Robustness of digital filters with respect to limit-cycle behavior under coefficient perturbations

Kulasekere, Ernst, Premaratne, Kamal, Bauer, Peter

Ernst C. Kulasekere, Kamal Premaratne, Peter H. Bauer, "Robustness of digital filters with respect to limit-cycle behavior under coefficient perturbations," Proc. SPIE 2750, Digital Signal Processing Technology, (7 June 1996); doi: 10.1117/12.241995

SPIE.

Event: Aerospace/Defense Sensing and Controls, 1996, Orlando, FL, United States

Robustness of Digital Filters With Respect to Limit Cycle Behavior Under Coefficient Perturbations

E. C. Kulasekere

Department of Electrical and Computer Engineering
University of Miami

K. Premaratne

Department of Electrical and Computer Engineering
University of Miami

P. H. Bauer

Department of Electrical Engineering
University of Notre Dame

ABSTRACT

A digital filter which has been designed to be limit cycle free may exhibit limit cycles at the implementation stage. This is due to the inability to implement filter coefficients exactly in hardware when they are quantized to satisfy available wordlength requirements. Given a digital filter which is limit cycle free under zero input conditions, the work below presents an algorithm which finds a region in the coefficient space, about the nominal filter coefficient values, wherein the filter remains limit cycle free. Furthermore the results of the algorithm will also indicate the availability of other machine representable numbers for the coefficients that fall within this robustness region. Hence one may even choose shorter wordlength registers for coefficient storage if the corresponding grid falls within the constructed robustness region.

Keywords : Limit Cycles, Robustness, Coefficient sensitivity.

1 INTRODUCTION

Over the years many papers have been published regarding limit cycle properties of digital filters[1-5]. Granular limit cycles is the subject of many of these publications, in particular limit cycles due to zero or low input conditions are addressed. This is especially critical since the occurrence of limit cycle oscillations at zero or low

0-8194-2131-6/96/$6.00

input conditions significantly impair the signal to noise ratio at the output. An extensive listing of limit cycle behavior due to different arithmetic types is given in the reference.[3]

A filter designer will always assume the transfer function coefficients can be accurately represented, *i.e* assume infinite wordlength, at the design stage. However at the implementation stage this system is represented with suitable digital hardware in a finite wordlength environment. Therefore the designer has to suitably approximate the infinite wordlength coefficients with a machine representable number using the available wordlength. This process will invariably introduce errors to the system since errors in coefficient representation introduces perturbations in the poles and zeros of the transfer function which in turn manifests themselves as errors in the frequency response.[7] Appreciable differences between the implemented and the designed coefficients will lead to a deviation of the frequency response from the nominal specifications. Ideally we would like to have both systems to have similar characteristics. Hence the knowledge of a robustness region would be the important link between the designed system and the practically realizable system.

An alternate method of finding the robustness region around a nominal parameter set by constructing parallel hyperplanes in the coefficient space is discussed in the reference.[2] Methods which exactly construct this region have been developed in[8] in the context of parameter estimation problem. This involves a complicated and time consuming procedure to converge to a realistic result. In contrast the algorithm to construct a robustness region given below is more general as the filter is assumed to be in its state-space form and the development of this region follows a methodical technique which is less time consuming and always converges to a result. Moreover, to reduce the computational burden, several crucial and novel notions have been incorporated.

There are several motivating reasons for undertaking such a study, they are:

If an exhaustive search method is carried out to determine stability[1,2] of the digital system, the linear stability region will be covered by a finite grid before the search algorithm is applied to grid points to determine stability. Hence a robustness region about the nominal filter would conclude whether filters that were not captured by this grid remain limit cycle free. It can also provide a suitable grid size *a priori*.

Once a robustness region about a nominal filter is available. By superimposing different grid sizes on this region, one may choose different machine representable numbers that fall within the robustness region to determine the shortest wordlength possible to represent the coefficient values and still preserve the system characteristics. This will enable the hardware designer to choose the optimum bit length for coefficient storage. Especially for systems with large number of coefficients this procedure will yield a considerable saving in hardware.

Due to coefficient representation error introduced via finite precision effects, the filter being checked is different than the designed filter. If a robustness region is available, one may determine whether the latter is indeed limit cycle free.

2 NOMENCLATURE

The following notation will be used throughout the paper.

$\Re^{m\times n}$, $\mathcal{Z}^{m\times n}$	Set of matrices of size $m \times n$ over the reals and integers.
a_{ij}	(i,j)-th element of the matrix $A = \{a_{ij}\}$.
I, $\mathbf{0}$	Identity matrix and null matrix of appropriate sizes.
E_{jj}, $\hat{E}_{jj}$	A square zero matrix with '1' at its (j,j)-th position, $\hat{E}_{jj} = I - E_{jj}$.
$\mathbf{x}(k)$	Filter state vector at instant k.
$x_i(k)$	i-th component of the state vector $\mathbf{x}(k)$.
$\hat{M}_i$	Upper bound for absolute value of amplitude of $x_i(k)$, $k \in \mathcal{Z}_+$.
$\mathcal{Q}[\cdot]$	Quantization nonlinearity operator.

$\mathcal{S}^{(0)}$	Set of state vectors satisfying the upper bound $\hat{M}_i$ such that $\lvert x_i\rvert \leq \hat{M}_i \; \forall i$.
conv[·]	Convex hull of set [·].

3 ROBUSTNESS REGION CONSTRUCTION

Consider a filter that has been identified as limit cycle free by an appropriate algorithm.[1] To determine the robustness region for this filter all orbits traversed by the nominal filter for a given set of initial conditions has to be computed[1] before the algorithm can be applied to compute the robustness region.

The general realization of a filter state propagation represented in state-space form under zero input conditions with the appearance of a pertinent quantization nonlinearity ($\mathcal{Q}$) can be modeled as

$$\mathbf{x}(k+1) = \mathcal{Q}[A \cdot \mathbf{x}(k)] \tag{1}$$

Where $A \in \Re^{m\times m}$ and $\mathbf{x} \in \Re^m$.
Then it can be concluded[1] that

$$|x_i(k)| \leq \hat{M}_i \quad \forall\, i \quad \text{where } \hat{M}_i \in \mathcal{Z}^+ \tag{2}$$

Where $\hat{M}_i$ is the upper bound or the maximum occurable number for the state x_i with respect to the given filter. Hence from (2) the maximum period of a limit cycle is bounded by

$$T \leq \prod_{i=1}^{m}(2\hat{M}_i + 1) = T_{max} \tag{3}$$

The set of initial conditions satisfying (2) can be represented by the set

$$\mathcal{S}^{(0)} = \left\{\mathbf{x}(k) \in \mathcal{Z}^m \,\middle|\, |x_i(k)| \leq \hat{M}_i \quad i = 1, 2, \ldots m\right\} \tag{4}$$

We use the notation $\mathcal{O}_A$ to denote all orbits being traversed by the nominal system matrix A for all initial conditions given by (4).

Consider a small perturbation Δa_{ij} of each coefficient about its nominal value a_{ij}. Due to quantization, it is possible that

$$\mathbf{x}(k+1) = \mathcal{Q}[(A + \Delta A) \cdot \mathbf{x}(k)] = \mathcal{Q}[A \cdot \mathbf{x}(k)], \tag{5}$$

where $\Delta A = \{\Delta a_{ij}\} \in \Re^{m\times m}$ and

$$\Delta \mathbf{a}_i \doteq [\Delta a_{i1}, \Delta a_{i2}, l \ldots, \Delta a_{im}] \in \Re^m,\; i = 1, 2, \ldots, m. \tag{6}$$

We now make an important observation: Given the nominal filter A that has already been verified to be limit cycle free, we only consider those filters $A + \Delta A$ that follow identical orbits as A. Hence, with identical initial conditions, $\mathcal{O}_A = \mathcal{O}_{A+\Delta A}$. Note that, there are a maximum of $T_{\max} - 1$ different orbits in $\mathcal{O}_A$, and we enumerate them as

$$\mathcal{O}_A = \left\{\mathcal{O}_A^\ell | \ell = 1, \ldots, \tau,\; \tau \leq T_{\max} - 1, \right\}. \tag{7}$$

where

$$\mathcal{O}_A^\ell = \{\mathbf{x}^{(\ell)}(0), \mathbf{x}^{(\ell)}(1), \ldots, \mathbf{x}^{(\ell)}(T_{\max} - 1)\},\; \ell = 1, \ldots, T_{\max} - 1. \tag{8}$$

Hence, from (2) and (4), we see that the upper bounds and $\mathcal{S}^{(0)}$ corresponding to A and $A + \Delta A$ are identical. From a designer's point of view, the fact that we restrict ourselves to those filters in the robustness region that possess identical impulse responses, we believe, is more sensible.

Next, recall the following quantization characteristics:
For Sign-Magnitude Roundoff (SMR) quantization,

$$\mathcal{Q}_{SMR}[a] = x \Rightarrow \begin{cases} x - \frac{1}{2} \leq a < x + \frac{1}{2}, & x > 0 \\ x - \frac{1}{2} < a \leq x + \frac{1}{2}, & x < 0 \\ -\frac{1}{2} < a < +\frac{1}{2}, & x = 0. \end{cases} \tag{9}$$

For Two's Complement Truncation (TCT) quantization,

$$\mathcal{Q}_{TCT}[a] = x \Rightarrow \{ \; x \leq a < x + 1, \quad \forall x. \tag{10}$$

For Sign-Magnitude Truncation (SMT) quantization,

$$\mathcal{Q}_{SMT}[a] = x \Rightarrow \begin{cases} x \leq a < x + 1, & x > 0 \\ x - 1 < a \leq x, & x < 0 \\ -1 < a < 1, & x = 0. \end{cases} \tag{11}$$

Double-Length Accumulator Case
In this case, (5) is interpreted as follows: For two consecutive vectors $\mathbf{x}(k)$, $\mathbf{x}(k+1) \in \mathcal{O}_A^\ell$

$$\mathcal{Q}\left[\sum_{j=1}^{m}(a_{ij} + \Delta a_{ij}) \cdot x_j(k)\right] = \mathcal{Q}\left[\sum_{j=1}^{m} a_{ij} \cdot x_j(k)\right] = x_j(k+1). \tag{12}$$

Substituting (12) into the quantization scheme being used, $\Delta \mathbf{a}_i$ that satisfy (5) takes the following forms:
For SMR quantization, for $k = 0, 1, \ldots, T_{\max} - 1$,

$$\mathcal{G}^{(i)}_{\mathbf{x}(k)} = \begin{cases} \Delta a_{ij}: & x_j(k+1) - \frac{1}{2} - \sum_{j=1}^{m} a_{ij}x_j(k) \leq \Delta \mathbf{a}_i \cdot \mathbf{x}(k) < x_j(k+1) + \frac{1}{2} - \sum_{j=1}^{m} a_{ij}x_j(k) \\ & \text{for } x_j(k+1) > 0 \\ \Delta a_{ij}: & x_j(k+1) - \frac{1}{2} - \sum_{j=1}^{m} a_{ij}x_j(k) < \Delta \mathbf{a}_i \cdot \mathbf{x}(k) \leq x_j(k+1) + \frac{1}{2} - \sum_{j=1}^{m} a_{ij}x_j(k) \\ & \text{for } x_j(k+1) < 0 \\ \Delta a_{ij}: & -\frac{1}{2} - \sum_{j=1}^{m} a_{ij}x_j(k) < \Delta \mathbf{a}_i \cdot \mathbf{x}(k) < \frac{1}{2} - \sum_{j=1}^{m} a_{ij}x_j(k) \\ & \text{for } x_j(k+1) = 0. \end{cases} \tag{13}$$

For TCT quantization, for $k = 0, 1, \ldots, T_{\max} - 1$,

$$\mathcal{G}^{(i)}_{\mathbf{x}(k)} = \begin{cases} \Delta a_{ij}: & x_j(k+1) - \sum_{j=1}^{m} a_{ij}x_j(k) \leq \Delta \mathbf{a}_i \cdot \mathbf{x}(k) < x_j(k+1) + 1 - \sum_{j=1}^{m} a_{ij}x_j(k) \\ & \forall x_j(k+1). \end{cases} \tag{14}$$

For SMT quantization, for $k = 0, 1, \ldots, T_{\max} - 1$,

$$\mathcal{G}^{(i)}_{\mathbf{x}(k)} = \begin{cases} \Delta a_{ij}: & x_j(k+1) - \sum_{j=1}^{m} a_{ij}x_j(k) \leq \Delta \mathbf{a}_i \cdot \mathbf{x}(k) < x_j(k+1) + 1 - \sum_{j=1}^{m} a_{ij}x_j(k) \\ & \text{for } x_j(k+1) > 0 \\ \Delta a_{ij}: & x_j(k+1) - 1 - \sum_{j=1}^{m} a_{ij}x_j(k) < \Delta \mathbf{a}_i \cdot \mathbf{x}(k) \leq x_j(k+1) - \sum_{j=1}^{m} a_{ij}x_j(k) \\ & \text{for } x_j(k+1) < 0 \\ \Delta a_{ij}: & -1 - \sum_{j=1}^{m} a_{ij}x_j(k) < \Delta \mathbf{a}_i \cdot \mathbf{x}(k) < 1 - \sum_{j=1}^{m} a_{ij}x_j(k) \\ & \text{for } x_j(k+1) = 0. \end{cases} \tag{15}$$

Note that, each of these inequalities are of the following form:

$$\alpha_k \lhd \Delta \mathbf{a}_i \cdot \mathbf{x}^{(\ell)}(k) \lhd \beta_k, \; k = 0, \ldots, T_{\max} - 1, \tag{16}$$

where, to remain concise yet general, we have used the notation $\lhd$ to denote either '$\leq$' or '$<$' (depending on the quantization scheme being used and value of $\mathbf{x}^{(\ell)}(k+1)$—see (13- 15)). Note that, $\alpha_k = \alpha_k(\mathbf{x}^{(\ell)}(k+1), \mathbf{x}^{(\ell)}(k), a_{ij})$ and $\beta_k = \beta_k(\mathbf{x}^{(\ell)}(k+1), \mathbf{x}^{(\ell)}(k), a_{ij})$. Since A is limit cycle free, $\mathbf{x}^{(\ell)}(T_{\max}) = \mathbf{0}$.

Let

$$\mathcal{G}_i^{(\ell)} \doteq \{\Delta\mathbf{a}_i \in \Re^m : \Delta\mathbf{a}_i \text{ satisfies (16)}\}.$$

Recall that, $\mathbf{x}^{(\ell)}$ are known from the orbit $\mathcal{O}_A^{\ell}$ of nominal filter; a_{ij} of course are its coefficients. Hence, α_k, β_k, $i = 0, \ldots, T_{\max} - 1$, are all known quantities. We are seeking the set of all perturbations $\Delta\mathbf{a}_i$ of row i that belong to

$$\mathcal{G}_i \doteq \left\{ \Delta\mathbf{a}_i \in \Re^m : \Delta\mathbf{a}_i \in \bigcap_{\ell=1,\ldots,T_{\max}-1} \mathcal{G}_i^{(\ell)} \right\}.$$

Clearly, each $\Delta\mathbf{a}_i \in \mathcal{G}_i$ may be described by a set of inequalities of the form

$$\alpha_k \lhd \Delta\mathbf{a}_i \mathbf{x}^{(\ell)}(k) \lhd \beta_k, \; k = 0, \ldots, T_{\max} - 1; \; \ell = 1, \ldots, T_{\max} - 1. \tag{17}$$

Each inequality being linear in $\Delta\mathbf{a}_i$, $\mathcal{G}_i$ is in fact a convex hull generated by a finite number of vertices. This observation is crucial in the development of the following procedure which constructs $\mathcal{G}_i$. Let E_{jj} denote the zero matrix of size $m \times m$ with '1' at its (j,j)-th position; also, $\hat{E}_{jj} = I - E_{jj}$. A procedure to construct the robustness region is now proposed:

I. Let $r = 0$. For each $j = 1, \ldots, m$, substitute $\Delta\mathbf{a}_i E_{jj} = [0, \ldots, 0, \Delta a_{ij}, 0, \ldots, 0]$ instead of $\Delta\mathbf{a}_i$ and solve the inequality set in (17). For each j, this results in a single inequality of the form

$$\underline{v}_{ij} \lhd \Delta a_{ij} \lhd \overline{v}_{ij}. \tag{18}$$

Clearly,

$$\Omega_i^{(0)} \subseteq \mathcal{G}_i, \quad \text{where} \quad \Omega_i^{(0)} \doteq \text{conv}\left[\bigcap_{j=1,\ldots,m} \mathcal{G}_i \Big|_{\Delta\mathbf{a}_i E_{jj}} \right].$$

Here, conv[·] denotes the convex hull of set [·]. Note that, each $\Omega_i^{(0)}$ is a hyperdiamond in m-D space (of coefficient perturbations) with $2m$ generators. Those of $\Omega_i^{(0)}$ are $\underline{v}_{ij}, \overline{v}_{ij}$, $j = 1, \ldots, m$; its principal axes are $[\underline{v}_{ij}, \overline{v}_{ij}]$, $j = 1, \ldots, m$.

II. Let $r = 1$. Note that, $\Omega_i^{(0)} \Big|_{\Delta a_{ik}=0}$ is a $(m-1)$-D plane. Consider the $(m-1)$-D hyperrectangle with 1-D edges $[\underline{v}_{ij}, \overline{v}_{ij}]$, $j \neq k$. Create a grid by dividing each segment $[\underline{v}_{ij}, 0]$ and $[0, \overline{v}_{ij}]$ into 2^r equal subdivisions. This procedure creates a grid of $(2^{r+1}+1)^{m-1}$ points in the $(m-1)$-D plane $\Omega_i^{(0)} \Big|_{\Delta a_{ik}=0}$. Let us denote this set of points by

$$G^{(r)} = \{\mathbf{g}_k^{(r)} \in \Re^{m-1}, \; k = 1, \ldots, (2^{r+1}+1)^{m-1}\}.$$

Clearly,

$$\Omega_i^{(r-1)} \subseteq \Omega_i^{(r)} \subseteq \mathcal{G}_i, \quad \text{where} \quad \Omega_i^{(r)} \doteq \text{conv}\left[\bigcap_{\substack{j=1,\ldots,m \\ \mathbf{g}_k^{(r)} \in G^{(r)}}} \mathcal{G}_i \Big|_{\Delta\mathbf{a}_i E_{jj} + \mathbf{g}_k^{(r)} \hat{E}_{jj}} \right].$$

Each $\Omega_i^{(r)}$ has $2m(2^{r+1}+1)^{m-1}$ generators.

III. Repeat Step II with $r+1$ until growth of $\Omega_i^{(r)}$ is insignificant.

Remarks.

1. At each iteration step, grid $G^{(r)}$ is created by bisection of the 1-D edges $[\underline{v}_{ij}, 0]$ and $[0, \overline{v}_{ij}]$. This allows a better approximate of $\mathcal{G}_i$.
2. In practice, there was no significant growth in $\Omega_i^{(r)}$ after $r = 3$. Hence, this procedure is quite fast.
3. Computational speed of an actual implementation of the above may be significantly increased by incorporating the following:
   (a) At each step, all grid points in $G^{(r)}$ need not be checked; some have already been checked in prior steps.
   (b) Certain orbits are partially overlapped; these portions need not be repeatedly checked.

Single-Length Accumulator Case

In this case, (5) is interpreted as follows: For two consecutive vectors $\mathbf{x}(k), \mathbf{x}(k+1) \in \mathcal{O}_A^{\ell}$,

$$\sum_{j=1}^{m} \mathcal{Q}\left[(a_{ij} + \Delta a_{ij}) \cdot x_j(k)\right] = \sum_{j=1}^{m} \mathcal{Q}\left[a_{ij} \cdot x_j(k)\right] = x_j(k+1). \tag{19}$$

A sufficient condition for (19) to be valid is

$$\mathcal{Q}\left[(a_{ij} + \Delta a_{ij}) \cdot x_j(k)\right] = \mathcal{Q}\left[a_{ij} \cdot x_j(k)\right], \; i, j = 1, 2, \ldots, m. \tag{20}$$

It is this condition that we utilize to obtain the robustness region. Hence, unlike the double-length accumulator case, the region obtained here would be conservative.

Substituting (20) into the quantization scheme being used, $\Delta \mathbf{a}_i$ that satisfy (5) takes the following forms: For SMR quantization, for $k = 0, 1, \ldots, T_{\max} - 1$,

$$\mathcal{G}^{(ij)}_{\mathbf{x}(k)} = \begin{cases} \Delta a_{ij}: & \bar{x}_j(k+1) - \frac{1}{2} - a_{ij} \cdot x_j(k) \le \Delta a_{ij} \cdot x_j(k) < \bar{x}_j(k+1) + \frac{1}{2} - a_{ij} \cdot x_j(k) \\ & \text{for } \bar{x}_j(k+1) > 0 \\ \Delta a_{ij}: & \bar{x}_j(k+1) - \frac{1}{2} - a_{ij} \cdot x_j(k) < \Delta a_{ij} \cdot x_j(k) \le \bar{x}_j(k+1) + \frac{1}{2} - a_{ij} \cdot x_j(k) \\ & \text{for } \bar{x}_j(k+1) < 0 \\ \Delta a_{ij}: & -\frac{1}{2} - a_{ij} \cdot x_j(k) < \Delta a_{ij} \cdot x_j(k) < \frac{1}{2} - a_{ij} \cdot x_j(k) \\ & \text{for } \bar{x}_j(k+1) = 0. \end{cases} \tag{21}$$

For TCT quantization, for $k = 0, 1, \ldots, T_{\max} - 1$,

$$\mathcal{G}^{(ij)}_{\mathbf{x}(k)} = \begin{cases} \Delta a_{ij}: & \bar{x}_j(k+1) - a_{ij} \cdot x_j(k) \le \Delta a_{ij} \cdot x_j(k) < \bar{x}_j(k+1) + 1 - a_{ij} \cdot x_j(k) \\ & \forall \bar{x}_j(k+1). \end{cases} \tag{22}$$

For SMT quantization, for $k = 0, 1, \ldots, T_{\max} - 1$,

$$\mathcal{G}^{(ij)}_{\mathbf{x}(k)} = \begin{cases} \Delta a_{ij}: & \bar{x}_j(k+1) - a_{ij} \cdot x_j(k) \le \Delta a_{ij} \cdot x_j(k) < \bar{x}_j(k+1) + 1 - a_{ij} \cdot x_j(k) \\ & \text{for } \bar{x}_j(k+1) > 0 \\ \Delta a_{ij}: & \bar{x}_j(k+1) - 1 - a_{ij} \cdot x_j(k) < \Delta a_{ij} \cdot x_j(k) \le \bar{x}_j(k+1) - a_{ij} \cdot x_j(k) \\ & \text{for } \bar{x}_j(k+1) < 0 \\ \Delta a_{ij}: & -1 - a_{ij} \cdot x_j(k) < \Delta a_{ij} \cdot x_j(k) < 1 - a_{ij} \cdot x_j(k) \\ & \text{for } \bar{x}_j(k+1) = 0. \end{cases} \tag{23}$$

Here, $\bar{x}_j(k+1) \doteq \mathcal{Q}[a_{ij} \cdot x_j(k)]$. Note that, each of these inequalities are of the following form:

$$\alpha_{kj} \lhd \Delta \mathbf{a}_i \cdot x_j^{(\ell)}(k) \lhd \beta_{kj}, \; k = 0, \ldots, T_{\max} - 1, \; j = 1, 2, \ldots, m. \tag{24}$$

Also, $\alpha_{kj} = \alpha_{kj}(x_j^{(\ell)}(k+1), x_j^{(\ell)}(k), a_{ij})$ and $\beta_{kj} = \beta_{kj}(x_j^{(\ell)}(k+1), x_j^{(\ell)}(k), a_{ij})$.

Let

$$\mathcal{G}_{ij}^{(\ell)} \doteq \{\Delta a_{ij} \in \Re : \Delta a_{ij} \text{ satisfies (24)}\}. \tag{25}$$

As before, α_{kj}, β_{kj}, $k = 0, \ldots, T_{\max} - 1$, $j = 1, \ldots, m$, are all known quantities. We are seeking the set of all perturbations Δa_{ij} that belong to

$$\mathcal{G}_{ij} \doteq \left\{ \Delta a_{ij} \in \Re : \Delta a_{ij} \in \bigcap_{\ell=1,\ldots,T_{\max}-1} \mathcal{G}_{ij}^{(\ell)} \right\}. \tag{26}$$

Proceeding as before, we observe the following: The robustness region for each Δa_{ij}, that is, $\mathcal{G}_{ij}$, is a hyperrectangle in the m-D coefficient space; its 1-D edges are parallel to the corresponding axes.

4 NORMAL FORM REPRESENTATION WITH TCT

Claim : For a normal form filter with two's complement truncation quantization scheme and double length accumulator environment, the limit cycle free region extends on the 45^0 line with zero robustness.

A formal proof for the above claim is given below. Consider a general representation of a normal form matrix on the 45^0 line,

$$A = \begin{bmatrix} -a & a \\ -a & -a \end{bmatrix} \tag{27}$$

For the above matrix global asymptotic stability is already proven[6] for values $0 \leq a \leq 0.5$. Hence we will only consider the section given by $0.5 < a < \frac{1}{\sqrt{2}}$. The analysis for the values $-\frac{1}{\sqrt{2}} < a < -0.5$ will be identical. Starting from $\mathbf{0}$ and mapping in the reverse[1] direction we identify all integer initial conditions vectors that converge to the zero vector by consecutive iterations. Consider all initial conditions $[x_1\ x_2]^t$, $x_1, x_2 \in \mathcal{Z}$ that converges to $\mathbf{0}$ in one iteration of the following equation

$$\mathcal{Q}\left[\begin{pmatrix} -a & a \\ -a & -a \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{28}$$

where $0.5 < a < \frac{1}{\sqrt{2}}$. From (10), (28) can be interpreted as the following two inequalities

$$0 \leq -ax_1 + ax_2 < 1 \tag{29}$$

$$0 \leq -ax_1 - ax_2 < 1 \tag{30}$$

Each inequality given in (29) and (30) consists of two parallel lines enclosing a region where one side is open and the other side is . closed The only possible x_1 value for the given variation in parameter a is -1, and in the x_2 direction the only integer included in the region is 0. Hence for $0.5 < a < \frac{1}{\sqrt{2}}$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \leftarrow \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

Likewise we consider all initial condition vectors in $\mathcal{Z}$ that converge to $[-1\ 0]^t$ for $0.5 < a < \frac{1}{\sqrt{2}}$. After consecutive reverse operations of this form we obtain the following orbit for $0.5 < a < \frac{1}{\sqrt{2}}$.

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \leftarrow \begin{bmatrix} -1 \\ 0 \end{bmatrix} \leftarrow \begin{bmatrix} 0 \\ -1 \end{bmatrix} \leftarrow \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Consider the mapping

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} \longleftarrow \begin{bmatrix} x_7 \\ x_8 \end{bmatrix} \quad x_7, x_8 \in \mathcal{Z}$$

The inequalities for the above iteration is,

$$0 \leq -ax_7 + ax_8 < 1 \tag{31}$$

$$1 \leq -ax_7 - ax_8 < 2 \tag{32}$$

When $0.5 < a < \frac{1}{\sqrt{2}}$ there are two vectors satisfying (31) and (32) that converge to $[0 \quad 1]^t$, they are

$$\left\{ \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix} \right\}$$

It is observed that only one vector converge to $[0 \quad 1]^t$ in the entire region given by $0.5 < a < \frac{1}{\sqrt{2}}$ namely $[-1 \quad -1]^t$. As a increases the area inside the diamond formed by (31) and (32) grows smaller, and at one point the vector $[-2 \quad -1]^t$ falls outside the boundary of the diamond. The edge where this crossing occurs is given by (32).

$$-ax_7 - ax_8 < 2$$

If $[-2 \quad -1]^t$ is just satisfied

$$-a(-2) - a(-1) < 2 \quad \rightarrow \quad a < \frac{2}{3}$$

Therefore for $0.5 < a < \frac{2}{3}$

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} \longleftarrow \left\{ \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix} \right\}$$

and for $\frac{2}{3} \leq a < \frac{1}{\sqrt{2}}$

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} \longleftarrow \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

For future calculations let us only consider the parameter values $\frac{2}{3} \leq a < \frac{1}{\sqrt{2}}$. Following the earlier procedure the orbit is constructed for $\frac{2}{3} \leq a < \frac{1}{\sqrt{2}}$

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} \longleftarrow \begin{bmatrix} -1 \\ -1 \end{bmatrix} \longleftarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix} \longleftarrow \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Consider the mapping

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} \longleftarrow \begin{bmatrix} x_{13} \\ x_{14} \end{bmatrix} \quad x_{13}, x_{14} \in \mathcal{Z}$$

The above mapping consists of the two inequalities given by,

$$-1 \leq -ax_{13} + ax_{14} < 0 \tag{33}$$

$$-1 \leq -ax_{13} - ax_{14} < 0 \tag{34}$$

It is seen that apart from the two initial conditions $[-1 \quad -1]$ and $[-1 \quad -2]$ which lie on the open boundaries there are no integers inside the region. Hence we can conclude that in the region given by $\frac{2}{3} \leq a < \frac{1}{\sqrt{2}}$ there are no integers that converge to $[-1 \quad 1]$. That is for $\frac{2}{3} \leq a < \frac{1}{\sqrt{2}}$ the orbits that reach $\mathbf{0}$ will only have vectors that were obtained in Steps 1-6. The upper bound[1] for the matrix given in (27) is given by

$$\frac{2a+1}{1-2a^2}$$

Therefore for the variation $\frac{2}{3} \leq a < \frac{1}{\sqrt{2}}$ the upperbound variation is given by

$$21 \leq M < \infty$$

Note that both components of a vector will be bounded by the same value. This implies that in the region considered not all vectors in the set $\mathcal{S}^{(0)}$ converge to zero. Hence it can be concluded that for a filter given by (27) with $\frac{2}{3} \leq a < \frac{1}{\sqrt{2}}$ under two's complement truncation and double length accumulator environment will have limit cycles.

5 RESULTS

The results are listed for a digital filter in a 10 Bit environment. Since most industrial applications are made out of second order blocks the results are given for a second order filter. Consider a second order digital filter subsection with the transfer function given by (35).

$$H(z) = \frac{1}{1 + 0.8z^{-1} + 0.32z^{-2}} \tag{35}$$

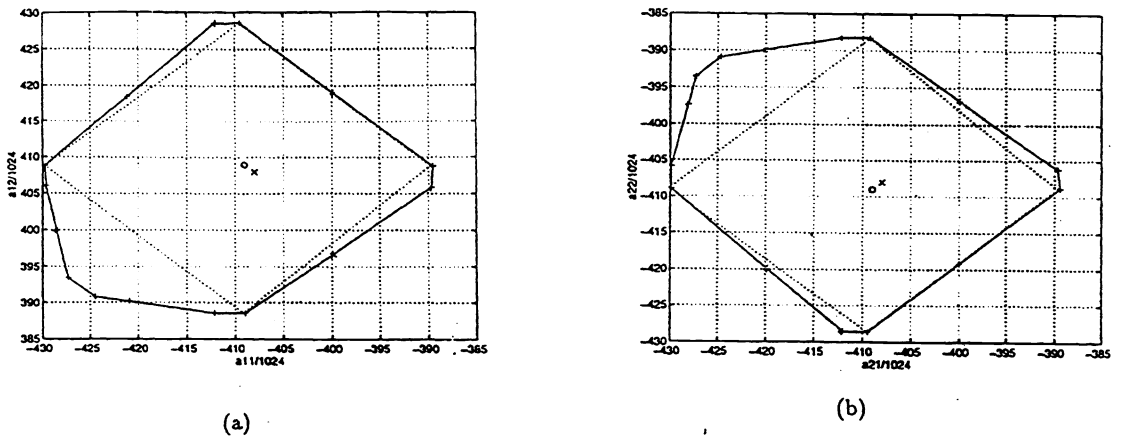
The poles of the above transfer function, if infinite wordlength is assumed, are located at $0.4 \pm j0.4$. If (35) is converted into state-space form the normal form coefficient matrix is

$$A_{N,\infty} = \begin{bmatrix} -0.4000 & 0.4000 \\ -0.4000 & -0.4000 \end{bmatrix}. \tag{36}$$

The closest 10-bit machine representable form of A_N is as follows:

$$A_{N,10} = \begin{bmatrix} -\frac{409}{1024} & \frac{409}{1024} \\ -\frac{409}{1024} & -\frac{409}{1024} \end{bmatrix}. \tag{37}$$

This matrix is limit cycle free; its robustness region, together with $\Omega_i^{(0)}$, $\Omega_i^{(1)}$, and $\Omega_i^{(2)}$, are depicted in Fig. (1). Notice that,



(a) (b)

Figure 1: Robustness region for (37)

$$\hat{A}_{N,10} = \begin{bmatrix} -\frac{408}{1024} & \frac{408}{1024} \\ -\frac{408}{1024} & -\frac{408}{1024} \end{bmatrix} = \begin{bmatrix} -\frac{102}{256} & \frac{102}{256} \\ -\frac{102}{256} & -\frac{102}{256} \end{bmatrix} \tag{38}$$

falls on this 10-bit grid as well as the 8-bit grid within the robustness region. Hence, being certain of no limit cycles as the robustness region indicates, one may represent the filter coefficients with only 8-bit registers! In fact, there are several choices for the designer; for example, a filter with coefficients $\pm 416/1024 = \pm 104/256$ also falls on both grids. The impulse responses of (36) and (37) are plotted in Fig. (2). We notice that they are identical hence the designed system and the implemented system will have similar characteristics. In high order filter implementations, savings accrued via shorter coefficient registers can be substantial, especially in high speed applications.

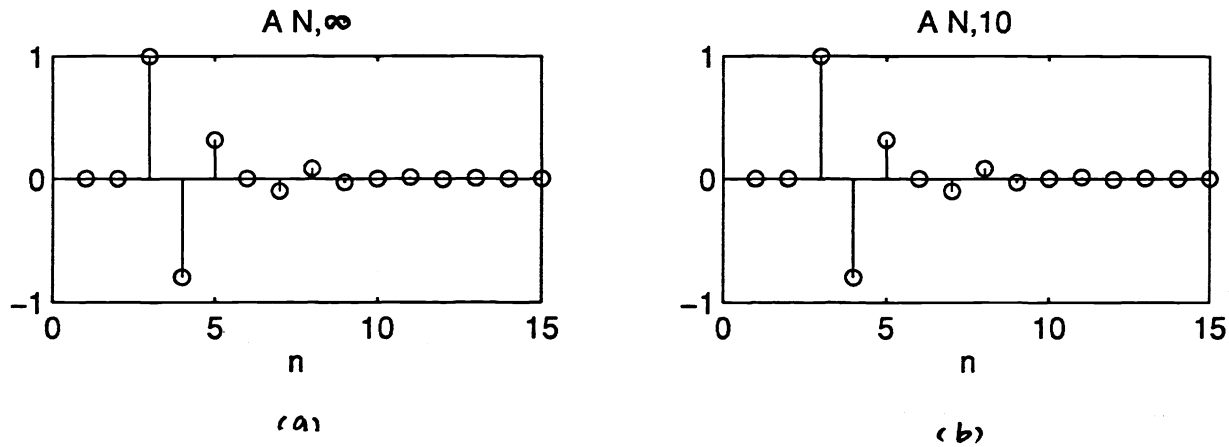


Figure 2: Impulse response of, (a) (36) and (b) (37).

Zero Robustness Results Consider the following normal form matrix on the 45^0 line,

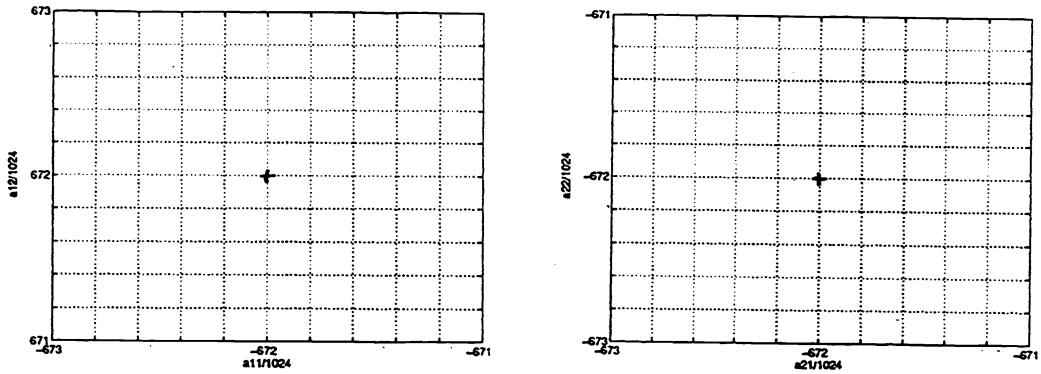
$$A_{N,10} = \begin{bmatrix} -\frac{672}{1024} & \frac{672}{1024} \\ -\frac{672}{1024} & -\frac{672}{1024} \end{bmatrix}. \tag{39}$$

The robustness region for the above matrix is plotted in Fig. (3), it is seen that it has zero robustness thus supporting the proof in section 4. A series of stable points will extend on the 45^0 line up to $a < \frac{2}{3}$.

6 CONCLUSION

A novel technique for constructing the robustness region for a given digital filter was presented. This technique explicitly constructs the region to the required degree of accuracy under the assumption of identical orbits. The algorithm is highly versatile since the digital filter is assumed to be in its state-space form. It is also applicable to any arithmetic type, order, and for single as well as double length accumulator lengths. The complexity of the computations increases with the filter order.

The constructed region can be used to optimize the bit length allocation for the filter coefficients. A shorter bit length for for the coefficient storage will yield cost effective designs when the filter considered has a large



(a) (b)

Figure 3: Robustness region for (39).

number of coefficients. In conclusion this algorithm provides a valuable design tool for filters having low or zero input conditions. degree of accuracy

7 REFERENCES

[1] K. Premaratne, E. Kulasekere, P. Bauer and L. Leclerc "An exhaustive search algorithm for checking limit cycles of digital filters," *Proc. 1995 IEEE Int. Symp., Circ. Syst.*, Vol. 3 pp. 2035-2038, Apr. 1995.

[2] P.H. Bauer and L.J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed-point digital filters," *IEEE Trans. Sig. Proc.,* vol. 39. no. 11, pp. 2400-2409, Nov. 1991.

[3] A. C. M. Claasen, F. G. Mecklenbräuker and J. B. H. Peek, "Effects of quantization and overflow in recursive digital filters," *IEEE Trans. Acoust., Speech, Sig. Processing,* vol. ASSP-24, no. 6, pp. 517-529, Dec 1976.

[4] S. R. Parker and S. F. Hess, "Limit cycle oscillations in digital filters," *IEEE Trans. Circ. Theory,* vol. CT-18, no. 10. pp. 687-697, Nov. 1971.

[5] Young. Lim, "Prediction coding for FIR filter wordlength reduction," *IEEE Tran. Circ. Syst.,* vol. CAS-32, no. 4, pp. 365-372, Apr. 1985.

[6] T. Bose, "Stability of digital filters implemented with Two's complement truncation quantization," *IEEE Trans. Sig. Proc.,* vol. 40. no. 1, pp. 24-31, Jan. 1992.

[7] Digital Filters:Analysis and Design, Andreas Antoniou, ©*McGraw-Hill, Inc.,* 1979.

[8] E. Walter and H. Piet-Lahanier, "Exact and recursive description of the feasible parameter set for bounded error models," *Proc. 26th Conf. Decision Contr.,* (Los Angeles, CA). pp. 1921-1922, Jul. 1988.