Hankel-norm approximation of FIR filters: A descriptor-systems based approach

George Halikias*, Vasilis Tsoulkas†, Athanasios Pantelous‡ and Efstathios Milonidis*

1. Abstract

We propose a new method for approximating a matrix Finite Impulse Response (FIR) filter by an Infinite Impulse Response (IIR) filter of lower McMillan degree. This is based on a technique for approximating discrete-time descriptor systems and requires only standard linear algebraic routines, while avoiding altogether the solution of two matrix Lyapunov equations which is computationally expensive. Both the optimal and the suboptimal cases are addressed using a unified treatment. A detailed solution is developed in state-space or polynomial form, using only the Markov parameters of the FIR filter which is approximated. The method is finally applied to the design of scalar Infinite Impulse Response (IIR) filters with specified magnitude frequency-response tolerances and approximately linear phase characteristics. A-priori bounds on the magnitude and phase errors are obtained which may be used to select the reduced-order IIR filter order which satisfies the specified design tolerances. The effectiveness of the method is illustrated with a numerical example. Additional applications of the method are also briefly discussed.

Keywords: Hankel-norm approximation, descriptor systems, FIR filters, linear phase response.

2. Introduction and Notation

In this paper we apply Hankel-norm approximation techniques to (matrix) FIR filters. Our technique is based on results involving approximations of discrete-time descriptor systems (see [1], [2], [10]). This allows us to treat systems with poles at the origin and has been applied successfully in the context of mixed $\mathcal{H}_\infty/\mathcal{H}_2$ optimization [11]. Our results apply both to the γ-suboptimal and the strictly optimal problem, although in the later case the resulting state-space realization is non-minimal. Using an all-pass matrix dilation technique, a state-space parametrization of the complete family of solutions can be obtained, in the form of a linear fractional map of the ball of unstable contractions [1], [2], [9]. The solution is derived entirely in terms of the Markov parameters of the FIR filter which is approximated and can be expressed in either state-space or transfer-function form. Additional advantages of our method over parallel techniques [3], [5], [6], [12], [17] include computational efficiency and a unified treatment for the optimal and the sub-optimal case. The most recent reference in the area is [6] which develops an elegant self-contained approach for the solution of the optimal Hankel norm approximation problem of FIR filters in the scalar case. Our paper develops an alternative procedure for the solution of this problem using the descriptor-based approach of [1], [2], [9]. The solution, initially developed for the matrix-valued version of the problem and subsequently specialized to the scalar case, is obtained in terms of the Markov parameters of the filter which is approximated. This may prove helpful in establishing links

*Control Engineering Research Centre, School of Engineering and Mathematical Sciences, City University, Northampton Square, London EC1V 0HB, United Kingdom, e-mails: G.Halikias@city.ac.uk, e.milonidis@city.ac.uk

†General Secretariat for Research and Technology and Department of Mathematics, University of Athens, Athens, Greece, e-mail: btsu@gsrt.gr

‡Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZL, United Kingdom, e-mail: A.Pantelous@liverpool.ac.uk

between model reduction, statistical identification and partial realization theory (Markov parameters are often obtained by direct measurement in identification experiments).

In the second part of the paper the results are applied to design IIR filters with almost linear phase characteristics. Finite Impulse Response (FIR) filters are widely used in many digital signal processing applications, especially in the field of communications, because they can exhibit *linear* phase characteristics. Unfortunately, in many cases the order of such filters is prohibitively high for practical implementation. In general, the number of delay elements and multipliers needed for an FIR design tends to be much higher compared to similar Infinite Impulse Response (IIR) implementations. This is especially true for filters with sharp cut-off characteristics which have a long impulse response [14], [18]. It is therefore natural to ask whether a linear-phase FIR filter $H(z)$ can be approximated by a low-order IIR filter, without degrading significantly its magnitude and phase characteristics. In this work the approach adopted consists of two steps. First, a linear phase FIR filter is designed subject to pre-specified magnitude-response tolerances using optimization methods. This is subsequently approximated in the second step of the design by a reduced-order IIR filter which has approximately linear phase-response characteristics. Magnitude and phase error bounds of the approximation error can be obtained in terms of the Hankel singular values of $H(z)$, which allows the a-priori determination of the IIR filter order satisfying specified magnitude ("ripple") specifications and an acceptable phase deviation from linearity. The effectiveness of the method is illustrated via a numerical example, involving the approximation of a high-order linear-phase FIR filter, designed using linear programming techniques [16]. Finally, extensions and further applications of the method are briefly discussed.

Most of the notation used is standard and is reproduced here for completeness. $\mathcal{D}$ denotes the open unit disc $\mathcal{D} = \{\zeta \in \mathcal{C} : |\zeta| < 1\}$, with $\bar{\mathcal{D}}$ and $\partial\mathcal{D}$ its closure and boundary respectively. $\mathcal{L}_2(\partial\mathcal{D})$ denotes the Hilbert space of all matrix-valued functions F defined on the unit circle such that

$$\int_0^{2\pi} \text{trace}[F^*(e^{j\omega})F(e^{j\omega})]d\omega < \infty$$

where $(\cdot)^*$ denotes the complex conjugate transpose of a matrix. The corresponding inner product of two $\mathcal{L}_2(\partial\mathcal{D})$ functions F and G of compatible dimensions is given as:

$$\langle F, G\rangle = \frac{1}{2\pi}\int_0^{2\pi} \text{trace}[F^*(e^{j\omega})G(e^{j\omega})]d\omega$$

$\mathcal{H}_2(\partial\mathcal{D})$ and $\mathcal{H}_2^\perp(\partial\mathcal{D})$ are the closed subspaces of $\mathcal{L}_2$ consisting of all functions analytic in $\mathcal{C} \setminus \bar{\mathcal{D}}$ and $\mathcal{D}$, respectively. $\mathcal{L}_\infty(\partial\mathcal{D})$ is the space of all uniformly-bounded matrix-valued functions in $\partial\mathcal{D}$, i.e. all functions defined on the unit circle whose norm:

$$\|F\|_\infty = \sup_{\omega\in[0,2\pi)} \bar{\sigma}[F(e^{j\omega})]$$

is finite. Here $\bar{\sigma}(\cdot)$ denotes the largest singular value of a matrix. $\mathcal{H}_\infty(\partial\mathcal{D})$ and $\mathcal{H}_\infty^-(\partial\mathcal{D})$ denote the closed subspaces of $\mathcal{L}_\infty(\partial\mathcal{D})$ consisting of all functions analytic in $\mathcal{C} \setminus \bar{\mathcal{D}}$ and $\mathcal{D}$, respectively, while $\mathcal{H}_\infty^{-,k}(\partial\mathcal{D})$ is the set of all functions in $\mathcal{L}_\infty(\partial\mathcal{D})$ with no more than k poles in $\mathcal{D}$. Spaces of real-rational functions will be indicated by the suffix $\mathcal{R}$ before the corresponding space symbol. The *unit ball* of $\mathcal{H}_\infty^-(\partial\mathcal{D})$ is the set $\mathcal{B}\mathcal{H}_\infty^-(\partial\mathcal{D}) = \{U \in \mathcal{H}_\infty^-(\partial\mathcal{D}) : \|U\|_\infty \le 1\}$. If $G \in \mathcal{L}_\infty(\partial\mathcal{D})$, then the Hankel operator with symbol G is defined as:

$$\Gamma_G : \mathcal{L}_2(\partial\mathcal{D}) \to \mathcal{H}_2(\partial\mathcal{D}),\ \ \Gamma_G = \Pi_+ M_G|_{\mathcal{H}_2^\perp(\partial\mathcal{D})}$$

in which M_G denotes the multiplication operator $M_G : \mathcal{L}_2(\partial\mathcal{D}) \to \mathcal{L}_2(\partial\mathcal{D})$, $M_G f = Gf$ and Π_+ denotes the orthogonal projection $\Pi_+ : \mathcal{L}_2(\partial\mathcal{D}) \to \mathcal{H}_2(\partial\mathcal{D})$. If

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \in \mathcal{R}\mathcal{L}_\infty^{(p_1+p_2)\times(m_1+m_2)}(\partial\mathcal{D})$$

with $H_{ij} \in \mathcal{RL}_\infty^{p_i \times m_j}(\partial\mathcal{D})$, $i = 1, 2$, $j = 1, 2$ and $U \in \mathcal{RL}_\infty^{m_2 \times p_2}(\partial\mathcal{D})$, we define the *lower linear fractional map* of H and U as $\mathcal{F}_l(H, U) = H_{11} + H_{12}U(I - H_{22}U)^{-1}H_{21}$, provided that $\det(I - H_{22}(\infty)U(\infty)) \neq 0$. If $\mathcal{U}$ is a set of $m_2 \times p_2$ matrix functions, then $\mathcal{F}_l(H,\mathcal{U})$ denotes the set $\{\mathcal{F}_l(H, U) : U \in \mathcal{U}\}$.

3. Hankel-norm approximation of FIR filters

In this section we propose a Hankel-norm method for approximating FIR filters via lower-order IIR filters. The method is based on a recent result involving model-reduction of descriptor discrete-time systems [1], [2].

The main advantage of Hankel-norm approximation methods over other model-reduction techniques is that they offer tight bounds on the infinity-norm of the approximation error; in particular, it is shown in [9] how to construct k-th order approximations $X(z) \in \mathcal{RH}_\infty(\partial\mathcal{D})$ of $H(z) \in \mathcal{RH}_\infty(\partial\mathcal{D})$ with $\deg X(z) = k < \deg H(z) = n$, such that:

$$\|H(z) + X(z)\|_\infty \leq \sum_{i=k+1}^{n} \sigma_i(\Gamma_H) \tag{1}$$

where $\sigma_i(\Gamma_H)$ denotes the i-th singular value of Γ_H, indexed in non-increasing order of magnitude. This inequality can be used to determine a-priori the order k of the low-order system $X(z)$ which satisfies magnitude error specifications. The method was extended by [8] to discrete-time systems, under the assumption that the system which is approximated does not have poles at the origin. This assumption is not valid for the problem under consideration in this work where FIR filters are considered. Although the assumption is technical and can be easily removed (e.g. via bilinear transformations) the more general descriptor approximation framework of [1], [2] is more appropriate to our purposes.

Theorem 2 below parametrises all solutions $X(z)$ to the Hankel-norm approximation problem $\|\Gamma_H + \Gamma_X\| \leq \gamma$, in which $H(z)$ is the matrix FIR filter

$$H(z) = H_0 + H_1 z^{-1} + \ldots + H_n z^{-n}$$

and $X(z)$ is a matrix IIR filter of degree $\deg X(z) \leq k$. The parametrisation is given in descriptor form and hence applies both in the sub-optimal case ($\sigma_{k+1}(\Gamma_H) < \gamma < \sigma_k(\Gamma_H)$) and the optimal case ($\gamma = \sigma_{k+1}(\Gamma_H)$). Before stating this theorem, however, the following preliminary result is needed:

Theorem 1: Let

$$\tilde{H}(z) = H(z) - H_0 = H_1 z^{-1} + H_2 z^{-2} + \ldots + H_n z^{-n}$$

with $H_i \in \mathcal{R}^{p \times l}$ for $i = 1, 2, \ldots, n$. Then:

1. The singular values of $\Gamma_{\tilde{H}}$, $\sigma_i(\Gamma_{\tilde{H}})$ (indexed in decreasing order of magnitude) are the singular values of the (Hankel) matrix:

$$R_1 = \begin{pmatrix} H_1 & H_2 & \ldots & H_{n-1} & H_n \\ H_2 & H_3 & \ldots & H_n & 0 \\ H_3 & H_4 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ H_{n-1} & H_n & \ldots & 0 & 0 \\ H_n & 0 & \ldots & 0 & 0 \end{pmatrix}$$

2. A state-space realisation of $\tilde{H}(z)$ is given by $\tilde{H}(z) = C(zI - A)^{-1}B$ with:

$$A = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ I_p & 0 & 0 & \dots & 0 & 0 \\ 0 & I_p & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & I_p & 0 \end{pmatrix} \in \mathcal{R}^{np \times np}$$

$$B = \begin{pmatrix} H_n^t & H_{n-1}^t & \dots & H_2^t & H_1^t \end{pmatrix}^t \in \mathcal{R}^{np \times l}$$

$$C = \begin{pmatrix} 0 & 0 & \dots & 0 & I_p \end{pmatrix} \in \mathcal{R}^{p \times np}$$

Further, the realisation is output balanced, i.e. the unique solution of the Lyapunov equation $Q = A^tQA + C^tC$ is given by $Q = I_{np}$.

3. The controllability grammian of the realisation in part 2, i.e. the unique solution of the Lyapunov equation $P = APA^t + BB^t$, can be factored as $P = R_2R_2^t$ where:

$$R_2 = \begin{pmatrix} H_n & 0 & \dots & 0 \\ H_{n-1} & H_n & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ H_1 & H_2 & \dots & H_n \end{pmatrix}$$

Equivalently P can be block-partitioned as:

$$(P_{ij})_{i=1,2,\dots n}^{j=1,2,\dots n} \text{ where } P_{ij} = \sum_{k=1}^{\min(i,j)} H_{n-i+k}H_{n-j+k}^t$$

and $\sigma_i^2(\Gamma_{\tilde{H}}) = \lambda_i(P)$ for each i.

Proof: Straightforward and hence omitted. □

The following Theorem gives a complete parametrization of all $k - th$ order γ-suboptimal Hankel-norm approximations of matrix FIR filters in state-space form. All optimal k-th order approximations can also be obtained from the result given in the Theorem by setting $\gamma = \sigma_{k+1}(\Gamma_{\tilde{H}})$.

Theorem 2: Let $\tilde{H}(z) = C(zI_{np} - A)^{-1}B$ be the realization of the matrix FIR filter defined in Theorem 1. Let γ satisfy $\sigma_{k+1}(\Gamma_{\tilde{H}}) \leq \gamma < \sigma_k(\Gamma_{\tilde{H}})$ and assume that the matrix $A_0 := P - \gamma^2 ZA^tZ^{-t}$ is non-singular, where P is defined in Theorem 1 and $Z := I_{np} + A$. Then all $X(z) \in \mathcal{H}_\infty^{-,k}(\partial\mathcal{D})$ such that $\|\tilde{H}(z) + X(z)\|_\infty \leq \gamma$ are generated via the lower linear fractional transformation:

$$\mathcal{X} = \{\mathcal{F}_l(H_a, U) : \gamma U \in \mathcal{B}\mathcal{H}_\infty^-(\partial\mathcal{D})\}$$

where:

$$H_a(z) = D_H + C_H(zI - A_H)^{-1}B_H$$

with

$$A_H = -I_{pn} - (\gamma^2 I_{pn} - P)A_0^{-1}Z, \quad B_H = -(\gamma^2 I_n - P)A_0^{-1}\begin{pmatrix} B & B_0 \end{pmatrix}$$

$$C_H = -\begin{pmatrix} CZ^{-1}P \\ C_0 \end{pmatrix} A_0^{-1}Z, \quad D_H = \begin{pmatrix} CZ^{-1}B - CZ^{-1}PA_0^{-1}B & \gamma I - CZ^{-1}PA_0^{-1}B_0 \\ \gamma I - C_0A_0^{-1}B & -C_0A_0^{-1}B_0 \end{pmatrix}$$

where $B_0 := \gamma ZZ^{-t}C^t$ and $C_0 = \gamma B^tZ^{-t}$.

Proof: Let $\tilde{H}(z)$ be the matrix FIR filter defined in Theorem 1 and assume that $\sigma_{k+1}(\Gamma_{\tilde{H}}) \leq \gamma < \sigma_k(\Gamma_{\tilde{H}})$. Then, it follows from [1], [2] that all $X(z) \in \mathcal{H}_\infty^{-,k}(\partial\mathcal{D})$ such that $\|\tilde{H}(z) + X(z)\|_\infty \leq \gamma$ are generated via the lower linear fractional transformation:

$$\mathcal{X} = \{\mathcal{F}_l(S_a, U) : \gamma U \in \mathcal{BH}_\infty^-(\partial\mathcal{D})\} = S_{11} + \gamma^{-1} S_{12}\mathcal{BH}_\infty^-(\partial\mathcal{D})(I - \gamma^{-1} S_{22}\mathcal{BH}_\infty^-(\partial\mathcal{D}))^{-1} S_{21}$$

where:

$$S_a(z) = \begin{pmatrix} S_{11}(z) & S_{12}(z) \\ S_{21}(z) & S_{22}(z) \end{pmatrix} = D_S + C_S(zE_S - A_S)^{-1}B_S$$

with $S_{11}(z) \in \mathcal{RL}^{p\times l}(\partial\mathcal{D})$, $S_{12}(z) \in \mathcal{RL}^{p\times p}(\partial\mathcal{D})$, $S_{21}(z) \in \mathcal{RL}^{l\times l}(\partial\mathcal{D})$ and $S_{22}(z) \in \mathcal{RL}^{l\times l}(\partial\mathcal{D})$; further,

$$E_S = \begin{pmatrix} I_{np} & 0 \\ 0 & 0 \end{pmatrix}, \quad A_S = \begin{pmatrix} -I_{np} & \gamma^2 I_{np} - P \\ Z & A_0 \end{pmatrix}$$
$$B_S = \begin{pmatrix} 0 & 0 \\ B & B_0 \end{pmatrix}, \quad C_S = \begin{pmatrix} 0 & CZ^{-1}P \\ 0 & C_0 \end{pmatrix}$$

and

$$D_s = \begin{pmatrix} CZ^{-1}B & \gamma I_p \\ \gamma I_l & 0 \end{pmatrix}$$

where A, B, C and P are defined in Theorem 1, and $Z = I_{np} + A$, $A_0 = P - \gamma^2 Z A^t Z^{-t}$, $B_0 = \gamma Z Z^{-t} C^t$ and $C_0 = \gamma B^t Z^{-t}$. The equivalent state-space realization of $H_a(z)$, the generator of all k-th order Hankel-norm approximations of $\tilde{H}(z)$ given in the Theorem, follows from a standard procedure of transforming descriptor realisations to state-space form. □

The proof of Theorem 2 follows by specialising the results of [1] and [2] to the matrix FIR case considered here. Note that the generator of all Hankel norm approximations $S_a(z)$ defined in the proof of the Theorem is given in descriptor form; this is subsequently converted into state-space form given by the indicated realisation of $H_a(z)$ using the standard approach. In particular, it is always possible to obtain a state-space realisation of $S_a(z)$ of order $2np - \text{Rank}(A_0)$; if A_0 is non-singular, we obtain a state-space description of order np. Note also, that although Theorem 2 is still valid in the optimal case $\gamma = \sigma_{k+1}(\Gamma_H)$, the resulting realisation is non-minimal. In this case, a minimal realisation can be obtained in closed form via a singular perturbation argument (see [2] for details).

It is clear from Theorem 2 that the solution of a pair of Lyapunov equations used to obtain a balanced realisation of the system which is approximated is completely avoided in our framework; in fact the generator $H_a(z)$ of all Hankel-norm approximations $\tilde{H}(z)$ is completely defined by the Markov parameters of $\tilde{H}(z)$ (and parameter γ defining the level of sub-optimality). This dependence is made explicit in Theorem 3 below which specialises the parametrisation of Theorem 2 to the scalar case. In particular, Theorem 3 below shows that in the state-space realization of the generator $H_a(z)$ of all Hankel-norm approximations defined in Theorem 2 above has a special structure. In particular, the state-matrix of the realization is in companion form, which allows us to obtain an analytic expression for the transfer function of the so-called "central approximation" (obtained by setting $U(z) = 0$ in the linear fractional transformation which defines the parametrisation of solutions).

Theorem 3: Let all variables be defined as in Theorems 1-2 above, set $p = m = 1$ and assume that A_0 is non-singular and that $\sigma_{k+1}(\Gamma_{\tilde{h}}) < \gamma < \sigma_{k+1}(\Gamma_{\tilde{h}})$ (sub-optimal case). Then the realization of $H_a(z) =$

(A_H, B_H, C_H, D_H) defined in Theorem 2 is of the form:

$$A_H = \begin{pmatrix} a_1 & a_2 & a_3 & \dots & a_{n-1} & a_n \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}, \quad B_H = \begin{pmatrix} b_1 & b_2 \\ h_{n-1} & 0 \\ h_{n-2} & 0 \\ \vdots & \vdots \\ h_1 & 0 \end{pmatrix}$$

$$C_H = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & -1 \\ c_1 & c_2 & c_3 & \dots & c_{n-1} & c_n \end{pmatrix}, \quad D_H = \begin{pmatrix} 0 & 0 \\ 0 & d_{22} \end{pmatrix}$$

where $\tilde{h}(z) = h_1 z^{-1} + h_2 z^{-2} + \ldots + h_n z^{-n}$ and

$$a_i = \frac{\gamma^2(\theta_i + \theta_j)}{1 - \gamma^2\theta_1} \;\; (i = 1, 2, \ldots, n-1), \quad a_n = \frac{\gamma^2\theta_n}{1 - \gamma^2\theta_1}$$
$$b_1 = h_n + \frac{\gamma^2}{1 - \gamma^2\theta_1} \sum_{i=1}^{n} \theta_i h_{n-i+1}, \quad b_2 = -\frac{\gamma}{1 - \gamma^2\theta_1}$$

in which

$$\theta_j = \sum_{i=1}^{n} (-1)^i \hat{P}_{ij}, \;\; (j = 1, 2, \ldots, n), \quad \hat{P} = (P - \gamma^2 I_n)^{-1}$$

In particular, the transfer function of the central approximation ($U(z) = 0$) is given as:

$$X(z) = -\frac{h_1 z^{n-1} + (h_2 - h_1 a_1) z^{n-2} + \ldots + (h_{n-1} - h_1 a_{n-2}) z + (b_1 - h_1 a_{n-1})}{z^n - a_1 z^{n-1} - a_2 z^{n-2} - \ldots - a_{n-1} z - a_n}$$

Proof: First note that since $\sigma_{k+1}(\Gamma_{\tilde{h}}) < \gamma < \sigma_k(\Gamma_{\tilde{h}})$, $P - \gamma^2 I_n$ is nonsingular. It is also easy to see that:

$$Z = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \end{pmatrix}, \quad Z^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 1 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ (-1)^{n+1} & (-1)^n & (-1)^{n+1} & \dots & -1 & 1 \end{pmatrix}$$

Thus,

$$P - \gamma^2 I_n - A_0 = \gamma^2 (Z A^t Z^{-t} - I_n) = \begin{pmatrix} \gamma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} -1 & 1 & \dots & (-1)^n \end{pmatrix} := uv^t$$

and hence, using the matrix inversion lemma,

$$A_0^{-1} = (P - \gamma^2 I_n - uv^t)^{-1} = (P - \gamma^2 I_n)^{-1} + \lambda (P - \gamma^2 I_n)^{-1} uv^t (P - \gamma^2 I_n)^{-1}$$

where we have defined

$$\lambda = \frac{1}{1 - v^t (P - \gamma^2 I_n)^{-1} u} = \frac{1}{1 - v^t \hat{P} u}$$

Setting

$$\theta^t = \begin{pmatrix} \theta_1 & \theta_2 & \dots & \theta_n \end{pmatrix} = v^t (P - \gamma^2 I_n)^{-1}$$

shows that

$$(\gamma^2 I_n - P) A_0^{-1} = -I_n - \lambda u \theta^t$$

and hence

$$A_H = -I_n - (\gamma^2 I_n - P)A_0^{-1}Z = \begin{pmatrix} \lambda\gamma^2(\theta_1+\theta_2) & \lambda\gamma^2(\theta_2+\theta_3) & \dots & \lambda\gamma^2(\theta_{n-1}+\theta_n) & \lambda\gamma^2\theta_n \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

which is of the required form on noting that

$$\lambda = \frac{1}{1-\theta^t u} = \frac{1}{1-\gamma^2\theta_1} = \frac{1}{1-\gamma^2\sum_{i=1}^n (-1)^i \hat{P}_{i1}}$$

Next consider B_H. Using the fact that

$$(\gamma^2 I_n - P)A_0^{-1} = -\begin{pmatrix} 1+\lambda\gamma^2\theta_1 & \lambda\gamma^2\theta_2 & \dots & \lambda\gamma^2\theta_{n-1} & \lambda\gamma^2\theta_n \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

the first column of B_H can be written as:

$$-(\gamma^2 I_n - P)A_0^{-1}B = \begin{pmatrix} h_n + \frac{\gamma^2}{1-\gamma^2\theta_1}\sum_{i=1}^n \theta_i h_{n-i+1} \\ h_{n-1} \\ \vdots \\ h_2 \\ h_1 \end{pmatrix}$$

Similarly, since

$$B_0 = \gamma Z Z^{-t} C^t = \gamma \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ -1 \\ \vdots \\ (-1)^n \end{pmatrix} = -\gamma \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

the second column of B_H can be written as:

$$-(\gamma^2 I_n - P)A_0^{-1}B_0 = \begin{pmatrix} 1+\lambda\gamma^2\theta_1 & \lambda\gamma^2\theta_2 & \dots & \lambda\gamma^2\theta_{n-1} & \lambda\gamma^2\theta_n \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} -\gamma \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{\gamma}{1-\gamma^2\theta_1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

as required. Next consider the first row of C_H; this is given by:

$$\begin{aligned} -CZ^{-1}PA_0^{-1}Z &= -CZ^{-1}P(P-\gamma^2 ZA^tZ^{-t})^{-1}Z \\ &= -CZ^{-1}(P-\gamma^2 ZA^tZ^{-t}+\gamma^2 ZA^tZ^{-t})(P-\gamma^2 ZA^tZ^{-t})^{-1}Z \\ &= -C-\gamma^2 CA^tZ^{-t}(P-\gamma^2 ZA^tZ^{-t})^{-1}Z = -C \end{aligned}$$

since $CA^t = 0$. Finally, consider D_H. Using a similar argument to the one above, its $(1,1)$ element is:

$$\begin{aligned} D_{H,11} &= CZ^{-1}B - CZ^{-1}P(P-\gamma^2 ZA^tZ^{-t})^{-1}B \\ &= CZ^{-1}B - CZ^{-1}B - \gamma^2 CA^tZ^{-t}(P-\gamma^2 ZA^tZ^{-t})^{-1}B = 0 \end{aligned}$$

Similarly,

$$\begin{aligned} D_{H,12} &= \gamma - \gamma C Z^{-1} P (P - \gamma^2 Z A^t Z^{-t})^{-1} Z Z^{-t} C^t \\ &= \gamma - \gamma C Z^{-t} C^t - \gamma^3 C A^t Z^{-t} (P - \gamma^2 Z A^t Z^{-t})^{-1} Z Z^{-t} C^t \\ &= \gamma - \gamma C C^t = 0 \end{aligned}$$

Finally consider the $(2,1)$ element of D_H, $D_{H,21} = \gamma - C_0 A_0^{-1} B$. We first show that $A_0^{-1}B$ is a zero column vector, except from its first element which is equal to h_n^{-1}. Write

$$A_0 = P - \gamma^2 Z A^t Z^{-t} = \begin{pmatrix} x_1 & X_2 \end{pmatrix}$$

where x_1 is a column vector and set

$$A_0^{-1} B = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$$

where ϕ_1 is a scalar. Then:

$$\phi_1 x_1 + X_2 \phi_2 = B$$

It is easy that the first column of $Z A^t Z^{-t}$ is zero, and hence x_1 is the first column of P which is given by

$$x_1^t = \begin{pmatrix} h_n & h_{n-1} & \cdots & h_1 \end{pmatrix} h_n$$

Thus x_1 is a multiple of B, from which it follows that $\phi_1 = h_n^{-1}$ and $\phi_2 = 0$, so that

$$A_0^{-1} B = \begin{pmatrix} h_n^{-1} & 0 & \cdots & 0 \end{pmatrix}^t$$

Thus,

$$D_{H,21} = \gamma - \gamma B^t Z^{-t} A_0^{-1} B = \gamma - \gamma h_n^{-1} \begin{pmatrix} h_n & h_{n-1} & \cdots & h_1 \end{pmatrix} \begin{pmatrix} 1 & -1 & \cdots & (-1)^n & (-1)^{n+1} \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} = 0$$

The transfer function of the central approximation follows by routine calculations since A_H is in companion form. □

4. Application: Design of IIR filters with approximately linear phase

The fact that FIR filters can have a linear phase response has been used extensively in many digital signal processing applications. A filter with a nonlinear phase characteristic causes distortion to its input signal, since the various frequency components in the signal will be delayed, in general, by amounts which are not proportional to their frequency, thus altering their mutual harmonic relationships. A distortion of this form is undesirable in many practical digital signal processing applications such as music, video, data transmission and biomedicine and can be avoided by using filters with linear phase over the frequency range of interest [18]. One of the earlier and simpler methods for designing FIR filters with linear phase-response characteristics [16] is outlined in the next few paragraphs.

Suppose that the unit pulse response of an FIR filter is given by $\{h(0), h(1), h(2), \ldots, h(N)\}$. For this filter to have a *linear* phase response, say $\theta(\omega) = -\alpha\omega$ for α constant, it must have an impulse response with *positive* symmetry, i.e. $h(n) = h(N-n-1)$, where n varies as $n = 0, 1, 2, \ldots, \frac{N-1}{2}$ for N odd and $n = 0, 1, 2, \ldots, \frac{N}{2} - 1$ for N even. Such a filter has identical phase and group delay and these are independent of frequency, i.e.

$$T_p = -\frac{\theta(\omega)}{\omega} = T_g = -\frac{d\theta(\omega)}{d\omega} = \alpha$$

An FIR filter whose impulse response has *negative* symmetry, i.e. $h(n) = -h(N-n-1)$, has an *affine* phase-response characteristic of the form $\theta(\omega) = \beta - \alpha\omega$ with α and β constant and its group delay $T_g = \alpha$ is independent of ω.

The simple relationship between the impulse response of an FIR filter and its frequency response has been exploited to design filters of this type via a variety of optimisation methods [4], [7], [15], [16], [13], [20], [19]. One of the earliest and simplest approaches is [16], in which a linear programming procedure is proposed for designing linear phase FIR filters with specified magnitude-response characteristics (low-pass, high-pass, etc). As an example, consider the filter

$$H(z) = h(0) + h(1)z^{-1} + \ldots + h(p-1)z^{-(p-1)} + h(p)z^{-p} + h(p+1)z^{-(p+1)} + \ldots + h(2p)z^{-2p}$$

in which positive symmetry of the impulse response is enforced around the central coefficient $h(p)$ by setting $h(0) = h(2p), h(1) = h(2p-1), \ldots, h(p-1) = h(p+1)$. The frequency response of the filter may be written as:

$$H(e^{j\omega}) = e^{j\omega p} M(\omega)$$

where $M(\omega)$ is a real linear function of the filter's coefficients:

$$M(\omega) = (2\cos(p\omega) \;\; 2\cos((p-1)\omega) \;\; \ldots \;\; 1) \begin{pmatrix} h(0) \\ h(1) \\ \vdots \\ h(p) \end{pmatrix}$$

Suppose now that we want to satisfy the following specifications: (a) $1-\delta \leq |H(e^{j\omega})| \leq 1+\delta$ for all frequencies in the pass-band $\omega \in [0, \omega_p]$, (b) $|H(e^{j\omega})| \leq \delta$ for all frequencies in the stop-band $\omega \in [\omega_s, \pi]$ where $\omega_p < \omega_s$, and (c) minimise the "ripple" δ subject to constraints (a) and (b). Specifications (a) and (b) can enforced by discretising the pass-band and stop-band frequency intervals using n frequencies, say, i.e.

$$0 = \omega_1 < \omega_2 < \ldots < \omega_n = \omega_p$$

and

$$\omega_s = \omega_{n+1} < \omega_{n+2} < \ldots < \omega_{2n} = \pi$$

and enforcing the specifications via the inequalities:

$$1 - \delta \leq M(\omega_i) \leq 1 + \delta, \quad i = 1, 2, \ldots, n \tag{2}$$

and

$$-\delta \leq M(\omega_i) \leq \delta, \quad i = n+1, n+2, \ldots, 2n \tag{3}$$

The optimisation problem: $\min \delta$ subject to (1) and (2), can now be formulated as a linear programme in the standard form:

$$\min c^t x \quad \text{s.t.} \quad Ax \leq b$$

where $x^t = (h(0)\; h(1) \ldots h(p)\; \delta)$, $c^t = (0 \ldots 0\; 1)$, b is a $4n$-dimensional vector and A is a $4n \times (p+2)$-dimensional matrix. This can be solved efficiently using standard techniques, e.g. the simplex algorithm or interior point methods.

Once a linear-phase FIR filter has been designed, the next step is to approximate it with a low-order IIR filter using the Hankel-norm approximation method outlined earlier. In the remaining part of this section we derive a bound on the phase error due to the approximation. This can be used to determine the minimum degree of the IIR filter for which the deviation of the phase response from linearity does not exceed a specified tolerance. Note first, that in the scalar case, the solution to the Hankel-norm approximation problem is, in general, unique only in the optimal case. The so-called "central solution" is obtained by setting $U(z) = 0$. Having obtained

the central approximation, one next needs to remove its anti-stable component. The extraction of the stable projection can be performed either by factoring the denominator polynomial of $X(z)$,

$$p(z) = z^n - a_1 z^{n-1} - a_2 z^{n-2} - \cdots - a_{n-1} z - a_n$$

which is guaranteed to have k roots (strictly) inside the unit circle and $n-k$ roots (strictly) outside the unit circle. Alternatively, the stable projection can be extracted from the state-space realization of $X(z)$ by transforming the system to block-Schur form using an appropriate orthogonal state-space transformation and ordering the eigenvalues of A_H in ascending order of magnitude; decoupling of the stable and anti-stable parts then requires the solution of a matrix Sylvester equation.

Reference [9] shows that having solved the k-th order Hankel-norm approximation problem, it is always possible to choose a constant term x_0 so that the approximation error satisfies the bound:

$$\|h(z) + (x(z) + x_0)\|_\infty \leq \sum_{i=k+1}^{n} \sigma_i(\Gamma_h) \tag{4}$$

Since this bound applies uniformly in frequency, it can be used to give an immediate bound on the approximation phase error:

Theorem 4: Let $\hat{x}(z) = x(z) + x_0$ be a k-th order optimal Hankel norm approximation of the scalar FIR filter $h(z)$ such that (4) holds. Then, if $\phi(\omega) = \arg(h(e^{j\omega})) + \arg(\hat{x}(e^{j\omega}))$, we have that:

$$|\sin(\phi(\omega))| \leq \frac{\sum_{i=k+1}^{n} \sigma_i(\Gamma_h)}{|h(e^{j\omega})|} \tag{5}$$

for every $\omega \in [0, \pi)$. In particular, if the frequency interval $[\omega_1 \ \omega_2]$ lies in the filter's passband in which $1 - \delta \leq |h(e^{j\omega})| \leq 1 + \delta$, then

$$|\sin(\phi(\omega))| \leq \frac{\sum_{i=k+1}^{n} \sigma_i(\Gamma_h)}{1 - \delta} \tag{6}$$

for every $\omega \in [\omega_1 \ \omega_2]$.

Proof: Straightforward and hence omitted. □

The phase error bound given in Theorem 4 can be used to select the minimum order of approximation k consistent with a worst-case phase-error specification; If $h(z)$ is a linear-phase FIR filter, then the RHS of (5) or (6) quantifies the deviation in the phase of the IIR approximation of $h(z)$ from linearity. If the magnitude tolerances resulting from the optimization procedure are tight, the specifications need to be tightened to account for the approximation error.

5. Example

In this section some of the results presented in the paper are illustrated by means of a computer example. First, a linear phase FIR filter of order $n = 21$ was designed using the linear programming procedure outlined in section 3. The design specifications were defined as: $\omega_p = 1$ rad/sample, $\omega_s = 1.5$ rads/sample and the frequency intervals $[0, \omega_p]$ and $[\omega_s, \pi]$ were each discretised using 50 equally-spaced frequencies. The linear programme was then set up and solved using Matlab's function *linprog.m*. The minimum ripple was obtained as $\delta = 0.0232$; the first 11 optimal impulse response coefficients of the resulting filter $h(z)$ are tabulated below (the last 10 coefficients are symmetric and are not included):

i	$h(i)$	i	$h(i)$	i	$h(i)$
0	0.0017	4	0.0358	8	0.0919
1	-0.0212	5	-0.0015	9	0.2995
2	-0.0123	6	-0.0662	10	0.3980
3	0.0178	7	-0.0561		

Table 1: Impulse response coefficients

The magnitude frequency response of the filter is shown in Figure 1 below. Note that the response in the passband and the stopband lies within the desired bounds $1 \pm \delta$.

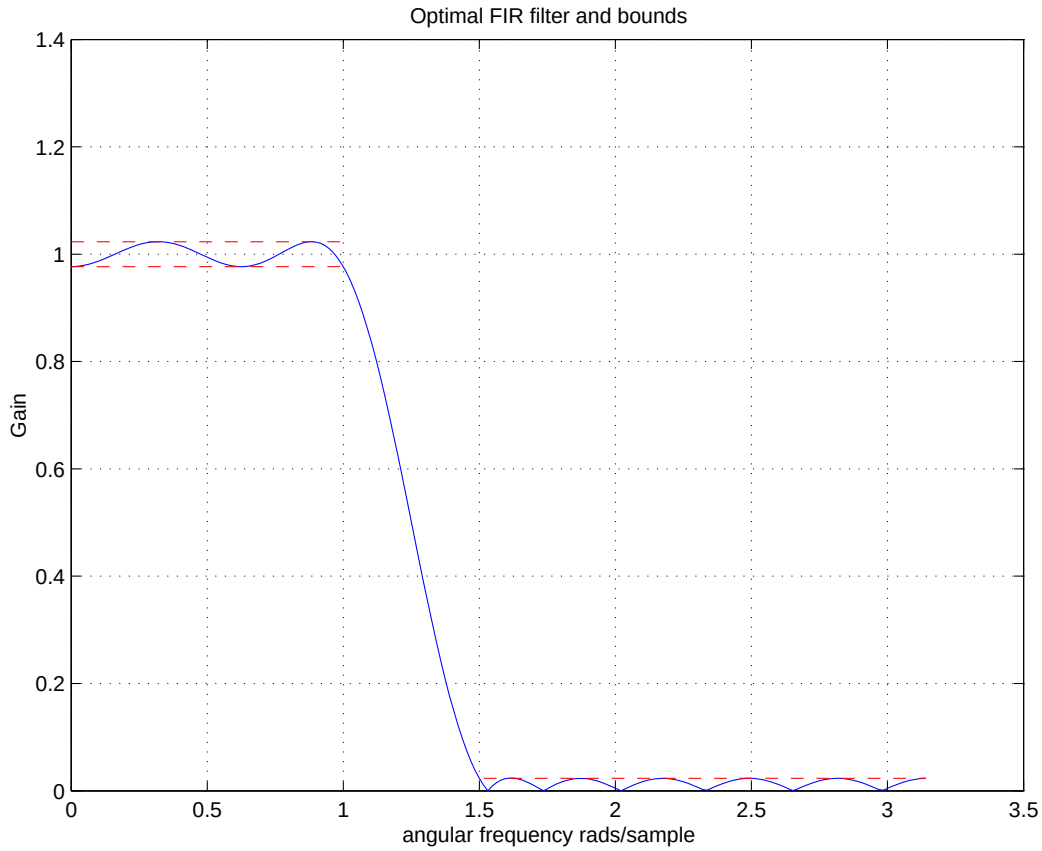

Figure 1: *Magnitude frequency response*

Next the Hankel-norm approximation method outlined in Theorem 2 was applied to the FIR filter $h(z)$. The Hankel singular values of $h(z)$ are shown in Table 2. Parameter γ was chosen as $\gamma = 0.03$ resulting in a 7-th order (sub-optimal) Hankel-norm approximation. The seven stable poles of the approximant were obtained as: 0.7467, $0.2841 \pm 0.8152j$, $0.4977 \pm 0.6347j$ and $0.6886 \pm 0.3363j$.

i	σ_i	i	σ_i	i	σ_i	i	σ_i
1	1.0000	6	0.1765	11	0.0117	16	0.0116
2	0.9973	7	0.0602	12	0.0117	17	0.0002
3	0.9563	8	0.0232	13	0.0117	18	0.0002
4	0.7791	9	0.0135	14	0.0116	19	0.0000
5	0.4344	10	0.0118	15	0.0116	20	0.0000

Table 2: Hankel singular values

The magnitude response of $h(z)$ and its IIR approximation is shown in Figure 2. It can be seen that the IIR filter has a slightly larger ripple in the pass-band. Figure 3 shows the impulse responses of the two filters.

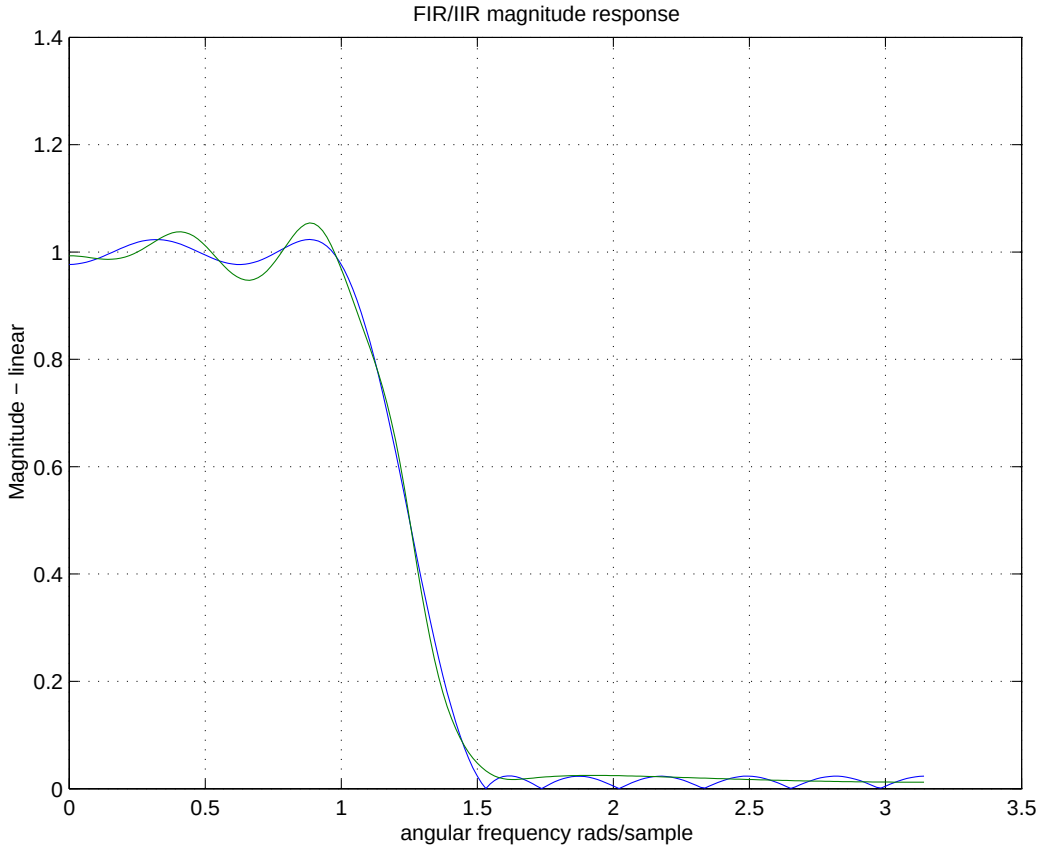

Figure 2: *FIR/IIR magnitude frequency response*

Finally, Figure 4 shows the phase error in the passband $0 \leq \omega \leq 1$ rad/sample arising from the approximation (i.e. the deviation of the phase of the IIR filter from linearity), while Figure 5 shows the pole/zero pattern of the IIR filter (there is an additional zero at $z = -15.6521$ not shown in the figure).

6. Conclusions

The paper has presented an algorithm for approximating discrete-time matrix FIR filters by reduced order IIR filters using Hankel-norm approximation methods. The algorithm is relies only on the Markov parameters of the filter which is approximated, is computationally efficient and is based on Hankel-norm approximation techniques of discrete-time descriptor systems [1], [2]. The method can be used to define a systematic approach for designing low-order IIR filters with approximately linear phase response characteristics. The definition of a-priori error bounds for the magnitude and phase errors due to the Hankel-norm approximation allows the designer to choose the minimal IIR filter order consistent with the design specifications. A low-order example has illustrated the effectiveness of the method.

There are a number of issues related to the proposed technique that we intend to pursue in the future. These include: (i) Derivation of tighter error bounds than those applying in general for the Hankel-norm approach. This seems possible given the special structure of FIR filters (all poles lie at the origin). (ii) Investigation of

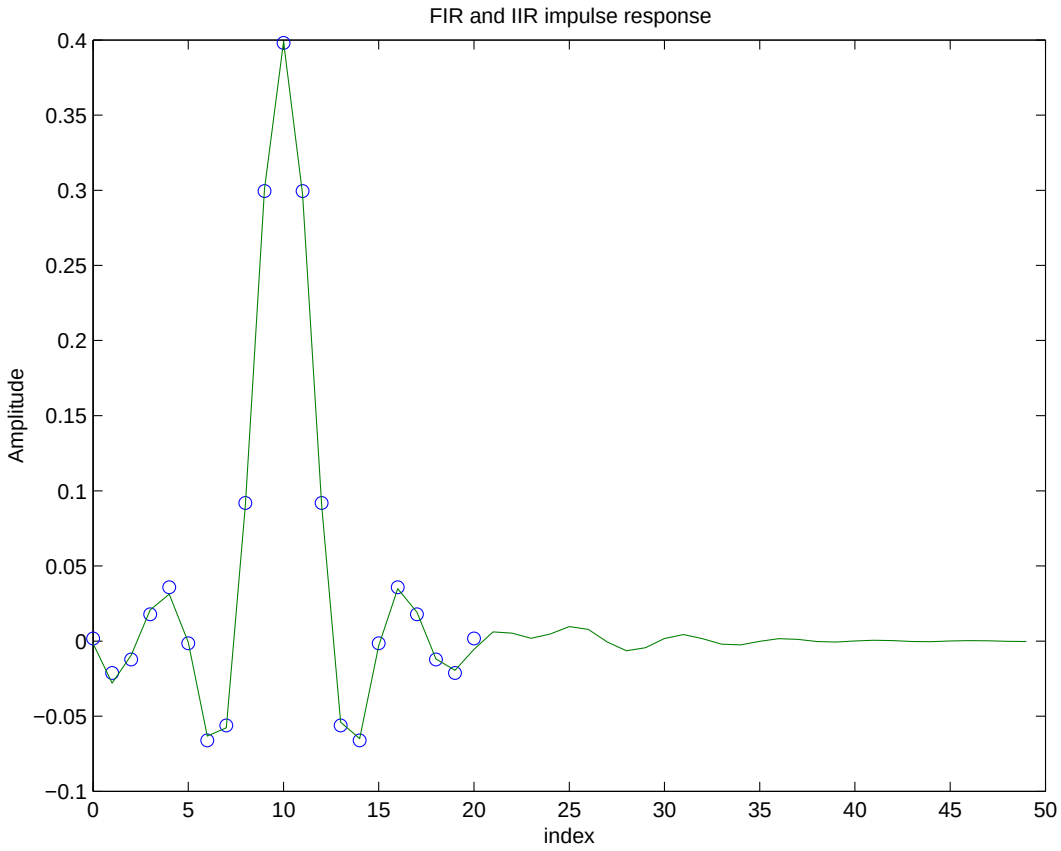

Figure 3: *FIR/IIR impulse response*

the sensitivity properties of the proposed design under finite-precision implementation. (iii) Generalization of the method to other types of approximation, e.g. relative-error and general frequency weighted approximations. (iv) Extension of the method to more general application domains in the fields of Signal Processing, Systems Identification and Control (e.g. multi-dimensional systems, model order selection techniques).

7. Acknowledgement

The authors wish to thank an anonymous reviewer for bringing to their attention reference [6].

References

[1] M. Al-Husari, I.M. Jaimoukha and D.J.N. Limebeer (1993), A descriptor approach for the solution of the one-block discrete distance problem, *Proc. IFAC World Congress*, Sydney, Australia.

[2] M. Al-Husari, I.M. Jaimoukha and D.J.N. Limebeer (1997), A descriptor solution to a class of discrete distance problems, *IEEE Trans. Auto. Control*, 42(11):1558-1564.

[3] R.W. Aldhaheri (2000), Design of linear-phase IIR digital filters using singular perturbational model reduction, *IEE Proc.-Vis. Image Signal Process.*, 147(5):409-414.

[4] X. Chen and T.W. Parks (1987), Design of FIR filters in the complex domain, *IEEE Trans. Acoust. Speech and Signal Processing*, 35:144-153.

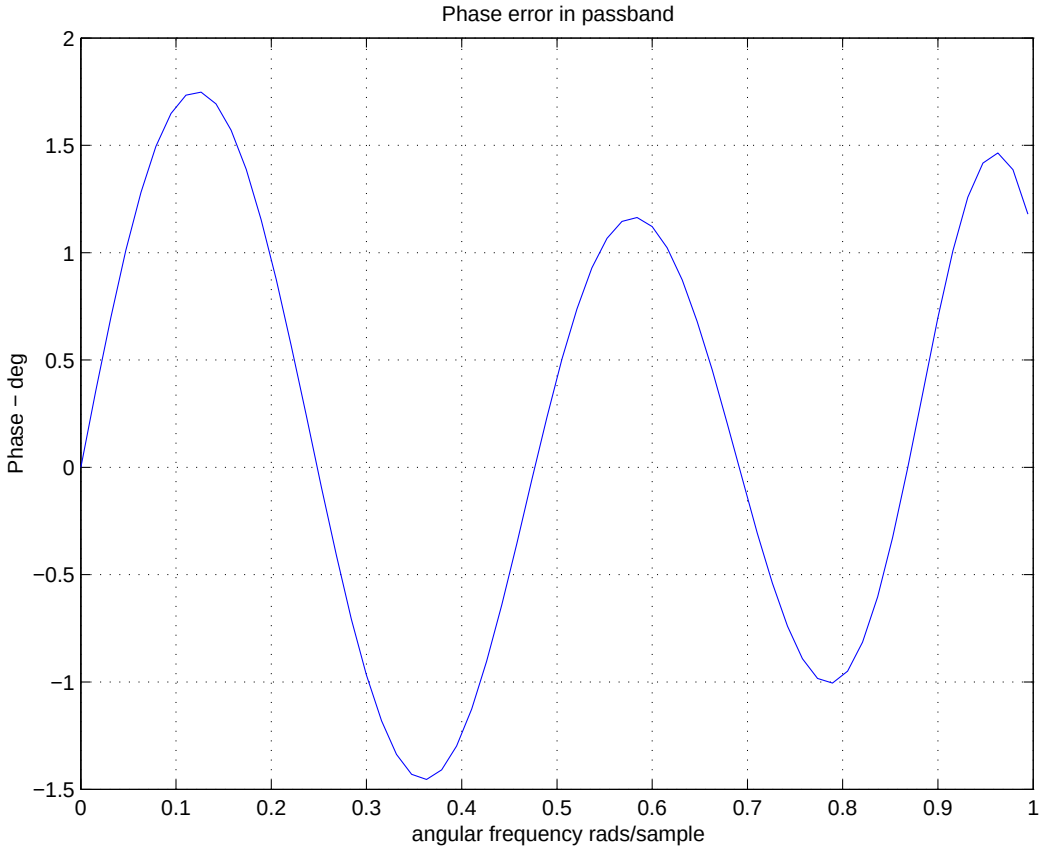

Figure 4: *Phase error in passband (deg)*

[5] B.S. Chen, S.C. Peng and B.W. Chiou (1992), IIR filter design via optimal Hankel-norm approximation, *IEE Proceedings-G*, 139(5):586-590.

[6] N. Deng, S. Shao and G. Gu (2006), A system approach to the design of linear phase IIR filters via optimal Hankel-norm criterion, *Proceesdings of the 3rd Int. Conf. Impulse Dynamic Syst. and Appl.*, 984-988.

[7] Y. Genin, Y. Hachez, Yu. Nesterov and P. Van Dooren (2000), Convex optimization over positive polynomials and filter design, *Proceedings of the 2000 UKACC International Conference on Control*, Cambridge University, UK.

[8] G. Gu (2005), All optimal Hankel norm approximations and their $\mathcal{L}_\infty$ bounds in discrete-time, *International Journal of Control*, 78(6):408-423.

[9] K. Glover (1984), All optimal Hankel-norm approximations of linear multivariable systems and their $\mathcal{L}_\infty$ error bounds, *Int. J. Control*, 39:1115-1193.

[10] G.D. Halikias and I.M. Jaimoukha (2003), Design of approximately linear-phase Infinite Impulse Response filters via optimal Hankel-norm approximation, *Proc. European Control Conference, ECC03*, Cambridge University, UK.

[11] G.D. Halikias, I.M. Jaimoukha and D.A. Wilson (1997), A numerical solution to the matrix $\mathcal{H}_\infty/\mathcal{H}_2$ optimal control problem, *Int. J. Robust and Nonlinear Control*, 7(7):711-726.

[12] I. Kale, J. Gryka, G.D. Cain and B. Beliczynski (1994), FIR filter order reduction: balanced model truncation and Hankel-norm optimal approximation *IEE Proc-Vis. Image Signal Process.*, 141(3):168-174.

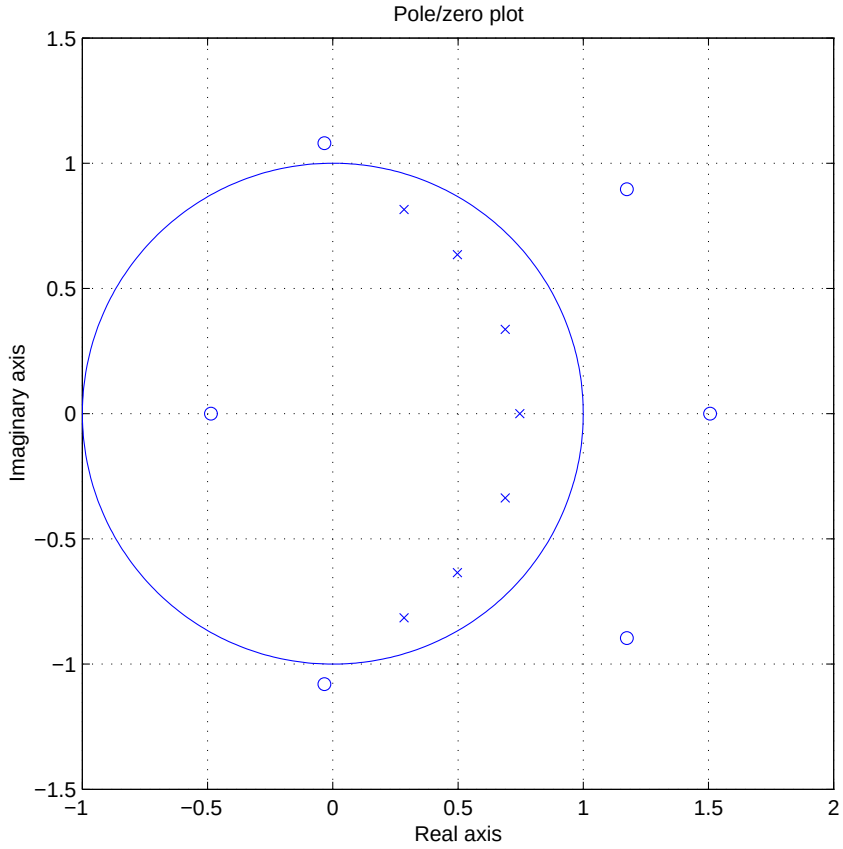

Figure 5: *Pole/zero pattern of IIR filter*

[13] J.H. McClellan and T.W. Parks (1973), A unified approach to the design of optimal FIR linear-phase filters, *IEEE Trans. Circuit Theory*, 20:697-701.

[14] A.V. Oppenheim and R.W. Schafer (1975), Digital Signal Processing, *Prentice Hall, Inc.,*, Englewood Cliffs, NJ.

[15] K. Preuss (1989), On the design of FIR filters by complex Chebychev approximations, *IEEE Trans. Acoust. Speech and Signal Processing*, 37:702-712.

[16] L.R. Rabiner (1972), Linear Program Design of Finite Impulse Response (FIR) Digital Filters, *IEEE Tran. Audio and Electroacoustics*, AU20(4):280-288.

[17] V. Sreeram and P. Agathoklis (1992), Design of Linear-Phase IIR Filters via Impulse-Response Gramians, *IEEE Transactions on Signal Processing*, 40(2):389-394.

[18] H. Stark and F.B. Tuteur (1979), Modern Electrical Communications, *Prentice Hall, Inc.,*, Englewood Cliffs, NJ.

[19] L. Vandenberghe and S. Boyd (1996), Semidefinite programming, *SIAM Review*, 38(1):49-95.

[20] S.P. Wu, S. Boyd and L. Vandenberghe (1997), FIR filter design via Spectral factorization and Convex Optimization, *Applied Computational Control*, Signal and Comunications, Biswa Datta ed., Birkhauser.