsymbol{s}^T(\omega) & \hat{\boldsymbol{s}}^T(\omega) \end{bmatrix}, \\ \boldsymbol{c}_k(\omega) &= \begin{bmatrix} -A_d(\omega)(\cos\kappa\omega + \boldsymbol{c}^T(\omega)\boldsymbol{a}_k) + \hat{\boldsymbol{c}}^T(\omega)\boldsymbol{b}_k \\ -A_d(\omega)(\sin\kappa\omega + \boldsymbol{s}^T(\omega)\boldsymbol{a}_k) + \hat{\boldsymbol{s}}^T(\omega)\boldsymbol{b}_k \end{bmatrix}, \\ \boldsymbol{p}_k(\omega) &= \begin{bmatrix} |a_k(e^{j\omega})| \\ \beta_k \nabla_a |a_k(e^{j\omega})| \end{bmatrix},\ q_k(\omega) = |a_k(e^{j\omega})|\beta_k \end{aligned}$$

For each $\omega \in \Omega$, (7) defines a second-order cone (SOC) which is known to be convex [18].

IV. Linear stability constraints based on enhanced Rouché theorem

A. Enhanced Rouch theorem

We consider

$$\tilde{a}(z) = z^n + a_1 z^{n-1} + \cdots + a_n$$

which is equivalent to $a(z)$ in (1) but $\tilde{a}(z)$ is a polynomial in z instead of z^{-1}, and we call $\tilde{a}(z)$ a Schur polynomial if all zeros of $\tilde{a}(z)$ are strictly inside the unit circle $\{z : |z| = 1\}$. Suppose $\tilde{a}(z)$ is a Schur polynomial and we want to examine the stability of polynomial $f(z) = \tilde{a}(z) + \tilde{\delta}(z)$ with $\tilde{\delta}(z)$ a perturbing polynomial. Rouché's theorem [16] says that as long as $\tilde{\delta}(z)$ obeys

$$|\tilde{\delta}(z)| < |\tilde{a}(z)| \text{ for all } z \text{ on the unit circle} \tag{8}$$

$f(z)$ remains to be Schur. Rouché's theorem is well known in complex analysis, what is not so well known is an enhanced variant of the theorem, which is stated below.

Theorem 1 If functions $f(z)$ and $\tilde{a}(z)$ are analytic inside and on a closed contour $\mathcal{C}$ and

$$|f(z) - \tilde{a}(z)| < |f(z)| + |\tilde{a}(z)| \tag{9}$$

holds true for every point z on C, then $f(z)$ and $\tilde{a}(z)$ have the same number of zeros inside $\mathcal{C}$. Theorem 1 and its proof can be found in [17, p. 156]. Interpreting the theorem in the context of IIR filters, we let $f(z) = \tilde{a}(z) + \tilde{\delta}(z)$ and take C to be the unit circle so that (9) becomes

$$|\tilde{\delta}(z)| < |\tilde{a}(z) + \tilde{\delta}(z)| + |\tilde{a}(z)| \text{ for all } z \text{ on the unit circle} \tag{10}$$

Clearly, any pair $\{\tilde{a}(z), \tilde{\delta}(z)\}$ satisfying (8) also satisfies (10) but not vice versa, and this explains why a stability region based on (10) is expected to be larger than that deduced from (8).

B. Stability constraints based on (10)

For z on the unit circle $\mathcal{C} = \{z : z = e^{j\omega},\ 0 \le \omega \le 2\pi\}$ we write

$$\begin{aligned} \tilde{a}(z) &= (\cos n\omega + \tilde{\boldsymbol{c}}(\omega)^T\boldsymbol{a}) + j(\sin n\omega + \tilde{\boldsymbol{s}}(\omega)^T\boldsymbol{a}) \\ &\equiv\ \ u(\omega,\boldsymbol{a}) + jv(\omega,\boldsymbol{a}) \qquad (11a) \\ \tilde{\delta}(z) &= \tilde{\boldsymbol{c}}(\omega)^T\boldsymbol{\delta}_a + j\tilde{\boldsymbol{s}}(\omega)^T\boldsymbol{\delta}_a \qquad (11b) \end{aligned}$$

hence

$$f(z) = [u(\omega,\boldsymbol{a}) + \tilde{\boldsymbol{c}}(\omega)^T\boldsymbol{\delta}_a] + j[v(\omega,\boldsymbol{a}) + \tilde{\boldsymbol{s}}(\omega)^T\boldsymbol{\delta}_a] \tag{11c}$$

where $\tilde{\boldsymbol{c}}(\omega) = [\cos(n-1)\omega\ \cdots\ \cos\omega\ 1]^T$ and $\tilde{\boldsymbol{s}}(\omega) = [\sin(n-1)\omega\ \cdots\ \sin\omega\ 0]^T$. Note that (10) is equivalent to

$$|\tilde{\delta}(z)|^2 < |\tilde{a}(z)+\tilde{\delta}(z)|^2 + |\tilde{a}(z)|^2 + 2|\tilde{a}(z)|\cdot|\tilde{a}(z)+\tilde{\delta}(z)| \tag{12}$$

The set of $\boldsymbol{\delta}_a$ satisfying (12) is nonconvex, and a convexification of (12) is needed in order to construct convex constraints on $\boldsymbol{\delta}_a$ so as for $f(z)$ to remain to be a Schur polynomial. This can be accomplished by dropping out the third term on the right-hand side of (12), which yields

$$|\tilde{\delta}(z)|^2 < |\tilde{a}(z) + \tilde{\delta}(z)|^2 + |\tilde{a}(z)|^2 \tag{13}$$

Because $2|\tilde{a}(z)| \cdot |\tilde{a}(z) + \tilde{\delta}(z)|$ is nonnegative, a pair $\{\tilde{a}(z), \tilde{\delta}(z)\}$ satisfying (13) also satisfies (12), hence (13) is a relaxation of (12). To show the convexity of constraint (13), we follow (11a)-(11c) to write $|\tilde{a}(z)|^2 = u(\omega,\boldsymbol{a})^2 + v(\omega,\boldsymbol{a})^2$, $|\tilde{\delta}(z)|^2 = \boldsymbol{\delta}_a^T\boldsymbol{V}(\omega)\boldsymbol{\delta}_a$ with $\boldsymbol{V}(\omega) = \tilde{\boldsymbol{c}}(\omega)\tilde{\boldsymbol{c}}(\omega)^T + \tilde{\boldsymbol{s}}(\omega)\tilde{\boldsymbol{s}}(\omega)^T \succeq \boldsymbol{0}$; and $|\tilde{a}(z) + \tilde{\delta}(z)|^2 = u(\omega,\boldsymbol{a})^2 + v(\omega,\boldsymbol{a})^2 + 2[u(\omega,\boldsymbol{a})\tilde{\boldsymbol{c}}(\omega) + v(\omega,\boldsymbol{a})\tilde{\boldsymbol{s}}(\omega)]^T\boldsymbol{\delta}_a + \boldsymbol{\delta}_a^T\boldsymbol{V}(\omega)\boldsymbol{\delta}_a$. Hence (13) becomes

$$-[u(\omega,\boldsymbol{a})\tilde{c}(\omega) + v(\omega,\boldsymbol{a})\tilde{\boldsymbol{s}}(\omega)]^T\boldsymbol{\delta}_a < u(\omega,\boldsymbol{a})^2 + v(\omega,\boldsymbol{a})^2 \tag{14}$$

To construct a *closed* convex region, (14) is further relaxed to

$$-[u(\omega,\boldsymbol{a})\tilde{\boldsymbol{c}}(\omega) + v(\omega,\boldsymbol{a})\tilde{\boldsymbol{s}}(\omega)]^T\boldsymbol{\delta}_a \le u(\omega,\boldsymbol{a})^2 + v(\omega,\boldsymbol{a})^2 - \varepsilon \tag{15}$$

with $\varepsilon > 0$ an arbitrarily small constant. For practical purposes, constraint (15) is imposed over a set of dense frequency grids $\tilde{\Omega} = \{\omega_i, i = 1,\ 2,\ \ldots,\ K\}$ and this leads (15) to

$$\boldsymbol{A}\boldsymbol{\delta}_a \leq \boldsymbol{p} \tag{16}$$

where $\boldsymbol{A} \in R^{K\times n}$ with $-[u(\omega_i,\boldsymbol{a})\tilde{\boldsymbol{c}}(\omega_i)+v(\omega_i,\boldsymbol{a})\tilde{\boldsymbol{s}}(\omega_i)]^T$ as its ith row and $\boldsymbol{p} \in R^{K\times 1}$ with $u(\omega_i,\boldsymbol{a})^2+v(\omega_i,\boldsymbol{a})^2-\varepsilon$ as ith entry. With a given Schur polynomial $\tilde{a}(z)$ and set $\tilde{\Omega}$, both matrix $\boldsymbol{A}$ and vector $\boldsymbol{p}$ are constant, hence (16) contains K linear inequalities that define a region $\mathcal{S} = \{\boldsymbol{\delta}_a : \boldsymbol{A}\boldsymbol{\delta}_a \leq \boldsymbol{p}\}$ so that $f(z) = \tilde{a}(z) + \tilde{\delta}(z)$ is Schur as long as $\boldsymbol{\delta}_a \in \mathcal{S}$.

C. Linear constraints for robust stability

By applying the enhanced Rouché theorem to a contour which is the circle with center at the origin and radius r, namely $\mathcal{C}_r = \{z : z = r\boldsymbol{e}^{j\omega},\ 0 \leq \omega \leq 2\pi\}$, for some fixed $r < 1$, an analysis similar to that given above can be carried out to construct linear constraints

$$\boldsymbol{A}_r\boldsymbol{\delta}_a \leq \boldsymbol{p}_r \tag{17}$$

where $\boldsymbol{A}_r \in R^{K\times n}$ with $-[u(\omega_i,\boldsymbol{a})\tilde{\boldsymbol{c}}_r(\omega_i)+v(\omega_i,\boldsymbol{a})\tilde{\boldsymbol{s}}_r(\omega_i)]^T$ as its ith row, $\boldsymbol{p}_r \in R^{K\times 1}$ with $u_r(\omega_i,\boldsymbol{a})^2+v_r(\omega_i,\boldsymbol{a})^2-\varepsilon$ as its ith entry, $\tilde{\boldsymbol{c}}_r(\omega) = [r^{n-1}\cos(n-1)\omega\ \ \cdots\ \ r\cos\omega\ \ 1]^T$, $\tilde{\boldsymbol{s}}_r(\omega) = [r^{n-1}\sin(n-1)\omega\ \ \cdots\ \ r\sin\omega\ \ 0]^T$, $u_r(\omega,\boldsymbol{a}) = r^n\cos n\omega + \tilde{\boldsymbol{c}}_r(\omega)^T\boldsymbol{a}$, and $v_r(\omega,\boldsymbol{a}) = r^n\sin n\omega + \tilde{\boldsymbol{s}}_r(\omega)^T\boldsymbol{a}$. Due to space limitation the details involved in deriving (17) are omitted. The linear constraints in (17) define a region $\mathcal{S}_r = \{\boldsymbol{\delta}_a : \boldsymbol{A}_r\boldsymbol{\delta}_a \leq \boldsymbol{p}_r\}$ to ensure that the largest magnitude of the zeros of $f(z) = \tilde{a}(z) + \tilde{\delta}(z)$ is no greater that r as long as $\tilde{a}(z)$ is r-stable (i.e., its zeros reside inside $\mathcal{C}_r$) and $\boldsymbol{\delta}_a \in \mathcal{S}_r$.

D. Relation of (13) to Strictly-Positive-Realness based Stability Criterion [11]

It has been shown [11] that $f(z) = \tilde{a}(z) + \tilde{\delta}(z)$ is a Schur polynomial if

$$G(z) = 1 + \frac{\tilde{\delta}(z)}{\tilde{a}(z)}$$

is *strictly positive real* (SPR), that is, $\tilde{a}(z)$ is a Schur polynomial and $\mathrm{Re}[G(e^{j\omega})] > 0$ for any $\omega \in [0,\ 2\pi]$. Note that on contour $\mathcal{C}$ we have $z = e^{j\omega}$ and the convex relaxation in (13) becomes $|\tilde{\delta}(e^{j\omega})|^2 < |\tilde{a}(e^{j\omega})+\tilde{\delta}(e^{j\omega})|^2+|\tilde{a}(e^{j\omega})|^2$ from which we deduce

$$\begin{aligned}
|\tilde{\delta}(e^{j\omega})|^2 &< |\tilde{a}(e^{j\omega})+\tilde{\delta}(e^{j\omega})|^2+|\tilde{a}(e^{j\omega})|^2\\
\leftrightarrow\quad & \left|\frac{\tilde{\delta}(e^{j\omega})}{\tilde{a}(e^{j\omega})}\right|^2 < \left|1+\frac{\tilde{\delta}(e^{j\omega})}{\tilde{a}(e^{j\omega})}\right|^2+1\\
\leftrightarrow\quad & \left|\frac{\tilde{\delta}(e^{j\omega})}{\tilde{a}(e^{j\omega})}\right|^2 < 2+\left|\frac{\tilde{\delta}(e^{j\omega})}{\tilde{a}(e^{j\omega})}\right|^2+2\mathrm{Re}\left[\frac{\tilde{\delta}(e^{j\omega})}{\tilde{a}(e^{j\omega})}\right]\\
\leftrightarrow\quad & \mathrm{Re}\left[1+\frac{\tilde{\delta}(e^{j\omega})}{\tilde{a}(e^{j\omega})}\right] > 0 \leftrightarrow \mathrm{Re}\left[G(e^{j\omega})\right] > 0
\end{aligned}$$

Therefore, given a Schur polynomial $\tilde{a}(z)$, (13) (hence (16)) is equivalent to the SPR of $G(z)$. Consequently, the stability region defined by (16) is expected to be practically the same as that characterized based on the SPR of $G(z)$ [11]. The above analysis also motivates an alternative derivation of (16) based on the SPR of $G(z)$. This can be done by deducing an explicit version of the SPR of $G(z)$ in term of vectors $\boldsymbol{a}$ and $\boldsymbol{\delta}_a$ as follows.

$$\begin{aligned}
&\mathrm{Re}\left[1+\frac{\tilde{\delta}(e^{j\omega})}{\tilde{a}(e^{j\omega})}\right] = \mathrm{Re}\left[1+\frac{\tilde{\boldsymbol{c}}(\omega)^T\boldsymbol{\delta}_a + j\tilde{\boldsymbol{s}}(\omega)^T\boldsymbol{\delta}_a}{u(\omega,\boldsymbol{a})+jv(\omega,\boldsymbol{a})}\right]\\
&= \mathrm{Re}\left[1+\frac{[u(\omega,\boldsymbol{a})-jv(\omega,\boldsymbol{a})][\tilde{\boldsymbol{c}}(\omega)^T\boldsymbol{\delta}_a + j\tilde{\boldsymbol{s}}(\omega)^T\boldsymbol{\delta}_a]}{u(\omega,\boldsymbol{a})^2+v(\omega,\boldsymbol{a})^2}\right]\\
&= 1+\frac{[u(\omega,\boldsymbol{a})\tilde{\boldsymbol{c}}(\omega)+v(\omega,\boldsymbol{a})\tilde{\boldsymbol{s}}(\omega)]^T\boldsymbol{\delta}_a}{u(\omega,\boldsymbol{a})^2+v(\omega,\boldsymbol{a})^2}
\end{aligned}$$

Hence $\mathrm{Re}\big[G(e^{j\omega})\big] > 0$ if and only if

$$u(\omega,\boldsymbol{a})^2+v(\omega,\boldsymbol{a})^2+[u(\omega,\boldsymbol{a})\tilde{\boldsymbol{c}}(\omega)+v(\omega,\boldsymbol{a})\tilde{\boldsymbol{s}}(\omega)]^T\boldsymbol{\delta}_a > 0$$

which coincides with (14), hence implies the constraints in (16). For illustration, suppose $\tilde{a}(z) = z^2 + 0.6z + 0.7$ which is a Schur polynomial. With $\boldsymbol{a} = [0.6\ \ 0.7]^T$, $\boldsymbol{\delta}_a = [\delta_1\ \ \delta_2]^T$, $K = 100$ and $\varepsilon = 0.01$, the linear constraints in (16) defines a convex stability region for polynomial $\tilde{a}(z) = z^2 + (0.6 + \delta_1)z + (0.7 + \delta_2)$, which is shown in Fig. 1 in magenta. For comparison, the stability region constructed based on Rouché theorem [7] is depicted in green. Note that there is a fine curve in black compassing the magenta region, which was produced as the boundary of the stability region based on SPR [11]. With a reduced value of ε the region defined by (16) is found to be practically the same as that produced by SPR.

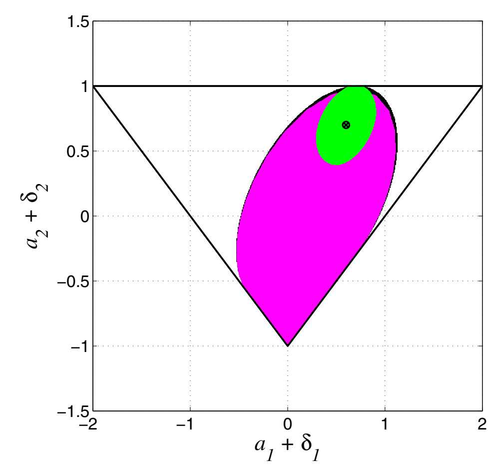

Fig. 1. Stability region for $\tilde{a}(z)$ given by (16) with $K = 100$ and $\varepsilon = 0.01$ in magenta versus that from Rouché theorem [7] in green.

V. ALGORITHM AND DESIGN EXAMPLES

Applying the results obtained from Secs. III and IV, the nonconvex problem in (3) is relaxed to

$$\begin{aligned}
&\text{minimize} && \delta_\beta && \text{(18a)}\\
&\text{subject to:} && \text{constraint (7) for } \omega \in \Omega_d && \text{(18b)}\\
& && \boldsymbol{A}_r\boldsymbol{\delta}_a \leq \boldsymbol{p}_r \text{in (17) for } \omega \in \tilde{\Omega} && \text{(18c)}
\end{aligned}$$

where $\Omega_d = \{\omega_l, l = 1,\ 2,\ \dots,\ L\} \subset [0, \omega_p] \bigcup [\omega_a, \pi]$ and $\tilde{\Omega} = \{\omega_i,\ i = 1,\ 2,\ \dots,\ K\} \subset [0, 2\pi]$ are dense frequency grids that are evenly distributed in the respective frequency ranges. The objective function as well as constraints (18c) are linear, while the constraints in (18b) are second-order convex cones. Hence (18) is a convex (known as SOCP) problem [18]. Based on this, an algorithm for the design of stable IIR digital filter of order (n, m) can be outlined as follows.

Algorithm 1

Step 1

Input desired $H_d(\omega)$, filter order (n, m), frequency region of interest Ω, stable initial point $\{\boldsymbol{a}_0, \boldsymbol{b}_0\}$, weight $W(\omega)$, parameters L, K, r, ε, and tolerance ε_c. Set $k = 0$. Compute

$$\beta_0 = \max_{\omega \in \Omega} W(\omega)|H_0(e^{j\omega}) - H_d(\omega)|;$$

Step 2

Solve (18) to update $\{\boldsymbol{a}_k, \boldsymbol{b}_k, \beta_k\}$ to $\{\boldsymbol{a}_{k+1} = \boldsymbol{a}_k + \boldsymbol{\delta}_a,\ \boldsymbol{b}_{k+1} = \boldsymbol{b}_k + \boldsymbol{\delta}_b,\ \beta_{k+1} = \beta_k + \delta_\beta\}$;

Step 3

If $|\delta_\beta| < \varepsilon_c$, stop and output $\{\boldsymbol{a}_{k+1}, \boldsymbol{b}_{k+1}\}$ as the solution; otherwise set $k = k + 1$ and repeat from Step 2.

The algorithm converges due to the fact that objective function in (18a) monotonically decreases as the iteration proceeds while is bounded from below because $\delta_\beta \geq 0$. On the other hand, the design produced from the algorithm is only a local solution because the original problem in (3) is nonconvex. Below we present two examples to illustrate the proposed design method.

Example 1 Algorithm 1 was applied to design a lowpass filter of order $(n, m) = (4, 15)$ with normalized passband edge $\omega_p = 0.4\pi$, stopband edge $\omega_a = 0.56\pi$, and passband group delay $\kappa = 12$. The initial design $\{\boldsymbol{a}_0, \boldsymbol{b}_0\}$ was taken to be $\boldsymbol{a}_0 = \mathbf{0}$ and $\boldsymbol{b}_0 =$ the impulse response of a least-squares linear-phase FIR filter of order 15. The weight $W(\omega)$ was a piecewise constant function being unity over the passband and constant w over the stopband, where the value of w was determined by trial-and-error to equalize the approximation errors in passband and stopband. With $L = 400$, $K = 1000$, $r = 0.9$, $w = 1.01$, $\varepsilon = 0.002$, and $\varepsilon_c = 2 \times 10^{-9}$, it took the algorithm 23 iterations to converge to a solution. The maximum amplitude of the poles was 0.8596. The amplitude response of the IIR filter obtained is depicted in Fig. 2.

The largest deviation in passband amplitude over that passband was 0.045 dB that corresponds to an error of 0.005121 in passband gain; the minimum stopband attenuation was -45.7722 dB that corresponds to a stopband approximation error of 0.005145. The design specifications used in this example were identical to Example 3 in [15] as well as Example 1 in [14], and the approximation errors in the passband and stopband were 0.005177 and 0.005181 (with 69 iterations) in [15], and 0.005127 and 0.005211 in [14].

Example 2 Algorithm 1 was applied to design a lowpass filter of order $(n, m) = (6, 12)$ with normalized passband edge $\omega_p = 0.5\pi$, stopband edge $\omega_a = 0.6\pi$, and passband group delay $\kappa = 9$. The initial design $\{\boldsymbol{a}_0, \boldsymbol{b}_0\}$ was taken to be $\boldsymbol{a}_0 = \mathbf{0}$ and $\boldsymbol{b}_0 =$ the impulse response of a least-squares linear-phase FIR filter of order 12.

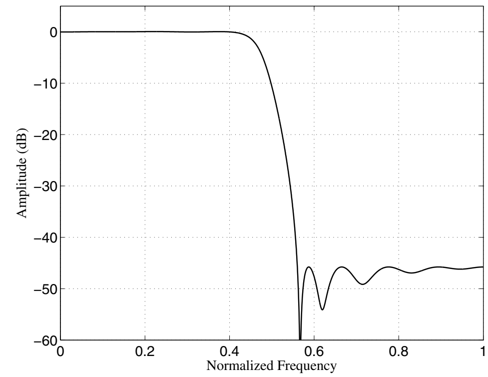

Fig. 2. Amplitude response of the IIR filter in Example 1.

With $L = 400$, $K = 1000$, $r = 0.975$, $w = 1.112$, $\varepsilon = 0.002$, and $\varepsilon_c = 8 \times 10^{-5}$, it took the algorithm 29 iterations to converge to a solution. The maximum amplitude of the poles was 0.9422. The amplitude response of the IIR filter obtained is depicted in Fig. 3.

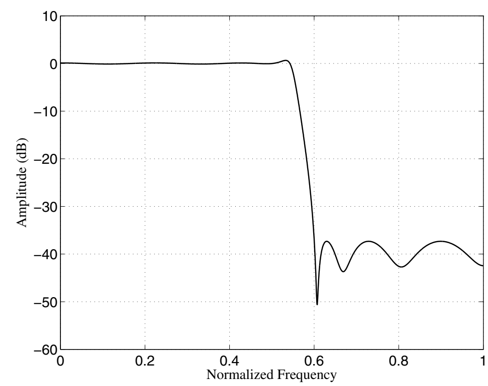

Fig. 3. Amplitude response of the IIR filter in Example 2.

The largest deviation in passband amplitude over that passband was 0.1192 dB that corresponds to an error of 0.01363 in passband gain; the minimum stopband attenuation was -37.3177 dB that corresponds to a stopband approximation error of 0.01362. The design specifications used in this example were identical to Example 5 in [15] as well as Example 1 in [10], and the approximation errors in the passband and stopband were 0.0147 and 0.0149 (with 84 iterations) in [15] and 0.0155 and 0.0155 in [10].

VI. Conclusion

An enhanced SM procedure, that performs first-order convex relaxation for the denominator polynomial, has been proposed for the design of IIR filters. A stronger variant of Rouché's theorem is reviewed and shown to be equivalent to that based on SPR. Design examples are presented to evaluate the proposed design method relative to several existing ones in the literature.

References

[1] A. Antoniou, *Digital Filters: Analysis, Design, and Applications*, 2nd ed., McGraw Hill, 1993.

[2] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete Time Signal Processing*, 2nd ed., Prentice Hall, 1999.

[3] S. K. Mitra, *Digital Signal Processing A Computer-Based Approach*, 3rd ed., McGraw Hill, 2006.

[4] A. T. Chottera and G. A. Jullien, "A linear programming approach to recursive digital filter design with linear phase," *IEEE Trans. Circuits Syst.*, vol. 29, pp. 139–149, Mar. 1982.

[5] Y. C. Lim, J.-H. Lee, C. K. Chen, and R.-H. Yang, "A weighted least-squares algorithm for quasi-equiripple FIR and IIR digital filter design," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 551–558, Mar. 1998.

[6] W.-S. Lu, S.-C. Pei, and C.-C. Tseng, "A weighted least-squares method for the design of 1-D and 2-D IIR digital filters," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 1–10, Jan. 1998.

[7] M. C. Lang, "Least-squares design of IIR filters with prescribed magnitude and phase responses and a pole radius constraint," *IEEE Trans. Signal Process.*, vol. 48, no. 11, pp. 3109–3121, Nov. 2000.

[8] C.-C. Tseng, "Stable IIR digital differentiator design using iterative quadratic programming approach," *Signal Process.*, vol. 80, pp. 857–866, 2000.

[9] W.-S. Lu, "Design of stable minimax IIR digital filters using semidefinite programming," *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 1, pp. 355–358, Geneva, May 2000.

[10] R. Niemistö and B. Dumitrescu, "Simplified procedures for quasi-equiripple IIR filter design," *IEEE Signal Process. Lett.*, vol. 11, pp. 308–311, Mar. 2004.

[11] B. Dumitrescu and R. Niemistö, "Multistage IIR filter design using convex stability domains defined by positive realness," *IEEE Trans. Signal Process.*, vol. 52, no. 4, pp. 962–974, April 2004.

[12] W.-S. Lu, "An argument-principle based stability criterion and application to the design of IIR digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 4431–4434, Island of Kos, Greece, May 2006.

[13] A. Jiang and H. K. Kwan, "IIR digital filter design with new stability constraint based on argument principle," *IEEE Trans. Circuits Syst.*, I, vol. 56, no. 3, pp. 583–593, Mar. 2009.

[14] A. Jiang and H. K. Kwan, "Minimax design of IIR digital filters using SDP relaxation technique," *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 477–480, Taipei, May 2009.

[15] X. Lai and Z. Lin, "Minimax design of IIR digital filters using a sequential constrained least squares method," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3901–3906, July 2010.

[16] E. C. Titchmarsh, *The Theory of Functions*, 2nd ed., Oxford University Press, 1979.

[17] T. Estermann, *Complex Numbers and Functions*, University of London, Athlone Press, 1962.

[18] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*, Springer, 2007.

[19] A. Jiang, H. K. Kwan, Y. Zhu, N. Xu, and X. Liu, "Minimax design of IIR digital filters using partial second-order factor updates," *Proc. IEEE Conf. DSP*, July 2015.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE