Optimum structure design for finite precision implementation of digital systems

Hao, Jinxin

2006

Hao, J. (2006). Optimum structure design for finite precision implementation of digital systems. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/3476>

<https://doi.org/10.32657/10356/3476>

Nanyang Technological University

Downloaded on 26 Apr 2023 15:57:27 SGT

Optimum Structure Design for Finite Precision Implementation of Digital Systems

Hao Jinxin

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2006

*To my parents and Ran,
for their encouragement and love.*

Acknowledgments

First of all, I would like to express sincere thanks to Dr. Gang Li, my supervisor, for his professional and outstanding guidance during my research towards the Ph.D. degree. When starting this Ph.D. program three years ago, I, a fresh Bachelor holder, wondered if I could complete it three years later. Dr. Li's invaluable support and continuous encouragement greatly increased my confidence and made the whole research process thoroughly enjoyable.

I am also very grateful to School of EEE, Nanyang Technological University for funding my scholarship.

This work is dedicated to my parents, Yi Hao and Xiaohong Yan, who have provided me with the ideal environment to grow up in. Their persistent emotional support always gives me the strength to pass through all kinds of difficulties. Particularly, I want to thank my grandma for her very helpful enlightenment in my early childhood education.

I would like to extend my special thanks to Ran Duan, my wonderful girl friend, for her sweet love, unbelievable mild temper and the good luck she always brings to me. Many thanks go to all my faithful friends, for the kind help they have ever offered to me and every great time we have shared together. The names are not listed here but the friendship will last for long.

Abstract

It has long been noted that the actual performance of a well-designed digital system may degrade greatly when it is implemented with a digital device due to the finite word length (FWL) effects. These effects have been considered as one of the most important issues in digital system implementation. It is well known that a system can be realized in many different structures and that different structures possess different capabilities against the FWL errors. Therefore, the optimal FWL structure design is to search for those structures that minimize the FWL effects in a certain performance criterion.

This thesis covers a series of optimal FWL structure problems for digital filters and controllers with stability robustness and roundoff noise consideration.

Firstly, based on a pole modulus sensitivity approach, a new stability-related measure is proposed. The optimal FWL state-space realization problems in terms of maximizing this measure are solved for digital filters and controllers, respectively. It is found the optimal realizations yield a better stability performance than the canonical shift operator-based realizations, but are usually fully parametrized, which increase the implementation complexity.

Secondly, to improve the implementation efficiency, two sparse structures are developed using a set of first-order polynomial operators. Each structure

contains some free parameters, with which the structure can be optimized to achieve nice stability performance. How to choose the free parameters to enhance the stability robustness is investigated, which on the one hand, helps to accelerate the searching process for the optimal structures, while on the other hand, reflects some limitations of first-order polynomial operator-based structures. Then, two improved sparse structures are exploited adopting a set of second-order operators. These structures have more degree of freedom, with which the performance can be further optimized for a wide range of systems. The optimal sparse structure problems are solved in terms of maximizing the stability robustness with practical considerations. It is shown that the second-order operator-based structures exhibit better stability performance than those using first-order operators, and beat the fully parametrized realizations greatly in terms of both stability robustness and implementation efficiency.

Thirdly, the effect of roundoff noise in a digital controller is investigated for a discrete-time control system. The optimal roundoff noise structure problems are solved for both fully parametrized realizations and the second-order operator-based sparse structures. It is shown that the corresponding optimized sparse structures exhibit excellent performance in terms of reducing the roundoff noise and the implementation complexity.

Finally, the pole deviation is analyzed based on a second-order perturbation theory. A new measure is then defined to evaluate the system stability, which is a very good approximation of the unsolved classical stability measure and is more accurate than the approach based on the first-order approximation. The study of second-order pole sensitivity provides some perspectives for future research on the stability issue of digital systems.

Contents

Acknowledgments	i
Abstract	ii
List of Figures	viii
List of Tables	x
Notations	xi
1 Introduction	1
1.1 Motivation and objectives	1
1.2 A literature review	4
1.3 Major contributions	16
1.4 Outline of the thesis	18
2 Preliminaries	21
2.1 Digital systems, parametrizations and structures	21
2.2 Stability measures	25
2.2.1 A rigorous stability measure	27

2.2.2	A practical pole sensitivity-based measure	28
2.3	Roundoff noise gain	31
2.4	Summary	34
3	A New Stability-Related Measure	35
3.1	A pole modulus sensitivity-based measure	35
3.2	Optimal stability realizations	38
3.2.1	Open-loop systems - digital filters	39
3.2.2	Closed-loop systems - digital controllers	44
3.3	Numerical examples and simulations	46
3.4	Summary	51
4	Polynomial Operator-Based Sparse Structures	53
4.1	Polynomial parametrization	55
4.2	Two sparse structures	56
4.2.1	Generalized direct-form II transposed structures	56
4.2.2	Equivalent state-space realizations	58
4.3	Stability analysis and structure optimization	59
4.3.1	Optimal ρ DFIIt structures	60
4.3.2	Optimal ρ -realizations	66
4.4	Numerical examples	67
4.5	Summary	79
5	Improved Sparse Structures Using Second-Order Operators	82

5.1	Two improved sparse structures	83
5.1.1	The improved ρ DFIIt structure	83
5.1.2	The improved ρ -realization	88
5.2	Performance analysis and optimized structures	90
5.2.1	Optimized ρ IDFIIIt structures	90
5.2.2	Optimized ρ_I -realizations	96
5.3	Numerical examples	97
5.4	Summary	109
6	Optimal Roundoff Noise Controller Structure Design	110
6.1	Optimal roundoff noise controller realizations	111
6.1.1	Derivation of roundoff noise gain	113
6.1.2	Optimal fully parametrized state-space realizations	118
6.2	Sparse controller structures with minimum roundoff noise gain .	121
6.2.1	Roundoff noise analysis	123
6.2.2	Structure optimization	127
6.3	A numerical example	130
6.4	Summary	136
7	Pole Deviation Analysis Based on Second-Order Perturbation Theory	137
7.1	The first- and second-order approximation	138
7.2	Computing the second-order pole sensitivity	139
7.3	A more accurate measure to approximate μ_0	143

<i>Contents</i>	vii
7.4 A numerical example	145
7.5 Summary	147
8 Conclusions and Future Work	148
8.1 Conclusions	148
8.2 Recommendations for future work	151
A A Brief Description of GA to Solve Optimization Problems	154
Author's Publications	158
Bibliography	160

List of Figures

2.1	Block diagram of the direct-form II structure	22
2.2	General state-space description of a linear time-invariant system	24
2.3	Block diagram of a discrete-time feedback system	26
4.1	A realization of $\rho_k^{-1}(z)$ defined in (4.5)	58
4.2	Block diagram of the ρ DFIIt structure	58
4.3	Pole distributions of different state-space realizations (Example 4.2)	75
4.4	Pole distributions of different DFIIt structures (Example 4.2) . .	76
4.5	Pole distributions of different state-space realizations (Example 4.3)	79
4.6	Pole distributions of different DFIIt structures (Example 4.3) . .	80
5.1	Block diagram of the proposed ρ IDFIIt structure	87
5.2	Implementation of the m th block in the ρ IDFIIt structure, where $\bar{m} \triangleq K - 2p + 2m$ and $w_{K+1}(n) = 0$	87
5.3	Pole distributions of R_ρ^{opt} and $R_{\rho_I}^{opt}$	101
5.4	Pole distributions of ρ DFIIt and ρ IDFIIt	101

5.5	Unit impulse responses of the closed-loop system for ideal controller (-), 10-bit implemented R_c (...), and 8-bit implemented R_f^{opt} (- -).	107
5.6	Unit impulse responses of the closed-loop system for ideal controller (-), 4-bit implemented R_ρ^{opt} (- -), and 4-bit implemented $R_{\rho_I}^{opt}$ (-.-).	108
5.7	Unit impulse responses of the closed-loop system for ideal controller (-), 10-bit implemented $zDFIIt$ (...), 4-bit implemented $\rho DFIIt$ (- -), and 4-bit implemented $\rho IDFIIt$ (-.-).	108
6.1	Unit impulse responses of the closed-loop system, where the solid line is for the ideal response, while the dotted, dashed and dash-dot lines are for $R_c^{l_2}$, R_ρ and R_{ρ_I} , respectively, with $B_s = 8$ bits.	135
6.2	Unit impulse responses of the closed-loop system, where the solid line is for the ideal response, while the dotted, dashed and dash-dot lines are for $zDFIIt$, $\rho DFIIt$ and $\rho IDFIIt$, respectively, with $B_s = 8$ bits.	135
A.1	Flowchart of the GA to solve optimization problems	157

List of Tables

3.1	Stability robustness comparison of R_0 and R_n	48
3.2	Stability robustness comparison of R_c and R_f^{opt}	51
4.1	Comparison of different state-space realizations (Example 4.1) .	69
4.2	Comparison of different DFIIIt structures (Example 4.1)	70
4.3	Comparison of different state-space realizations (Example 4.2) .	73
4.4	Comparison of different DFIIIt structures (Example 4.2)	73
4.5	Comparison of different state-space realizations (Example 4.3) .	78
4.6	Comparison of different DFIIIt structures (Example 4.3)	78
5.1	Comparison of two types of optimal sparse structures	100
5.2	Comparison of different state-space realizations (Example 5.2) .	104
5.3	Comparison of different DFIIIt structures (Example 5.2)	105
6.1	Comparison of different state-space realizations	132
6.2	Comparison of different DFIIIt structures	133
7.1	Comparison of two stability measures	146

Notations

- $x(n)$: a discrete-time signal x
- z^{-1} : the backward shift operator such that $z^{-1}x(n) = x(n-1)$
- $\lambda_k(A)$: the k th eigenvalue of matrix $A = \{a_{ij}\} \in \mathcal{R}^{K \times K}$
- A^{-1}, A^T : inverse and transpose of matrix A , respectively
- A^H : complex conjugate and transpose of matrix A
- $Q[x]$: the quantized version of a variable x
- $M^{1/2} \geq 0$: the square root matrix of $M \geq 0$, $M^{1/2}M^{1/2} = M$
- M_c^r, M_c^i : the real and imaginary parts of a complex valued matrix M_c , respectively
- $\mathbf{0} / I$: zero / identity matrix of proper dimension
- $\mathbf{0}_m / I_m$: the zero / identity matrix of dimension m
- e_k : the k th elementary (column) vector, whose elements are all zero except the k th one which is 1
- $E[\cdot]$: the statistical average operation
- $\|\cdot\|_F$: the *Frobenius* norm
- $tr(\cdot)$: the trace operation
- $Vec(\cdot)$: the column stacking operator

Chapter 1

Introduction

1.1 Motivation and objectives

The design of digital systems¹ has been a very important field of many applications. See [1] - [3] for digital filter design, and [4], [5] for controller design. In real-time applications, well-designed filters and controllers are usually implemented with a digital device such as a digital signal processor (DSP) or a general purpose microprocessor. Nowadays, the rapid developments and improvements in processor technology have greatly decreased the cost of implementing digital systems. Meanwhile, the main digital system design requirements, such as the hardware implementation issues and relevant parametrization concepts (i.e. the design of the actual system realization parameters), have become increasingly important in the design cycle.

When an infinite precision filter or controller is implemented with a digital device, which is always of finite precision, the actually implemented system will

¹In this thesis, we consider two types of digital systems: digital filters and digital controllers.

be different from the ideal one due to the so-called finite word length (FWL) effects. The effects of the FWL errors have been considered as one of the most serious problems in digital system implementation [6], [7]. Generally speaking, there are mainly two types of FWL errors in the digital system. The first one is the perturbation of the system parameters implemented with FWL and the second one is the rounding error that occurs in arithmetic operations. In practice, these two FWL effects occur jointly, and both result in a performance degradation of the digital system compared with its ideal behavior.

It is well known that a linear system can be represented with many different parametrizations and implemented with different structures. Theoretically, these parametrizations/structures are totally equivalent in infinite precision since they represent one and the same system. The important point, however, is that different parametrizations/structures have different numerical properties, which lead to different system performance degradations when implemented with FWL. Given this fundamental observation, it is then natural for one to go through a second stage of design, the *a posteriori* structure design to overcome the FWL effects. In the early days, the digital system was implemented with direct forms and the structure problem was not seriously considered. Later, the poor performance of a system implemented in these forms made one realize the importance of the structure issue, which leads to the so-called *optimal FWL structure design*. Since different structures have different robustness against the FWL effects, the optimal FWL structure design is to search for those structures that minimize the FWL effects in terms of some specified design performance criteria. These performance criteria are related to the two main types of FWL errors. The most commonly used performance criteria are a sensitivity measure and a roundoff noise gain.

The optimal FWL state-space realization design for digital filters has been an active research field during the last thirty years or so, where the optimal realizations are defined as those that minimize a given cost function such as a transfer function sensitivity measure and a roundoff noise gain, while the stability issue is rarely touched. Stability is one of the essential design requirements. It should be noted that a well-designed stable filter may become unstable due to the perturbation of its parameters implemented with FWL. To avoid this, it is desired that the filter be implemented with a structure that has a very large stability robustness against the FWL effects. *So the first objective of this thesis is to investigate the problem of optimal filter realizations which can achieve very good stability performance.*

The effects of FWL errors have been much studied by the digital signal processing community, while these effects have received much less attention in control problems despite the widespread use of microprocessors in control applications. The numerical problems in digital control are just as important as those in digital signal processing, particularly with the recent trend towards the use of fast sampling control. *The second objective of this thesis is to investigate the optimal structures for digital controllers in terms of maximizing the stability robustness and minimizing the roundoff noise gain of the closed-loop systems.*

It has been noted that the optimal realizations are usually fully parametrized, for which all the parameters in the realization are nontrivial². This would increase the complexity of implementation and hence slow down the processing speed. In practice, it is desired that the actually implemented systems have a nice performance against the FWL effects as well as a very simple structure

²By *trivial parameters*, we mean those that are 0 and ± 1 . Other parameters are, therefore, referred to *nontrivial parameters*.

that possesses many trivial parameters which can be implemented exactly and produce no FWL errors. Exploiting sparse structures of high performance is one of the trends in digital system structure design. *The third objective of this thesis is to develop sparse structures based on the polynomial parametrization concept, which are not only efficient for implementation, but also exhibit excellent performance against the FWL effects.*

1.2 A literature review

New technologies in the past 20 years have reduced the cost of digital hardware, and its speed has increased to such an extent that digital signal processing has replaced a great deal of analog signal processing. One of the basic problems in replacing analog signals and processors is the finite precision problem. In implementing a system with digital processors, the finite precision aspects are mainly twofold: the FWL representation of system parameters and the finite precision arithmetic operations in which roundoff occurs. Traditionally, finite precision analysis is separated into sensitivity and roundoff noise studies, which will be elaborated further below.

Much attention has been paid to the numerical problems caused by the FWL effects in digital filter implementation for more than two decades. The optimal FWL state-space realization design has been considered as one of the most effective methods to minimize the effects of FWL errors on the performance of digital filters.

It turns out that the first significant results on the design of state-space filter realizations minimizing FWL effects are results on the minimization of the

roundoff noise gain rather than on the minimization of transfer function sensitivity measures. The pioneering work in exploring the use of different realizations in actual filter implementations can be found in Mullis and Roberts' seminal 1976 paper [8]. In their work, similarity transformations were explored to identify the state-space realizations that minimize the output error variance due to arithmetic roundoff errors. Independently, Hwang considered the same problem and gave a constructive procedure to determine the optimal transformations or realizations [9]. Since then the optimal FWL state-space realization design has become one of the attractive subjects in digital filter implementation. Early studies of roundoff noise minimization for two dimensional state-space filters can be found in [10] - [13]. Many available results have been collected in [6], [7].

Error feedback [14] - [27] has been known as another effective technique for reducing the roundoff noise at the filter output. This is achieved by extracting the quantization error after multiplication and addition, and then feeding the error signal back through simple circuits. It can be applied to digital filters that are described by either external or internal models for roundoff noise minimization without affecting the filter's input-output relationship. In [22], Hinamoto presented a number of results for computing the optimal error feedback coefficients for any given state-space realization.

The design of state-space realizations minimizing transfer function sensitivity measures was first studied by Tavsanoglu and Thiele [28], [29]. They defined a rather sophisticated frequency independent sensitivity measure of a transfer function with respect to the parameters of a corresponding realization (A, B, C) . This measure contained a mixture of an L_1 -norm on the sensitivity function with respect to the matrix A and an L_2 -norm with respect to the vectors B and

C , which is therefore called an L_1/L_2 sensitivity measure. The success of this method stems from the fact that an upper bound of this sensitivity measure can be obtained and easily minimized. In [30], Thiele showed that not only the upper bound but also the sensitivity measure itself is minimized by all realizations whose controllability and observability Gramian matrices are identical. In [31], the same conclusion was achieved using a geometric approach. These results have since been extended to multiple input and multiple output systems [32]. In the subsequent research, it is found that the combination of an L_1 -norm with respect to some parameters and an L_2 -norm with respect to others is opportunistic rather than logical. In [7], a more reasonable transfer function sensitivity measure was proposed, which is based purely on L_2 -norm for all the sensitivity functions instead of a mixture of L_1 - and L_2 -norms. The corresponding optimal sensitivity problem has been solved [7], [33]. An improved L_2 sensitivity measure was proposed in [34], which distinguishes the trivial parameters from other coefficients. Minimization of L_2 sensitivity for two dimensional system realizations was investigated in [35] - [39].

As discussed above, the optimal realization problem has been well studied for digital filters in terms of minimizing the roundoff noise gain and the transfer function sensitivity measures. During the past years, a lot of attention has been paid to the sensitivity of the transfer function with respect to the realization coefficients, but less to the problem of pole sensitivity with respect to the system parameters. In [40], Mantey examined how to use a similarity transformation to obtain realizations that minimize some measure of the pole sensitivity. By comparing a few special realizations, the author argued that in order to minimize the pole sensitivity, the realizations that have a block diagonal form are desired. Clearly, this conclusion falls of short of generality. In [41], Skelton and

Wagie concluded that the realizations whose matrix A is normal have a minimal pole sensitivity, but the problem of how to transform an arbitrary realization into a normal one was not addressed. Gevers and Li provided a solution to this problem and obtained a number of results in zero sensitivity [7]. It has been noticed [42] the realizations that have a minimal pole sensitivity do not guarantee a minimal zero sensitivity and vice versa. The pole and zero sensitivity simultaneous minimization problem was investigated by Li in [43], where an iterative algorithm was proposed to find the optimal realizations in terms of minimizing both the pole and zero sensitivity.

The pole sensitivity study plays an very important role in digital system design. It is well known that a digital filter or a closed-loop system is said to be stable if the poles of the system are inside the unit circle. When the digital system is implemented with FWL, the perturbation of the system parameters leads to the deviation of the system poles and the stability may be lost. The pole sensitivity study is, to some extent, an effective way to evaluate the stability robustness of digital systems. The optimal stability realizations can be defined as those minimizing the pole sensitivities. As far as we know, the stability issue for digital filter implementation with FWL is rarely touched, while for the digital controller, many studies have addressed the problem of finding optimal controller structures based on various FWL stability measures, in which the pole sensitivity-based measures are the most popular ones for the optimal controller state-space realization design in terms of maximizing the stability robustness [44], [45]. In Chapter 3 of this thesis, a new stability-related measure is proposed based on a pole modulus sensitivity approach. This measure can be applied to the optimal structure design for both digital filters and controllers with stability consideration.

The results discussed so far have been on the minimization of FWL effects in the realization of digital filters. By comparison, the FWL effects have received much less attention in the control literature. This is also one of the reasons why more strength has been put into this thesis on the investigation of FWL effects in digital controller structure design.

It was not until the late 1980s that the problem of optimal digital controller realizations minimizing the roundoff noise gain was addressed. The roundoff noise gain was derived for a control system with a state-estimate feedback controller and the corresponding optimal realization problem was solved in [46], while the roundoff error effect on the linear quadratic regulation (LQG) performance was investigated in [47] and the optimal solution was obtained by Liu *et al* [48]. The problem of finding the optimum roundoff noise structures of digital controllers in a sampled-data system has been investigated in [49].

The other type of FWL errors in digital controllers, that is, the controller parameter errors due to FWL, are classically studied with a transfer function sensitivity measure. In [7], [46], the analysis of FWL errors in controllers on the transfer function was performed based on the discrete-time counterpart of the sampled-data system. The sensitivity measure is the same as that proposed in [29] but the results are an extension of [29], [32] to the feedback control problem. Later, Madievski *et al* [50] derived a sensitivity measure based on a hybrid operator transfer function of the sampled-data system and solved the corresponding optimal realization problem.

Stability is a much more important issue in control system than in an open-loop system. In fact, it is the premier concern in control system design [51]. The stability of the closed-loop control system may be lost due to the FWL

errors of the digital controller parameters. Recently, the effects of the controller parameter errors have been investigated with some stability robustness-related measures such as the one based on the complex stability radius [52], [53] and those based on pole sensitivity [54] - [58]. In [53], Fialho and Georgiou used the complex stability radius measure to formulate an optimal FWL controller realization problem that can be represented as a special H_∞ -norm minimization problem and formulated as a linear matrix inequality [5]. The pole sensitivity approach became widely used in studying FWL effects on the stability behavior of the closed-loop system since 1998, when Li proposed a tractable pole sensitivity-based measure to evaluate the closed-loop stability and to find optimal controller structures with maximum stability robustness [54]. Later, some alternative pole sensitivity-based measures have been achieved by several researchers [55] - [58], in which the l_1 -norm- and l_2 -norm-based measures are the most popularly used ones. In [55], an l_1 -norm-based measure was derived to estimate the close-loop stability robustness. An improved version of this measure was later proposed in [56], using the pole modulus sensitivity, and the optimal structure problem was solved by maximizing the proposed measure. This improved measure is more reasonable than that in [55] due to the use of pole modulus and the consideration of the system structure details. The measure in [57], by nature, is the same as the one in [56] but incorporates a unification of different representation schemes and considers the dynamical range and precision requirements. We shall see in Chapter 3 that our newly proposed stability measure is in fact an improved version of the measure in [54]. The new measure is derived using l_2 -norm based on the pole modulus sensitivity approach. With this measure, the optimal controller structure problem is formulated and solved. It has been found that the optimal structure problems can be carried

out more easily with the l_2 -norm-based measures than the l_1 -norm-based measures. In this sense, the newly proposed measure combines the advantages of two prevailing measures addressed in [54] and [56].

It has been noted that the optimal realizations obtained with the above design methods are usually fully parametrized. This means that for a digital system of order K , the optimal realizations contain $(K + 1)^2$ nontrivial parameters, which, consequently, increase the implementation complexity and hence slow down the processing speed. In practice, it is desirable to implement the digital system with such a structure that not only yields a very good performance against the FWL effects but also possesses many trivial parameters, which can be implemented exactly and produce no FWL errors. Implementation using such a sparse structure can reduce the storage requirements and the computation time. Nowadays, exploiting sparse structures of high performance becomes one of the trends in digital system structure design.

Much effort has been made to achieve sparse optimal or quasi-optimal realizations by several researchers. For digital filters, since the degrees of freedom in the optimal roundoff noise realizations are very limited, sophisticated numerical algorithms were developed to obtain the so-called quasi-optimal realizations that have a sparse structure [59], [60]. Besides the numerical difficulty involved in those algorithms, the position of each trivial parameter in the obtained structure is not predictable. In fact, the quasi-optimal sparse structure for one digital filter may be very different from that for another. For minimal sensitivity realizations, even though the number of nontrivial parameters can be much reduced using the degrees of freedom in the optimal realizations, the amount of nontrivial parameters is still proportional to K^2 [61], [62]. In [63], based on a matrix

factorization an interesting structure was derived, which possesses $5K - 1$ non-trivial parameters and the position of each parameter is always fixed.

As far as we know, few results have been published on the sparseness issue for the controller structure design. One exception is [54], in which Li investigated the sparse controller structures that maximize the stability robustness of the closed-loop system. By extending the algorithms adopted in [59], [60], a sophisticated numerical algorithm was used to transform a fully parametrized realization into a sparse one. However, the transformation procedure was performed at the cost of sacrificing the stability robustness and with a lot of computations involved due to the fact that the number of variables in the corresponding optimization is proportional to K^2 . Besides, the positions of trivial parameters in the resultant sparse realizations were not predictable, which is a drawback for implementation.

In the digital signal processing area, it is well known that though having poor numerical properties, the conventional shift operator-based direct forms such as direct-form II (DFII) and direct-form II transposed (DFII_t) structures are the simplest structures. Recently, the direct forms in delta operator have been studied by researchers (see, e.g., [64] - [69]) to improve the poor FWL performance of those direct forms in shift operator. An extensive comparative study of different direct forms in the delta operator was carried out in [66] for IIR filters, where the transfer function is cascaded into second-order sections and each section is implemented with a direct form in delta operator. It was shown there that among all the direct forms, the delta DFII_t structure (δ DFII_t) has the lowest quantization noise level at output. In [69], the δ DFII_t structure was investigated for an arbitrary order IIR filter, where the concept of different

coupling coefficients at different branch nodes is utilized for better roundoff noise gain suppression.

The use of delta operator, defined as $\delta = \frac{z-1}{T_s}$ with T_s the sampling period, was first promoted by Peterka [70] and Middleton and Goodwin [71] in estimation and control applications. Two major advantages are claimed for the use of this operator: a theoretically interesting unified formation of continuous-time and discrete-time filtering and control theory, and a range of practically interesting numerical advantages connected with FWL effects. Later on, the numerical properties of the delta operator, where T_s is replaced by a positive factor Δ , were investigated by Li from a pure algebraic point of view [7], [72]. Recently, the use of delta operator has also been promoted in optimal FWL controller structure design [73], [74]. In [75], a comparative study on optimizing closed-loop stability bounds of FWL controller structures with shift and delta operators was carried out for a sampled-data control system.

Throughout the previous studies, it was found that one can make the transfer function in the delta operator have better numerical properties in the case where the poles of the transfer function are closer to $z = +1$ than $z = 0$, i.e. the type of low-pass narrow band filters and those closed-loop systems whose poles are clustered around $z = +1$ (see [7], [72], [73]). This means that the delta operator-based structures (including the δ DFIIt structure) may not yield a satisfactory performance if some of poles of the transfer function are far away from $z = +1$.

Noting the above limitation of using delta operator, one question we ask ourselves is what can be achieved if the system is parametrized with more generalized operators. This is the motivation for us to study the generalized polynomial parametrization problem. The main objective in Chapter 4 of the thesis

is to develop a set of special polynomial operators, called ρ -operators, based on the concept of polynomial parametrization proposed in [7]. Applying the ρ -operators into the DFII structure, a generalized DFII structure, denoted as ρ DFII, is derived. The equivalent state-space realization of the ρ DFII structure, called ρ -realization, is also obtained. These two structures have a pre-determined sparse form and are very efficient in terms of implementation. More interestingly, each of the two structures contains $2K$ free parameters, with which the structure can be optimized to achieve nice stability performance. The performance analysis and optimization procedures for both structures are elaborated in Chapter 4.

When we analyze the stability behavior of the ρ -operator-based structures (the ρ DFII structure and ρ -realization), the expression for the stability-related measure is specified in terms of the free parameters and the system poles. It is found that when the polynomials are chosen such that their roots are close to the system poles, the pole sensitivities can be much reduced and hence the stability robustness increases. This expression gives some insights to help search for the optimal sparse structures in a more efficient way. Meanwhile, it also reflects a limitation of the ρ -operator-based structures, that is, the polynomials using ρ -operators have real roots only while the system poles are generally complex. For those systems whose poles are far away from the real axis, the ρ -operator-based structures would not be competent. Then a question comes out of our mind: can the ρ -operator be extended to a more generalized operator which leads to a set of polynomials having complex roots? If the answer is yes, we conjecture that the structures based on the new generalized operator may yield further better FWL performance.

In the subsequent research, our conjecture turns into reality. A set of second-order operators are developed in Chapter 5, with which two new efficient structures are obtained. One of them can be considered as an improved version of the ρ DFIIt structure proposed in Chapter 4 and is denoted as ρ IDFIIt for convenience. The other one is the equivalent state-space realization of the ρ IDFIIt structure, called ρ_I -realization. Compared with the ρ DFIIt structure, the ρ IDFIIt structure has some more free parameters, which can be utilized to minimize the FWL effects. The problems of finding optimized sparse structures are investigated with stability consideration. It is shown that the stability robustness for the ρ IDFIIt structure and ρ_I -realization can be much enhanced when the roots of the corresponding polynomials are close to the dominant system poles. With the extra degree of freedom introduced, the new set of polynomials can have any assigned roots (both the real roots and complex ones) and consequently, the stability robustness of the second-order operator-based structures can be further maximized.

The sparseness issue for optimal controller structure design with minimum closed-loop roundoff noise gain seems to be a totally new topic. As we mentioned before, roundoff noise analysis received much less attention in the control system design than the design of digital filters. However, since there exist some parallels between filter and controller implementation, some of the existing concepts on filter design can be applied to control problems with necessary adaptations. Chapter 6 presents a study on the effect of roundoff noise in a digital controller for a discrete-time control system. It has been shown in Chapter 5 that the second-order operator-based structures (the ρ IDFIIt structure and ρ_I -realization) can be optimized to exhibit very nice stability behavior. Now the question is whether these two structures can be optimized to achieve minimum

roundoff noise gain performance. The answer to this question constitutes the major objective of Chapter 6 of this thesis. In Chapter 6, only investigations on the optimal roundoff noise control structures are provided for closed-loop systems, since similar results can be obtained for digital filters by simplifying the analysis for the closed-loop system into that for the open-loop case.

Before ending this section, we turn back to the stability issue of the digital systems. The classical stability robustness measure μ_0 defined in [52], though best quantifying the FWL stability character of a system realization, has not been computed explicitly so far. The prevailing pole sensitivity-based measures [54] - [58], including the one proposed in Chapter 3, are the alternative tractable measures to approximate μ_0 . It should be noted that the derivation of all these measures is based on a first-order approximation, where the pole deviation is proportional to the pole sensitivity. We shall, in Chapter 7, analyze the pole deviation based on a second-order perturbation theory. A new measure is then defined to evaluate the system stability, which is a very good approximation of μ_0 and is more accurate than the approach based on the first-order approximation. Second-order pole sensitivity study provides us a potential space to investigate the stability issue of digital systems. Developing new tractable stability measure using this approach and applying the measure into the optimal structure design are promising future works to be accomplished.

Investigation on the optimal digital system structures with maximum stability robustness or minimum roundoff noise gain is the core work of this thesis. The two FWL criteria, say the stability robustness and the roundoff noise gain, are investigated with respect to the two main FWL effects, the parameter perturbation and the roundoff noise, respectively. There exist other types of FWL

effects, such as limit cycles and overflow, which were well studied in [76] - [84]. For a comprehensive understanding the framework of finite word length effects, one can refer to [85].

The design of efficient digital system structures with high performance finds a lot of applications in many different areas, such as military, automotive, consumer electronic, medical and safety critical systems. For example, for a mass-produced product such as a portable CD player, the implementation cost is a critical issue. Thus it is very important to reduce the word length, memory requirements and the number of arithmetic operations. This may be achieved by adopting efficient digital filter structures with high performance for implementations. In [86], the controller structures with sparse coefficient realizations were applied to the area of telesurgery control systems. This seminal work addresses the need for amalgamation of the finite precision efficient controller studies in biomedical control systems and in particular for different telemedical applications that require mobility in their control system hardware together with the necessary good performance and stability margins. With the recent advances in the hardware design of the relevant processing architectures, the future remains open for more research in developing efficient digital system structures to fulfill the high performance, better stability, and cost-effective requirements of many important applications.

1.3 Major contributions

The contribution in this thesis covers a series of optimal FWL structure problems for digital filters and controllers with stability robustness and roundoff

noise consideration.

Specific research contributions of this thesis include:

- A new stability-related measure has been proposed. The corresponding optimal FWL state-space realization problem in terms of maximizing this measure has been solved analytically for digital filter case. As to the digital controller case, the expression of the measure for the closed-loop system has been achieved and the corresponding optimal structure problem has been attacked using a standard optimization algorithm.
- Based on a polynomial operator approach, two sparse structures have been derived. With the newly proposed stability measure, the stability behavior of each structure has been analyzed. How to choose the free parameters in the structures to enhance the stability robustness has been discussed. Each of the two sparse structures has been optimized to achieve a very nice stability performance, which is very close to (in some cases, better than) that of the fully parametrized optimal realizations.
- Two improved sparse structures have been developed with a set of second-order polynomial operators. These structures have some more free parameters, which can be utilized to further maximize the stability robustness, while the price paid for that is only a few more nontrivial parameters. The corresponding optimized sparse structures exhibit excellent stability performance and computation efficiency.
- The effect of roundoff noise in a digital controller has been investigated for a discrete-time control system. The optimal roundoff noise structure problems have been solved for both fully parametrized realizations and the

improved sparse structures. It has been shown that the optimized sparse structures beat the fully parametrized optimal realizations in terms of both roundoff noise gain and implementation efficiency.

- The pole deviation analysis has been conducted based on a second-order perturbation theory. A new measure has been defined to evaluate the system stability, which is a very good approximation of the classical stability measure and is more accurate than the approach based on the first-order approximation.

1.4 Outline of the thesis

In this thesis, the optimal FWL structure problems are investigated focusing on the digital system implementation in terms of the stability robustness and the roundoff noise gain. The outline of this thesis is given as follows.

We start in Chapter 2 with preliminaries that consist of some fundamentals in system parametrizations and structures, and a few classical results of commonly used criteria for optimal FWL structure design, including two stability robustness measures for digital systems, and a roundoff noise gain for state-space digital filters.

The next three chapters present the main results of our studies on the optimal FWL structure problem with stability consideration.

In Chapter 3, a new stability robustness-related measure is derived based on a pole modulus sensitivity approach. The optimal fully parametrized realization problems are solved in terms of maximizing our proposed measure for digital filters and controllers, respectively. Chapter 4 is devoted to exploiting

sparse structures with high stability performance. Based on the polynomial parametrization concept, two efficient structures are obtained by reparametrizing the transfer function with a set of special polynomial operators, called ρ -operators. Each of the proposed structures has a pre-determined sparse form and contains some free parameters with which the structure can be optimized to achieve nice stability performance. The stability behavior of each structure is analyzed and the optimal structure problem is solved with practical considerations using a genetic algorithm. In Chapter 5, we conduct further research for sparse structures with higher degree of stability robustness. Two improved sparse structures are developed with a set of second-order operators. These structures have some more free parameters, which can be utilized to further maximize the stability robustness, while the price paid for that is very slight. Through these chapters, numerical examples for digital filters and controllers are given to illustrate the design procedure and to compare the stability performance of different structures. Simulations are also performed to confirm our theoretical results.

The subject of Chapter 6 is to investigate the effect of roundoff noise in a digital controller for a discrete-time control system. An analytical expression of the roundoff noise gain is obtained for state-space controller realizations, with which the problem of identifying minimum roundoff noise structures is solved on the set of fully parametrized realizations. The performance of the second-order operator-based structures proposed in Chapter 5 is analyzed by deriving the corresponding expressions of roundoff noise gain and the problems of finding optimized sparse structures are solved. Through a numerical example and the simulations, the excellent performance of the second-order operator-based sparse structures is demonstrated in terms of reducing roundoff noise and

implementation complexity.

Chapter 7 takes up some further studies on the pole deviation analysis based on a second-order perturbation theory. A new measure is then defined to evaluate the system stability, which is a very good approximation of the unsolved classical stability measure and is more accurate than the approach based on the first-order approximation. This part of research is just a beginning and detailed framework needs further investigation.

Finally in Chapter 8, some concluding remarks are given and possible directions for future work are discussed.

Chapter 2

Preliminaries

This chapter is devoted to providing some preliminaries that are needed in the sequel of the thesis, including some fundamentals in system parametrizations and structures, and a few classical results of commonly used criteria for optimal FWL structure design, such as two stability robustness measures for digital systems, and a roundoff noise gain for state-space digital filters.

2.1 Digital systems, parametrizations and structures

Consider the following discrete time-invariant linear system $H(z)$, represented by its transfer function

$$H(z) = \frac{\sum_{k=0}^K \zeta_k z^{K-k}}{z^K + \sum_{k=1}^K \xi_k z^{K-k}} \triangleq \frac{N(z)}{D(z)}. \quad (2.1)$$

This system can be implemented with its input-output relationship:

$$y(n) = - \sum_{k=1}^K \xi_k y(n-k) + \sum_{k=0}^K \zeta_k u(n-k), \quad (2.2)$$

where $u(n)$ and $y(n)$ are the input and output signals of the system, respectively. (2.2) is usually referred as a general description of the direct-form structures¹, such as direct-form I (DFI) structure and direct-form II (DFII) structure. The DFII structure is more efficient than the DFI in terms of memory requirements and is used extensively in practical applications. The block diagram of the DFII structure is shown in Figure 2.1.

Figure 2.1: Block diagram of the direct-form II structure

The transfer function and the direct-form structures are characterized by

¹A *structure* means a way that the output of the system is computed with an input signal given.

the following parameter vector formed with coefficients of the numerator $N(z)$ and denominator $D(z)$ of $H(z)$:

$$\theta_d \triangleq \begin{bmatrix} \xi_1 & \xi_2 & \cdots & \xi_K & \zeta_0 & \zeta_1 & \cdots & \zeta_K \end{bmatrix}^T, \quad (2.3)$$

where $\mathcal{T}$ denotes the transpose operator. θ_d is called the *parametrization* of $H(z)$ in the shift operator z .

It is well known that a linear system can be characterized with different parametrizations and that for a given parametrization, it can be implemented with many different structures. Different parametrizations/structures, though theoretically equivalent, have different numerical properties.

Besides the direct-forms, another class of structures is the so-called *state-space realizations*:

$$\begin{cases} x(n+1) = Ax(n) + Bu(n) \\ y(n) = Cx(n) + du(n), \end{cases} \quad (2.4)$$

where $x(n) \in \mathcal{R}^{K \times 1}$ is the state variable vector, and $R \triangleq (A, B, C, d)$ is called a realization of $H(z)$ with $A = \{a_{ij}\} \in \mathcal{R}^{K \times K}$, $B = \{b_i\} \in \mathcal{R}^{K \times 1}$, $C = \{c_j\} \in \mathcal{R}^{1 \times K}$ and $d \in \mathcal{R}$, satisfying

$$H(z) = d + C(zI - A)^{-1}B. \quad (2.5)$$

We refer to (2.4) as the linear time-invariant state-space model, which can be represented by the simple block diagram in Figure 2.2.

A realization that can be obtained directly from the parametrization θ_d is

Figure 2.2: General state-space description of a linear time-invariant system

$R_c \triangleq (A_c, B_c, C_c, d)$, called *canonical* realization²:

$$\left\{ \begin{array}{l} A_c = \begin{bmatrix} -\xi_1 & 1 & 0 & 0 & \dots & 0 \\ -\xi_2 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & \vdots & \vdots & \\ -\xi_{K-1} & 0 & 0 & 0 & \dots & 1 \\ -\xi_K & 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \\ B_c = \begin{bmatrix} b_1 & b_2 & b_3 & \dots & b_K \end{bmatrix}^T, \quad b_k = \zeta_k - \zeta_0 \xi_k, \quad \forall k, \\ C_c = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad d = \zeta_0. \end{array} \right. \quad (2.6)$$

(2.5) means that the transfer function is parametrized with the realization (A, B, C, d) . It is interesting to note that there exists an infinite set of realizations. In fact, denote S_H as the set of all the realizations:

$$S_H \triangleq \{(A, B, C, d) : H(z) = d + C(zI - A)^{-1}B\}. \quad (2.7)$$

²In this thesis, we regard the realization (2.6) as a typical canonical realization popularly used in digital system implementation. This realization is taken as initial realization for the optimization procedures in the following chapters. There are other canonical realizations (See, e.g. [87]) which can also be used in a similar way.

It can be shown that S_H can be characterized by

$$A = T^{-1}A_0T \quad B = T^{-1}B_0 \quad C = C_0T \quad (2.8)$$

where (A_0, B_0, C_0, d) is an initial realization and $T \in \mathcal{R}^{K \times K}$ is *any* nonsingular matrix.

As mentioned above, a digital system can be implemented with many different structures, such as the direct forms and state-space realizations, which are totally equivalent in infinite precision implementation. In practice, the designed systems have to be implemented with a digital device, which is always of finite precision, and consequently, the actual performance of the system may degrade greatly due to the FWL effects. The key point is that different structures of the same system yield different performance when implemented with finite word length. The optimal FWL structure design is to search for those structures that optimize a certain FWL performance criterion (or measure). The most commonly used performance criteria are the stability-related measure and the roundoff noise gain. Now we present some earlier classical stability measures and roundoff noise gain, as well as the available results concerning their optimization. We start with several stability measures.

2.2 Stability measures

In this thesis, we consider two types of systems: digital filters and digital controllers. Usually, a digital filter is referred as an *open loop* system, while the digital controller is used to achieve a certain performance of the *closed-loop* system by controlling the plant.

For open-loop systems (digital filters), $H(z)$ is said stable if its poles, i.e., the eigenvalues of A (of a realization), are all inside the unit circle $|z| = 1$. Parallely, let us consider the discrete-time feedback system depicted in Figure 2.3, where $P(z)$ is the discrete-time plant to be controlled by the digital controller $H(z)$.

Figure 2.3: Block diagram of a discrete-time feedback system

Suppose the plant $P(z)$ has a realization (A_p, B_p, C_p, d_p) and is strictly proper, which means that $d_p = 0$. Let the controller $H(z)$ be implemented with (2.4) and $(\bar{A}, \bar{B}, \bar{C}, 0)$ be the corresponding realization of the closed-loop system (with $r(n)$ and $w(n)$ as input and output, respectively). It can be shown that

$$\begin{aligned}
 \bar{A} &= \begin{bmatrix} A_p + dB_pC_p & B_pC \\ BC_p & A \end{bmatrix} \\
 &= \begin{bmatrix} A_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} B_p & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} d & C \\ B & A \end{bmatrix} \begin{bmatrix} C_p & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \\
 &\triangleq M_0 + M_1 X M_2 \triangleq \bar{A}(X),
 \end{aligned} \tag{2.9}$$

where X is the system matrix of the controller $H(z)$ and all the $\mathbf{0}$ s / I s are zero / identity matrices of appropriate dimension. The digital controller has to be designed such that the closed-loop system is stable, which means that the

eigenvalues of $\bar{A}$, denoted by $\{\lambda_k = \lambda_k(\bar{A})\}$, are all inside the unit circle $|z| = 1$.

For open loop systems (digital filters), we define $\bar{A} = X = A$. In the sequel, $X = A$ when $H(z)$ is a filter and X is given by (2.9) when $H(z)$ is a controller.

2.2.1 A rigorous stability measure

In real-time applications, the well-designed system has to be implemented in a given structure, which can be either a direct-form or a state-space realization, and is denoted as S_τ in general. Let $\{\tau_k\}$ be the set of the parameters contained in the structure S_τ , which need to be implemented with some arithmetic. In this thesis, we are confined to *fixed-point* implementations for which the FWL effects are more serious. There exist other implementation schemes such as the *floating-point* based ones (see [57], [88]) but implementations with a fixed-point arithmetic have advantages such as lower cost and higher speed.

For a fixed-point implementation of B_τ bits, it is desired that the parameters to be implemented in S_τ be all absolutely not bigger than one, otherwise, a proper scaling is needed, and that every parameter in S_τ has to be truncated or rounded into a B_τ -bit format except if it is equal to ± 1 . Therefore, a parameter τ_k will produce an FWL error unless it belongs to the following space

$$S_{FWL} \triangleq \{-1, 1\} \cup \left\{ \pm \sum_{m=1}^{B_\tau} c_m 2^{-m}, c_m = 0, 1, \forall m \right\}, \quad (2.10)$$

which is a *discrete space* containing $2^{B_\tau+1} + 1$ elements.

With the FWL errors, the actually implemented parameters are $\{\tau_k + \Delta\tau_k\}$

instead of the ideal $\{\tau_k\}$, where all $\Delta\tau_k$ are bounded by $\epsilon/2$,

$$\eta_\tau \triangleq \max_k |\Delta\tau_k| \leq \epsilon/2 \quad (2.11)$$

with $\epsilon \triangleq 2^{-B_\tau}$. Therefore, $\{\Delta\tau_k\}$ may make the system poles outside the unit circle, causing instability of the filter or the closed-loop system. The stability behavior of the structure S_τ can be evaluated with the following stability robustness measure [52]:

$$\begin{aligned} \mu_0(S_\tau) \triangleq \inf\{\eta_\tau : \text{there exist } \{\Delta\tau_k\} \text{ in (2.11) that make} \\ \text{the filter or closed-loop system unstable}\}. \end{aligned} \quad (2.12)$$

From the viewpoint of definition, this measure may be the best to quantify the FWL stability character of a given structure. However, computing explicitly the value for $\mu_0(S_\tau)$ seems very hard, and is still an open problem. This stability robustness is also a function of structure S_τ . The optimal structure problem is to find those structures that maximize $\mu_0(S_\tau)$, which is of great importance. In fact, the digital system implemented with such an optimal structure implies that it has the best stability robustness against the FWL errors and can be implemented with a low cost. Like the well-known *real stability radius* [89], $\mu_0(S_\tau)$ is not a tractable function of S_τ and the corresponding optimal structure problem can not be solved so far.

2.2.2 A practical pole sensitivity-based measure

Noting the difficulty in computing $\mu_0(S_\tau)$ and the intractability of this measure, several researchers proposed various tractable pole sensitivity-based stability

measures to replace $\mu_0(S_\tau)$ in some senses [54] - [58]. In what follows, let us review an important pole sensitivity-based measure, which was derived for the state-space controller realizations using the l_2 -norm. For state-space controller realizations, S_τ is referred to the system matrix X of the controller, defined in (2.9).

Let $\lambda_k = \lambda_k(\bar{A})$ denote the k th eigenvalue of $\bar{A}$. In [54], based on a *first-order approximation*, i.e., sensitivity approach, the following measure was proposed:

$$\tilde{\mu}_2(X) \triangleq \min_k \frac{1 - |\lambda_k|}{\sqrt{N}\Psi_k}, \quad (2.13)$$

where N is the number of non-zero elements in $\Delta X \triangleq \{\Delta x_{ij}\}$ with Δx_{ij} the perturbation of the (i, j) th element of X , and Ψ_k is called the *partial pole sensitivity measure*:

$$\Psi_k \triangleq \sum_{i,j=1} \delta(x_{ij}) \left| \frac{\partial \lambda_k}{\partial x_{ij}} \right|^2, \quad (2.14)$$

where $\delta(x_{ij}) = 0$ for $x_{ij} \in S_{FWL}$, otherwise $\delta(x_{ij}) = 1$, which is used to take care of the structure details.

The detailed derivation of the above measure can be found in [54]. For fully parametrized realizations, we have $\delta(x_{ij}) = 1$ and $N = (1 + K)^2$ with K the order of the controller. In that case, it has been shown that $\Psi_k \geq \alpha > 0$, where α is a constant. See Chapter 8 by Li in [51] for details. Then it follows from (2.13) that $0 < \tilde{\mu}_2 \leq \frac{1 - \max_k |\lambda_k|}{(1+K)\sqrt{\alpha}}$. From this upper bound of $\tilde{\mu}_2$, one can see that when the poles λ_k are close to the unit circle or the order of the controller is high, the closed-loop stability robustness will decrease.

In order to evaluate $\tilde{\mu}_2(X)$, one needs the pole sensitivities that can be computed with the following technical lemma [54] and Theorem 1.

Lemma 1. Let $f(M) \in C$ be a differentiable function of a matrix $M \triangleq \{m_{ij}\}$ and $M_k, \forall k$ be constant matrices. Let $M = g(X)$, where X is a matrix of appropriate dimension. Denote $\frac{\partial f(M)}{\partial M}$ as the sensitivity matrix of $f(M)$ with respect to M , whose (i, j) th element is $\{\frac{\partial f(M)}{\partial m_{ij}}\}$, and $F(X) = f(g(X))$.

- If $M = M_0 + M_1 X M_2$, then

$$\frac{\partial F(X)}{\partial X} = M_1^T \frac{\partial f(M)}{\partial M} M_2^T. \quad (2.15)$$

- If $M = M_0 + M_1 X^{-1} M_2$, then

$$\frac{\partial F(X)}{\partial X} = -(M_1 X^{-1})^T \frac{\partial f(M)}{\partial M} (X^{-1} M_2)^T. \quad (2.16)$$

Applying the above technical lemma to λ_k , which is a function of $\bar{A} = M_0 + M_1 X M_2$ defined by (2.9), we can now compute the k th pole sensitivity of the closed-loop system with respect to the system matrix X :

$$\frac{\partial \lambda_k}{\partial X} = M_1^T \frac{\partial \lambda_k}{\partial \bar{A}} M_2^T, \quad (2.17)$$

where $\frac{\partial \lambda_k}{\partial \bar{A}}$ can be computed using the following theorem, which can be found from many textbooks (see [90], [91]).

Theorem 1. Let $\bar{A} \in \mathcal{R}^{m \times m}$ be a matrix having m linearly independent eigenvectors and $\{\lambda_k\} = \lambda(\bar{A})$ be its eigenvalues. Denote $M_x \triangleq \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix}$, where x_k is a right eigenvector of $\bar{A}$ corresponding to the eigenvalue λ_k , and $M_y = \begin{bmatrix} y_1 & y_2 & \dots & y_m \end{bmatrix} \triangleq M_x^{-\mathcal{H}}$. Then

$$\left(\frac{\partial \lambda_k}{\partial \bar{A}}\right)^T = x_k y_k^{\mathcal{H}}, \quad k = 1, \dots, m, \quad (2.18)$$

where y_k is called the reciprocal left eigenvector corresponding to x_k and $\mathcal{H}$ denotes the transpose and conjugate operation.

In Theorem 1, $\bar{A}$ is assumed to be non-defective. If $\bar{A}$ is defective (e.g. having repeated poles), the sensitivity of the defective eigenvalue is infinite [91], [92]. That is why such a situation should be avoided in system design, especially when a FWL implementation has to be carried out. In the sequel, $\bar{A}$ is assumed to be non-defective.

It is interesting to note that different realizations (leading to different X) yield different $\tilde{\mu}_2$. The optimal realization problem is to solve the following optimization

$$X_{opt} : \max_X \tilde{\mu}_2(X). \quad (2.19)$$

This problem was well solved for a discrete-time control system in [54].

2.3 Roundoff noise gain

In digital system implementation, when the signals processed by the system are encoded in finite word length, roundoff must occur at every arithmetic operation and the input-output behavior of the system is different from what it would be in infinite precision. The error caused by the roundoff operation is called roundoff noise. Usually, two implementations exist depending on whether the signals are rounded off before or after multiplication. Assuming rounding occurs after multiplication (RAM), a variable, say x , computed with a multiplication, has to be replaced by its quantized version, denoted as $Q[x]$, in the

ideal computation model. The difference $Q[x] - x$ is the corresponding round-off noise, which is usually modelled as a white noise sequence and statistically independent of those produced by other sources [7] - [9]. Besides the signal-plus-white-noise technique, some tools such as the ergodic theory of dynamic systems have been developed to analyze the long-term statistical behavior of the closed-loop trajectories [93], where the long-term behavior is quite complicated and is not assessed accurately by standard difference equation stability theory or by signal-plus-noise analysis. In this thesis, white noise models are adopted, which, as to be seen later, is convenient for us to use the effective l_2 -scaling scheme in designing structures for digital system implementation.

Let (A, B, C, d) be a realization used in (2.5). W_c , W_o are the controllability and observability Gramians of this realization, respectively, which are the solutions of the following Lyapunov equations

$$W_c = AW_cA^T + BB^T, \quad W_o = A^TW_oA + C^TC.$$

When the filter is implemented with (2.4), it can be shown in [7] - [9] that the variance of the output error due to the roundoff of the states is given by

$$\sigma_{\Delta y}^2 = [tr(W_oQ) + m_{K+1}]\sigma_0^2$$

where $tr(\cdot)$ denotes the trace operation, σ_0^2 is a constant, determined by the word-length used for representing the states, $Q = diag(m_1, \dots, m_k, \dots, m_K)$ with m_k the number of nontrivial parameters in the k th row of $[A \ B]$ for $k = 1, \dots, K$, and m_{K+1} is the number of nontrivial parameters in the row vector $[C \ d]$. Then

the roundoff noise gain for a state-space digital filter is defined as:

$$G \triangleq \frac{\sigma_{\Delta y}^2}{\sigma_0^2} = \text{tr}(W_o Q) + m_{K+1}. \quad (2.20)$$

For a detailed derivation and computation of the above gain, please refer to [7] - [9].

It is well known (see, e.g., [6], [69]) that in an implementation system, all the signals should be sustained within a certain dynamic range in order to avoid overflow. Under the assumption that the input $u(n)$ and the output $y(n)$ of the filter are properly pre-scaled, the only signals which may have overflow are the elements of the state vector $x(n)$, which, therefore, have to be scaled. There exist different scaling schemes for preventing variables from overflow. The popularly used ones are the l_2 - and l_∞ -scalings. In what follows, (when the roundoff noise is analyzed for a digital system), we will concentrate on the l_2 -scaling scheme.

The l_2 -scaling means that each state variable should have a unit variance when the input is a white noise with a unit variance. This can be achieved if (see [8], [9])

$$W_c(i, i) = 1, \quad \forall i. \quad (2.21)$$

Equation (2.21) defines a subset, denoted by $S_H^{l_2}$, of the realization set S_H defined in (2.7). The optimal roundoff noise realization problem is to find those realizations in $S_H^{l_2}$ that minimize G given by (2.20)

$$R_{opt} : \min_{R \in S_H^{l_2}} G. \quad (2.22)$$

This problem seems very difficult and is still an open problem, due to the fact

that Q and m_{K+1} depend on the number of nontrivial elements in the realization. In [60], a numerical algorithm was proposed to solve a suboptimal problem. We note that if the realization is fully parametrized (with all parameters nontrivial), that is $m_k = K + 1, \forall k$, one has

$$G = [\text{tr}(W_o) + 1](K + 1). \quad (2.23)$$

With G given by (2.23), the corresponding optimal realization problem (2.22) was solved in [8] and [9] independently.

2.4 Summary

In this chapter, we have introduced the concept of system parametrizations and structures in the traditional shift operator, and have reviewed some existing results on optimal FWL structure design with several FWL performance criteria, such as the stability robustness maximization and roundoff noise gain minimization. Our attention has been focused on a practical pole sensitivity-based stability measure for digital systems, and a well-accepted roundoff noise gain for state-space filter realizations. In the remainder of this thesis we will extend these classical techniques to a range of new FWL performance-related measures and consider not only the set of the state-space realizations but also the direct-form II transposed (DFIIt) structures. The investigations on the optimal FWL design problems are not just for digital filters but for control problems as well. Now it is time to get into some new materials. Let us move on to the next chapter, where a new stability-related measure will be proposed.

Chapter 3

A New Stability-Related Measure

In Chapter 2, a rigorous stability robustness measure μ_0 is formulated in (2.12). However, computing the value for μ_0 is very hard and remains as an open problem. Recently, this measure has been replaced by several alternative measures which are approximation of μ_0 [54] - [57]. In this chapter, a new stability-related measure is derived based on a pole modulus sensitivity approach and the corresponding optimal state-space realization problems are analyzed for digital filters and controllers, respectively.

3.1 A pole modulus sensitivity-based measure

Let $\{\lambda_k\}$ be the set of the ideal system poles, which are function of the parameters $\{\tau_m\}$ in the structure S_τ . With the FWL errors $\{\Delta\tau_m\}$, the actual poles are $\{\tilde{\lambda}_k\}$, which may differ from $\{\lambda_k\}$. To overcome the difficulty of tractability,

a pole sensitivity-based measure, denoted as $\tilde{\mu}_2$, was derived in [54]. This measure is based on the fact that $\tilde{\lambda}_k$ is inside the unit circle if $|\tilde{\lambda}_k - \lambda_k| < 1 - |\lambda_k|$. However, from the triangle inequality $|\tilde{\lambda}_k| - |\lambda_k| \leq |\tilde{\lambda}_k - \lambda_k|$, one can see that $|\tilde{\lambda}_k - \lambda_k| > 1 - |\lambda_k|$ does not necessarily imply $|\tilde{\lambda}_k| - |\lambda_k| > 1 - |\lambda_k|$ and hence $|\tilde{\lambda}_k| > 1$. In fact, whether $\tilde{\lambda}_k$ is inside $|z| = 1$ or not is more related to $|\tilde{\lambda}_k| - |\lambda_k|$ than $\tilde{\lambda}_k - \lambda_k$.

Based on this observation and using the same procedure as in [54], one has the following pole modulus sensitivity-based stability measure:

$$\mu_2 \triangleq \min_k \frac{1 - |\lambda_k|}{\sqrt{N\Phi_k}} \quad (3.1)$$

where N is the number of those parameters in $\{\tau_m\}$, which do not belong to S_{FWL} defined before, and

$$\Phi_k \triangleq \sum_{m=1}^{N_\tau} \delta(\tau_m) \left| \frac{\partial |\lambda_k|}{\partial \tau_m} \right|^2, \quad \forall k \quad (3.2)$$

is the partial pole modulus sensitivity with N_τ the number of parameters in S_τ and $\delta(\tau_m) = 0$ if $\tau_m \in S_{FWL}$, otherwise $\delta(\tau_m) = 1$, which is used to take care of the structure details.

Denote

$$P_{\Delta\tau} \triangleq \{ \{ \Delta\tau_m \} : |\tilde{\lambda}_k| - |\lambda_k| < \eta_\tau \sqrt{N\Phi_k}, \quad \forall k \}$$

where η_τ is defined in (2.11). μ_2 has the following property over $P_{\Delta\tau}$: all the $\tilde{\lambda}_k$ are inside the unit circle if $\eta_\tau < \mu_2$ for any $\{ \Delta\tau_m \} \in P_{\Delta\tau}$. The proof of this property is as follows. First of all, one has

$$|\tilde{\lambda}_k| \leq |\lambda_k| + ||\tilde{\lambda}_k| - |\lambda_k||.$$

For any $\{\Delta\tau_m\} \in P_{\Delta\tau}$, if $\eta_\tau < \mu_2$, the above inequality yields

$$\begin{aligned} |\tilde{\lambda}_k| &\leq |\lambda_k| + \eta_\tau \sqrt{N\Phi_k} \\ &< |\lambda_k| + \mu_2 \sqrt{N\Phi_k} \\ &\leq |\lambda_k| + \frac{1 - |\lambda_k|}{\sqrt{N\Phi_k}} \sqrt{N\Phi_k} = 1, \quad \forall k. \end{aligned}$$

It should be pointed out that μ_2 , like all other first-order approximation-based stability-related measures (see, e.g., [54] - [57]), has no rigorous connection with μ_0 since it is derived based on the first-order perturbation theory. This measure, however, is reasonable for studying the system structure problem due to the fact that it can be computed easily for a given structure and as to be seen later, is tractable for optimal realization searching. It is argued in [90] that the estimates obtained from first-order perturbation theory are often more realistic than rigorous bounds obtained by other means. Since no tractable rigorous measure is available so far, this measure is in fact a good tradeoff between rigor and computational tractability.

The measure μ_2 can be computed for a given structure S_τ using (3.1) and (3.2) as long as the pole modulus sensitivities $\partial|\lambda_k|/\partial\tau_m$ are known. It can be shown that

$$\frac{\partial|\lambda_k|}{\partial\tau_m} = \frac{1}{2}(\lambda_k\lambda_k^*)^{-1/2} \left[\frac{\partial\lambda_k}{\partial\tau_m} \lambda_k^* + \frac{\partial\lambda_k^*}{\partial\tau_m} \lambda_k \right], \quad \forall k \quad (3.3)$$

where $*$ denotes the conjugate operation.

It is easy to understand that different structures yield different μ_2 . The

optimal structure problem is to solve the following optimization

$$\max_{S_\tau} \mu_2. \quad (3.4)$$

At the end of this section, we compare μ_2 with some of the existing measures. First of all, we point out that the new measure μ_2 , like other existing ones [54] - [57], is a first-order approximation of μ_0 . The closeness to μ_0 of these measures is system and structure dependent. Compared to $\tilde{\mu}_2$ proposed in [54], the improved version μ_2 is more reasonable due to the fact that μ_2 adopts the pole modulus sensitivity approach and the system stability depends only on the moduli of its eigenvalues. Besides the l_2 -norm-based measures (say $\tilde{\mu}_2, \mu_2$), there are alternative measures using l_1 -norm in defining the pole sensitivity. The pole modulus sensitivity-based measure using l_1 -norm was initially proposed in [56]. This measure has the advantage over that in [54] due to its use of the pole modulus sensitivity. As to the l_2 -norm-based measures, they have the advantage over the l_1 -norm-based measures in terms of tractability and the searching complexity for the optimal realizations, which will be shown in the next section. Therefore, the newly proposed measure μ_2 combines the advantages of two prevailing measures addressed in [54] and [56].

3.2 Optimal stability realizations

In this section, we discuss how to compute the optimal state-space realizations that maximize the stability-related measure μ_2 , and present some results in fully parametrized optimal realizations for digital filters and controllers, respectively.

It is noted that the problem defined by (3.4) is, generally speaking, extremely

hard to solve due to the structure function $\delta(\tau_m)$ involved. In what follows, we solve the problem on the set of fully parameterized state-space realizations, for which $\delta(\tau_m) = 1$, $\forall m$ and hence N in (3.1) is the total number of elements in the structure.

3.2.1 Open-loop systems - digital filters

For open-loop systems, as mentioned in Section 2.2, we have $X = A, M_0 = \mathbf{0}, M_1 = I$ and $M_2 = I$ and hence $\bar{A}(X) = A$. Under the assumption that A matrix is fully parameterized, the pole modulus sensitivity measure can be rewritten as

$$\Phi_k \triangleq \left\| \frac{\partial |\lambda_k|}{\partial A} \right\|_F^2 = \text{tr} \left[\left(\frac{\partial |\lambda_k|}{\partial A} \right)^T \frac{\partial |\lambda_k|}{\partial A} \right] \quad (3.5)$$

where $\|\cdot\|_F$ denotes the *Frobenius* norm.

It follows from (3.3) and Theorem 1 that

$$\left(\frac{\partial |\lambda_k|}{\partial A} \right)^T = \frac{1}{2|\lambda_k|} [\lambda_k^* x_k y_k^H + \lambda_k x_k^* y_k^T] \triangleq Q_k \quad (3.6)$$

where x_k and y_k are the right and left eigenvectors of A corresponding to λ_k , respectively.

Let $R \triangleq (A, B, C, d)$ be the realization obtained from an initial realization, say $R_0 \triangleq (A_0, B_0, C_0, d) \in S_H$, with similarity transformation matrix T , and x_k^0 be a right eigenvector and y_k^0 , the reciprocal left eigenvector, of A_0 , corresponding to λ_k . One then has

$$x_k = T^{-1} x_k^0, \quad y_k = T^T y_k^0$$

and hence $Q_k = T^{-1}Q_k^0T$, where Q_k^0 is given by (3.6) but corresponding to x_k^0 and y_k^0 , which leads to

$$\Phi_k = \text{tr}(P^{-1}Q_k^0PQ_k^{0T}), \quad P \triangleq TT^T. \quad (3.7)$$

This clearly shows that different realizations have different Φ_k and hence different μ_2 . Our first main result is to show that μ_2 is maximized if and only if the realizations are normal.¹

Let M_x^0 be the right eigenvector matrix of an initial realization R_0 , all the similarity transformations that convert R_0 into a normal realization are characterized with (see [43])

$$T_n = (M_x^0 \Sigma M_x^{0H})^{1/2} U, \quad (3.8)$$

where $M^{1/2} \geq 0$ denotes the square root matrix of $M \geq 0$: $M^{1/2}M^{1/2} = M$, $\Sigma > 0$ is an arbitrary diagonal matrix, and U is any orthogonal matrix. Clearly, if two matrices have the same right eigenvector matrix, then they have the same set of similarity transformations that make them normal.

In order to present our first main result, we need the following technical lemma.

Lemma 2. *Let M_x be a right eigenvector matrix of a diagonalizable $A \in \mathcal{R}^{K \times K}$ having an eigenvalue set $\lambda(A) = \{\lambda_k\}$, and M_y , the reciprocal left eigenvector matrix, all as defined in Theorem 1. Denote M_c^r and M_c^i as the real and imaginary parts of a complex valued matrix M_c , respectively. Then*

¹A realization (A, B, C, d) is called a normal realization if A is normal, i.e., $AA^T = A^T A$.

- if λ_k is complex, then the corresponding x_k and y_k satisfy

$$x_k^{rT} y_k^r = x_k^{iT} y_k^i = \frac{1}{2}, \quad x_k^{rT} y_k^i = x_k^{iT} y_k^r = 0. \quad (3.9)$$

- M_x is also a right eigenvector matrix of Q_k defined by (3.6) and the eigenvalues of Q_k are (i) $\{\frac{\lambda_k}{|\lambda_k|}, 0, \dots, 0\}$ if λ_k is real, and (ii) $\{\frac{\lambda_k}{2|\lambda_k|}, \frac{\lambda_k^*}{2|\lambda_k|}, 0, \dots, 0\}$ if λ_k is complex.

Proof. First of all, it follows from $M_x^H M_y = I$ that $x_k^H y_k = 1$, $\forall k$ and $x_l^H y_k = 0$, $\forall l \neq k$. Therefore,

$$x_k^{rT} y_k^r + x_k^{iT} y_k^i = 1, \quad x_k^{rT} y_k^i - x_k^{iT} y_k^r = 0.$$

Since A is real, there exists $l \neq k$ such that $x_l = x_k^* = x_k^r - jx_k^i$ when λ_k is complex. It then follows from $x_l^H y_k = 0$ that

$$x_k^{rT} y_k^r - x_k^{iT} y_k^i = 0, \quad x_k^{rT} y_k^i + x_k^{iT} y_k^r = 0.$$

(3.9) follows directly from these equations.

Now let us look at the second part of the lemma. It is easy to see from $M_y^H M_x = I$ that $y_k^H x_k = 1$, $\forall k$ and $y_l^H x_k = 0$, $\forall l \neq k$. This part of the lemma follows from a direct verification noting that for a real A , x_k and hence y_k can be taken real if λ_k is real and there exists a $l \neq k$ such that $y_k^* = y_l$ if λ_k is complex. $\square$

Our first main result is given in the following theorem.

Theorem 2. *For digital filters, normal realizations are the unique solutions to (3.4) on the space of fully parameterized realizations, for which*

$$\Phi_k = \begin{cases} 1, & \text{if } \lambda_k \text{ is real} \\ \frac{1}{2}, & \text{if } \lambda_k \text{ is complex} \end{cases} \quad (3.10)$$

Proof. First of all, without loss of generality let us take any normal realization as the initial realization. It is well known [94] that for a normal matrix, say A_0 , M_x^0 can be taken unitary and hence $y_k^0 = x_k^0$, $\forall k$. In order to avoid complicated notations, in the proof we drop the superscript “0”, which indicates the initial realization.

Here, the optimal stability realization problem for state-space filters can be formulated as

$$\max_{P>0} \mu_2, \quad (3.11)$$

which is equivalent to the minimization problem as below

$$\min_{P>0, k} \Phi_k. \quad (3.12)$$

First let us consider the following minimization:

$$\min_{P>0} \Phi_k. \quad (3.13)$$

It is easy to understand that if (3.13) has a solution P such that all the Φ_k can be minimized simultaneously, then P is also a solution of the problem (3.12) and hence a solution of (3.11).

In (3.13), we note that Φ_k is a differentiable function of $P > 0$. Its minimum

points should satisfy $\frac{\partial \Phi_k}{\partial P} = 0$, which leads to

$$Q_k^T P^{-1} Q_k - P^{-1} Q_k P Q_k^T P^{-1} = 0. \quad (3.14)$$

It is easy to see that the theorem can be proved if one can show that the partial pole modulus sensitivity measures $\{\Phi_k\}$ are simultaneously minimized if and only if the realizations are normal and that (3.10) holds for normal realizations.

First of all, with $y_k = x_k$ (for the reason mentioned earlier) it can be shown that

$$Q_k = \frac{\lambda_k}{|\lambda_k|} x_k x_k^T$$

if λ_k is real, and

$$Q_k = |\lambda_k|^{-1} \{ \lambda_k^r [x_k^r x_k^{rT} + x_k^i x_k^{iT}] + \lambda_k^i [x_k^i x_k^{rT} - x_k^r x_k^{iT}] \}$$

if λ_k is complex.

With some direct manipulations using Lemma 2, one can show that $Q_k Q_k^T = Q_k^T Q_k$, which implies that $P = I$ is a solution to (3.14), and $\Phi_k = \text{tr}(Q_k Q_k^T) = 1$ if λ_k is real and $\Phi_k = \frac{1}{2}$ if λ_k is complex.

Now, suppose $P > 0$ is a solution to (3.14). Denoting $\bar{Q}_k \triangleq P^{-\frac{1}{2}} Q_k P^{\frac{1}{2}}$, (3.14) can be rewritten as

$$\bar{Q}_k^T \bar{Q}_k = \bar{Q}_k \bar{Q}_k^T,$$

which means that $\bar{Q}_k$ is normal, transformed from Q_k with the similarity transformation $T_n = P^{1/2}$. According to Lemma 2 and (3.8), Q_k and A_0 have the same right eigenvector matrix and hence the same set of similarity transformations (to normality). Since $T_n = P^{1/2}$ belongs to this set, it should make

$A = P^{-\frac{1}{2}} A_0 P^{\frac{1}{2}}$ normal. That completes the proof. $\square$

Therefore, we conclude that for digital filters, normal realizations are the unique solutions to the optimal structure problem (3.4) on the set of fully parametrized state-space realizations. Next, let us see how to solve the optimal structure problem (3.4) for fully parametrized controller realizations in the closed-loop system.

3.2.2 Closed-loop systems - digital controllers

For control system, $\bar{A}(X) = M_0 + M_1 X M_2$ where X is the system matrix of the digital controller to be implemented. When X has no parameters in S_{FWL} , $N = K^2 + 2K + 1 \triangleq N_c$ and it then follows from (3.2) that

$$\Phi_k = \left\| \frac{\partial |\lambda_k|}{\partial A} \right\|_F^2 + \left\| \frac{\partial |\lambda_k|}{\partial B} \right\|_F^2 + \left\| \frac{\partial |\lambda_k|}{\partial C} \right\|_F^2 + \left\| \frac{\partial |\lambda_k|}{\partial d} \right\|_F^2. \quad (3.15)$$

Let $x_k = \begin{bmatrix} x_{1k}^T & x_{2k}^T \end{bmatrix}^T$ be a right eigenvector of $\bar{A}$ for pole λ_k and $y_k = \begin{bmatrix} y_{1k}^T & y_{2k}^T \end{bmatrix}^T$ be the corresponding left eigenvector with the partitions corresponding to the block partitioned structure of $\bar{A}$ in (2.9). It follows from (2.17) and (2.18) with M_1 and M_2 defined in (2.9) that

$$\begin{cases} \left(\frac{\partial \lambda_k}{\partial A} \right)^T = x_{2k} y_{2k}^H & \left(\frac{\partial \lambda_k}{\partial B} \right)^T = C_p x_{1k} y_{2k}^H \\ \left(\frac{\partial \lambda_k}{\partial C} \right)^T = x_{2k} y_{1k}^H B_p & \left(\frac{\partial \lambda_k}{\partial d} \right)^T = C_p x_{1k} y_{1k}^H B_p. \end{cases} \quad (3.16)$$

With (2.8), the closed-loop transition matrix $\bar{A}$ given by (2.9) can be ex-

pressed as

$$\bar{A} = \bar{T}^{-1} \bar{A}_0 \bar{T}, \quad \bar{T} \triangleq \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}, \quad (3.17)$$

where $\bar{A}_0$ is the closed-loop transition matrix corresponding to the initial realization R_0 .

Now, let $x_k^0 = \begin{bmatrix} x_{1k}^{0T} & x_{2k}^{0T} \end{bmatrix}^T$ be a right eigenvector of $\bar{A}_0$ corresponding to the eigenvalue λ_k , and let $y_k^0 = \begin{bmatrix} y_{1k}^{0T} & y_{2k}^{0T} \end{bmatrix}^T$ be the corresponding left eigenvector. It is easy to see that

$$x_{1k} = x_{1k}^0, \quad y_{1k} = y_{1k}^0, \quad x_{2k} = T^{-1} x_{2k}^0, \quad y_{2k} = T^T y_{2k}^0.$$

With some manipulations, one has

$$\Phi_k = \text{tr}(P^{-1} Q_k^a P Q_k^{aT}) + \text{tr}(Q_k^{bT} Q_k^b P) + \text{tr}(P^{-1} Q_k^c Q_k^{cT}) + \text{tr}(Q_k^d Q_k^{dT}), \quad (3.18)$$

where $P = T T^T$, and

$$\begin{cases} Q_k^a = [\frac{\lambda_k^*}{|\lambda_k|} x_{2k}^0 y_{2k}^{0H}]^r, & Q_k^b = [\frac{\lambda_k^*}{|\lambda_k|} C_p x_{1k}^0 y_{2k}^{0H}]^r \\ Q_k^c = [\frac{\lambda_k^*}{|\lambda_k|} x_{2k}^0 y_{1k}^{0H} B_p]^r, & Q_k^d = [\frac{\lambda_k^*}{|\lambda_k|} C_p x_{1k}^0 y_{1k}^{0H} B_p]^r \end{cases} \quad (3.19)$$

with M_c^r , as defined before, denoting the real part of a complex valued matrix M_c .

Therefore, Φ_k and hence μ_2 are a well-behaved function of P . The optimal controller structure problem is to maximize μ_2 with respect to P , which is

equivalent to

$$\min_{P>0} \max_k \frac{N_c}{(1 - |\lambda_k|)^2} \Phi_k. \quad (3.20)$$

This is a highly nonlinear optimization problem and no analytical solutions have been found so far. This problem, however, can be attacked with one of the existing standard numerical algorithms. Here, we adopt the algorithm *fminimax.m* in MATLAB optimization toolbox, which uses a sequential quadratic programming method and is designed to solve the problems of the same type as (3.20). Generally speaking, this algorithm yields a local minimum. The convergence of this algorithm can be speeded up with the following gradient function [95]

$$\begin{aligned} \frac{\partial \Phi_k}{\partial P} = & Q_k^{aT} P^{-1} Q_k^a - P^{-1} Q_k^a P Q_k^{aT} P^{-1} + Q_k^{bT} Q_k^b \\ & - P^{-1} Q_k^c Q_k^{cT} P^{-1}. \end{aligned} \quad (3.21)$$

Suppose P_{opt} is a solution obtained with the algorithm. The corresponding optimal similarity transformation matrices can be characterized as

$$T_{opt} = P_{opt}^{1/2} V, \quad (3.22)$$

where V is an arbitrary orthogonal matrix. The optimal controller realization, denoted by R_f^{opt} , can then be obtained with (2.8) where $T = T_{opt}$.

3.3 Numerical examples and simulations

In this section, we illustrate our theoretical results and the design procedure with two numerical examples. First we present an example to show the optimal

realization design for digital filters.

Example 3.1: This is a low-pass fourth-order Butterworth filter, generated with MATLAB command $[V_q, V_p] = \text{butter}(4, 0.05)$, where V_p and V_q are the coefficient vectors of the denominator and numerator of the transfer function, respectively. The corresponding poles are located at $\lambda_{1,2} = 0.8630 \pm j0.0523$, $\lambda_{3,4} = 0.9319 \pm j0.1364$ with $|\lambda_{1,2}| = 0.8646$, $|\lambda_{3,4}| = 0.9418$.

The filter can be described in the following realization, denoted as $R_0 \triangleq (A_0, B_0, C_0, d)$:

$$\begin{aligned} A_0 &= \begin{bmatrix} 3.5897 & 1 & 0 & 0 \\ -4.8513 & 0 & 1 & 0 \\ 2.9241 & 0 & 0 & 1 \\ -0.6630 & 0 & 0 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0.2371 \\ 0.0359 \\ 0.2163 \\ 0.0105 \end{bmatrix} \times 10^{-3} \\ C_0 &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \quad d = 3.1239 \times 10^{-5}. \end{aligned}$$

R_0 is actually the canonical realization in shift operator. We point out that the numbers in the realization R_0 are presented using their first four fractional digits only. This is assumed in the sequel.

Taking R_0 as the initial realization, we can compute a similarity transformation matrix T_n with (3.8), where Σ is taken identity matrix. A fully param-

Table 3.1: Stability robustness comparison of R_0 and R_n

	$\tilde{\mu}_2$	μ_2	μ_0^*
R_0	9.37×10^{-5}	1.08×10^{-4}	1.25×10^{-4}
R_n	1.45×10^{-2}	2.06×10^{-2}	4.21×10^{-2}

eterized normal realization, denoted as R_n , is obtained:

$$\begin{aligned}
 A_n &= \begin{bmatrix} 0.9040 & 0.1076 & -0.0379 & 0.0174 \\ -0.1093 & 0.9308 & 0.0805 & -0.0015 \\ -0.0296 & -0.0813 & 0.8908 & 0.0529 \\ -0.0222 & -0.0156 & -0.0487 & 0.8642 \end{bmatrix}, \quad B_n = \begin{bmatrix} 0.4343 \\ 0.4741 \\ 0.5144 \\ 0.5545 \end{bmatrix} \\
 C_n &= \begin{bmatrix} 0.1533 & -0.3777 & 0.3134 & -0.0875 \end{bmatrix}.
 \end{aligned}$$

For both realizations, μ_2 is computed and μ_0 is estimated using simulations. The estimate of the μ_0 , denoted as μ_0^* , is determined as follows. Let θ_k be the parameter vector formed with all the coefficients that do not belong to S_{FWL} in a given structure with $k = 1, 2$ corresponding to R_0, R_n , respectively. For each θ_k we generate a series of 200,000 different perturbations, denoted as $\{\Delta\theta_k(i), i = 1, 2, \dots, 200,000\}$ using MATLAB command *randn.m*, where all the elements of $\Delta\theta_k(i)$ are bounded by a constant σ . Denote $A_k(i)$ as the A -matrix corresponding to $\theta_k + \Delta\theta_k(i)$. μ_0^* is the maximal value for σ to keep all the poles of $A_k(i)$ within the unit circle for $i = 1, 2, \dots, 200,000$. This μ_0^* , as an estimate of μ_0 by simulations, can be used to verify the validity of μ_2 .

The computation results are presented in Table 3.1.

Comment 3.1: From this example, one can see that

- for each structure μ_2 is a good approximation of μ_0^* which is an estimate of μ_0 by simulations as mentioned above,
- μ_2 is a better approximation of μ_0^* than $\tilde{\mu}_2$,
- the shift operator-based canonical realization R_0 is very poor in terms of stability robustness, while the normal realization R_n has a much better stability behavior.

Now we present an example to illustrate the design procedure for optimal controller realizations.

Example 3.2: This example was used by Chen and Francis [96] for studying the input-output stability of sampled-data systems, where it was shown that the corresponding sampled-data system is stable if the sampling frequency $f_s > 1$. With $f_s = 5$, we obtain the following transfer functions for the plant

$$P(z) = \frac{0.0018z^4 + 0.0003z^3 - 0.0163z^2 + 0.0011z + 0.0016}{z^5 - 3.7860z^4 + 6.3299z^3 - 6.0926z^2 + 3.3395z - 0.7908},$$

and for the controller

$$H(z) = 0.0460 + \frac{0.2119z^5 - 0.8081z^4 + 1.3570z^3 - 1.3108z^2 + 0.7220z - 0.1720}{z^6 - 4.2600z^5 + 8.2537z^4 - 9.4410z^3 + 6.6649z^2 - 2.6885z + 0.4709}.$$

The poles of the ideal closed-loop system are $\{0.4837 \pm j0.8557, 0.4814 \pm j0.8536, 0.9999 \pm j0.0004, 0.8397 \pm j0.1651, 0.8088 \pm j0.1203, 0.8190\}$.

Let $R_c \triangleq (A_c, B_c, C_c, d)$ be the canonical controller realization of $H(z)$, which

is given by

$$A_c = \begin{bmatrix} 4.2600 & 1 & 0 & 0 & 0 & 0 \\ -8.2537 & 0 & 1 & 0 & 0 & 0 \\ 9.4410 & 0 & 0 & 1 & 0 & 0 \\ -6.6649 & 0 & 0 & 0 & 1 & 0 \\ 2.6885 & 0 & 0 & 0 & 0 & 1 \\ -0.4709 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_c = \begin{bmatrix} 0.2119 \\ -0.8081 \\ 1.3570 \\ -1.3108 \\ 0.7220 \\ -0.1720 \end{bmatrix}$$

$$C_c = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad d = 0.0460.$$

Taking R_c as the initial structure, we solve (3.20) using the MATLAB command *fminimax.m* and eventually obtain a fully parameterized optimal controller realization, denote as R_f^{opt} :

$$A_f^{opt} = \begin{bmatrix} -5.6994 & -1.3980 & 5.8857 & -1.6080 & 0.6582 & -0.0781 \\ 9.2574 & 2.4223 & -8.1375 & 2.2637 & -1.1064 & 0.3532 \\ -3.9034 & -1.1741 & 3.2986 & -0.9488 & 0.3146 & -0.0840 \\ 3.2054 & 0.0745 & -3.8660 & 1.7748 & -0.4699 & 0.1119 \\ -2.6943 & -0.0535 & 3.7287 & -0.8162 & 1.3213 & 0.0243 \\ 12.3064 & 1.0724 & -14.1246 & 3.3280 & -1.4615 & 1.1423 \end{bmatrix}$$

$$B_f^{opt} = \begin{bmatrix} -7.1101 & 10.0534 & -3.1831 & 4.4044 & -4.1725 & 16.3281 \end{bmatrix}^T$$

$$C_f^{opt} = \begin{bmatrix} 0.0520 & 0.0192 & -0.0454 & 0.0037 & -0.0103 & 0.01130 \end{bmatrix}.$$

For each realization, μ_2 is computed and the estimate μ_0^* is obtained with 200,000 different perturbations, which are presented in Table 3.2.

Table 3.2: Stability robustness comparison of R_c and R_f^{opt}

	$\tilde{\mu}_2$	μ_2	μ_0^*
R_c	2.41×10^{-10}	3.59×10^{-8}	1.15×10^{-8}
R_f^{opt}	2.64×10^{-6}	9.20×10^{-6}	2.52×10^{-5}

Comment 3.2: From this example, it is easy to see that

- the stability performance of R_f^{opt} is much better than that of R_c ,
- μ_2 is a very good approximation of μ_0^* for both realizations and μ_2 is better than $\tilde{\mu}_2$ to approximate μ_0^* .

3.4 Summary

In this chapter, we have addressed the optimal FWL structure problem with stability consideration. Based on a pole modulus sensitivity approach, a new stability-related measure has been proposed. The corresponding optimal state-space realization problem in terms of maximizing this measure has been solved analytically for digital filter case. It is shown that a fully parametrized filter realization is optimal if and only if it is normal. As to the digital controller case, the expression for the measure of the closed-loop system has been obtained and the corresponding optimal controller realization problem has been attacked on the set of fully parametrized state-space realizations using a standard numerical algorithm. Two numerical examples have been given to confirm the validity of the proposed measure and the theoretical results achieved, which show that the optimal realizations (for both digital filter and controller cases) yield a

much better stability performance than those shift operator-based canonical realizations.

It should be pointed out that the optimal realizations obtained in this chapter are usually fully parameterized. From a practical point of view, it would be better to implement the digital system with such a structure that not only yields a large stability robustness but also possesses as few parameters to be implemented as possible. In the next chapter, two sparse structures will be derived based on a polynomial operator approach. These structures are very efficient in terms of implementation and contain some free parameters which can be used to optimize the structure performance. The optimized sparse structures even have a better stability behavior than those fully parametrized ones.

Chapter 4

Polynomial Operator-Based Sparse Structures

The optimal realizations discussed in the previous chapters are usually fully parametrized. This means that they increase the implementation complexity and slow down the arithmetic processing. In practice, a sparse structure, which not only possesses many trivial parameters but also yields very good FWL performance, is more desirable for implementation. Such a sparse structure can reduce the storage requirements and the computation time. Therefore, exploiting sparse structures with high degree of FWL performance is the main topic of this chapter.

It is well known that the direct-form II-based structures in conventional shift operator are the most efficient ones in terms of computation and memory storage and that these structures, however, have a very poor performance against the FWL effects in many situations [6], [7]. Recently, the direct forms in delta operator have been studied by researchers (see [64] - [69]) to improve the poor

FWL performance of those direct forms in shift operator.

The use of delta operator, defined as $\delta = \frac{z-1}{T_s}$ with T_s the sampling period, was first promoted by Peterka [70] and Middleton and Goodwin [71] in estimation and control applications. Later, the numerical properties of the delta operator, where T_s is replaced by a positive factor Δ , were investigated by Li from a pure algebraic point of view [7]. It was found that one can make the transfer function in the delta operator have better numerical properties in the case where the poles of the transfer function are closer to $z = +1$ than $z = 0$ (see, e.g., [7], [72], [73]). This means that the delta operator-based structures may not yield a satisfactory performance if some of poles of the transfer function are far away from $z = +1$.

Noting the above limitation of using delta operator, one question we ask ourselves is what can be achieved if the system is parametrized with more generalized operators. This is the motivation for us to study the generalized polynomial parametrization problem.

The main objective of this chapter is to develop a set of special polynomial operators, called ρ -operators, based on the concept of polynomial parametrization proposed in [7]. Applying the ρ -operators into the direct-form II transposed (DFIIt) structure, a generalized DFIIt structure, denoted as ρ DFIIt, is derived. The equivalent state-space realization of the ρ DFIIt structure, called ρ -realization, is also obtained. These two structures have a pre-determined sparse form and are very efficient in terms of implementation. More interestingly, each of the two structures contains some free parameters, with which the structure can be optimized to achieve very nice stability performance.

Now, we start by introducing the concept of polynomial parametrization.

4.1 Polynomial parametrization

Denoting

$$\bar{z} \triangleq \begin{bmatrix} z^K \\ \vdots \\ z \\ 1 \end{bmatrix}, \quad V_\xi \triangleq \begin{bmatrix} 1 \\ \xi_1 \\ \vdots \\ \xi_K \end{bmatrix}, \quad V_\zeta \triangleq \begin{bmatrix} \zeta_0 \\ \zeta_1 \\ \vdots \\ \zeta_K \end{bmatrix}, \quad (4.1)$$

(2.1) can be rewritten as

$$H(z) = \frac{N(z)}{D(z)} = \frac{V_\zeta^T \bar{z}}{V_\xi^T \bar{z}} = \frac{V_\zeta^T T_p^{-1} T_p \bar{z}}{V_\xi^T T_p^{-1} T_p \bar{z}} \triangleq \frac{V_\beta^T \bar{p}(z)}{V_\alpha^T \bar{p}(z)},$$

where T_p is a non-singular matrix of dimension $(K+1)$ with

$$\bar{p}(z) \triangleq \begin{bmatrix} p_0(z) & p_1(z) & \cdots & p_K(z) \end{bmatrix}^T = T_p \bar{z}$$

and

$$\begin{cases} V_\alpha \triangleq \begin{bmatrix} 1 & \alpha_1 & \cdots & \alpha_K \end{bmatrix}^T = \kappa^{-1} T_p^{-T} V_\xi \\ V_\beta \triangleq \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_K \end{bmatrix}^T = \kappa^{-1} T_p^{-T} V_\zeta \end{cases} \quad (4.2)$$

with κ a certain (normalization) constant such that $V_\alpha(1) = 1$. Therefore, one has

$$H(z) = \frac{\beta_0 p_0(z) + \beta_1 p_1(z) + \cdots + \beta_K p_K(z)}{p_0(z) + \alpha_1 p_1(z) + \cdots + \alpha_K p_K(z)}, \quad (4.3)$$

where $\{p_k(z)\}$ are called *polynomial operators*.

So, (4.3) implies that the system $H(z)$ is reparametrized with $\{\alpha_k\}$ and $\{\beta_k\}$ in the polynomial operators $\{p_k(z)\}$. It is easy to see that (4.3) returns to the transfer function in the classical shift operator z (i.e., (2.1)) with $p_k(z) = z^{K-k}$, $\forall k \in \{0, 1, 2, \dots, K\}$, and the one in delta operator with

$$p_k(z) = \left(\frac{z-1}{\Delta}\right)^{K-k}, \quad \forall k \in \{0, 1, 2, \dots, K\}.$$

The transfer function (4.3) parametrized in the polynomial operators can be implemented using the following input-output relationship¹

$$y(n) = - \sum_{k=1}^K \alpha_k \frac{p_k(z)}{p_0(z)} y(n) + \sum_{k=0}^K \beta_k \frac{p_k(z)}{p_0(z)} u(n). \quad (4.4)$$

4.2 Two sparse structures

In (4.3), we reparametrize the transfer function $H(z)$ with a set of polynomial operators $\{p_k(z)\}$. The choice of the polynomial operators depends on the applications. In [7], the problem of how to choose these polynomials was investigated for system identification and parameter estimation. In this section, we will consider a special set of the polynomial operators, based on which two sparse structures are developed for digital system implementation.

4.2.1 Generalized direct-form II transposed structures

Define

$$\rho_k(z) \triangleq \frac{z - \gamma_k}{\Delta_k}, \quad k = 1, 2, \dots, K \quad (4.5)$$

where $\{\gamma_k\}$ and $\{\Delta_k > 0\}$ are two sets of coefficients to be discussed later. Let us consider the following choice of polynomial operators:

$$p_k(z) \triangleq \prod_{m=k+1}^K \rho_m(z), \quad k = 0, 1, \dots, K-1, \quad p_K(z) \triangleq 1. \quad (4.6)$$

¹In the sequel, we use $w(z)u(n)$ to denote the output of the system with transfer function $w(z)$, when excited with the input signal $u(n)$.

The denominator and numerator of the transfer function $H(z)$ given by (2.1) can be rewritten as

$$D(z) = \kappa[p_0(z) + \sum_{i=1}^K \alpha_i p_i(z)], \quad N(z) = \kappa \sum_{i=0}^K \beta_i p_i(z) \quad (4.7)$$

with $\kappa = \prod_{k=1}^K \Delta_k$. The corresponding transformation $T_p \in \mathcal{R}^{(K+1) \times (K+1)}$ in (4.2) is an upper triangular matrix whose k th row is determined by the coefficients of the polynomial $\prod_{m=k}^K \rho_m(z)$ for $k = 1, 2, \dots, K$ and $T_p(K+1, K+1) = 1$.

It is easy to see that the output of $H(z)$ can be computed with the following equations:

$$\begin{cases} y(n) = \beta_0 u(n) + w_1(n) \\ \vdots \\ w_k(n) = \rho_k^{-1}(z)[\beta_k u(n) - \alpha_k y(n) + w_{k+1}(n)], \quad k = 2, \dots, K-1 \\ \vdots \\ w_K(n) = \rho_K^{-1}(z)[\beta_K u(n) - \alpha_K y(n)] \end{cases} \quad (4.8)$$

where $w_k(n)$ is the output of $\rho_k^{-1}(z)$ and can be computed with the structure depicted in Figure 4.1. Figure 4.2 shows the corresponding structure to (4.8) parametrized with a new parameter set $\{\alpha_k, \beta_k, \gamma_k, \Delta_k\}$, in which there are $2K$ free parameters $\{\gamma_k\}$ and $\{\Delta_k\}$. For a given digital system $H(z)$, there exists a class of such structures, depending on the space within which $\{\gamma_k\}$ and $\{\Delta_k\}$ take values. Clearly, when $\gamma_k = 0$, $\Delta_k = 1$, $\forall k$, Figure 4.2 is the conventional z DFIIt structure, while with $\gamma_k = 1$, $\forall k$, one gets the δ DFIIt structure studied in [69]. For convenience, a structure defined by Figures 4.1 and 4.2 is called a generalized DFIIt structure, denoted as ρ DFIIt.

Figure 4.1: A realization of $\rho_k^{-1}(z)$ defined in (4.5) Figure 4.2: Block diagram of the ρ DFII structure

4.2.2 Equivalent state-space realizations

With $\{x_k(n)\}$ indicated in Figure 4.1 as the state variables and $x(n)$ denoting the state vector, it can be shown that a ρ DFII structure is equivalent to the following state-space realization

$$\begin{cases} x(n+1) = A_\rho x(n) + B_\rho u(n) \\ y(n) = C_\rho x(n) + d u(n) \end{cases} \quad (4.9)$$

where

$$\left\{ \begin{array}{l} A_\rho \triangleq \begin{bmatrix} a_{11} & \Delta_2 & 0 & \cdots & 0 & 0 \\ a_{21} & \gamma_2 & \Delta_3 & \cdots & 0 & 0 \\ & & \vdots & & & \\ a_{(K-1)1} & 0 & 0 & \cdots & \gamma_{K-1} & \Delta_K \\ a_{K1} & 0 & 0 & \cdots & 0 & \gamma_K \end{bmatrix}, \quad B_\rho \triangleq \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{K-1} \\ b_K \end{bmatrix} \\ C_\rho \triangleq \begin{bmatrix} \Delta_1 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad d \triangleq \beta_0 \end{array} \right. \quad (4.10)$$

with

$$\begin{cases} a_{11} = \gamma_1 - \Delta_1 \alpha_1, & a_{k1} = -\Delta_1 \alpha_k, \quad k \in \{2, 3, \dots, K\} \\ b_k = \beta_k - \beta_0 \alpha_k, & k \in \{1, 2, \dots, K\}. \end{cases}$$

This equivalent state-space realization is called ρ -realization in the sequel and is denoted as $R_\rho \triangleq (A_\rho, B_\rho, C_\rho, d)$ for convenience. Clearly, R_ρ is a very sparse realization possessing $4K$ parameters to be implemented. It is easy to understand that for a given digital system, there exists a class of such ρ -realizations depending on the space within which $\{\gamma_k\}$ and $\{\Delta_k\}$ take values.

For a given digital system $H(z)$ and a specified set $\{\gamma_k, \Delta_k\}$, one can compute the parameters $\{\alpha_k, \beta_k\}$ and hence implement $H(z)$ with the ρ DFII structure as shown in Figures 4.1 and 4.2. One can also implement $H(z)$ with the equivalent state-space realization $(A_\rho, B_\rho, C_\rho, d)$ (ρ -realization) with (4.9). Both structures are very sparse, which will increase the implementation efficiency. In the next section, we will analyze the stability performance of the two structures and solve the corresponding optimal sparse structure problems for both digital filters and controllers.

4.3 Stability analysis and structure optimization

Now, for both digital filters and controllers, we analyze the stability behavior of the two sparse structures with the stability-related measure defined in Section 3.1 and show how the stability performance can be optimized with respect to the free parameters $\{\gamma_k\}$ and $\{\Delta_k\}$ for each structure.

For any $\{\gamma_k, \Delta_k\}$ given, one can always find the corresponding $\{\alpha_k, \beta_k\}$ and hence the ρ DFIIt structure. Noting that $\{\gamma_k\}$ and $\{\Delta_k\}$ are parameters to be implemented in the structure, it is desired that they are chosen from S_{FWL} defined before such that they produce no FWL errors at all. The same argument also applies to the ρ -realizations. Denote S_γ and S_Δ as two sets from which $\{\gamma_k\}$ and $\{\Delta_k\}$ take values, respectively. It is assumed in the sequel that $S_\gamma, S_\Delta \subset S_{FWL}$, which means that all γ_k and Δ_k are of exact B_γ bit- and B_Δ bit-formats, respectively, with both B_γ and B_Δ smaller than B_τ .

4.3.1 Optimal ρ DFIIt structures

Consider a digital filter implemented with a ρ DFIIt structure. Noting that $\{\gamma_k, \Delta_k\}$ can be implemented exactly, the poles are only affected by the FWL errors of the parameters $\{\alpha_k\}$. So, it follows from (3.1) and (3.2) that the corresponding stability measure for ρ DFIIt filter is given by

$$\mu_2 = \min_k \frac{1 - |\lambda_k|}{\sqrt{K \sum_{i=1}^K \left| \frac{\partial |\lambda_k|}{\partial \alpha_i} \right|^2}}, \quad (4.11)$$

where $\{\partial |\lambda_k| / \partial \alpha_i\}$ can be computed using (3.3) with

$$\frac{\partial \lambda_k}{\partial \alpha_i} = -e_i^T \frac{\partial \lambda_k}{\partial A_\rho} e_1 \Delta_1, \quad \forall i$$

where e_k denotes the k th elementary (column) vector whose elements are all zero except the k th one which is 1.

When the digital controller is implemented with a ρ DFIIt structure, the poles of the closed-loop system are only affected by the FWL errors due to

the parameters $\{\alpha_k, \beta_k\}$ since $\{\gamma_k, \Delta_k\}$ can be implemented exactly, as assumed above. It then follows from (3.1) and (3.2) that the corresponding closed-loop stability measure for this structure is given by

$$\mu_2 = \min_k \frac{1 - |\lambda_k|}{\sqrt{(2K+1) \left\{ \sum_{i=1}^K \left[\left| \frac{\partial |\lambda_k|}{\partial \alpha_i} \right|^2 + \left| \frac{\partial |\lambda_k|}{\partial \beta_i} \right|^2 \right] + \left| \frac{\partial |\lambda_k|}{\partial \beta_0} \right|^2 \right\}}}. \quad (4.12)$$

To evaluate the above μ_2 , one has to compute $\{\partial |\lambda_k| / \partial \alpha_i\}$ and $\{\partial |\lambda_k| / \partial \beta_i\}$, which can be obtained using (3.3) once $\{\partial \lambda_k / \partial \alpha_i\}$ and $\{\partial \lambda_k / \partial \beta_i\}$ are known.

The equivalent system matrix $X_\rho \triangleq \begin{bmatrix} d & C_\rho \\ B_\rho & A_\rho \end{bmatrix}$ of a ρ DFII structure can be expressed as

$$X_\rho = \begin{bmatrix} \beta_0 & \Delta_1 & 0 & 0 & \cdots & 0 & 0 \\ \beta_1 - \beta_0 \alpha_1 & \gamma_1 - \Delta_1 \alpha_1 & \Delta_2 & 0 & \cdots & 0 & 0 \\ \beta_2 - \beta_0 \alpha_2 & -\Delta_1 \alpha_2 & \gamma_2 & \Delta_3 & \cdots & 0 & 0 \\ & & & & \vdots & & \\ \beta_{K-1} - \beta_0 \alpha_{K-1} & -\Delta_1 \alpha_{K-1} & 0 & 0 & \cdots & \gamma_{K-1} & \Delta_K \\ \beta_K - \beta_0 \alpha_K & -\Delta_1 \alpha_K & 0 & 0 & \cdots & 0 & \gamma_K \end{bmatrix}. \quad (4.13)$$

With some manipulations, it can be shown that

$$\begin{cases} \frac{\partial \lambda_k}{\partial \alpha_i} = -e_{i+1}^T \frac{\partial \lambda_k}{\partial X_\rho} (\beta_0 e_1 + \Delta_1 e_2), \quad \forall i \\ \frac{\partial \lambda_k}{\partial \beta_0} = \begin{bmatrix} 1 & -\alpha_1 & -\alpha_2 & \cdots & -\alpha_K \end{bmatrix} \frac{\partial \lambda_k}{\partial X_\rho} e_1 \\ \frac{\partial \lambda_k}{\partial \beta_i} = e_{i+1}^T \frac{\partial \lambda_k}{\partial X_\rho} e_1, \quad \forall i \neq 0, \end{cases} \quad (4.14)$$

where $\partial \lambda_k / \partial X_\rho$ can be computed with (2.17) and (2.18). Then μ_2 in (4.12) can be obtained easily.

We have analyzed the stability behavior of the ρ DFIIt structure for digital filters and controllers. Now, we move on to see how the structure can be optimized to achieve nice stability performance. The following structure optimization procedure can be applied to both digital filters and controllers implemented with a ρ DFIIt structure. However, it should be pointed out that if the optimal filter structure problem is considered, μ_2 is computed with (4.11), and if the optimal structure problem is for closed-loop systems, μ_2 is evaluated by (4.12).

Since different sets of $\{\gamma_k, \Delta_k\}$ yield different μ_2 , the interesting problem is to maximize μ_2 with respect to these free parameters. Another point we should mention is that the dynamical range of the parameters $\{\alpha_k, \beta_k\}$ is determined by the two sets of free parameters. Noting that for a fixed-point implementation, it is desired that all the parameters $\{\alpha_k, \beta_k, \gamma_k, \Delta_k\}$ in the ρ DFIIt structure be absolutely bounded by one, which implies

$$|\alpha_k| \leq 1, \quad |\beta_k| \leq 1, \forall k \quad (4.15)$$

with $\{\alpha_k, \beta_k\}$ computed using (4.2). Therefore, the optimal ρ DFIIt structure problem is to find a solution to the following problem:

$$\max_{\{\gamma_k: \gamma_k \in S_\gamma\}, \{\Delta_k: \Delta_k \in S_\Delta\}} \mu_2, \quad \text{subject to (4.15)}. \quad (4.16)$$

This problem is highly nonlinear in terms of $\{\gamma_k\}$ and $\{\Delta_k\}$, and no analytical solutions have been found so far.

Noting the fact that the search spaces S_γ, S_Δ for the problem (4.16) are finite, the true optimal structures can definitely be achieved with exhaustive searching. Suppose γ_k and Δ_k are of exact B_γ bit- and B_Δ bit-formats (including one bit

for the sign), S_γ and S_Δ contain $2^{B_\gamma} + 1$ and $2^{B_\Delta - 1}$ elements, respectively. For a digital system of order K , there will be $[(2^{B_\gamma} + 1)(2^{B_\Delta - 1})]^K$ combinations for exhaustive searching when we solve the optimization problem in (4.16). So, it is easy to see that this optimization method may not be suitable when the dimension of the system is very large.

Alternatively, one can attack such a problem using the existing optimization techniques such as the adaptive simulated annealing (ASA) [97] and the genetic algorithm (GA) [98], which are popularly utilized for nonlinear optimization problems nowadays.

GA was first studied by Holland and his colleagues at the University of Michigan in 1975 [99] and has been applied to a number of design and optimization issues for both digital filters and controllers in the past few years [100] - [106]. It attempts to mimic (in an over-simplified manner) the natural evolution of species in order to solve difficult optimization problems. Appendix A gives a brief description about how to apply GA to solve some of the optimization problems in this thesis. For the numerical examples in the next section, both GA and the exhaustive searching are used to solve (4.16). It is found that GA is much more efficient and yields a structure which is almost the same as the optimal one obtained using the exhaustive searching.

In what follows, we will make an important analysis on how to choose the free parameters $\{\gamma_k\}$, $\{\Delta_k\}$ to maximize the stability robustness (minimize the pole sensitivities).

For digital filter case, it follows from (4.7) that

$$D(z) = \kappa[p_0(z) + \sum_{i=1}^K \alpha_i p_i(z)] = \prod_{k=1}^K (z - \lambda_k) \quad (4.17)$$

where $\{\lambda_k\}$ are the poles of the filter. According to the *implicit function theorem* [107], one has

$$\frac{\partial \lambda_k}{\partial \alpha_i} = -\frac{\kappa p_i(\lambda_k)}{\prod_{k \neq j}^K (\lambda_k - \lambda_j)}, \quad \forall i. \quad (4.18)$$

With $\kappa = \prod_{k=1}^K \Delta_k$ and (4.6), it then can be shown that

$$\begin{cases} \frac{\partial \lambda_k}{\partial \alpha_i} = -\frac{\prod_{m=1}^i \Delta_m \prod_{n=i+1}^K (\lambda_k - \gamma_n)}{\prod_{k \neq j}^K (\lambda_k - \lambda_j)}, & \forall i \neq K \\ \frac{\partial \lambda_k}{\partial \alpha_K} = -\frac{\prod_{m=1}^K \Delta_m}{\prod_{k \neq j}^K (\lambda_k - \lambda_j)}. \end{cases} \quad (4.19)$$

For digital controller case, the transfer function of the feedback control system depicted in Figure 2.3 is

$$H_{cl}(z) = \frac{P(z)}{1 - P(z)H(z)} = \frac{D(z)N_p(z)}{D_p(z)D(z) - N_p(z)N(z)} \triangleq \frac{N_{cl}(z)}{D_{cl}(z)} \quad (4.20)$$

where $D_p(z)$ and $N_p(z)$ are the denominator and numerator of the plant $P(z)$.

It then follows from (4.7) that

$$\begin{aligned} D_{cl}(z) &= \kappa D_p(z)[p_0(z) + \sum_{i=1}^K \alpha_i p_i(z)] - \kappa N_p(z) \sum_{i=0}^K \beta_i p_i(z) \\ &= \prod_{k=1}^{N_0} (z - \lambda_k) \end{aligned} \quad (4.21)$$

where $\{\lambda_k\}$ are the poles of the closed-loop system and $N_0 \triangleq K + J$ with J the order of the plant. According to the *implicit function theorem*, one has

$$\begin{cases} \frac{\partial \lambda_k}{\partial \alpha_i} = -\frac{\kappa p_i(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} D_p(\lambda_k), & \forall i \\ \frac{\partial \lambda_k}{\partial \beta_i} = \frac{\kappa p_i(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} N_p(\lambda_k), & \forall i. \end{cases} \quad (4.22)$$

Noting $\kappa = \prod_{k=1}^K \Delta_k$ and (4.6), one has

$$\left\{ \begin{array}{l} \frac{\partial \lambda_k}{\partial \alpha_i} = -\frac{\prod_{m=1}^i \Delta_m \prod_{n=i+1}^K (\lambda_k - \gamma_n)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} D_p(\lambda_k), \quad \forall i \neq K \\ \frac{\partial \lambda_k}{\partial \alpha_K} = -\frac{\kappa D_p(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} \\ \frac{\partial \lambda_k}{\partial \beta_i} = \frac{\prod_{m=1}^i \Delta_m \prod_{n=i+1}^K (\lambda_k - \gamma_n)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} N_p(\lambda_k), \quad \forall i \neq 0, K \\ \frac{\partial \lambda_k}{\partial \beta_0} = \frac{\prod_{n=1}^K (\lambda_k - \gamma_n)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} N_p(\lambda_k) \\ \frac{\partial \lambda_k}{\partial \beta_K} = \frac{\kappa N_p(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)}. \end{array} \right. \quad (4.23)$$

Some useful results can be obtained through analyzing the numerators of the equations (4.19) and (4.23) for a ρ DFII structure. Here we put $\partial \lambda_k / \partial \alpha_i$ as an example. The value of $\partial \lambda_k / \partial \alpha_i$ is mostly affected by two parts $\prod_{m=1}^i \Delta_m$ and $\prod_{n=i+1}^K (\lambda_k - \gamma_n)$ in the numerators. From the first part, one can see that the value of $\partial \lambda_k / \partial \alpha_1$ depends on Δ_1 and the value of $\partial \lambda_k / \partial \alpha_i$ depends on $\Delta_1, \Delta_2, \dots, \Delta_i$. So the first observation is that of all the $\{\Delta_k\}$, Δ_1 affects $\{\partial \lambda_k / \partial \alpha_i, \partial \lambda_k / \partial \beta_i\}$ the most and the effect of each Δ_k on the pole sensitivities decreases one by one. From the second part, one can see that if $\prod_{n=i+1}^K (\lambda_k - \gamma_n) \approx 0$, then $\partial \lambda_k / \partial \alpha_i \approx 0$. So the second observation is that when $\{\gamma_k\}$ are chosen such that they are close to the system poles, the pole sensitivities can be much reduced.

These observations give some insights on how to choose $\{\gamma_k\}$ and $\{\Delta_k\}$ as good initial conditions to speed up the algorithm (i.e., GA) used for attacking the optimal structure problem (4.16) or to reduce the search space to make the exhaustive searching for the optimal structures more efficiently.

4.3.2 Optimal ρ -realizations

In this case, the only parameters to be implemented in (4.10) are $\{a_{k1}\}$, $\{b_k\}$, d , $\{\gamma_k\}$ and $\{\Delta_k\}$. Noting that $\{\gamma_k\}$ and $\{\Delta_k\}$ are assumed to be of an exact FWL format, it follows from (3.2) that the pole modulus sensitivity measure Φ_k of a ρ -realization can be computed with

$$\Phi_k = \sum_{i=1}^K \left| \frac{\partial |\lambda_k|}{\partial a_{i1}} \right|^2 + \varepsilon \left(\left| \frac{\partial |\lambda_k|}{\partial d} \right|^2 + \sum_{i=1}^K \left| \frac{\partial |\lambda_k|}{\partial b_i} \right|^2 \right), \quad (4.24)$$

where $\varepsilon = 0$ for digital filters, and $\varepsilon = 1$ for closed-loop systems. Then one can compute the corresponding μ_2 using (3.1), where $N = K$ when $\varepsilon = 0$ while $N = 2K + 1$ when $\varepsilon = 1$.

Similarly, one has the following dynamical range constraints

$$|a_{k1}| \leq 1, \quad |b_k| \leq 1, \quad \forall k \quad (4.25)$$

and hence the optimal ρ -realization problem is

$$\max_{\{\gamma_k: \gamma_k \in S_\gamma\}, \{\Delta_k: \Delta_k \in S_\Delta\}} \mu_2, \quad \text{subject to (4.25)}. \quad (4.26)$$

This problem can be solved using the same way as the one for the optimal ρ DFII structure problem (4.16).

With some manipulations, one has the following relationship of pole sensitivities between the two sparse structures

$$\begin{cases} \frac{\partial \lambda_k}{\partial \alpha_i} &= -\Delta_1 \frac{\partial \lambda_k}{\partial a_{i1}} - \beta_0 \frac{\partial \lambda_k}{\partial b_i}, \quad \forall i \\ \frac{\partial \lambda_k}{\partial \beta_0} &= -\sum_{i=1}^K \frac{\partial \lambda_k}{\partial b_i} \alpha_i + \frac{\partial \lambda_k}{\partial d} \\ \frac{\partial \lambda_k}{\partial \beta_i} &= \frac{\partial \lambda_k}{\partial b_i}, \quad \forall i \neq 0. \end{cases} \quad (4.27)$$

From above equations and (4.19), (4.23), one can see that when $\{\gamma_k\}$ are chosen such that they are close to the system poles, the stability performance of the ρ -realizations can increase, which is the same as for the ρ DFII structures. There is, however, no direct relationship between the stability measures of the two sparse structures.

4.4 Numerical examples

Now we present several numerical examples to confirm our theoretical results and to illustrate the stability performance of different structures. In these examples, it is assumed that $S_\gamma = \{\pm 1, \pm(2^{-1} + 2^{-2}), \pm 2^{-1}, \pm 2^{-2}, 0\}$ and $S_\Delta = \{1, 2^{-1} + 2^{-2}, 2^{-1}, 2^{-2}\}$. The elements in these two sets are of exact 3-bit format (including one bit for the sign). Other formats with more bits have been used and the stability performance of each structure, as expected, can be further improved but the improvement is not significant for these examples.

A genetic algorithm (GA) is used to solve the optimal structure problems (4.16) and (4.26), and the results are confirmed with the exhaustive searching. For each problem using GA, the free parameters are chosen such that the roots of the polynomials are close to the system poles. Such free parameters are encoded to binary strings to generate the initial population. Then we run the GA with 20 generations to yield a solution, which is the best for this run. This solution is used to generate the initial population for the next run and the algorithm is performed again. After 3 to 4 runs, the result cannot be improved further, thus the GA yields the optimal solution. For the details of the flow of GA, please refer to Appendix A.

First, let us see two examples for digital filters.

Example 4.1: This is the low-pass fourth-order Butterworth filter already presented in Example 3.1, where we computed and compared the stability robustness measure μ_2 for the initial filter realization R_0 in shift operator and the normal realization R_n , which is a fully parametrized optimal realization.

Using the procedures in the previous sections of this chapter, another two realizations are obtained. R_δ is the classical delta operator-based realization (the equivalent state-space realization of δ DFIIt structure) with $\Delta_k = 1, \gamma_k = 1, \forall k$. R_ρ^{opt} is the optimal ρ -realization corresponding to $\{\gamma_k\} = \{\gamma_1 = \gamma_3 = 0.75, \gamma_2 = \gamma_4 = 1\}$ and $\{\Delta_k\} = \{\Delta_k = 0.25, \forall k\}$, which is the solution to (4.26) and is given by:

$$\begin{aligned} A_\rho^{opt} &= \begin{bmatrix} 0.8397 & 0.25 & 0 & 0 \\ -0.1680 & 1 & 0.25 & 0 \\ 0.0193 & 0 & 0.75 & 0.25 \\ -0.0320 & 0 & 0 & 1 \end{bmatrix}, \quad B_\rho^{opt} = \begin{bmatrix} 0.0009 \\ 0.0110 \\ 0.0530 \\ 0.1280 \end{bmatrix} \\ C_\rho^{opt} &= \begin{bmatrix} 0.25 & 0 & 0 & 0 \end{bmatrix}, \quad d = 3.1239 \times 10^{-5}. \end{aligned}$$

For both R_δ and R_ρ^{opt} , μ_2 are computed and μ_0^* are estimated using simulations (see Section 3.3 for details). The computation results, together with those in Example 3.1, are presented in Table 4.1, where N_p is the number of nontrivial parameters in each structure and DR means the dynamical range such that no parameter in the structure is absolutely bigger than one.

Comment 4.1: From this example, one can see that

- for each structure μ_2 is a lower bound of μ_0^* which is an estimate of μ_0 by simulations as mentioned before,

Table 4.1: Comparison of different state-space realizations (Example 4.1)

	R_0	R_δ	R_n	R_ρ^{opt}
μ_2	1.08×10^{-4}	2.22×10^{-4}	2.06×10^{-2}	1.31×10^{-2}
μ_0^*	1.25×10^{-4}	4.99×10^{-4}	4.21×10^{-2}	1.99×10^{-2}
N_p	9	9	25	14
DR	No	Yes	Yes	Yes

- R_0 and R_δ , though each having only 9 nontrivial parameters, are very poor in terms of stability robustness, while R_n and R_ρ^{opt} have a much better stability behavior,
- the normal realization R_n yields a larger μ_2 than the ρ -realization R_ρ^{opt} , but it has $(K+1)^2$ parameters to implement while the amount of parameters in R_ρ^{opt} for implementation is just proportional to K ,
- for R_ρ^{opt} , the optimal $\{\gamma_k\}$ are close to the poles of the filter, which confirms our argument on how to choose γ_k to enhance μ_2 ,
- the delta operator-based R_δ ($\gamma_k = 1, \forall k$) is better than the shift operator-based R_0 ($\gamma_k = 0, \forall k$) due to the fact that the poles of this low-pass narrow band filter are distributed around $z = +1$.

Now we compare the performance of different DFII structures. Solving (4.16), one obtains an optimal ρ DFII structure, denoted as ρ DFII, which can be implemented with the structure form depicted in Figures 4.1 and 4.2, where

Table 4.2: Comparison of different DFIIIt structures (Example 4.1)

	$z\text{DFIIIt}$	δDFIIIt	ρDFIIIt
μ_2	1.08×10^{-4}	2.22×10^{-4}	5.22×10^{-2}
μ_0^*	1.29×10^{-4}	4.99×10^{-4}	7.97×10^{-2}
N_p	9	9	15
DR	No	No	Yes

the corresponding parameter sets are given by

$$\{\alpha_k\} = \{1, -0.3589, 0.6721, -0.0771, 0.1280\},$$

$$\{\beta_k\} = \{0.0000, 0.0009, 0.0110, 0.0530, 0.1280\},$$

$$\{\gamma_k\} = \{0.75, 1, 0.75, 1\},$$

$$\{\Delta_k\} = \{0.25, 0.25, 0.25, 0.25\}.$$

$z\text{DFIIIt}$ and δDFIIIt are two classical DFIIIt structures, corresponding to the shift- and delta-operators, which are obtained with $\{\Delta_k = 1, \forall k\}$, while $\gamma_k = 0, \forall k$ and $\gamma_k = 1, \forall k$, respectively. Table 4.2 shows the computation results of different DFIIIt structures.

Comment 4.2: From this example, we observe that

- ρDFIIIt yields a much better stability performance than $z\text{DFIIIt}$ and δDFIIIt , and beats R_n greatly in terms of implementation efficiency,
- the optimal $\{\gamma_k\}$ for ρDFIIIt are close to the poles of the filter, which leads to a large μ_2 ,
- ρDFIIIt also satisfies the parameter dynamical range constraints, which is desired for a fixed-point implementation.

Example 4.2: This is a band-pass Butterworth filter of order six, generated with MATLAB command $[V_q, V_p] = \text{butter}(3, [0.65, 0.85])$. The corresponding poles are $\lambda_{1,2} = -0.4122 \pm j0.7106$, $\lambda_{3,4} = -0.7962 \pm j0.4181$, $\lambda_{5,6} = -0.5612 \pm j0.4412$. We note that the real parts of the poles are around $z = -0.5$ and $z = -0.75$.

This filter can be described in the shift operator-based canonical realization R_0 :

$$A_0 = \begin{bmatrix} -3.5391 & 1 & 0 & 0 & 0 & 0 \\ -6.0181 & 0 & 1 & 0 & 0 & 0 \\ -6.1109 & 0 & 0 & 1 & 0 & 0 \\ -3.9247 & 0 & 0 & 0 & 1 & 0 \\ -1.4997 & 0 & 0 & 0 & 0 & 1 \\ -0.2781 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} -0.0641 \\ -0.1632 \\ -0.1106 \\ -0.0167 \\ -0.0271 \\ -0.0231 \end{bmatrix}$$

$$C_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad d = 0.0181.$$

Taking R_0 as the initial realization, we can obtain a normal realization R_n

using the same way as in Example 3.1. R_n is given by

$$\begin{aligned} A_n &= \begin{bmatrix} -0.7516 & 0.4014 & -0.1253 & 0.0877 & 0.0950 & 0.1458 \\ -0.3762 & -0.5857 & 0.4455 & -0.0559 & -0.0248 & -0.0669 \\ -0.0653 & -0.3460 & -0.4890 & 0.5288 & 0.0433 & 0.1204 \\ -0.1212 & -0.2567 & -0.4163 & -0.5778 & 0.3363 & -0.1621 \\ -0.0161 & -0.0375 & -0.1330 & -0.3205 & -0.5301 & 0.3831 \\ -0.2325 & -0.0077 & -0.1221 & -0.0092 & -0.3797 & -0.6049 \end{bmatrix} \\ B_n &= \begin{bmatrix} 0.6023 & -0.7068 & -0.8187 & 2.2923 & -1.9234 & -0.9715 \end{bmatrix}^T \\ C_n &= \begin{bmatrix} 0.1272 & 0.2342 & 0.2349 & 0.1312 & 0.0425 & 0.0019 \end{bmatrix}. \end{aligned}$$

Solving (4.26) leads to the optimal ρ -realization R_ρ^{opt} , for which $\{\gamma_k\} = \{-0.5, -0.75, -0.5, -0.75, -0.5, -0.75\}$ and $\{\Delta_k\} = \{0.5, 1, 0.25, 1, 0.25, 0.75\}$.

The corresponding realization is

$$\begin{aligned} A_\rho^{opt} &= \begin{bmatrix} -0.2891 & 1 & 0 & 0 & 0 & 0 \\ -0.8912 & -0.75 & 0.25 & 0 & 0 & 0 \\ -0.0027 & 0 & -0.5 & 1 & 0 & 0 \\ -0.9859 & 0 & 0 & -0.75 & 0.25 & 0 \\ -0.0943 & 0 & 0 & 0 & -0.5 & 0.75 \\ -0.5380 & 0 & 0 & 0 & 0 & -0.75 \end{bmatrix} \\ B_\rho^{opt} &= \begin{bmatrix} -0.1281 & 0.0899 & 0.3619 & -0.3202 & -0.7868 & -0.0841 \end{bmatrix}^T \\ C_\rho^{opt} &= \begin{bmatrix} 0.5 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Like Example 4.1, we compare different state-space realizations in terms

Table 4.3: Comparison of different state-space realizations (Example 4.2)

	R_0	R_δ	R_n	R_ρ^{opt}
μ_2	2.75×10^{-3}	1.76×10^{-4}	2.37×10^{-2}	6.38×10^{-2}
N_p	13	13	49	22
DR	No	No	No	Yes

Table 4.4: Comparison of different DFIIIt structures (Example 4.2)

	$z\text{DFIIIt}$	δDFIIIt	ρDFIIIt
μ_2	2.75×10^{-3}	1.76×10^{-4}	6.47×10^{-2}
N_p	13	13	22
DR	No	No	Yes

of stability robustness, number of nontrivial parameters, and dynamical range constraints, which are presented in Table 4.3.

With the same procedure as in Example 4.1, the stability measures of different DFIIIt structures are computed, in which the optimal ρDFIIIt structure is obtained with $\{\gamma_k\} = \{-0.75, -0.5, -0.5, -0.5, -0.5, -0.75\}$ and $\{\Delta_k\} = \{1, 1, 0.25, 1, 0.25, 0.5\}$. The computation results are shown in Table 4.4.

Through this example, we have the following observations for the two classes of structures (state-space realizations and DFIIIt structures).

- The ρ -operator-based structures (R_ρ^{opt} , ρDFIIIt) have a much better stability behavior than those structures in shift operator (R_0 , $z\text{DFIIIt}$) and the ones in delta operator (R_δ , δDFIIIt).
- Both R_ρ^{opt} and ρDFIIIt are very sparse, each having 22 nontrivial param-

eters, while R_n has 49. The two sparse structures even beat R_n in terms of stability performance.

- For both R_ρ^{opt} and ρ DFIIt, the corresponding optimal $\{\gamma_k\}$ are close to the poles of the filter.
- The delta operator-based structures perform very badly in this example. This is mainly due to the fact that the poles of the filter are far away from $z = +1$ (i.e. band-pass and high-pass filters).

Some simulations are also performed to confirm our theoretical results. Using a similar way as we determine the μ_0^* , first we generate a same succession of ten different perturbations on θ_k for each state-space realization. Let $\{\Delta\theta_k(i), i = 1, 2, \dots, 10\}$ be the perturbation matrices added to θ_k for the i th perturbation, where all the elements of $\Delta\theta_k(i)$ are bounded by some constant σ . Then we compute the poles of each perturbed realization which are the eigenvalues of $A_k(i)$, the corresponding A -matrix to $\theta_k + \Delta\theta_k(i)$, for $i = 1, 2, \dots, 10$. If any one of the poles of a perturbed realization goes out of the unit circle, it means that this perturbed realization is not stable. Similar simulations are also performed for DFIIt structures.

Figures 4.3 and 4.4 represent the pole distributions for the different state-space realizations and DFIIt structures, respectively, with $\sigma = 2 \times 10^{-5}$. In each figure, the poles of the exact realization are represented by a '+' while the poles of the ten perturbed realizations are represented by 'x'. Since the poles, if complex, appear in conjugate pair, only those of non-negative imaginary part are shown in the figures.

On the one hand, we note that some of the perturbed R_0 and z DFIIt lead to

Figure 4.3: Pole distributions of different state-space realizations (Example 4.2)

the instability of the filter. The performance of the perturbed R_δ and δ DFIIt are even worse, in which most of the perturbed structures are unstable. On the other hand, the figures clearly indicate that the errors on the poles are exceedingly small in the realization R_n and are even smaller in R_ρ^{opt} and ρ DFIIt, which shows the nice stability performance of these optimal structures.

Now we go on to see an example of the discrete-time closed-loop control system.

Figure 4.4: Pole distributions of different DFII structures (Example 4.2)

Example 4.3: Here we return to the Example 3.2 of Chapter 3 to illustrate the effectiveness of the optimal sparse structures for digital controllers. In that example, we have computed the fully parametrized optimal realization R_f^{opt} , which has a much larger μ_2 than the canonical realization R_c .

Solving (4.26) with μ_2 computed by (3.1) and Φ_k given by (4.24), one obtains the optimal ρ -realization R_ρ^{opt} , for which $\{\gamma_k\} = \{1, 0.5, 0.75, 0.75, 1, 1\}$ and

$\{\Delta_k\} = \{0.25, 1, 0.75, 0.25, 0.25, 0.25\}$. The corresponding realization is:

$$\begin{aligned} A_\rho^{opt} &= \begin{bmatrix} 0.2600 & 1 & 0 & 0 & 0 & 0 \\ -0.9013 & 0.5 & 0.75 & 0 & 0 & 0 \\ -0.3479 & 0 & 0.75 & 0.25 & 0 & 0 \\ -0.6385 & 0 & 0 & 0.75 & 0.25 & 0 \\ -0.7099 & 0 & 0 & 0 & 1 & 0.25 \\ -0.0000 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\ B_\rho^{opt} &= \begin{bmatrix} 0.8475 & 0.1577 & 0.8404 & -0.3815 & -0.1301 & 0.0000 \end{bmatrix}^T \\ C_\rho^{opt} &= \begin{bmatrix} 0.25 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

With μ_2 evaluated by (4.12), we solve (4.16) and then get the optimal ρ DFIIt structure, denoted as ρ DFIIt, which can be implemented using the structure form depicted in Figures 4.1 and 4.2. The corresponding optimal free parameters and structure coefficients are

$$\begin{aligned} \{\gamma_k\} &= \{1, 0.75, 0.75, 0.75, 0.75, 1\}, \\ \{\Delta_k\} &= \{1, 1, 0.25, 0.25, 0.25, 0.25\}, \\ \{\alpha_k\} &= \{1, 0.7400, 0.8388, 0.0199, 0.8718, 0.2140, 0.0000\}, \\ \{\beta_k\} &= \{0.0460, 0.2459, 0.0780, 0.5388, -0.8763, 0.1984, 0.0000\}. \end{aligned}$$

We also obtain R_δ , z DFIIt and δ DFIIt as defined before. Tables 4.5 and 4.6 compare the performance of the different state-space realizations and the different DFIIt structures, respectively.

The results are self-explanatory. Particularly, one can observe that both

Table 4.5: Comparison of different state-space realizations (Example 4.3)

	R_c	R_δ	R_f^{opt}	R_ρ^{opt}
μ_2	3.59×10^{-8}	7.72×10^{-8}	9.20×10^{-6}	2.24×10^{-5}
N_p	13	13	49	21
DR	No	No	No	Yes

Table 4.6: Comparison of different DFIIIt structures (Example 4.3)

	z DFIIIt	δ DFIIIt	ρ DFIIIt
μ_2	3.42×10^{-8}	7.73×10^{-8}	1.84×10^{-5}
N_p	13	13	21
DR	No	No	Yes

R_ρ^{opt} and ρ DFIIIt beat the fully parametrized realization R_f^{opt} in terms of not only structure sparseness but also stability performance. These two sparse structures also satisfy the parameter dynamical range constraints, which is desired for a fixed-point implementation. The optimal $\{\gamma_k\}$ for R_ρ^{opt} and ρ DFIIIt are close to the dominant poles of the close-loop system, which confirms our theoretical analysis.

We also perform some simulations to test the pole distributions of the closed-loop system when the controller is implemented with different structures under some certain perturbations. The procedure is similar with that in Example 4.2. Figures 4.5 and 4.6 represent the pole distributions for the different state-space realizations and DFIIIt structures, respectively, with $\sigma = 8 \times 10^{-6}$.

The figures clearly show the errors on the closed-loop poles for each state-space realization and DFIIIt structure, which reflects the different stability ro-

(a) Pole distribution for R_c (b) Pole distribution for R_δ (c) Pole distribution for R_f^{opt} (d) Pole distribution for R_ρ^{opt}

Figure 4.5: Pole distributions of different state-space realizations (Example 4.3)

bustness of the different structures. Therefore, observations from the figures and the simulation outcomes have further confirmed our theoretical results.

4.5 Summary

In this chapter, we have investigated the sparse structure problem subject to minimizing the FWL effects. Based on the polynomial parametrization concept,

(a) Pole distribution for z DFIIt (b) Pole distribution for δ DFIIt (c) Pole distribution for ρ DFIIt

Figure 4.6: Pole distributions of different DFIIt structures (Example 4.3)

two new sparse structures have been developed by reparametrizing the transfer function with a set of special polynomial operators, called ρ -operators. Both structures are very efficient in terms of implementation and contain some free parameters with which the structure can be optimized to achieve nice stability performance. The stability behavior of each structure has been analyzed with the stability robustness measure proposed in Chapter 3 and the corresponding optimal sparse structure problem has been solved using a genetic algorithm with practical considerations.

Through design examples and the corresponding simulations, the stability performance of different structures has been compared for both digital filters and controllers. It is shown that the optimal sparse structures have a stability performance that is very close to (in some cases, better than) that of the fully parametrized optimal realizations and is always better than those of the structures based on the traditional shift operator and the prevailing delta operator.

For the ρ -operator-based structures (ρ DFIIt structure and ρ -realization), the expression of the stability-related measure has been specified in terms of the free parameters and the system poles. It has been found that when the polynomials are chosen such that their roots (say $\{\gamma_k\}$) are close to the system poles, the pole sensitivities can be much reduced and hence the stability robustness increases. This expression gives some insights to help search for the optimal sparse structures in a more efficient way. Meanwhile, it also reflects a limitation of the ρ -operator-based structures, that is, the polynomials using ρ -operators have real roots only while the system poles are generally complex. For those systems whose poles are far away from the real axis, the ρ -operator-based structures would not be competent.

Then several questions come out of our mind: can the ρ -operator be extended to a more generalized operator which leads to a set of polynomials having complex roots? If so, is there any structures that can be developed using the more generalized operator and be optimized to achieve further better FWL performance?

The answers to these questions will soon be revealed as we open the door to the next chapter.

Chapter 5

Improved Sparse Structures Using Second-Order Operators

It has been noted that for the ρ -operator-based structures in Chapter 4, when the roots of the polynomials, say $\{\gamma_k\}$, are chosen such that they are close to the system poles, the pole sensitivities can be much reduced and hence the stability robustness increases. Due to the fact that $\{\gamma_k\}$ is a real set while the system poles are generally complex, the degree of maximizing the stability robustness for the ρ -operator-based structures is limited. It is this limitation that motivates us to conduct further research for sparse structures with higher degree of stability performance.

In this chapter, we will develop new polynomials using a set of second-order operators, with which two new efficient structures are obtained. One of them can be considered as an improved version of the ρ DFIIt structure proposed in Chapter 4 and is denoted as ρ IDFIIt for convenience. The other one is the equivalent state-space realization of the ρ IDFIIt structure, called ρ_I -realization.

Compared with the ρ -operator-based structures, the improved sparse structures have some more free parameters, which can be utilized to minimize the FWL effects. The problems of finding optimized sparse structures are investigated with stability consideration. It is shown that the stability robustness for the ρ IDFIIIt structure and ρ_I -realization can be much enhanced when the roots of the corresponding polynomials are close to the dominant system poles. With the extra degree of freedom introduced, the new set of polynomials can have both real and complex roots. As a result of that, the stability robustness of the second-order operator-based structures can be further maximized than that of the ρ -operator-based ones, while the price paid for that is very slight.

5.1 Two improved sparse structures

In Section 4.2, we reparametrized the transfer function $H(z)$ with a set of special polynomials $\{p_k(z)\}$ in the ρ -operator and derived two sparse structures for digital system implementation. Now let us consider another special set of the polynomials using the second-order operators, with which two new sparse structures will be developed, which can be regarded as the improved version of those structures proposed in Section 4.2.

5.1.1 The improved ρ DFIIIt structure

Define

$$p \triangleq \begin{cases} K/2 & \text{when } K \text{ is even} \\ (K+1)/2 & \text{when } K \text{ is odd} \end{cases}$$

and the first-order ρ -operator (see (4.5))

$$\rho_k(z) \triangleq \frac{z - \gamma_k}{\Delta_k}, \quad k = 1, 2, \dots, K-1, K, \quad (5.1)$$

where $\{\gamma_k\}$ and $\{\Delta_k > 0\}$ are two sets of free parameters. One then has a new set of polynomials $\{q_k(z), k = 0, 1, \dots, K\}$ obtained with the following recursive equations

$$\begin{cases} q_K(z) = 1, & q_{K-1}(z) = \rho_K(z) \\ q_{K-2m}(z) = [\rho_{K-2m+1}(z)\rho_{K-2m+2}(z) + \eta_{p-m+1}]q_{K-2m+2}(z) \\ \quad \triangleq \bar{\rho}_{p-m+1}(z)q_{K-2m+2}(z) \\ q_{K-2m-1}(z) = \rho_{K-2m}(z)q_{K-2m}(z) \end{cases} \quad (5.2)$$

for all $m = 1, 2, \dots, p$, where the second-order operator $\bar{\rho}_m(z)$ is defined as

$$\begin{cases} \bar{\rho}_{p-m+1}(z) = \rho_{K-2m+1}(z)\rho_{K-2m+2}(z) + \eta_{p-m+1} \\ \eta_{p-m+1} \triangleq \frac{\bar{\eta}_{p-m+1}}{\Delta_{K-2m+1}\Delta_{K-2m+2}} \end{cases} \quad (5.3)$$

with $\{\bar{\eta}_m\}$ a set of free parameters to be discussed later. It should be pointed out that when K is odd, $\rho_0(z) \triangleq 1$, $\bar{\eta}_1 = 0$ are assumed, which leads to $\bar{\rho}_1(z) = \rho_1(z)$ and hence $q_0(z) = \bar{\rho}_1(z)q_1(z)$.

Here we recall that the polynomials $p_k(z)$ defined in (4.6) have only real roots, which are $\{\gamma_k\}$. The reason why we choose the new polynomials $q_k(z)$ as in (5.2) and (5.3) is to formulate a set of second-order polynomials $q_k(z)$, which, as to be seen later, can have both real and complex roots and lead to a larger stability robustness of the second-order operator-based structures.

Denote

$$\begin{aligned} V_x &\triangleq \begin{bmatrix} x_0 & x_1 & \cdots & x_k & \cdots & x_K \end{bmatrix}^T, \quad x = \xi, \zeta, \quad \xi_0 = 1 \\ \bar{z} &\triangleq \begin{bmatrix} z^K & z^{K-1} & \cdots & z^{K-k} & \cdots & 1 \end{bmatrix}^T. \end{aligned}$$

Like (4.7), $D(z)$ and $N(z)$ can be rewritten as

$$D(z) = \kappa[q_0(z) + \sum_{k=1}^K \alpha_k q_k(z)], \quad N(z) = \kappa \sum_{k=0}^K \beta_k q_k(z) \quad (5.4)$$

with $\kappa = \prod_{k=1}^K \Delta_k$, which leads to

$$\begin{aligned} H(z) &= \frac{\sum_{k=0}^K \beta_k q_k(z)}{q_0(z) + \sum_{k=1}^K \alpha_k q_k(z)} \\ &= \frac{\beta_0 q_0(z) + \sum_{m=0}^{p-1} B_{p-m}(z) q_{K-2m}(z)}{q_0(z) + \sum_{m=0}^{p-1} A_{p-m}(z) q_{K-2m}(z)} \end{aligned} \quad (5.5)$$

where

$$\begin{cases} B_{p-m}(z) \triangleq \beta_{K-2m-1} \rho_{K-2m}(z) + \beta_{K-2m} \\ A_{p-m}(z) \triangleq \alpha_{K-2m-1} \rho_{K-2m}(z) + \alpha_{K-2m} \end{cases}$$

for all possible m except $m = p - 1$ when K is odd, for which, $A_1(z) = \alpha_1$, $B_1(z) = \beta_1$, and

$$\begin{cases} V_\alpha \triangleq \begin{bmatrix} 1 & \alpha_1 & \cdots & \alpha_k & \cdots & \alpha_K \end{bmatrix}^T = \kappa^{-1} T_q^{-T} V_\xi \\ V_\beta \triangleq \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_k & \cdots & \beta_K \end{bmatrix}^T = \kappa^{-1} T_q^{-T} V_\zeta \end{cases} \quad (5.6)$$

with $T_q \in \mathcal{R}^{(K+1) \times (K+1)}$ an upper triangular matrix satisfying

$$\begin{bmatrix} q_0(z) & q_1(z) & \cdots & q_{K-k}(z) & \cdots & q_K(z) \end{bmatrix}^T = T_q \bar{z}.$$

Equation (5.5) implies that the transfer function $H(z)$ is reparametrized with $\{\alpha_k\}$ and $\{\beta_k\}$ in the new set of polynomial operators $\{q_k(z)\}$.

Now, let us consider how to implement the digital system using the corresponding polynomial parametrized transfer function (5.5). First of all, we have the following input-output relationship

$$\begin{aligned} y(n) &= \beta_0 u(n) + \sum_{m=0}^{p-1} [B_{p-m}(z) \frac{q_{K-2m}(z)}{q_0(z)} u(n) - A_{p-m}(z) \frac{q_{K-2m}(z)}{q_0(z)} y(n)] \\ &\triangleq \beta_0 u(n) + w_1(n). \end{aligned} \quad (5.7)$$

Noting $q_{K-2m}(z) = \bar{\rho}_{p-m+1}(z) q_{K-2m+2}$ and $q_0(z) = \prod_{m=1}^p \bar{\rho}_m(z)$ (for any K), one can show that

$$\begin{aligned} w_1(z) &= \bar{\rho}_1^{-1}(z) \{ B_1(z) u(n) - A_1(z) y(n) + \sum_{m=0}^{p-2} [B_{p-m}(z) \frac{q_{K-2m}(z)}{\prod_{m=2}^p \bar{\rho}_m(z)} u(n) \\ &\quad - A_{p-m}(z) \frac{q_{K-2m}(z)}{\prod_{m=2}^p \bar{\rho}_m(z)} y(n)] \} \\ &\triangleq \bar{\rho}_1^{-1}(z) [B_1(z) u(n) - A_1(z) y(n) + w_2(n)]. \end{aligned}$$

Using the same procedure, one obtains that with $w_{p+1}(n) = 0$

$$w_m(z) = \bar{\rho}_m^{-1}(z) [B_m(z) u(n) - A_m(z) y(n) + w_{m+1}(n)], \quad \forall m. \quad (5.8)$$

Figure 5.1 shows the corresponding block diagram to (5.7) - (5.8), where the output of the m th block, i.e., $w_m(n)$, can be computed with the structure

Figure 5.1: Block diagram of the proposed ρ IDFIIt structure Figure 5.2: Implementation of the m th block in the ρ IDFIIt structure, where $\bar{m} \triangleq K - 2p + 2m$ and $w_{K+1}(n) = 0$

depicted in Figure 5.2. The structure shown in Figures 5.1 and 5.2 is called an improved ρ DFIIt structure, denoted as ρ IDFIIt, which possesses $\{\alpha_k, \beta_k\}$ and three sets of free parameters $\{\gamma_k\}$, $\{\Delta_k\}$ and $\{\eta_m\}$. For a given digital system $H(z)$, there exists a class of such structures, depending on the space within which $\{\gamma_k\}$, $\{\Delta_k\}$ and $\{\bar{\eta}_m\}$ take values. It is interesting to note that when $\bar{\eta}_m = 0, \forall m$, the ρ IDFIIt structure returns to the ρ DFIIt structure proposed in Section 4.2.

5.1.2 The improved ρ -realization

We choose $\{x_k(n)\}$ indicated in Figure 5.2 as the state variables and denote $x(n)$ as the state vector. It can be shown that a ρ IDFIIt structure is equivalent to the following state-space equations

$$\begin{cases} x(n+1) &= A_{\rho_I}x(n) + B_{\rho_I}u(n) \\ y(n) &= C_{\rho_I}x(n) + du(n) \end{cases}$$

with the equivalent realization $R_{\rho_I} \triangleq (A_{\rho_I}, B_{\rho_I}, C_{\rho_I}, d)$ satisfying

$$H(z) = d + C_{\rho_I}(zI - A_{\rho_I})^{-1}B_{\rho_I} \quad (5.9)$$

and given by

$$\begin{aligned} A_{\rho_I} &= D_\gamma + [A_o - A_\eta]D_\Delta, \quad B_{\rho_I} = \bar{V}_\beta - \beta_0\bar{V}_\alpha \\ C_{\rho_I} &= e_1^T D_\Delta, \quad d = \beta_0 \end{aligned} \quad (5.10)$$

where e_k is the k th elementary vector as defined before, $D_\gamma \triangleq \text{diag}(\gamma_1, \dots, \gamma_K)$, $D_\Delta \triangleq \text{diag}(\Delta_1, \dots, \Delta_K)$ two diagonal matrices, and

$$A_o \triangleq \begin{bmatrix} -\alpha_1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ -\alpha_2 & 0 & 1 & \cdots & 0 & 0 & 0 \\ -\alpha_3 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\alpha_4 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & & & & & & \\ -\alpha_{K-1} & 0 & 0 & \cdots & 0 & 0 & 1 \\ -\alpha_K & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}, \quad A_\eta \triangleq \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \eta_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \eta_2 & \cdots & 0 & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \eta_p & 0 \end{bmatrix}$$

and

$$\begin{cases} \bar{V}_\beta \triangleq \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_{K-1} & \beta_K \end{bmatrix}^T \\ \bar{V}_\alpha \triangleq \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{K-1} & \alpha_K \end{bmatrix}^T \end{cases} \quad (5.11)$$

It should be pointed out that the A_η presented here is for K is even, that is $K = 2p$. When K is odd, $A_\eta(2m-1, 2m-2) = \eta_m$ for $m = 2, 3, \dots, p$ with $\eta_1 = 0$. In what follows, it is assumed that K is even.

In Section 4.2, the equivalent state-space realization of the ρ DFIIt structure was obtained, which is called the ρ -realization. For convenience, R_{ρ_I} defined by (5.10) is called ρ_I -realization or the improved ρ -realization, which returns to the ρ -realization in (4.10) when $\bar{\eta}_m = 0, \forall m$. Clearly, the ρ_I -realization is very sparse and contains some free parameters. It is easy to understand that for a given digital system, there exists a class of such ρ_I -realizations depending on the space within which the free parameters $\{\gamma_k\}$, $\{\Delta_k\}$ and $\{\bar{\eta}_m\}$ take values.

It should be pointed out that the proposed ρ IDFIIt structure contains p more parameters $\{\eta_m\}$ than the ρ DFIIt structure in Section 4.2. However, it is these η_m that introduce extra degree of freedom, with which the ρ IDFIIt structure can be optimized to achieve better performance than the optimal ρ DFIIt structure, while the price paid for that is relatively small. The same argument also applies to the ρ_I -realization compared with the ρ -realization. The performance analysis and optimization procedure for the two proposed structures will be discussed in the next section.

5.2 Performance analysis and optimized structures

In this section, we will analyze the stability behavior of the two improved sparse structures. The problems of finding the optimized ρ IDFIIt structure and the optimized ρ_I -realization will be formulated and solved, respectively.

One notes that for a given digital system $H(z)$, there exists a class of ρ IDFIIt structures and ρ_I -realizations. Each class of the structures is determined by three spaces, denoted as S_γ , S_Δ and $S_{\bar{\eta}}$, from which the free parameters γ_k , Δ_k and $\bar{\eta}_m$ take values, respectively. Noting that $\{\gamma_k\}$ and $\{\Delta_k\}$ are parameters to be implemented in the structure, it is desired that they are chosen from S_{FWL} defined before such that they produce no FWL errors at all. Therefore, it is assumed in the sequel that $S_\gamma, S_\Delta \subset S_{FWL}$, which means that all γ_k and Δ_k are of exact B_γ bit- and B_Δ bit-formats, respectively, with both B_γ and B_Δ smaller than B_τ . As to $\bar{\eta}_m$, they are not the structure parameters and hence do not have to be in an FWL format. However, there exists a range for $\bar{\eta}_m$ to stay in and the expression of space $S_{\bar{\eta}}$ will be clarified later in this section.

5.2.1 Optimized ρ IDFIIt structures

For a digital filter implemented with a ρ IDFIIt structure, the poles are only affected by the FWL errors of the parameters $\{\alpha_k\}$ and $\{\eta_m\}$ since $\{\gamma_k, \Delta_k\}$ can be implemented exactly. So, the stability measure μ_2 for a ρ IDFIIt filter can be computed using (3.1) with the corresponding Φ_k given by

$$\Phi_k = \sum_{i=1}^K \left| \frac{\partial |\lambda_k|}{\partial \alpha_i} \right|^2 + \sum_{m=1}^p \delta(\eta_m) \left| \frac{\partial |\lambda_k|}{\partial \eta_m} \right|^2, \quad (5.12)$$

where $\{\partial|\lambda_k|/\partial\alpha_i\}$ and $\{\partial|\lambda_k|/\partial\eta_m\}$ can be computed using (3.3) with

$$\begin{cases} \frac{\partial\lambda_k}{\partial\alpha_i} = -e_i^T \frac{\partial\lambda_k}{\partial A_{\rho I}} e_1 \Delta_1, \quad \forall i \\ \frac{\partial\lambda_k}{\partial\eta_m} = -e_{2m}^T \frac{\partial\lambda_k}{\partial A_{\rho I}} e_{2m-1} \Delta_{2m-1}, \quad \forall m \end{cases}$$

where e_k is the k th elementary vector defined before.

For a digital controller implemented with a ρ IDFIIt structure, the poles of the closed-loop system are only affected by the FWL errors due to the parameters $\{\alpha_k, \beta_k\}$ and $\{\eta_m\}$. It then follows from (3.2) that

$$\Phi_k = \sum_{i=1}^K \left| \frac{\partial|\lambda_k|}{\partial\alpha_i} \right|^2 + \sum_{i=0}^K \left| \frac{\partial|\lambda_k|}{\partial\beta_i} \right|^2 + \sum_{m=1}^p \delta(\eta_m) \left| \frac{\partial|\lambda_k|}{\partial\eta_m} \right|^2 \quad (5.13)$$

where $\{\partial|\lambda_k|/\partial\alpha_i\}$, $\{\partial|\lambda_k|/\partial\beta_i\}$ and $\{\partial|\lambda_k|/\partial\eta_m\}$ can be obtained using (3.3) once $\{\partial\lambda_k/\partial\alpha_i\}$, $\{\partial\lambda_k/\partial\beta_i\}$ and $\{\partial\lambda_k/\partial\eta_m\}$ are known.

The equivalent system matrix $X_{\rho I} \triangleq \begin{bmatrix} d & C_{\rho I} \\ B_{\rho I} & A_{\rho I} \end{bmatrix}$ of a ρ IDFIIt structure can be expressed as

$$X_{\rho I} = \begin{bmatrix} \beta_0 & \Delta_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \beta_1 - \beta_0\alpha_1 & \gamma_1 - \Delta_1\alpha_1 & \Delta_2 & 0 & \cdots & 0 & 0 & 0 \\ \beta_2 - \beta_0\alpha_2 & -\Delta_1\alpha_2 & \gamma_2 & \Delta_3 & \cdots & 0 & 0 & 0 \\ \beta_3 - \beta_0\alpha_3 & -\Delta_1\alpha_3 & 0 & \gamma_3 & \cdots & 0 & 0 & 0 \\ \beta_4 - \beta_0\alpha_4 & -\Delta_1\alpha_4 & 0 & -\eta_2\Delta_3 & \cdots & 0 & 0 & 0 \\ \vdots & & & & & & & \\ \beta_{K-1} - \beta_0\alpha_{K-1} & -\Delta_1\alpha_{K-1} & 0 & 0 & \cdots & 0 & \gamma_{K-1} & \Delta_K \\ \beta_K - \beta_0\alpha_K & -\Delta_1\alpha_K & 0 & 0 & \cdots & 0 & -\eta_p\Delta_{2p-1} & \gamma_K \end{bmatrix}.$$

It can be shown that $\frac{\partial \lambda_k}{\partial x} = \frac{\partial \lambda_k}{\partial X_{\rho_I}} \frac{\partial X_{\rho_I}}{\partial x}$, $x = \alpha_i, \beta_i, \eta_m$, which leads to

$$\begin{cases} \frac{\partial \lambda_k}{\partial \alpha_i} = -e_{i+1}^T \frac{\partial \lambda_k}{\partial X_{\rho_I}} (\beta_0 e_1 + \Delta_1 e_2), \forall i \\ \frac{\partial \lambda_k}{\partial \beta_0} = \begin{bmatrix} 1 & -\alpha_1 & -\alpha_2 & \cdots & -\alpha_K \end{bmatrix} \frac{\partial \lambda_k}{\partial X_{\rho_I}} e_1 \\ \frac{\partial \lambda_k}{\partial \beta_i} = e_{i+1}^T \frac{\partial \lambda_k}{\partial X_{\rho_I}} e_1, \forall i \neq 0 \\ \frac{\partial \lambda_k}{\partial \eta_m} = -e_{2m+1}^T \frac{\partial \lambda_k}{\partial X_{\rho_I}} e_{2m} \Delta_{2m-1}, \forall m \end{cases} \quad (5.14)$$

with e_i defined before and $\partial \lambda_k / \partial X_{\rho_I}$ computed by (2.17) - (2.18). Then the corresponding closed-loop stability measure μ_2 for the ρ IDFIIIt structure can be obtained using (3.1) with Φ_k given by (5.13).

Since different sets of $\{\gamma_k, \Delta_k, \bar{\eta}_m\}$ yield different μ_2 , the interesting problem is to maximize μ_2 with respect to these free parameters. Another point we should mention is that the dynamical range of the parameters $\{\alpha_k, \beta_k, \eta_m\}$ is determined by the three sets of free parameters. Noting that for a fixed-point implementation, it is desired that all the parameters $\{\alpha_k, \beta_k, \gamma_k, \Delta_k, \eta_m\}$ in the ρ IDFIIIt structure be absolutely bounded by one, which implies

$$|\alpha_k| \leq 1, \quad |\beta_k| \leq 1, \quad |\eta_m| \leq 1, \quad \forall k, m \quad (5.15)$$

with $\{\alpha_k, \beta_k\}$ computed using (5.6) and $\{\eta_m\}$ given by (5.3). Therefore, the optimal ρ IDFIIIt structure problem can be defined as:

$$\max_{\gamma_k \in S_\gamma, \Delta_k \in S_\Delta} \left\{ \max_{\bar{\eta}_m \in S_{\bar{\eta}}} \mu_2 \right\}, \quad \text{subject to (5.15)}. \quad (5.16)$$

It seems impossible to obtain an analytical solution to this problem due to the high nonlinearity of μ_2 in $\{\gamma_k\}$, $\{\Delta_k\}$ and $\{\bar{\eta}_m\}$. However, noting that the number of elements in both S_γ and S_Δ is finite, the first maximization

can be done using exhaustive searching, while for the second one, a standard optimization algorithm, such as the command *fmincon.m* in MATLAB, can be used to find the corresponding optimal $\{\bar{\eta}_m\}$ to a given combination of $\{\gamma_k\}$ and $\{\Delta_k\}$.

Noting the drawback of the exhaustive searching explained in Section 4.3, we use GA to attack the problem (5.16). Please refer to Appendix A for details.

Now let us consider how to choose the three sets of free parameters $\{\gamma_k\}$, $\{\Delta_k\}$ and $\{\bar{\eta}_m\}$ to maximize the stability robustness (minimize the pole sensitivities) for ρ IDFII structures.

For a ρ IDFII filter, similar to (4.18) obtained for ρ DFII structure, one has

$$\frac{\partial \lambda_k}{\partial \alpha_i} = -\frac{\kappa q_i(\lambda_k)}{\prod_{k \neq j}^K (\lambda_k - \lambda_j)}, \quad \forall i. \quad (5.17)$$

It can be shown that

$$q_k(z) = \left[\prod_{l=k+1}^K \Delta_l^{-1} \right] \bar{q}_k(z), \quad \forall k \quad (5.18)$$

where all $\bar{q}_k(z)$ are obtained using (5.2) with $\Delta_k = 1$, $\forall k$. Noting $\kappa = \prod_{k=1}^K \Delta_k$ and (5.18), (5.17) can be rewritten as

$$\frac{\partial \lambda_k}{\partial \alpha_i} = -\left[\prod_{l=1}^i \Delta_l \right] \frac{\bar{q}_i(\lambda_k)}{\prod_{k \neq j}^K (\lambda_k - \lambda_j)}, \quad \forall i. \quad (5.19)$$

For digital controller case, similar to (4.22), one has

$$\begin{cases} \frac{\partial \lambda_k}{\partial \alpha_i} = -\frac{\kappa q_i(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} D_p(\lambda_k), & \forall i \\ \frac{\partial \lambda_k}{\partial \beta_i} = \frac{\kappa q_i(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} N_p(\lambda_k), & \forall i \end{cases} \quad (5.20)$$

where $D_p(z)$ and $N_p(z)$, as defined before, are the denominator and numerator of the plant $P(z)$. With $\kappa = \prod_{k=1}^K \Delta_k$ and (5.18), it can be shown that

$$\begin{cases} \frac{\partial \lambda_k}{\partial \alpha_i} = -[\prod_{l=1}^i \Delta_l] \frac{\bar{q}_i(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} D_p(\lambda_k), & \forall i \\ \frac{\partial \lambda_k}{\partial \beta_i} = [\prod_{l=1}^i \Delta_l] \frac{\bar{q}_i(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} N_p(\lambda_k), & \forall i \neq 0 \\ \frac{\partial \lambda_k}{\partial \beta_0} = \frac{\bar{q}_0(\lambda_k)}{\prod_{k \neq j}^{N_0} (\lambda_k - \lambda_j)} N_p(\lambda_k). \end{cases} \quad (5.21)$$

From (5.19) and (5.21), one can see that if $\bar{q}_i(\lambda_k) \approx 0$, then $\partial \lambda_k / \partial \alpha_i \approx 0$ and $\partial \lambda_k / \partial \beta_i \approx 0$. This means that for digital systems implemented with ρ IDFII structures, when the roots of $\bar{q}_k(z)$ are chosen such that they are close to the system poles, the pole sensitivities can be much reduced.

In fact, for the ρ IDFII structure

$$\bar{q}_{K-2m}(z) = [(z - \gamma_{K-2m+1})(z - \gamma_{K-2m+2}) + \bar{\eta}_{p-m+1}] \bar{q}_{K-2m+2}(z)$$

and $\bar{q}_{K-2m-1}(z) = [z - \gamma_{K-2m}] \bar{q}_{K-2m}(z)$. It is interesting to note that $\bar{q}_{K-2m}(z)$ and $\bar{q}_{K-2m-1}(z)$ can have any assigned roots (both the real roots and complex ones) by choosing the three free parameters γ_{K-2m+1} , γ_{K-2m+2} and $\bar{\eta}_{p-m+1}$. Let $z = re^{\pm j\theta}$ be a pair of poles to be assigned to $\bar{q}_{K-2m}(z)$. Clearly, this can be achieved if

$$\begin{cases} \gamma_{K-2m+1} + \gamma_{K-2m+2} = 2r \cos \theta \\ \bar{\eta}_{p-m+1} + \gamma_{K-2m+1} \gamma_{K-2m+2} = r^2. \end{cases} \quad (5.22)$$

One of the solutions to the above equations is

$$\begin{cases} \gamma_{K-2m+1} = \gamma_{K-2m+2} = r \cos \theta \\ \bar{\eta}_{p-m+1} = r^2 \sin^2 \theta. \end{cases} \quad (5.23)$$

With $0 \leq r \leq 1$ and $|\gamma_k| \leq 1$, it follows from (5.22) that the range for $\{\bar{\eta}_m\}$ can be shown as

$$S_{\bar{\eta}} = \{\bar{\eta}_m : -1 \leq \bar{\eta}_m \leq 2\}. \quad (5.24)$$

Now we see that the roots of $\bar{q}_k(z)$ (both real roots and complex ones) can be expressed by γ_k and $\bar{\eta}_m$. So based on the former analysis below (5.21), our first observation is that when $\{\gamma_k\}, \{\bar{\eta}_m\}$ are chosen such that the roots of the polynomials $\bar{q}_k(z)$ are close to the dominant system poles ($\bar{q}_i(\lambda_k) \approx 0$), the pole sensitivities can be much reduced and hence the stability robustness μ_2 can be much enhanced. One can also see from (5.19) and (5.21) that the values of $\partial\lambda_k/\partial\alpha_i$, $\partial\lambda_k/\partial\beta_i$ depend partly on the expression $\prod_{l=1}^i \Delta_l$. For example, the value of $\partial\lambda_k/\partial\alpha_1$ depends on Δ_1 and the value of $\partial\lambda_k/\partial\alpha_i$ depends on $\Delta_1, \Delta_2, \dots, \Delta_i$. Besides, small Δ_k leads to small $\partial\lambda_k/\partial\alpha_i$. So our second observation is that of all the $\{\Delta_k\}$, Δ_1 affects the pole sensitivities $\{\partial\lambda_k/\partial\alpha_i, \partial\lambda_k/\partial\beta_i\}$ the most and the effect of each Δ_k on the pole sensitivities decreases one by one. In addition, smaller $\{\Delta_k\}$ will lead to smaller pole sensitivities and larger μ_2 in a certain sense.

These observations give us some insights on how to choose $\{\gamma_k\}$, $\{\bar{\eta}_m\}$ and $\{\Delta_k\}$ to maximize the stability measure μ_2 of ρ IDFIIt structures, which can accelerate the exhaustive searching process for the optimal structures and provide a good initial population to speed up GA for attacking the optimal structure problem. The effectiveness of the above observations has been confirmed by many numerical examples, including the one to be presented in the next section.

It is observed that for the ρ DFIIt structure proposed in Section 4.2, the roots of the corresponding polynomials are from the set $\{\gamma_k\}$, which is real, while

the system poles are generally complex. Therefore, the degree of maximizing μ_2 for the ρ DFIIt structure is limited. With the extra free parameters $\{\bar{\eta}_m\}$ introduced, the polynomials $\bar{q}_k(z)$ for the ρ IDFIIt structure can have complex roots and consequently, the stability robustness of the ρ IDFIIt structure can be maximized more effectively.

5.2.2 Optimized ρ_I -realizations

Consider a digital system implemented with a ρ_I -realization, which can be rewritten in the following form

$$\left\{ \begin{array}{l} A_{\rho_I} \triangleq \begin{bmatrix} a_{11} & \Delta_2 & 0 & \cdots & 0 & 0 & 0 \\ a_{21} & \gamma_2 & \Delta_3 & \cdots & 0 & 0 & 0 \\ a_{31} & 0 & \gamma_3 & \cdots & 0 & 0 & 0 \\ a_{41} & 0 & a_{43} & \cdots & 0 & 0 & 0 \\ & & & \ddots & & & \\ a_{(K-1)1} & 0 & 0 & \cdots & 0 & \gamma_{K-1} & \Delta_K \\ a_{K1} & 0 & 0 & \cdots & 0 & a_{(2p)(2p-1)} & \gamma_K \end{bmatrix} \\ B_{\rho_I} \triangleq \begin{bmatrix} b_1 & b_2 & \cdots & b_{K-1} & b_K \end{bmatrix}^T \\ C_{\rho_I} \triangleq \begin{bmatrix} \Delta_1 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad d \triangleq \beta_0. \end{array} \right. \quad (5.25)$$

with

$$\left\{ \begin{array}{l} a_{11} = \gamma_1 - \Delta_1 \alpha_1 \\ a_{21} = -\Delta_1(\alpha_2 + \eta_1) \\ a_{k1} = -\Delta_1 \alpha_k, \quad k \in \{3, 4, \dots, K\} \\ a_{(2m)(2m-1)} = -\eta_m \Delta_{2m-1}, \quad m \in \{2, 3, \dots, p\} \\ b_k = \beta_k - \beta_0 \alpha_k, \quad k \in \{1, 2, \dots, K\}. \end{array} \right. \quad (5.26)$$

In this case, the only parameters to be implemented in (5.25) are $\{a_{k1}\}$, $\{a_{(2m)(2m-1)}\}$, $\{b_k\}$, d , $\{\gamma_k\}$ and $\{\Delta_k\}$. Noting that $\{\gamma_k\}$ and $\{\Delta_k\}$ are assumed to be of an exact FWL format, one can compute the corresponding μ_2 for a ρ_I -realization using (3.1), where the pole modulus sensitivity measure Φ_k is given by

$$\begin{aligned} \Phi_k = & \sum_{i=1}^K \left| \frac{\partial |\lambda_k|}{\partial a_{i1}} \right|^2 + \varepsilon \left(\sum_{i=1}^K \left| \frac{\partial |\lambda_k|}{\partial b_i} \right|^2 + \left| \frac{\partial |\lambda_k|}{\partial d} \right|^2 \right) \\ & + \sum_{m=2}^p \delta(a_{(2m)(2m-1)}) \left| \frac{\partial |\lambda_k|}{\partial a_{(2m)(2m-1)}} \right|^2 \end{aligned} \quad (5.27)$$

with $\varepsilon = 0$ for digital filters and $\varepsilon = 1$ for closed-loop systems.

Similarly, one has the following dynamical range constraints

$$|a_{k1}| \leq 1, \quad |a_{(2m)(2m-1)}| \leq 1, \quad |b_k| \leq 1, \quad \forall k, m \quad (5.28)$$

and hence the optimal ρ_I -realization problem is

$$\max_{\gamma_k \in S_\gamma, \Delta_k \in S_\Delta} \left\{ \max_{\bar{\eta}_m \in S_{\bar{\eta}}} \mu_2 \right\}, \quad \text{subject to (5.28)}. \quad (5.29)$$

This problem can be solved using the same way as the one for the optimal ρ IDFIIIt structure problem (5.16).

5.3 Numerical examples

In Section 4.4, numerical examples have demonstrated the very good stability performance of the optimized ρ -operator-based structures (R_ρ^{opt} , ρ DFIIIt). Each example also confirms the argument that the optimal $\{\gamma_k\}$ for these structures

are close to the system poles.

As we have discussed earlier in this chapter, the polynomials of ρ -operator-based structures have real roots (say $\{\gamma_k\}$) only while the system poles are generally complex. For those systems whose poles are far away from the real axis, the ρ -operator-based structures would not be effective. Due to this limitation, we have developed the second-order operator-based structures (R_{ρ_I} , ρ IDFIIt) in this chapter, for which the polynomials can have any assigned roots and the stability robustness can be further maximized.

We recall the examples used in Section 4.4. In Example 4.1, the filter poles are very close to the real axis. The same thing occurs for the dominant closed-loop poles in Example 4.3. In such cases, the ρ -operator-based structures are sufficient to enhance the stability robustness. The use of second-order operator-based structures leads to a very small (near to zero) improvement. However, for Example 4.2, the filter poles are far away from the real axis, then the ρ -operator-based structures, as to be seen later, will no longer be the best. The advantage of adopting second-order operator-based structures will be shown in this section by using Example 4.2 and a new example for closed-loop control system.

For the optimal structure problems (5.16) and (5.29), S_γ and S_Δ are the same as in Section 4.4 and $S_{\bar{\eta}}$ is defined by (5.24). A genetic algorithm (GA) is used to attack the problems and the results are confirmed with the exhaustive searching.

Example 5.1: Recall the six-order bandpass Butterworth filter presented in Example 4.2, where the optimized ρ -operator-based structures R_ρ^{opt} and ρ DFIIt have been examined. The former, having 22 nontrivial parameters, yields a

stability robustness of 6.38×10^{-2} , and the latter, possessing 22 nontrivial parameters, has a stability robustness of 6.47×10^{-2} . They both beat the fully parametrized optimal realization R_n in terms of stability performance and implementation efficiency.

Now we solve (5.29) with a GA program and then obtain an optimal ρ_I -realization, denoted by $R_{\rho_I}^{opt}$, for which $\{\gamma_k\} = \{-0.75, -0.5, -0.5, -0.5, -0.5, -1\}$, $\{\Delta_k\} = \{1, 0.25, 0.5, 0.5, 0.75, 0.25\}$, and $\{\bar{\eta}_m\} = \{0.2217, 0.5, 0.1875\}$. The corresponding optimal realization is given by

$$\begin{aligned} A_{\rho_I}^{opt} &= \begin{bmatrix} -0.5391 & 0.25 & 0 & 0 & 0 & 0 \\ -0.8537 & -0.5 & 0.5 & 0 & 0 & 0 \\ -0.3100 & 0 & -0.5 & 0.5 & 0 & 0 \\ -0.4733 & 0 & -1 & -0.5 & 0.75 & 0 \\ 0.2458 & 0 & 0 & 0 & -0.5 & 0.25 \\ -0.5888 & 0 & 0 & 0 & -0.75 & -1 \end{bmatrix} \\ B_{\rho_I}^{opt} &= \begin{bmatrix} -0.0641 & 0.1158 & 0.6821 & -0.9938 & -0.3824 & 0.8343 \end{bmatrix}^T \\ C_{\rho_I}^{opt} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Solving (5.16) leads to an optimal ρ IDFIIIt structure, denoted as ρ IDFIIIt. The corresponding optimal free parameters are

$$\begin{aligned} \{\gamma_k\} &= \{-0.75, -0.5, -0.5, -0.5, -0.75, -0.75\}, \\ \{\Delta_k\} &= \{0.25, 1, 0.5, 1, 0.25, 0.5\}, \\ \{\bar{\eta}_m\} &= \{0.25, 0.5, 0.125\}. \end{aligned}$$

Table 5.1: Comparison of two types of optimal sparse structures

	R_ρ^{opt}	$R_{\rho_I}^{opt}$	$\rho\text{DFII}t$	$\rho\text{IDFII}t$
μ_2	6.38×10^{-2}	2.16×10^{-1}	6.47×10^{-2}	2.47×10^{-1}
N_p	22	23	22	23
DR	Yes	Yes	Yes	Yes

The optimal $\rho\text{IDFII}t$ structure can be implemented using the structure form depicted in Figures 5.1 and 5.2 with $\{\gamma_k\}$, $\{\Delta_k\}$ above and

$$\begin{aligned}\{\alpha_k\} &= \{1.0000, -0.8437, -0.1463, 0.3100, 0.2366, -0.3687, 0.2572\}, \\ \{\beta_k\} &= \{0.0181, -0.2715, 0.1131, 0.6878, -0.4926, -0.5803, 0.3436\}, \\ \{\eta_m\} &= \{1, 1, 1\}.\end{aligned}$$

Computations show that for $R_{\rho_I}^{opt}$, the set of the roots for all the polynomials $\bar{q}_k(z)$ is $\{-0.5 \pm j0.7071, -0.6250 \pm j0.4540, -0.75 \pm j0.3536, -0.5, -1\}$, and for $\rho\text{IDFII}t$, the roots are $\{-0.5 \pm j0.7071, -0.6250 \pm j0.4841, -0.75 \pm 0.3536, -0.5, -0.75\}$. It is easy to see that both sets of the roots are very close to the poles of the filter $\{-0.4122 \pm j0.7106, -0.5612 \pm j0.4412, -0.7962 \pm 0.4181\}$, which consequently, leads to a very large stability robustness for the above two structures.

Table 5.1 compares the performance of the two types of optimal sparse structures, which are the ρ -operator-based structures (R_ρ^{opt} , $\rho\text{DFII}t$) and the second-order operator-based ones ($R_{\rho_I}^{opt}$, $\rho\text{IDFII}t$). It is shown that $R_{\rho_I}^{opt}$ and $\rho\text{IDFII}t$ yield much larger stability robustness than R_ρ^{opt} and $\rho\text{DFII}t$, respectively, while the price paid for that is just 1 more nontrivial parameters. The second-order

operator-based structures are also very efficient for implementation and satisfy the dynamical range constraints.

Figure 5.3: Pole distributions of R_{ρ}^{opt} and $R_{\rho_I}^{opt}$

Figure 5.4: Pole distributions of ρ_{DFIIt} and ρ_{IDFIIt}

With the same succession of ten perturbations as in Example 4.2, the pole distributions of perturbed structures for $R_{\rho_I}^{opt}$ and ρ_{IDFIIt} are obtained through simulations, which are presented in Figures 5.3 and 5.4, respectively. For comparison, the pole distributions for R_{ρ}^{opt} and ρ_{DFIIt} , already shown in Exam-

ple 4.2, are recalled here and joined into the figures. The advantages of $R_{\rho_I}^{opt}$ and ρ IDFIIt over R_{ρ}^{opt} and ρ DFIIt in terms of stability performance are clearly demonstrated by Figures 5.3 and 5.4, which confirms with the computation results shown in Table 5.1.

Example 5.2: Let us consider a discrete-time control system, where the digital plant $P(z)$ and controller $H(z)$ are given by the following canonical realizations, denoted as $R_p \triangleq (A_p, B_p, C_p)$ and $R_c \triangleq (A_c, B_c, C_c, d)$, respectively, with

$$A_p = \begin{bmatrix} 3.7174 & 1 & 0 & 0 & 0 \\ -5.7458 & 0 & 1 & 0 & 0 \\ 4.6673 & 0 & 0 & 1 & 0 \\ -2.0336 & 0 & 0 & 0 & 1 \\ 0.3953 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_p = \begin{bmatrix} 0.0018 \\ 0.0003 \\ -0.0163 \\ 0.0011 \\ 0.0016 \end{bmatrix}$$

$$C_p = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$A_c = \begin{bmatrix} 3.6172 & 1 & 0 & 0 & 0 & 0 \\ -5.9513 & 0 & 1 & 0 & 0 & 0 \\ 5.6335 & 0 & 0 & 1 & 0 & 0 \\ -3.2509 & 0 & 0 & 0 & 1 & 0 \\ 1.0895 & 0 & 0 & 0 & 0 & 1 \\ -0.1690 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_c = \begin{bmatrix} 0.2258 \\ -0.6588 \\ 0.8195 \\ -0.5320 \\ 0.1814 \\ -0.0234 \end{bmatrix}$$

$$C_c = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad d = 0.0577.$$

The corresponding poles of the closed-loop system are $\{0.4523 \pm j0.5315, 0.4837 \pm j0.4556, 0.6055 \pm j0.4108, 0.7814 \pm j0.3099, 0.8886 \pm j0.3326, 0.9113\}$.

Taking R_c as the initial realization, we solve (3.20) using the MATLAB algorithm *fminimax.m* and eventually obtain a fully parameterized optimal controller realization R_f^{opt} . Solving (5.29) with $\bar{\eta}_m = 0, \forall m$, one gets an optimal ρ -realization, denoted as R_ρ^{opt} , which is given by

$$\begin{aligned} A_\rho^{opt} &= \begin{bmatrix} 0.1172 & 0.75 & 0 & 0 & 0 & 0 \\ -0.8881 & 0.5 & 0.5 & 0 & 0 & 0 \\ -0.8297 & 0 & 0.75 & 0.5 & 0 & 0 \\ -0.7850 & 0 & 0 & 0.75 & 0.25 & 0 \\ -0.5211 & 0 & 0 & 0 & 0.75 & 0.25 \\ -0.5515 & 0 & 0 & 0 & 0 & 0.75 \end{bmatrix} \\ B_\rho^{opt} &= \begin{bmatrix} 0.4515 & 0.3503 & 0.6026 & 0.4355 & 0.4987 & 0.7221 \end{bmatrix}^T \\ C_\rho^{opt} &= \begin{bmatrix} 0.5 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

With $S_{\bar{\eta}}$ defined by (5.24), the optimization problem (5.29) is attacked and an optimal ρ_I -realization $R_{\rho_I}^{opt}$ is obtained:

$$\begin{aligned} A_{\rho_I}^{opt} &= \begin{bmatrix} -0.1328 & 0.75 & 0 & 0 & 0 & 0 \\ -0.8838 & 1 & 0.25 & 0 & 0 & 0 \\ -0.2061 & 0 & 0.5 & 0.25 & 0 & 0 \\ -0.5075 & 0 & -0.5716 & 0.75 & 0.75 & 0 \\ -0.1427 & 0 & 0 & 0 & 0.5 & 0.25 \\ -0.0764 & 0 & 0 & 0 & -0.5740 & 1 \end{bmatrix} \\ B_{\rho_I}^{opt} &= \begin{bmatrix} 0.4515 & 0.5008 & 0.3156 & 0.2918 & 0.2286 & 0.7193 \end{bmatrix}^T \\ C_{\rho_I}^{opt} &= \begin{bmatrix} 0.5 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Table 5.2: Comparison of different state-space realizations (Example 5.2)

	R_c	R_f^{opt}	R_ρ^{opt}	$R_{\rho_I}^{opt}$
μ_2	4.46×10^{-4}	7.38×10^{-3}	3.45×10^{-2}	6.86×10^{-2}
μ_0^*	1.31×10^{-3}	1.33×10^{-2}	9.76×10^{-2}	1.25×10^{-1}
N_p	13	49	24	24
DR	No	No	Yes	Yes

For each state-space realization obtained above, μ_2 is computed and the estimate μ_0^* of μ_0 is obtained via simulations in the same way as before. The computation results are presented in Table 5.2.

Similarly, solving (5.16), we get an optimal ρ IDFIIt structure, denoted as ρ IDFIIt, which can be implemented with the structure form depicted in Figures 5.1 and 5.2, where the corresponding parameter sets are given by

$$\begin{aligned}
\{\alpha_k\} &= \{1.0000, 0.8828, 0.9766, 0.7765, 0.1297, 0.1950, -0.7180\}, \\
\{\beta_k\} &= \{0.0577, 0.2767, 0.3820, 0.4216, 0.0673, 0.2309, -0.0692\}, \\
\{\gamma_k\} &= \{0.5, 1, 0.75, 0.5, 0.75, 1\}, \\
\{\Delta_k\} &= \{1, 0.75, 0.25, 0.5, 0.25, 0.5\}, \\
\{\eta_m\} &= \{0.25, 1, 1\}.
\end{aligned}$$

For comparison, the optimization problem (5.16) is solved with $\bar{\eta}_m = 0, \forall m$ and an optimal ρ DFIIt structure is obtained, for which the corresponding parameters

Table 5.3: Comparison of different DFIIIt structures (Example 5.2)

	$z\text{DFIIIt}$	δDFIIIt	ρDFIIIt	$\rho\text{IDFIIIt}$
μ_2	4.41×10^{-4}	9.48×10^{-4}	3.06×10^{-2}	5.34×10^{-2}
μ_0^*	1.29×10^{-3}	3.58×10^{-3}	7.47×10^{-2}	1.17×10^{-1}
N_p	13	13	23	23
DR	No	No	Yes	Yes

are:

$$\{\alpha_k\} = \{1.0000, 0.8828, 0.8881, 0.8297, 0.7850, 0.5211, 0.5515\},$$

$$\{\beta_k\} = \{0.0577, 0.2767, 0.2264, 0.3492, 0.2630, 0.2794, 0.3929\},$$

$$\{\gamma_k\} = \{1, 0.5, 0.75, 0.75, 0.75, 0.75\},$$

$$\{\Delta_k\} = \{1, 0.75, 0.5, 0.5, 0.25, 0.25\}.$$

We also compute $z\text{DFIIIt}$ and δDFIIIt , which are two classical DFIIIt structures defined before. Table 5.3 compares the performance of different DFIIIt structures.

Comment 5.1: From this example, one can observe that for each structure the measure μ_2 is a lower bound of μ_0^* which is an estimate of μ_0 by simulations as mentioned before. The canonical controller realization R_c and the classical DFIIIt structures $z\text{DFIIIt}$, δDFIIIt are the most sparse structures with only 13 nontrivial parameters, however, their stability performance are very poor. The optimal state-space realization R_f^{opt} has a much better stability behavior, but it is fully parametrized and all the 49 elements are nontrivial.

One can also see that both R_ρ^{opt} and $R_{\rho_I}^{\text{opt}}$ outperform R_f^{opt} greatly in terms of not only implementation efficiency but also stability robustness. These two

realizations also satisfy the parameter dynamical range constraints, which is desired for a fixed-point implementation. Moreover, it is important to note that, of all the four state-space realizations in Table 5.2, $R_{\rho_I}^{opt}$ yields the largest stability robustness which is nearly twice of that by R_{ρ}^{opt} , while the former has the same number of nontrivial parameters with the latter.

Comment 5.2: For the DFIIIt structures in Table 5.3, both ρ DFIIIt and ρ IDFIIIt yield a much better performance than R_f^{opt} and satisfy the dynamical range constraints. These two structures are also very sparse and each has only 23 nontrivial parameters which are less than half of those in R_f^{opt} . It is interesting to note that ρ IDFIIIt has a larger stability robustness than ρ DFIIIt without increasing any implementation complexity. This is due to the fact that some of the $\{\eta_m\}$ are trivial in the optimal ρ IDFIIIt structure.

Comment 5.3: Our argument on how to choose the polynomials with those free parameters to maximize the stability robustness is also confirmed by this example. It is interesting to note that for $R_{\rho_I}^{opt}$, the set of the roots for all the polynomials $\bar{q}_k(z)$ is $\{0.6250 \pm j0.3568, 0.75 \pm j0.2846, 0.8750 \pm j0.3394, 0.75, 1\}$, and for ρ IDFIIIt, the roots are $\{0.6250 \pm j0.3307, 0.75 \pm j0.3536, 0.8750 \pm 0.3307, 0.5, 1\}$. Each set of roots is very close to the dominant closed-loop poles $\{0.6055 \pm j0.4108, 0.7814 \pm j0.3099, 0.8886 \pm 0.3326\}$. As a result of that, the above two optimized sparse structures both exhibit high degree of stability in this example and thus validate our theoretical analysis.

Through the above comments for this example, one can see that our proposed structures outperform those traditional structures greatly in terms of both stability robustness and implementation efficiency. Generally speaking, this performance improvement is also applicable for many different examples

Figure 5.5: Unit impulse responses of the closed-loop system for ideal controller (-), 10-bit implemented R_c (...), and 8-bit implemented R_f^{opt} (- -).

and the degree of improvement is decided by the examples.

We also compute the unit impulse responses of the closed-loop control system for the ideal controller (the one implemented with infinite precision) and those FWL controllers implemented with various structures in a certain bits. The results are shown from Figures 5.5 to 5.7.

Firstly, Figure 5.5 shows that the closed-loop becomes unstable with a 10-bit implemented R_c , however, the closed-loop system remains stable and well-behaved with the 8-bit implemented R_f^{opt} . Secondly, from Figure 5.6, one can see that the closed-loop responses for 4-bit implemented R_ρ and R_{ρ_I} are very close to the ideal one and R_{ρ_I} yields a even better response than R_ρ . Finally, Figure 5.7 demonstrates that the closed-loop response corresponding to 10-bit implemented z DFIIt is very bad, while those for ρ DFIIt, ρ IDFIIt are far better, and ρ IDFIIt yields a response which is almost the same as the ideal one.

Figure 5.6: Unit impulse responses of the closed-loop system for ideal controller (-), 4-bit implemented R_{ρ}^{opt} (- -), and 4-bit implemented $R_{\rho_I}^{opt}$ (-.-).

Figure 5.7: Unit impulse responses of the closed-loop system for ideal controller (-), 10-bit implemented $zDFIIt$ (...), 4-bit implemented $\rho DFIIt$ (- -), and 4-bit implemented $\rho IDFIIt$ (-.-).

This example clearly demonstrates the effectiveness of our design procedure for optimal sparse controller structures with higher degree of stability and confirms our theoretical results.

5.4 Summary

In this chapter, two new sparse structures, called ρ IDFIIt structure and ρ_I -realization, have been developed, which can be considered as the improved version of the ρ -operator-based structures (ρ DFIIt structure and ρ -realization) derived in Chapter 4, where the first-order ρ -operators are replaced with a set of second-order operators. The stability behavior of each structure has been analyzed by deriving the corresponding expression of the pole modulus sensitivity-based stability measure proposed in Chapter 3. The optimal structure problem has been formulated and then investigated with the established relationship between the stability robustness measure, the system poles and the roots of polynomials. Although no analytical solution can be found due to the high nonlinearity involved in the stability robustness maximization, the problem has been effectively attacked under the parameter dynamical range constraints by using a genetic algorithm.

Two design examples have been given to verify the theoretical results and to illustrate the performance of different structures. It is shown that the optimized ρ IDFIIt structure and ρ_I -realization not only yield a much better stability performance than the optimal ρ DFIIt structure and ρ -realization, respectively, but also beat the fully parametrized optimal realization in terms of both stability robustness and computation efficiency.

Chapter 6

Optimal Roundoff Noise Controller Structure Design

The effects of roundoff noise have been well studied in digital signal processing, particularly in digital filter implementation, but have received much less attention in design of digital controllers for the closed-loop control system. Since there exist some parallels between filter and controller implementation, some of the existing concepts on roundoff noise analysis for filter design can be applied to control problems with necessary adaptations.

This chapter presents a study on the effect of roundoff noise in a digital controller for a discrete-time control system. We have two main objectives. The first one is to derive an analytical expression of the roundoff noise gain for state-space controller realizations and then solve the corresponding optimal roundoff noise realization problem. Noting the optimal realizations are usually fully parametrized, which increase the implementation complexity, we desire the controller be implemented with some sparse structures that also have very good

roundoff noise performance.

The sparseness issue for optimal controller structure design with minimum closed-loop roundoff noise gain seems to be a totally new topic. It has been shown in Chapter 5 that the second-order operator-based structures (the ρ IDFIIt structure and ρ_I -realization) can be optimized to exhibit very nice stability behavior. Now the question is whether these two structures can be optimized to achieve minimum roundoff noise gain performance. The answer to this question constitutes the second objective of this chapter.

In what follows, only investigations on the optimal roundoff noise control structures are provided for closed-loop systems, since similar results can be obtained for digital filters by simplifying the analysis for the closed-loop system into that for the open-loop case.

6.1 Optimal roundoff noise controller realizations

Consider a discrete-time feedback control system depicted in Figure 2.3, where $P(z)$ is the discrete-time plant and $H(z)$ is a digital controller implemented using the state-space equations defined by (2.4). It is well known that in an implementation system, all the signals should be sustained within a certain dynamic range in order to avoid overflow. Under the assumption that the input $r(n)$ and the output $u(n)$ (here $u(n) = w(n)$) of the closed-loop system are properly pre-scaled, the only signals which may have overflow are the elements of the controller state vector $x(n)$, which, therefore, have to be scaled.

Suppose the plant $P(z)$ is strictly proper and has a realization (A_p, B_p, C_p) . Noting that $u(n) = P(z)[r(n) + y(n)]$, one has

$$\begin{cases} v(n+1) &= A_p v(n) + B_p[r(n) + y(n)] \\ u(n) &= C_p v(n) \end{cases} \quad (6.1)$$

where $A_p \in \mathcal{R}^{J \times J}$, $B_p \in \mathcal{R}^{J \times 1}$, $C_p \in \mathcal{R}^{1 \times J}$. Denote

$$x_{cl}^T(n) \triangleq [v^T(n) \quad x^T(n)]$$

with $\mathcal{T}$ the transpose operator. It follows from (2.4) and (6.1) that

$$x_{cl}(n+1) = A_{cl}x_{cl}(n) + B_{cl}r(n)$$

where

$$A_{cl} = \begin{bmatrix} A_p + dB_pC_p & B_pC \\ BC_p & A \end{bmatrix}, \quad B_{cl} = \begin{bmatrix} B_p \\ \mathbf{0} \end{bmatrix} \quad (6.2)$$

with $\mathbf{0}$ denoting the zero vector of appropriate dimension.

There exist different scaling schemes for preventing variables from overflow. The popularly used ones are the l_2 - and l_∞ -scalings. In what follows, we will concentrate on the l_2 -scaling scheme. The l_2 -scaling means that each element of the controller state vector $x(n)$ should have a unit variance when the input $r(n)$ is a white noise with a unit variance, for which the probability distribution can be either Gaussian or uniform distributions. The l_2 -scaling can be achieved if

$$\bar{\mathcal{K}}(l, l) = 1, \quad l = J+1, J+2, \dots, J+K \quad (6.3)$$

where $\bar{\mathcal{K}}$ is given by

$$\bar{\mathcal{K}} = \sum_{k=0}^{+\infty} A_{cl}^k B_{cl} B_{cl}^T (A_{cl}^T)^k \quad (6.4)$$

satisfying

$$\bar{\mathcal{K}} = A_{cl} \bar{\mathcal{K}} A_{cl}^T + B_{cl} B_{cl}^T$$

with A_{cl} and B_{cl} given by (6.2). In the sequel, all the structures under discussion are assumed to be l_2 -scaled.

6.1.1 Derivation of roundoff noise gain

It should be pointed out that the state-space model (2.4) is the digital controller implemented with infinite precision. Though there exist different state-space realizations, they yield exactly the same performance - the desired one. In practice, however, a designed digital controller has to be implemented with finite precision and a rounding operation has to be applied if less-than-double precision fixed-point arithmetic is utilized.

Assuming rounding occurs after multiplication (RAM), as mentioned in Section 2.3, then a more practical digital controller model is

$$\begin{cases} x^*(n+1) &= Q[Ax^*(n)] + Q[Bu^*(n)] \\ y^*(n) &= Q[Cx^*(n)] + Q[du^*(n)] \end{cases} \quad (6.5)$$

where the coefficients in the realization (A, B, C, d) are assumed to be implemented exactly with B_τ bits and $Q[x]$ is the quantizer that rounds x to B_s bits in fractional part.

Let τ be a parameter in a controller structure and $Q[\tau s(n)]$ the quantized version of the product $\tau s(n)$. The roundoff noise due to the parameter τ can be defined as

$$\phi(\tau)\epsilon_\tau(n) \triangleq Q[\tau s(n)] - \tau s(n)$$

where $\phi(\tau) = 1$ if τ is nontrivial, otherwise, $\phi(\tau) = 0$. In fact, the function $\phi(\tau)$ is used for indicating the fact that τ produces no roundoff noise when it is trivial. Denote $\Delta u(n)$ as the corresponding output deviation of the closed-loop system to $\phi(\tau)\epsilon_\tau(n)$ and $F(z)$ as the transfer function between $\phi(\tau)\epsilon_\tau(n)$ and $\Delta u(n)$. It is well known (see, e.g., [7]) that $\Delta u(n)$ is a stationary process and the variance $E[(\Delta u(n))^2] = \phi(\tau)\|F(z)\|_2^2 E[\epsilon_\tau^2(n)]$. Then the roundoff noise gain for τ is defined as

$$G_\tau \triangleq \frac{E[(\Delta u(n))^2]}{E[\epsilon_\tau^2(n)]} = \phi(\tau)\|F(z)\|_2^2 \quad (6.6)$$

where $\|\cdot\|_2$ is the L_2 -norm:

$$\begin{aligned} \|F(z)\|_2 &\triangleq \left\{ \frac{1}{2\pi} \int_0^{2\pi} \sum_{i=1}^l \sum_{k=1}^m |f_{ik}(e^{j\omega})|^2 d\omega \right\}^{1/2} \\ &= \left\{ \text{tr} \left[\frac{1}{j2\pi} \oint_{|z|=1} F(z) F^H(z) z^{-1} dz \right] \right\}^{1/2} \end{aligned} \quad (6.7)$$

with $F(z) = \{f_{ik}(z)\} \in \mathcal{R}^{l \times m}$. Let $F(z) = D + \mathcal{L}(zI - \Phi)^{-1}\mathcal{J}$. It can be shown that

$$\|F(z)\|_2^2 = \text{tr}[DD^T + \mathcal{L}W_c\mathcal{L}^T] = \text{tr}[D^T D + \mathcal{J}^T W_o \mathcal{J}] \quad (6.8)$$

where W_c, W_o are the controllability and observability Gramians of the realization $(\Phi, \mathcal{J}, \mathcal{L}, D)$, respectively.

Now we consider a digital controller implemented with a state-space realization (A, B, C, d) . Let μ_{ik} be a parameter in the matrix $[A \ B]$ and ν_k be an

element in the row vector $[C \ d]$. It is easy to see that the roundoff noise due to μ_{ik} occurs only in the first equation of the state-space model, while the roundoff noise due to ν_k appears only in the second one. Therefore, in what follows, we compute the roundoff noise gain due to μ_{ik} and ν_k separately.

Firstly, let us look at the effect of roundoff noise due to μ_{ik} on the closed-loop output. In that case, (6.5) becomes

$$\begin{cases} x^*(n+1) &= Ax^*(n) + Bu^*(n) + e_i \phi(\mu_{ik}) \epsilon_{\mu_{ik}}(n) \\ y^*(n) &= Cx^*(n) + du^*(n) \end{cases} \quad (6.9)$$

where $\phi(\mu_{ik}) \epsilon_{\mu_{ik}}(n)$ is the roundoff noise due to μ_{ik} and $e_i \in \mathcal{R}^{K \times 1}$ denotes the i th elementary vector whose elements are all zero except the i th one which is 1. It then follows from (2.4) and (6.9) that

$$\begin{cases} \Delta x(n+1) &= A\Delta x(n) + B\Delta u(n) + e_i \phi(\mu_{ik}) \epsilon_{\mu_{ik}}(n) \\ \Delta y(n) &= C\Delta x(n) + d\Delta u(n) \end{cases}$$

where

$$\Delta u(n) = u^*(n) - u(n), \quad \Delta x(n) = x^*(n) - x(n), \quad \Delta y(n) = y^*(n) - y(n).$$

Noting $\Delta u(n) = P(z)\Delta y(n)$, one can show that

$$\begin{aligned} \Delta u(n) &= C_d[zI - A_{cl}]^{-1} \begin{bmatrix} \mathbf{0} \\ e_i \end{bmatrix} \phi(\mu_{ik}) \epsilon_{\mu_{ik}}(n) \\ &\triangleq H_i(z) \phi(\mu_{ik}) \epsilon_{\mu_{ik}}(n) \end{aligned}$$

where A_{cl} is given in (6.2) and

$$C_{cl} = \begin{bmatrix} C_p & \mathbf{0} \end{bmatrix}. \quad (6.10)$$

Therefore, it follows from (6.6) to (6.8) that the roundoff noise gain due to the parameter μ_{ik} is given by

$$G_{\mu_{ik}} = \phi(\mu_{ik}) \|H_i(z)\|_2^2 = \phi(\mu_{ik}) \begin{bmatrix} \mathbf{0} & e_i^T \end{bmatrix} \bar{W} \begin{bmatrix} \mathbf{0} \\ e_i \end{bmatrix}$$

where $\bar{W}$ is the observability Gramian of the realization (A_{cl}, B_{cl}, C_{cl}) , which is defined as

$$\bar{W} = \sum_{k=0}^{+\infty} (A_{cl}^T)^k C_{cl}^T C_{cl} A_{cl}^k \quad (6.11)$$

satisfying

$$\bar{W} = A_{cl}^T \bar{W} A_{cl} + C_{cl}^T C_{cl}.$$

Let

$$\bar{W} \triangleq \begin{bmatrix} U & W_{12} \\ W_{21} & W \end{bmatrix}$$

have the same partition as A_{cl} . Clearly, one has

$$G_{\mu_{ik}} = \phi(\mu_{ik}) e_i^T W e_i.$$

Similarly, one can analyze the roundoff noise gain due to the parameter ν_k .

Using the same procedure, one can show that

$$\begin{cases} \Delta x(n+1) &= A\Delta x(n) + B\Delta u(n) \\ \Delta y(n) &= C\Delta x(n) + d\Delta u(n) + \phi(\nu_k)\epsilon_{\nu_k}(n) \end{cases}$$

where $\phi(\nu_k)\epsilon_{\nu_k}(n)$ denotes the corresponding roundoff noise. With some manipulations, one has

$$\begin{aligned} \Delta u(n) &= C_{cl}[zI - A_{cl}]^{-1} \begin{bmatrix} B_p \\ \mathbf{0} \end{bmatrix} \phi(\nu_k)\epsilon_{\nu_k}(n) \\ &\triangleq H_\nu(z)\phi(\nu_k)\epsilon_{\nu_k}(n). \end{aligned}$$

Then the roundoff noise gain due to the parameter ν_k is

$$\begin{aligned} G_{\nu_k} &= \phi(\nu_k) \|H_\nu(z)\|_2^2 = \phi(\nu_k) \text{tr} \left(\begin{bmatrix} B_p^T & \mathbf{0} \end{bmatrix} \bar{W} \begin{bmatrix} B_p \\ \mathbf{0} \end{bmatrix} \right) \\ &= \phi(\nu_k) \text{tr}(B_p^T U B_p). \end{aligned}$$

Therefore, the total roundoff noise gain of the state-space controller realization is defined as

$$G_s \triangleq \sum_{i=1}^K \sum_{k=1}^{K+1} G_{\mu_{ik}} + \sum_{k=1}^{K+1} G_{\nu_k}.$$

It is easy to obtain that

$$G_s = \text{tr}(WQ) + m_{K+1} \text{tr}(B_p^T U B_p) \quad (6.12)$$

where $Q = \text{diag}(m_1, \dots, m_k, \dots, m_K)$ with m_k the number of nontrivial param-

ters in the k th row of $[A \ B]$ for $k = 1, \dots, K$, and m_{K+1} the number of nontrivial parameters in the row vector $[C \ d]$.

6.1.2 Optimal fully parametrized state-space realizations

Let (A_{cl}, B_{cl}, C_{cl}) and $(A_{cl}^0, B_{cl}^0, C_{cl}^0)$ be two realizations of the closed-loop system with A_{cl}, B_{cl} defined in (6.2) and C_{cl} in (6.10), corresponding to the two digital controller realizations $R \triangleq (A, B, C, d)$ and $R_0 \triangleq (A_0, B_0, C_0, d)$ which are related with (2.8), respectively. It can be shown that

$$\begin{cases} A_{cl} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-1} A_{cl}^0 \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix} \\ B_{cl} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-1} B_{cl}^0 \\ C_{cl} = C_{cl}^0 \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix} \end{cases} \quad (6.13)$$

It then follows from (6.13) that

$$\bar{W} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^T \bar{W}^0 \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}$$

where $\bar{W}^0$ is similar to $\bar{W}$ defined in (6.11) but corresponds to the controller realization R_0 .

$$\text{Let } \bar{W} \triangleq \begin{bmatrix} U & W_{12} \\ W_{21} & W \end{bmatrix} \text{ and } \bar{W}^0 \triangleq \begin{bmatrix} U_0 & W_{12}^0 \\ W_{21}^0 & W_0 \end{bmatrix} \text{ have the same partition}$$

as $\begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}$, it is easy to see that $W = T^T W_0 T$, $U = U_0$ and hence

$$G_s = \text{tr}(T^T W_0 T Q) + m_{K+1} \text{tr}(B_p^T U_0 B_p)$$

where W_0 and U_0 are independent of T .

Clearly, the roundoff noise gain G_s can be divided into two parts: one is a function of the controller realization, the other is a constant, having nothing to do with the controller structure. As mentioned before, the controller realization should be l_2 -scaled and this can be achieved if (6.3) is satisfied, where $\bar{\mathcal{K}}$ is actually the controllability Gramian of the closed-loop system. Similarly to $\bar{W}$, $\bar{\mathcal{K}}$ defined by (6.4) can be expressed as

$$\bar{\mathcal{K}} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-1} \bar{\mathcal{K}}_0 \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}^{-T}$$

where $\bar{\mathcal{K}}^0$ is similar to $\bar{\mathcal{K}}$ but corresponds to the controller realization R_0 . Let

$$\bar{\mathcal{K}} \triangleq \begin{bmatrix} \mathcal{K}_{11} & \mathcal{K}_{12} \\ \mathcal{K}_{21} & \mathcal{K} \end{bmatrix}, \quad \bar{\mathcal{K}}^0 \triangleq \begin{bmatrix} \mathcal{K}_{11}^0 & \mathcal{K}_{12}^0 \\ \mathcal{K}_{21}^0 & \mathcal{K}_0 \end{bmatrix} \quad (6.14)$$

have the same partition as $\begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}$, then

$$\mathcal{K} = T^{-1} \mathcal{K}_0 T^{-T} \quad (6.15)$$

where $\mathcal{K}_0$ is a positive-definite matrix independent of T .

It is easy to see from above equations that the l_2 -scaling constraint (6.3) can be satisfied if the diagonal elements of $\mathcal{K}$ are all equal to one, that is

$$\mathcal{K}(k, k) = 1, \forall k. \quad (6.16)$$

Therefore, the optimal roundoff noise realization problem can be defined as

$$\min_{R \in S_H} G_s \quad \text{subject to (6.16)}. \quad (6.17)$$

It should be pointed out that the optimization problem defined by (6.17) is, generally speaking, extremely hard to solve and is still an open problem. The difficulty of solving (6.17) is mainly due to the consideration of structure details with the factors Q and m_{K+1} involved in the roundoff noise gain computation. However, we note that if the controller realization is fully parametrized (with all parameters nontrivial), that is $m_k = K + 1, \forall k$, one has

$$G_s = [tr(T^T W_0 T) + tr(B_p^T U_0 B_p)](K + 1). \quad (6.18)$$

With G_s given by (6.18), the corresponding optimal realization problem (6.17) can be solved analytically. Using an existing procedure provided in [49] but corresponding to our newly defined roundoff noise gain, we solve the optimization problem (6.17) on the space of fully parameterized realizations. The corresponding optimal controller realization is denoted by R_f .

It is found that R_f yields a very good performance in terms of reducing FWL effects (the roundoff noise gain). It, however, possesses $(K + 1)^2$ nontrivial parameters, which will increase the implementation complexity and slow down the

processing. From a practical point of view, it would be better to implement the digital controller with such a structure that not only has a very robust performance against the FWL effects but also possesses as few nontrivial parameters as possible.

In Chapter 5, we have proposed two improved sparse structures (the so-called ρ IDFII structure and ρ_I -realization) based on the concept of polynomial operators. It has been shown that these structures are very efficient in terms of implementation and can be optimized to achieve large stability robustness, which in some cases, is even better than that of fully parametrized optimal realizations. Now the question is: can these two sparse structures be optimized to yield a minimum roundoff noise gain? With the roundoff noise analysis of these structures and the corresponding structure optimization solutions, which are provided in the next section, the answer to the above question comes out to be a positively yes.

6.2 Sparse controller structures with minimum roundoff noise gain

In this section, we will analyze the performance of the ρ IDFII structure and the ρ_I -realization in terms of roundoff noise gain for digital controllers in a closed-loop system. The problems of finding the optimized roundoff noise structures will then be formulated and investigated.

Consider the digital controller is implemented with a ρ IDFII structure, which is depicted by Figures 5.1 and 5.2, or a ρ_I -realization, which is given by (5.10). As mentioned in Section 6.1, the controller structure has to be properly

scaled in order to prevent the signals in the controller from overflow. For the ρ_I -realization, the l_2 -scaling can be achieved by choosing $\{\Delta_k\}$ properly.

Let $(A_{\rho_I}^0, B_{\rho_I}^0, C_{\rho_I}^0, \beta_0)$ be the equivalent state-space realization corresponding to $\Delta_k = 1, \forall k$. With (5.18), it can be shown that

$$A_{\rho_I} = T_{sc} A_{\rho_I}^0 T_{sc}^{-1}, \quad B_{\rho_I} = T_{sc} B_{\rho_I}^0, \quad C_{\rho_I} = C_{\rho_I}^0 T_{sc}^{-1}$$

where T_{sc}^{-1} is a diagonal scaling similarity transformation, and

$$T_{sc} = \text{diag}(d_1, d_2, \dots, d_K), \quad d_k = \prod_{l=1}^k \Delta_l^{-1}, \quad \forall k.$$

Denote $\bar{\mathcal{K}}_\rho$ and $\bar{\mathcal{K}}_\rho^0$ as the closed-loop controllability Gramians, corresponding to the controller realizations $(A_{\rho_I}, B_{\rho_I}, C_{\rho_I}, \beta_0)$ and $(A_{\rho_I}^0, B_{\rho_I}^0, C_{\rho_I}^0, \beta_0)$, respectively. Let $\mathcal{K}_\rho$ be the sub-matrix of $\bar{\mathcal{K}}_\rho$ with the partition defined in (6.14), then (6.15) becomes $\mathcal{K}_\rho = T_{sc} \mathcal{K}_\rho^0 T_{sc}^T$ with $\mathcal{K}_\rho^0$ the corresponding sub-matrix of $\bar{\mathcal{K}}_\rho^0$. It is easy to see that the l_2 -scaling can be achieved if $\mathcal{K}_\rho(k, k) = 1, \forall k$, or equivalently,

$$d_k^2 \mathcal{K}_\rho^0(k, k) = 1, \quad k = 1, 2, \dots, K$$

which leads to

$$\Delta_1 = \sqrt{\mathcal{K}_\rho^0(1, 1)}, \quad \Delta_k = \sqrt{\frac{\mathcal{K}_\rho^0(k, k)}{\mathcal{K}_\rho^0(k-1, k-1)}}, \quad k = 2, 3, \dots, K. \quad (6.19)$$

Clearly, the ρ IDFIIt structure will be l_2 -scaled if $\{\Delta_k\}$ are given by (6.19). Here we should note that $\{\Delta_k\}$ are referred as the coupling coefficients determined for l_2 -scaling, and are no longer what they used to be in Chapter 5, where they

are the free parameters used for optimizing the stability performance.

In the sequel, both the ρ IDFIIt structure and ρ_I -realization are assumed to have been l_2 -scaled. It is interesting to see that with (6.19), one has

$$\begin{cases} \kappa &= \sqrt{\mathcal{K}_\rho^0(K, K)} \\ q_0(z) &= \sqrt{\frac{1}{\mathcal{K}_\rho^0(K, K)}} \bar{q}_0(z), & q_K(z) = 1 \\ q_k(z) &= \sqrt{\frac{\mathcal{K}_\rho^0(k, k)}{\mathcal{K}_\rho^0(K, K)}} \bar{q}_k(z), & k = 1, 2, \dots, K-1 \end{cases} \quad (6.20)$$

One notes that for a given digital controller $H(z)$, there exists a class of l_2 -scaled ρ IDFIIt structures and ρ_I -realizations, which are determined by two spaces, denoted as S_γ and $S_{\bar{\eta}}$, from which the free parameters γ_k and $\bar{\eta}_m$ take values, respectively. It is easy to see that $\{\gamma_k\}$ are the parameters to be implemented directly in both structures. So it is desired that they are chosen from S_{FWL} defined before such that they produce no FWL errors. Therefore, one can choose $S_\gamma \subset S_{FWL}$, which means that all γ_k are of exact B_γ -bit format with $B_\gamma \leq B_\tau$. As to $\bar{\eta}_m$, they are not the structure parameters and hence do not have to be in an FWL format. The range for $\bar{\eta}_m$ has been specified in (5.24).

6.2.1 Roundoff noise analysis

First of all, consider a digital controller implemented with a ρ IDFIIt structure. We note that the parameters in a ρ IDFIIt structure are $\{\alpha_k\}$, $\{\beta_k\}$, $\{\Delta_k\}$, $\{\gamma_k\}$ and $\{\eta_m\}$. It follows from (5.5) that

$$y(n) = \beta_0 u(n) + \sum_{l=1}^K \left[\frac{q_l(z)}{q_0(z)} \beta_l u(n) - \frac{q_l(z)}{q_0(z)} \alpha_l y(n) \right]. \quad (6.21)$$

Let us first look at the effect of roundoff noise $\phi(\beta_0)\epsilon_{\beta_0}(n)$ due to β_0 on the closed-loop output. Let $u^*(n)$ and $y^*(n)$ be the corresponding outputs of the closed-loop system and the controller, respectively. Clearly, they obey (6.21) with $\beta_0 u^*(n)$ replaced by $\beta_0 u^*(n) + \phi(\beta_0)\epsilon_{\beta_0}(n)$. Denote $\Delta y(n) \triangleq y^*(n) - y(n)$. Then one can show that

$$\begin{aligned}\Delta y(n) &= [\beta_0 \Delta u(n) + \phi(\beta_0)\epsilon_{\beta_0}(n)] \\ &+ \sum_{l=1}^K \frac{q_l(z)}{q_0(z)} \beta_l \Delta u(n) - \sum_{l=1}^K \frac{q_l(z)}{q_0(z)} \alpha_l \Delta y(n)\end{aligned}\quad (6.22)$$

where $\Delta u(n) \triangleq u^*(n) - u(n)$, satisfying

$$\Delta u(n) = P(z)\Delta y(n). \quad (6.23)$$

Let $H_{cl}(z)$ be the transfer function of the closed-loop system, which is given by

$$H_{cl}(z) = \frac{P(z)}{1 - P(z)H(z)} \triangleq \frac{N_{cl}(z)}{D_{cl}(z)}$$

where $P(z)$ is the transfer function of plant and $H(z)$ the polynomial parametrized controller transfer function given by (5.5). It is easy to see that

$$H_{cl}(z) = D_{cl} + C_{cl}(zI - A_{cl})^{-1}B_{cl} \quad (6.24)$$

with $(A_{cl}, B_{cl}, C_{cl}, D_{cl})$ the realization of closed-loop system. It then follows from (6.22) and (6.23) that

$$\Delta u(n) = S_0(z)\phi(\beta_0)\epsilon_{\beta_0}(n)$$

where $S_0(z)$ is the transfer function between $\phi(\beta_0)\epsilon_{\beta_0}(n)$ and $\Delta u(n)$, which is given by

$$S_0(z) = H_{cl}(z)V_0(z)$$

with

$$V_0(z) \triangleq \frac{q_0(z)}{q_0(z) + \sum_{l=1}^K \alpha_l q_l(z)}.$$

Comparing $V_0(z)$ with (5.5), it follows from (5.9), (5.10) that

$$\begin{aligned} V_0(z) &= [\beta_0 + C_{\rho_I}(zI - A_{\rho_I})^{-1}B_{\rho_I}]|_{\beta_0=1, \bar{v}_\beta=\mathbf{0}} \\ &= 1 - C_{\rho_I}(zI - A_{\rho_I})^{-1}\bar{V}_\alpha. \end{aligned} \quad (6.25)$$

One observes that $S_0(z)$ is of the form $S_0(z) = [D_2 + C_2(zI_2 - A_2)^{-1}B_2][D_1 + C_1(zI_1 - A_1)^{-1}B_1]$, where $A_1 = A_{\rho_I}, B_1 = -\bar{V}_\alpha, C_1 = C_{\rho_I}, D_1 = 1, A_2 = A_{cl}, B_2 = B_{cl}, C_2 = C_{cl}, D_2 = D_{cl}$, and $I_k, k = 1, 2$ denotes the identity matrix of a proper dimension. It is easy to verify that

$$\begin{aligned} S_0(z) &= D_2D_1 + \begin{bmatrix} D_2C_1 & C_2 \end{bmatrix} (z \begin{bmatrix} I_1 & \mathbf{0} \\ \mathbf{0} & I_2 \end{bmatrix} - \begin{bmatrix} A_1 & \mathbf{0} \\ B_2C_1 & A_2 \end{bmatrix})^{-1} \begin{bmatrix} B_1 \\ B_2D_1 \end{bmatrix} \\ &\triangleq \tilde{D} + \tilde{C}(zI - \tilde{A})^{-1}\tilde{B}. \end{aligned}$$

According to (6.6) and (6.8), the roundoff noise gain due to parameter β_0 is given by

$$G_{\beta_0} = \phi(\beta_0) \|S_0(z)\|_2^2 = \phi(\beta_0) \text{tr}(\tilde{D}^T \tilde{D} + \tilde{B}^T \tilde{W} \tilde{B}) \triangleq \phi(\beta_0) G_0$$

where $\tilde{W}$ is the observability Gramian of the realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.

Using the same procedure, one can analyze the roundoff noise gain due to the parameter β_k . Let $\phi(\beta_k)\epsilon_{\beta_k}(n)$ be the corresponding roundoff noise. It can be shown that the transfer function from $\phi(\beta_k)\epsilon_{\beta_k}(n)$ to $\Delta u(n)$, denoted as $S_k(z)$, is

$$S_k(z) = H_{cl}(z)V_k(z)$$

with $H_{cl}(z)$ given by (6.24) and

$$V_k(z) = \frac{q_k(z)}{q_0(z) + \sum_{l=1}^K \alpha_l q_l(z)} = C_{\rho_I}(zI - A_{\rho_I})^{-1}e_k$$

for $k = 1, 2, \dots, K$, where e_k is the k th elementary vector as defined before.

Therefore,

$$G_{\beta_k} = \phi(\beta_k) \|S_k(z)\|_2^2 \triangleq \phi(\beta_k)G_k, \quad \forall k$$

with

$$G_k = \text{tr}(\tilde{D}_k^T \tilde{D}_k + \tilde{B}_k^T \tilde{W}_k \tilde{B}_k)$$

where $(\tilde{A}_k, \tilde{B}_k, \tilde{C}_k, \tilde{D}_k)$ is the realization of $S_k(z)$ and $\tilde{W}_k$ is the corresponding observability Gramian.

Comparing the positions of α_k, γ_k and Δ_{k+1} with that of β_k in Figure 5.2, one can see easily that

$$G_{\alpha_k} = \phi(\alpha_k)G_k, \quad G_{\gamma_k} = \phi(\gamma_k)G_k, \quad G_{\Delta_k} = \phi(\Delta_k)G_{k-1}$$

for $k = 1, 2, \dots, K$. Similarly, the positions of $\beta_{K-2p+2m}$ and η_m in Figure 5.2

suggest that for $m = 1, \dots, p$

$$G_{\eta_m} = \phi(\eta_m)G_{K-2p+2m}.$$

Therefore, the total roundoff noise gain of the ρ IDFIIt structure is

$$\begin{aligned} G_{\bar{\rho}} &\triangleq \sum_{k=1}^K [G_{\alpha_k} + G_{\gamma_k} + G_{\Delta_k}] + \sum_{k=0}^K G_{\beta_k} + \sum_{m=1}^p G_{\eta_m} \\ &\triangleq \sum_{k=0}^K v_k G_k \end{aligned} \quad (6.26)$$

where the coefficients v_k can be specified easily with the expressions, obtained above, of roundoff noise gain for all the parameters.

Now let us look at the equivalent state-space realization, that is, the ρ_I -realization. In Section 6.1, we have derived the expression of the roundoff noise gain for any given state-space controller realization in a close-loop control system (see (6.12)). Therefore, applying (6.12) to the ρ_I -realization, one can easily obtain the corresponding roundoff noise gain, which is denoted by $G_s^{\bar{\rho}}$.

6.2.2 Structure optimization

For a given digital controller $H(z)$ and any given free parameters $\{\gamma_k\}, \{\bar{\eta}_m\}$, one can obtain the l_2 -scaled ρ_I -realization and ρ IDFIIt structure with $\{\Delta_k\}$ given by (6.19). The roundoff noise gain $G_s^{\bar{\rho}}$ and $G_{\bar{\rho}}$ for the above two structures can then be evaluated with (6.12) and (6.26), respectively. Since different sets of $\{\gamma_k, \bar{\eta}_m\}$ yield different ρ IDFIIt structures and hence lead to different roundoff noise gain $G_{\bar{\rho}}$, an interesting problem is to minimize $G_{\bar{\rho}}$ with respect to these free

parameters, which leads to the following optimal ρ IDFIIt structure problem:

$$\min_{\gamma_k \in S_\gamma, \bar{\eta}_m \in S_{\bar{\eta}}} G_{\bar{\rho}}. \quad (6.27)$$

Similarly, the optimal ρ_I -realization problem in terms of minimizing the roundoff noise gain can be defined as:

$$\min_{\gamma_k \in S_\gamma, \bar{\eta}_m \in S_{\bar{\eta}}} G_s^{\bar{\rho}}. \quad (6.28)$$

It seems impossible to obtain analytical solutions to the problems defined in (6.27) and (6.28) due to the high nonlinearity of $G_{\bar{\rho}}$ and $G_s^{\bar{\rho}}$ in $\{\gamma_k\}$, $\{\bar{\eta}_m\}$. However, noting that S_γ is of finite elements, these problems can be solved using the exhaustive searching method combined with a standard optimization algorithm, such as the command *fmincon.m* in MATLAB. It is noted that this optimization method may take excessive computation time if there is a large number of enumerations (when the order of the controller is very high). Alternatively, the existing well-established optimization techniques, such as the genetic algorithm (GA), can be applied to attack these problems much more efficiently. In the numerical example, the optimal structure problems are attacked with both GA (see Appendix A for details) and the exhaustive searching.

In what follows, we will make an important analysis on the relationship between the choice of free parameters $\{\gamma_k\}$, $\{\bar{\eta}_m\}$ and the minimization of the roundoff noise gain. Based on this relationship, one can specify the space where the optimal $\{\gamma_k\}$, $\{\bar{\eta}_m\}$ lie in, which will not only enhance the efficiency of exhaustive searching but also construct a very good initial population to speed up GA to find an optimal solution.

One notes from (6.26) that $G_{\bar{\rho}}$ can be expressed in a linear combination of all $G_k = ||S_k(z)||_2^2$ with

$$S_k(z) = H_{cl}(z)V_k(z) = \frac{\kappa q_k(z)N_p(z)}{D_{cl}(z)}, \quad \forall k$$

where $N_p(z)$ is the numerator of the plant $P(z)$ and $D_{cl}(z)$ is the denominator of $H_{cl}(z)$ defined before. It then follows from (6.20) that

$$S_k(z) = \begin{cases} \frac{\bar{q}_0(z)N_p(z)}{D_{cl}(z)} & k = 0 \\ \sqrt{\mathcal{K}_{\rho}^0(k, k)} \frac{\bar{q}_k(z)N_p(z)}{D_{cl}(z)} & k = 1, 2, \dots, K-1 \\ \sqrt{\mathcal{K}_{\rho}^0(K, K)} \frac{N_p(z)}{D_{cl}(z)} & k = K \end{cases}$$

Let $D_{cl}(z) = \prod_{l=1}^{N_0} (z - \lambda_l)$ where $\{\lambda_l\}$ are the closed-loop poles and $N_0 \triangleq K + J$. To simplify the discussion, we assume that the closed-loop system has no repeated poles, that is $\lambda_l \neq \lambda_i, \forall l \neq i$, and no poles at $z = 0$. According to the Residue Theorem, it follows from (6.7) that

$$G_k = \begin{cases} \frac{\bar{q}_0(0)}{D_{cl}(0)} + \sum_{i=1}^{N_0} \frac{\bar{q}_0(\lambda_i)\bar{q}_0(\lambda_i^{-1})}{\prod_{l \neq i}^{N_0} (\lambda_i - \lambda_l)} \sigma_i & k = 0 \\ \mathcal{K}_{\rho}^0(k, k) \sum_{i=1}^{N_0} \frac{\bar{q}_k(\lambda_i)\bar{q}_k(\lambda_i^{-1})}{\prod_{l \neq i}^{N_0} (\lambda_i - \lambda_l)} \sigma_i & k = 1, 2, \dots, K-1 \\ \mathcal{K}_{\rho}^0(K, K) \sum_{i=1}^{N_0} \frac{1}{\prod_{l \neq i}^{N_0} (\lambda_i - \lambda_l)} \sigma_i & k = K \end{cases} \quad (6.29)$$

with

$$\sigma_i = \frac{N_p(\lambda_i)N_p(\lambda_i^{-1})}{D_{cl}(\lambda_i^{-1})} \lambda_i^{-1}, \quad \forall i.$$

One observes from (6.29) that G_k can be much reduced if $\bar{q}_k(\lambda_i) \approx 0$, which can be achieved by choosing the roots of $\bar{q}_k(z)$ close to $\{\lambda_i\}$. It has been shown in (5.22) and (5.23) that for ρ IDFII structures, the roots of $\bar{q}_k(z)$ can be expressed

by the free parameters $\{\gamma_k\}$ and $\{\bar{\eta}_m\}$. So when $\{\gamma_k\}, \{\bar{\eta}_m\}$ are chosen such that the roots of $\bar{q}_k(z)$ are close to the dominant poles of the closed-loop system, G_k can be much reduced, and hence $G_{\bar{\rho}}$, as the linear combination of G_k , can also be much reduced.

This observation gives us some insights on how to choose $\{\gamma_k\}, \{\bar{\eta}_m\}$ to minimize the roundoff noise gain of ρ IDFII structure and ρ_I -realization, which can accelerate the exhaustive searching process for the optimal roundoff noise structures. Suppose γ_k are of exact B_γ -bit format (including one bit for the sign) and take values from S_γ as defined before, then S_γ is a discrete space, containing $2^{B_\gamma} + 1$ elements. For a digital controller of order K , there will be $(2^{B_\gamma} + 1)^K$ combinations for exhaustive searching when we solve the optimization problems in (6.27) and (6.28). Since the optimal $\{\gamma_k\}$ and $\{\bar{\eta}_m\}$ can be searched within the space for which the roots of the corresponding $\bar{q}_k(z)$ are around the neighborhood of the dominant closed-loop poles, the quantity of the combinations can be much reduced, which will make the exhaustive searching for the optimal controller structures more efficiently. In addition, when the optimal structure problem is attacked with GA, a good initial population can be obtained by choosing the free parameters such that the roots of the polynomials are close to the dominant closed-loop poles. This initial population will speed up GA to find a very good optimal solution.

6.3 A numerical example

We now present a numerical example to illustrate the design procedure. In this example, it is assumed that all γ_k are of exact 3-bit format (including one bit

for the sign) and hence $S_\gamma = \{\pm 1, \pm(2^{-1} + 2^{-2}), \pm 2^{-1}, \pm 2^{-2}, 0\}$. $S_{\bar{\eta}}$ is given by (5.24) as defined in Chapter 5. The optimal roundoff noise structure problems (6.27) and (6.28) are attacked with the efficient genetic algorithm (GA) and the results are confirmed with the exhaustive searching.

It should be pointed out that the optimal structures obtained in this section are referred to the structures optimized in terms of minimizing the roundoff noise gain, which are different with those optimal structures optimized by maximizing the stability robustness presented in previous chapters.

Example 6.1: Here we use the same discrete-time control system as in Example 5.2. In that example, we have shown that the improved sparse structures R_{ρ_I} and ρIDFIIt can be optimized to achieve very nice stability performance, which is even better than that of R_ρ and ρDFIIt . Now we use this example to illustrate the effectiveness of these sparse structures with respect to the roundoff noise gain minimization.

With R_c in Example 5.2 as the initial controller realization, we solve (6.17) on the set of fully parameterized state-space realizations and eventually obtain the optimal realization R_f . Solving (6.28) with $\bar{\eta}_m = 0, \forall m$, one gets an optimized ρ -realization, denoted as R_ρ . With $S_{\bar{\eta}}$ defined by (5.24), the optimization problem

Table 6.1: Comparison of different state-space realizations

	$R_c^{l_2}$	R_f	R_ρ	R_{ρ_I}
G	1.2773×10^4	4.9919	2.2129	1.6424
N_p	19	49	24	25

(6.28) is attacked and an optimized ρ_I -realization R_{ρ_I} is obtained:

$$\begin{aligned}
 A_{\rho_I} &= \begin{bmatrix} 0.1172 & 0.2812 & 0 & 0 & 0 & 0 \\ -1.5617 & 0.75 & 0.7030 & 0 & 0 & 0 \\ -0.1111 & 0 & 0.50 & 0.4172 & 0 & 0 \\ 0.0721 & 0 & -0.2982 & 0.50 & 0.7093 & 0 \\ 0.2486 & 0 & 0 & 0 & 0.75 & 0.2850 \\ 0.6384 & 0 & 0 & 0 & -0.5790 & 1 \end{bmatrix} \\
 B_{\rho_I} &= \begin{bmatrix} 0.8150 & 1.6863 & 0.5280 & 0.1287 & -0.1359 & -0.4437 \end{bmatrix}^T \\
 C_{\rho_I} &= \begin{bmatrix} 0.2770 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad d = 0.0577.
 \end{aligned}$$

Computation shows that for R_{ρ_I} , the set of the roots for all the polynomials $\bar{q}_k(z)$ is $\{0.5 \pm j0.3527, 0.5, 0.75 \pm j0.3158, 0.75, 0.8750 \pm j0.3865, 1\}$. It is interesting to note that the roots are around the poles of the closed-loop system.

In Table 6.1, the comparison of different state-space realizations is presented, where $R_c^{l_2}$ is the l_2 -scaled R_c obtained with a diagonal similarity transformation, G is the roundoff noise gain and N_p is the number of nontrivial parameters in each realization.

Similarly, solving (6.27) leads to an optimized ρ IDFIIt structure, denoted as ρ IDFIIt, which can be implemented with the structure form depicted in

Table 6.2: Comparison of different DFIIIt structures

	$z\text{DFIIIt}$	δDFIIIt	ρDFIIIt	$\rho\text{IDFIIIt}$
G	1.5191×10^4	7.1763	1.0085	0.3747
N_p	19	19	24	26

Figures 5.1 and 5.2, where the corresponding parameter sets are given by

$$\begin{aligned}
\{\alpha_k\} &= \{3.1868, 5.1651, 0.4331, -0.0532, -0.8562, -2.3096\}, \\
\{\beta_k\} &= \{0.0577, 0.9989, 1.9844, 0.5881, 0.1894, -0.1702, -0.5698\}, \\
\{\gamma_k\} &= \{1, 0.75, 0.5, 0.5, 0.75, 1\}, \\
\{\Delta_k\} &= \{0.2770, 0.2812, 0.7481, 0.3974, 0.7060, 0.2817\}, \\
\{\eta_m\} &= \{0.6945, 0.3646, 0.8230\}.
\end{aligned}$$

Computation shows that for the $\rho\text{IDFIIIt}$, the roots for all $\bar{q}_k(z)$ are $\{0.5 \pm j0.3292, 0.5, 0.75, 0.8750 \pm j0.1962, 0.8750 \pm j0.3848, 1\}$, which are around the neighborhood of the closed-loop poles.

For comparison, the optimization problem (6.27) is solved with $\bar{\eta}_m = 0, \forall m$ and an optimized ρDFIIIt structure is obtained. $z\text{DFIIIt}$ and δDFIIIt are two special l_2 -scaled ρDFIIIt structures, corresponding to $\gamma_k = 0, \forall k$ and $\gamma_k = 1, \forall k$, respectively. Table 6.2 shows the computation results of G and N_p for each DFIIIt structure.

The results are self-explanatory. Particularly, one can observe that $\rho\text{IDFIIIt}$ and R_{ρ_I} outperform R_f greatly in terms of reducing roundoff noise and implementation complexity. Compared with ρDFIIIt and R_ρ , $\rho\text{IDFIIIt}$ and R_{ρ_I} yield an even smaller roundoff noise gain than their counterparts, while the slight

price paid for that is just one or two more nontrivial parameters. The roots of all the polynomials for $\rho\text{IDFIIIt}$ and R_{ρ_I} are found to be around the poles of the closed-loop system, as a result of that, these two structures both exhibit excellent roundoff noise performance and thus our argument on how to choose the polynomials with those free parameters to minimize the roundoff noise gain is supported by this example.

Although it is very hard to compute the total roundoff noise gain with the simulation data, it is expected that R_ρ or R_{ρ_I} should have a much smaller output error variance than $R_c^{l_2}$. Also ρDFIIIt and $\rho\text{IDFIIIt}$ should have a much smaller output error variance than $z\text{DFIIIt}$. To confirm these, some simulations have been conducted. In Figure 2.3, the input signal $r(n)$ is replaced with a white sequence of 10,000 points, generated with unit variance using the command *randn.m* in MATLAB. The digital controller can be implemented with either state-space realizations or DFIIIt structures, where each intermediate signal, after being multiplied with a nontrivial structure coefficient, is rounded into $B_s = 8$ bits on the fractional part. The variance of the error sequence between the ideal output and the actual one of the closed-loop system for each controller structure is computed with the same input sequence. For $R_c^{l_2}$, the variance is 1.7025×10^{-2} , while for R_ρ and R_{ρ_I} , we have 9.9526×10^{-5} and 9.7524×10^{-5} , respectively. $z\text{DFIIIt}$ yields a variance of 8.7102×10^{-2} , while the variances for ρDFIIIt and $\rho\text{IDFIIIt}$ are 5.7230×10^{-5} and 9.0488×10^{-6} , respectively.

Figure 6.1 shows the unit impulse responses of the closed-loop system, where the solid line is for the ideal response, while the dotted, dashed and dash-dot lines are for $R_c^{l_2}$, R_ρ and R_{ρ_I} , respectively. Clearly, the response corresponding to $R_c^{l_2}$ is far away from the ideal one, while those for R_ρ and R_{ρ_I} are very close

Figure 6.1: Unit impulse responses of the closed-loop system, where the solid line is for the ideal response, while the dotted, dashed and dash-dot lines are for $R_c^{l_2}$, R_ρ and R_{ρ_I} , respectively, with $B_s = 8$ bits.

Figure 6.2: Unit impulse responses of the closed-loop system, where the solid line is for the ideal response, while the dotted, dashed and dash-dot lines are for $z\text{DFIIt}$, ρDFIIt and ρIDFIIt , respectively, with $B_s = 8$ bits.

to the ideal response and R_{ρ_I} yields an even better response than R_ρ . Figure 6.2 compares the ideal unit impulse response of the closed-loop system with those where the controller is implemented with z DFIIt, ρ DFIIt and ρ IDFIIt, respectively. The advantages of ρ DFIIt and ρ IDFIIt over z DFIIt are obvious and ρ IDFIIt yields a response which is almost the same as the ideal one.

This example clearly demonstrates the effectiveness of our design procedure for optimal sparse controller structures with minimum roundoff noise gain.

6.4 Summary

In this chapter, we have addressed the optimal controller structure problem in a discrete-time control system with roundoff noise consideration.

Firstly, an analytical expression of the roundoff noise gain has been derived for digital controllers implemented with a state-space realization. The problem of identifying minimum roundoff noise realizations has been solved within the space of fully parametrized realizations. Secondly, we have analyzed the performance of the second-order operator-based sparse structures proposed in Chapter 5 by deriving the corresponding expressions of roundoff noise gain. With the established relationship between the closed-loop poles and optimal polynomial operators, the problems of finding optimized sparse structures have been efficiently attacked with a genetic algorithm. Finally, a numerical example has been presented to confirm the theoretical results and to illustrate the design procedure, with which the excellent performance of the second-order operator-based structures has been demonstrated in terms of minimizing the roundoff noise gain and reducing the implementation complexity.

Chapter 7

Pole Deviation Analysis Based on Second-Order Perturbation Theory

For this chapter, we return to the stability issue of the digital systems. As mentioned in Section 2.2, the classical FWL stability measure μ_0 addressed in [52] is the best measure quantifying the FWL stability character of a system realization. Unfortunately, how to calculate explicitly the value for μ_0 has not been solved so far. The prevailing pole (modulus) sensitivity-based measures used nowadays (see, e.g., [54], [56]), including the one proposed in Chapter 3, are the alternative tractable measures to approximate μ_0 in some senses. It should be noted that the derivation of all these measures is based on a first-order approximation, where the pole (modulus) deviation is proportional to the pole (modulus) sensitivity.

We shall, in this chapter, analyze the pole modulus deviation based on a

second-order perturbation theory. A new measure is then defined to evaluate the system stability, which is a very good approximation of μ_0 and is more accurate than the approach based on the first-order approximation.

7.1 The first- and second-order approximation

Let $\bar{A}$ and $X \triangleq \{x_{ij}\}$ be the matrices as defined in Section 2.2. Denote $\lambda(M)$ the eigenvalue set of a matrix M , then the system poles $\{\lambda_k\}$ are the eigenvalues of $\bar{A}$, $\{\lambda_k\} = \lambda(\bar{A})$. Let $\Delta|\lambda_k| \triangleq |\lambda_k(\bar{A}(X + \Delta X))| - |\lambda_k(\bar{A}(X))|$ be the pole modulus deviation due to the perturbation ΔX . With a first-order approximation, the deviation of each pole modulus is proportional to the pole modulus sensitivity and can be expressed as:

$$\Delta|\lambda_k| \approx \sum_{i,j} \frac{\partial |\lambda_k|}{\partial x_{ij}} \Delta x_{ij}.$$

Now we analyze the pole deviation based on a second-order perturbation theory. This approach is clearly more accurate than that based on the first-order approximation.

Denote

$$\begin{aligned} \mathcal{V} &= \begin{bmatrix} v_1 & \cdots & v_k & \cdots & v_N \end{bmatrix}^T \triangleq \text{Vec}(X) \\ \Delta \mathcal{V} &= \begin{bmatrix} \Delta v_1 & \cdots & \Delta v_k & \cdots & \Delta v_N \end{bmatrix}^T, \end{aligned} \quad (7.1)$$

where $\text{Vec}(\cdot)$ is the column stacking operator. The pole modulus deviation

$\Delta|\lambda_k|$ can be evaluated with the following second-order approximation:

$$\Delta|\lambda_k| \approx S_1^T(k)\Delta\mathcal{V} + (\Delta\mathcal{V})^T S_2(k)\Delta\mathcal{V}, \quad (7.2)$$

where $S_1(k) \triangleq \{\frac{\partial|\lambda_k|}{\partial v_i}\} \in \mathcal{R}^{N \times 1}$, $S_2(k) \triangleq \{\frac{\partial^2|\lambda_k|}{\partial v_i \partial v_j}\} \in \mathcal{R}^{N \times N}$. Here $\{\partial|\lambda_k|/\partial v_i\}$ and $\{\partial^2|\lambda_k|/\partial v_i \partial v_j\}$ are the first- and second-order pole modulus sensitivities, respectively. The former can be computed with (3.3), while for the latter, it can be shown that

$$\frac{\partial^2|\lambda_k|}{\partial v_i \partial v_j} = \frac{1}{|\lambda_k|} \{ \text{Re}[\lambda_k^* \frac{\partial^2 \lambda_k}{\partial v_i \partial v_j} + \frac{\partial \lambda_k}{\partial v_i} \frac{\partial \lambda_k^*}{\partial v_j}] - \frac{\partial|\lambda_k|}{\partial v_i} \frac{\partial|\lambda_k|}{\partial v_j} \} \quad (7.3)$$

in which the second-order pole sensitivities $\{\partial^2 \lambda_k / \partial v_i \partial v_j\}$, as far as we know, are not available in the literature.

7.2 Computing the second-order pole sensitivity

In what follows, we will derive an analytical expression for computing the second-order pole sensitivities. First, we need the following theorem, which is actually a generalized version of Theorem 1.

Theorem 3. Let $\bar{A} \in \mathcal{R}^{m \times m}$ be diagonalizable and a function of variable v . Denote $\lambda_k = \lambda_k(\bar{A})$ as the k th eigenvalue of $\bar{A}$. Let x_k be a right eigenvector of $\bar{A}$ corresponding to λ_k and y_k the reciprocal left eigenvector. Denote $M_x = \begin{bmatrix} x_1 & \cdots & x_k & \cdots & x_m \end{bmatrix}$ and $M_y = \begin{bmatrix} y_1 & \cdots & y_k & \cdots & y_m \end{bmatrix} \triangleq M_x^{-\mathcal{H}}$. Then

$$\frac{\partial \lambda_k}{\partial v} = y_k^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v} x_k, \quad \forall k. \quad (7.4)$$

140 Chapter 7. Pole Deviation Analysis Based on Second-Order Perturbation Theory

Proof. First of all, it follows from $M_x M_y^{\mathcal{H}} = I_m$, the identity matrix of dimension m , that

$$\frac{\partial M_y^{\mathcal{H}}}{\partial v} M_x + M_y^{\mathcal{H}} \frac{\partial M_x}{\partial v} = 0, \quad (7.5)$$

which implies $\frac{\partial y_k^{\mathcal{H}}}{\partial v} x_k + y_k^{\mathcal{H}} \frac{\partial x_k}{\partial v} = 0$. Noting $\bar{A}x_k = \lambda_k x_k$, one has $\lambda_k = y_k^{\mathcal{H}} \bar{A}x_k$ and hence

$$\frac{\partial \lambda_k}{\partial v} = \frac{\partial y_k^{\mathcal{H}}}{\partial v} \bar{A}x_k + y_k^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v} x_k + y_k^{\mathcal{H}} \bar{A} \frac{\partial x_k}{\partial v}. \quad (7.6)$$

(7.4) follows from (7.5) and $y_k^{\mathcal{H}} \bar{A} = \lambda_k y_k^{\mathcal{H}}$. $\square$

Clearly, with $\bar{A} \triangleq \{\bar{a}_{ij}\}$ and v replaced by $\bar{a}_{ij}$, it is easy to see that $\frac{\partial \bar{A}}{\partial \bar{a}_{ij}} = e_i e_j^T$, where e_k denotes the k th elementary vector whose elements are all zero except the k th one which is 1, and hence (7.4) yields (2.18). Noting $\bar{A} = M_0 + M_1 X M_2$ with X independent of M_0 , M_1 and M_2 , it can be shown by direct computation that (2.17) follows from (7.4).

Now, the following theorem provides a way to compute the second-order pole sensitivities $\frac{\partial^2 \lambda_k}{\partial v_i \partial v_j}$, $\forall i, j$.

Theorem 4. *With notations defined above in Theorem 3, the second-order pole sensitivities can be computed with*

$$\frac{\partial^2 \lambda_k}{\partial v_i \partial v_j} = y_k^{\mathcal{H}} \left[\frac{\partial \bar{A}}{\partial v_i} \frac{\partial x_k}{\partial v_j} + \left(\frac{\partial^2 \bar{A}}{\partial v_i \partial v_j} - \frac{\partial M_x}{\partial v_j} M_y^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_i} \right) x_k \right], \quad \forall i, j \quad (7.7)$$

where $\frac{\partial M_x}{\partial v_j} = \begin{bmatrix} \frac{\partial x_1}{\partial v_j} & \dots & \frac{\partial x_l}{\partial v_j} & \dots & \frac{\partial x_m}{\partial v_j} \end{bmatrix}$ with $\frac{\partial x_l}{\partial v_j}$ a solution to the following linear equations:

$$(\lambda_l I_m - \bar{A}) \frac{\partial x_l}{\partial v_j} = \left(\frac{\partial \bar{A}}{\partial v_j} - \frac{\partial \lambda_l}{\partial v_j} I_m \right) x_l, \quad \forall l, \quad (7.8)$$

in which $\frac{\partial \lambda_l}{\partial v_j}$ is given by (7.4).

Proof. It follows from (7.4) that

$$\frac{\partial^2 \lambda_k}{\partial v_i \partial v_j} = \frac{\partial y_k^{\mathcal{H}}}{\partial v_j} \frac{\partial \bar{A}}{\partial v_i} x_k + y_k^{\mathcal{H}} \frac{\partial^2 \bar{A}}{\partial v_i \partial v_j} x_k + y_k^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_i} \frac{\partial x_k}{\partial v_j}. \quad (7.9)$$

With (7.5), one has $\frac{\partial M_y^{\mathcal{H}}}{\partial v_j} = -M_y^{\mathcal{H}} \frac{\partial M_x}{\partial v_j} M_y^{\mathcal{H}}$ and hence

$$\frac{\partial y_k^{\mathcal{H}}}{\partial v_j} = -y_k^{\mathcal{H}} \frac{\partial M_x}{\partial v_j} M_y^{\mathcal{H}}. \quad (7.10)$$

(7.7) follows from combining (7.9) and (7.10). Noting $\bar{A}x_l = \lambda_l x_l$, one has

$$\frac{\partial \bar{A}}{\partial v_j} x_l + \bar{A} \frac{\partial x_l}{\partial v_j} = \frac{\partial \lambda_l}{\partial v_j} x_l + \lambda_l \frac{\partial x_l}{\partial v_j},$$

which leads to (7.8). □

It should be pointed out (7.8) always has solutions if $\bar{A}$ has no repeated eigenvalues. In fact, noting $\bar{A} = M_x \Sigma M_x^{-1}$ with $\Sigma \triangleq \text{diag}(\lambda_1, \dots, \lambda_l, \dots, \lambda_m)$, (7.8) can be rewritten as

$$(\lambda_l I_m - \Sigma) M_x^{-1} \frac{\partial x_l}{\partial v_j} = M_y^{\mathcal{H}} \left(\frac{\partial \bar{A}}{\partial v_j} - \frac{\partial \lambda_l}{\partial v_j} I_m \right) x_l. \quad (7.11)$$

It follows from $M_y^{\mathcal{H}} x_l = e_l$ and $y_l^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_j} x_l = \frac{\partial \lambda_l}{\partial v_j}$ (see Theorem 3) that the l th element of $M_y^{\mathcal{H}} \left(\frac{\partial \bar{A}}{\partial v_j} - \frac{\partial \lambda_l}{\partial v_j} I_m \right) x_l = M_y^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_j} x_l - \frac{\partial \lambda_l}{\partial v_j} e_l$ is zero. It then follows from the fact that $(\lambda_l I_m - \Sigma) M_x^{-1}$ has a rank of $(m - 1)$ with its l th row the zero (row) vector that (7.11) and hence (7.8) have solutions. Now, the question to be asked is: will different solutions to (7.8) yield one and the same $\frac{\partial^2 \lambda_k}{\partial v_i \partial v_j}$? The following lemma gives a yes-answer.

Chapter 7. Pole Deviation Analysis Based on Second-Order Perturbation Theory

Lemma 3. Assume that $\bar{A}$ has no repeated eigenvalues. Then,

- the solutions to (7.8) are characterized as

$$\frac{\partial M_x}{\partial v_j} = M_x D_\alpha + \Gamma(v_j), \quad (7.12)$$

where $D_\alpha \triangleq \text{diag}(\alpha_1, \dots, \alpha_l, \dots, \alpha_m)$ with α_l an arbitrary complex number and $\Gamma(v_j) \triangleq \begin{bmatrix} \bar{\gamma}_1(v_j) & \dots & \bar{\gamma}_l(v_j) & \dots & \bar{\gamma}_m(v_j) \end{bmatrix}$ with

$$\bar{\gamma}_l(v_j) \triangleq M_x D_l M_y^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_j} x_l \quad (7.13)$$

and $D_l \triangleq \text{diag}(d_1, \dots, d_i, \dots, d_m)$ with $d_i = \frac{1}{\lambda_l - \lambda_i}$ for $i \neq l$ while $d_i = 0$ for $i = l$.

- all $\frac{\partial M_x}{\partial v_j}$ by (7.12) yield one and the same $\frac{\partial^2 \lambda_k}{\partial v_i \partial v_j}$:

$$\frac{\partial^2 \lambda_k}{\partial v_i \partial v_j} = y_k^{\mathcal{H}} \left[\frac{\partial \bar{A}}{\partial v_i} \bar{\gamma}_k(v_j) + \left(\frac{\partial^2 \bar{A}}{\partial v_i \partial v_j} - \Gamma(v_j) M_y \frac{\partial \bar{A}}{\partial v_i} \right) x_k \right]. \quad (7.14)$$

Proof. The first part of this lemma follows steadily from (7.11). With some manipulations, it follows from (7.7) that

$$\begin{aligned} \frac{\partial^2 \lambda_k}{\partial v_i \partial v_j} &= (\alpha_k y_k^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_i} x_k - y_k^{\mathcal{H}} M_x D_\alpha M_y^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_i} x_k) \\ &+ y_k^{\mathcal{H}} \left[\frac{\partial \bar{A}}{\partial v_i} \bar{\gamma}_k(v_j) + \left(\frac{\partial^2 \bar{A}}{\partial v_i \partial v_j} - \Gamma(v_j) M_y^{\mathcal{H}} \frac{\partial \bar{A}}{\partial v_i} \right) x_k \right]. \end{aligned}$$

(7.14) follows from $y_k^{\mathcal{H}} M_x D_\alpha M_y^{\mathcal{H}} = \alpha_k y_k^{\mathcal{H}}$. □

Noting that when $\{v_i\}, \{v_j\}$ are elements of $\bar{A}$, one has $\frac{\partial^2 \bar{A}}{\partial v_i \partial v_j} = \mathbf{0}_m$, (7.14)

can then be rewritten as

$$\frac{\partial^2 \lambda_k}{\partial v_i \partial v_j} = y_k^H \left[\frac{\partial \bar{A}}{\partial v_i} \bar{\gamma}_k(v_j) - \Gamma(v_j) M_y \frac{\partial \bar{A}}{\partial v_i} x_k \right]. \quad (7.15)$$

In fact, for digital filters, $\bar{A} = A \triangleq \{a_{ij}\}$ and $\frac{\partial A}{\partial v_l} = \frac{\partial A}{\partial a_{l_1 l_2}} = e_{l_1} e_{l_2}^T$, and for controllers, $\frac{\partial \bar{A}}{\partial v_l} = \frac{\partial \bar{A}}{\partial x_{l_1 l_2}} = M_1 e_{l_1} e_{l_2}^T M_2$. With $\bar{\gamma}_k(v_j)$ and $\Gamma(v_j)$ computed by (7.13), the second-order pole sensitivities can be obtained with (7.15).

7.3 A more accurate measure to approximate

μ_0

In this section, we define a new measure to evaluate the system stability. This measure is expressed in an alternative form of the classical stability robustness measure μ_0 , where the pole modulus deviation is approximated with the results obtained in Section 7.2. The new measure, denoted by $\bar{\mu}_0$, as to be seen later, is a very good approximation of μ_0 .

Denoting $\lambda_k(M)$ as the k th eigenvalue of matrix M and

$$\mu(\Delta X) \triangleq \max_{i,j} |\Delta x_{ij}|,$$

we define

$$\varsigma_k \triangleq \inf \{ \mu(\Delta X) : |\lambda_k(\bar{A}(X + \Delta X))| \geq 1 \}, \quad \forall k. \quad (7.16)$$

It is easy to see that

$$\mu_0 = \min_k \varsigma_k. \quad (7.17)$$

With $\Delta\mathcal{V}$ denoted by (7.1), we define

$$\begin{aligned} F(\Delta\mathcal{V}) &= \left[|\Delta v_1| \quad \cdots \quad |\Delta v_k| \quad \cdots \quad |\Delta v_N| \right]^T \\ g_k(\Delta\mathcal{V}) &\triangleq 1 - |\lambda_k(\bar{A}(X + \Delta X))|, \end{aligned} \quad (7.18)$$

then (7.16) is equivalent to the following problem:

$$\begin{aligned} \min \max F(\Delta\mathcal{V}) &= \varsigma_k \\ \text{s. t. } g_k(\Delta\mathcal{V}) &\leq 0. \end{aligned} \quad (7.19)$$

Here the notation “minmax” is used to denote a “minmax” problem, which can be attacked by a standard optimization algorithm, such as the *fminimax.m* in MATLAB.

Combining (7.17) and (7.19), it comes to the classical stability robustness problem (2.12). In fact, if ς_k is obtained by solving (7.19), μ_0 can be computed with (7.17). However, since there exists no explicit expression for $g_k(\Delta\mathcal{V})$, the problem defined in (7.19) can not be solved so far.

Based on the second-order approximation of pole modulus deviation (7.2), one can use the following expression to approximate $g_k(\Delta\mathcal{V})$:

$$\begin{aligned} g_k(\Delta\mathcal{V}) &\approx 1 - |\lambda_k(\bar{A}(X))| - \\ &\quad [S_1^T(k)\Delta\mathcal{V} + (\Delta\mathcal{V})^T S_2(k)\Delta\mathcal{V}] \\ &\triangleq \bar{g}_k(\Delta\mathcal{V}), \end{aligned} \quad (7.20)$$

where $S_1(k)$ and $S_2(k)$ can be computed with (3.3) and (7.3), respectively.

A new measure, then, is defined as:

$$\bar{\mu}_0 = \min_k \bar{\varsigma}_k, \quad (7.21)$$

where

$$\begin{aligned} \bar{\varsigma}_k &= \min \max F(\Delta \mathcal{V}) \\ s. \ t. \quad &\bar{g}_k(\Delta \mathcal{V}) \leq 0. \end{aligned} \quad (7.22)$$

The problem defined in (7.22) is a nonlinear constrained optimization problem, which can be well attacked using one of the standard optimization algorithms, such as the *fminimax.m* in MATLAB. With $\bar{\varsigma}_k$ obtained by solving (7.22), $\bar{\mu}_0$ can be computed with (7.21).

7.4 A numerical example

Now we present a simple numerical example to verify our theoretical results.

Example 7.1: This is a second-order Butterworth filter, generated with MATLAB command $[V_q, V_p] = \text{butter}(2, 0.05)$, where V_p and V_q are the coefficient vectors of the denominator and numerator of the transfer function, respectively. The filter can be described in the following realization, denoted as $R_0 \triangleq (A_0, B_0, C_0, d)$:

$$\begin{aligned} A_0 &= \begin{bmatrix} 1.7786 & 1 \\ -0.8008 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0.0209 \\ 0.0011 \end{bmatrix} \\ C_0 &= \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad d = 0.0055. \end{aligned}$$

Table 7.1: Comparison of two stability measures

	μ_2	$\bar{\mu}_0$	μ_0^*
R_0	4.29×10^{-2}	1.71×10^{-2}	1.10×10^{-2}
R_n	7.43×10^{-2}	1.00×10^{-1}	9.35×10^{-2}

Taking R_0 as the initial realization, we can compute a similarity transformation matrix T_n with (3.8), where Σ is taken identity matrix. A fully parameterized normal realization, denoted as R_n , can be obtained. It is shown in Section 3.2 that for digital filters, normal realizations are a set of optimal realizations that maximize the stability measure μ_2 and have a much better stability performance than the canonical realization R_0 .

For both structures, μ_2 is computed with (3.1) and $\bar{\mu}_0$ is obtained by solving (7.22) and (7.21). To make a fair comparison, we have estimated μ_0 using simulations for each structure. The estimate of μ_0 , denoted as μ_0^* , is determined using the same method as that in Section 3.3. The computation results are presented in Table 7.1.

From this example, one can see that for each structure, $\bar{\mu}_0$ is much closer to μ_0^* than μ_2 , which indicates that the measure based on second-order approximation is more accurate. Clearly, when $\bar{\mu}_0$ is used to evaluate the stability behavior of each structure, it can be shown that the normal realization R_n has a much better performance than R_0 , which confirms with the result obtained in Section 3.2.

The validity and advantage of measure $\bar{\mu}_0$ are demonstrated clearly from this digital filter example. From the definition of the measure, it is easy to see that for a controller example, to compute $\bar{\mu}_0$ for a closed-loop system is very similar with that for the filter. So we just present a filter example here to avoid

repeating the similar things.

7.5 Summary

In this chapter, we have analyzed the pole deviation for digital systems based on a second-order perturbation theory. This approach is clearly more accurate than that based on the first-order approximation. An analytical expression for computing the second-order pole sensitivities has been derived, with which the pole modulus deviation has been approximated. Based on this approximation, a new measure has been defined to evaluate the system stability. The numerical example shows that this measure is a very good approximation of the unsolved classical stability measure μ_0 and is more accurate than the approach based on the first-order pole sensitivity study.

Second-order pole sensitivity study provides us a potential space to investigate the stability issue of digital systems. Developing new tractable stability measure using this approach and applying the measure into the optimal structure design are promising future works to be accomplished.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This thesis mainly investigates the optimal FWL structure problems for digital filters and controllers with stability robustness and roundoff noise consideration. The major results and contributions of the studies presented in this thesis are discussed below.

First, based on a pole modulus sensitivity approach, a new stability-related measure has been proposed in Chapter 3. The corresponding optimal state-space realization problem in terms of maximizing this measure has been solved analytically for digital filter case. It is shown that a fully parametrized filter realization is optimal if and only if it is normal. As to the digital controller case, the expression for the measure of the closed-loop system has been achieved and the corresponding optimal controller realization problem has been attacked on the set of fully parametrized state-space realizations using a standard optimization algorithm. Numerical examples have confirmed the validity of the

proposed measure and shown that the optimal realizations (for both digital filter and controller cases) yield a much better stability performance than those shift operator-based canonical realizations.

We note that the optimal realizations obtained in Chapter 3 are usually fully parameterized. From a practical point of view, it would be better to implement the digital systems with such a structure that not only yields a large stability robustness but also possesses as few parameters to be implemented as possible.

In Chapter 4, based on the polynomial parametrization concept, two new sparse structures (the ρ DFII structure and ρ -realization) have been developed by reparametrizing the transfer function with a set of special polynomial operators, called ρ -operators. Both structures are very efficient in terms of implementation and contain some free parameters with which the structure can be optimized to achieve nice stability performance. The stability behavior of each structure has been analyzed and the corresponding optimal sparse structure problem has been solved using a genetic algorithm with practical considerations. Through design examples and the corresponding simulations, it is shown that the optimized sparse structures have a stability performance that is very close to (in some cases, better than) that of the fully parametrized optimal realizations and is always better than those of the structures based on the traditional shift operator and the prevailing delta operator.

It has been noted that for the ρ -operator-based structures in Chapter 4, when the roots of the polynomials are chosen such that they are close to the system poles, the pole sensitivities can be much reduced and hence the stability robustness increases. Due to the fact that the polynomials of ρ -operator-based structures have real roots only while the system poles are generally complex, the

degree of maximizing the stability robustness for the ρ -operator-based structures is limited. It is this limitation that motivates us to conduct further research for sparse structures with higher degree of stability performance.

Then in Chapter 5, two improved sparse structures (the ρ IDFIIt structure and ρ_I -realization) have been developed with a set of second-order operators. These structures have some more free parameters, with which the polynomials can have any assigned roots and hence the stability robustness can be further maximized while the price paid for that is very slight. The stability behavior of each structure has been analyzed and the optimal structure problem has been effectively attacked under the parameter dynamical range constraints by using a genetic algorithm. Design examples and simulations have shown that the optimized ρ IDFIIt structure and ρ_I -realization not only yield a much better stability performance than the optimal ρ DFIIt structure and ρ -realization, respectively, but also beat the fully parametrized optimal realization in terms of both stability robustness and computation efficiency.

The effect of roundoff noise in a digital controller has been studied for a discrete-time control system in Chapter 6. An analytical expression of the roundoff noise gain has been obtained for state-space controller realizations, with which the problem of identifying minimum roundoff noise structures has been solved on the set of fully parametrized realizations. We have analyzed the performance of the second-order operator-based structures proposed in Chapter 5 by deriving the corresponding expressions of roundoff noise gain and the problems of finding optimized sparse structures have been solved. Through the numerical example and simulations, the excellent performance of the proposed structures is demonstrated in terms of reducing roundoff noise and implemen-

tation complexity.

Finally in Chapter 7, we have analyzed the pole deviation for digital systems based on a second-order perturbation theory. An analytical expression for computing the second-order pole sensitivities has been derived, with which the pole modulus deviation has been approximated. Based on this approximation, a new measure has been defined to evaluate the system stability. The numerical example shows that this measure is a very good approximation of the unsolved classical stability measure and is more accurate than the approach based on the first-order approximation. This part of research is just a beginning and detailed framework need further investigation.

8.2 Recommendations for future work

Some of the results presented in this thesis suggest a series of interesting topics for future research. They are summarized as below.

First, there remains a great deal of research potential in the studies of polynomial parametrization approach and its applications. In this thesis, we applied the concept of polynomial parametrization into the digital system structure design. By developing two sets of special polynomial operators, two types of sparse structures were obtained. These structures are very efficient in terms of implementation and can be optimized to achieve robust performance against the FWL effects. In the future, it is suggested to explore some new sets of special polynomial operators, and apply them to the optimal FWL structure design not just restricted to the state-space realizations and the DFII structures, but extended to some other interesting structures. New efficient structures with

enhanced stability and roundoff noise performance are expected through the further research.

Another application of the polynomial parametrization concept is in the fields of system identification and parameter estimation. In [7], the use of polynomial parametrization in estimation and identification had been examined by exhibiting the effects of the choice of parametrization on the information matrix, and hence on the convergence speed of parameter estimation algorithms. The parametrization techniques presented in this thesis, as we conjectured, are possible to be employed in the estimation problems. Whether these techniques can be utilized to improve the numerical properties of the information matrix for those problems is a question which is left to be answered by future studies.

The two types of FWL errors, due to coefficient truncation and arithmetic roundoff respectively, were studied separately in this thesis. For these two FWL effects, the optimal structure problems were solved by maximizing the stability robustness or minimizing the roundoff noise gain. Since parameter perturbation and signal roundoff exist simultaneously in actual implementations, it is of great interest to understand better the connections between the effects of these two types of FWL errors. It would be even more appealing to find a global measure that takes account of both effects in a coherent way, and with this measure, to find the structures that maximize the stability robustness and minimize the roundoff noise gain simultaneously or in a weighted manner. To our knowledge, no such unified measure has been given in the existing literature. In the future, this challenging topic is worthy of deep investigation.

Most of the pole sensitivity-based stability measures popularly used nowadays, including the one proposed in Chapter 3, are derived with a first-order

approximation. In Chapter 7 of this thesis, we are the first to have studied the pole deviation based on a second-order perturbation theory and presented a method to compute second-order pole sensitivities analytically. This part of research is just a beginning. But the framework of second-order pole sensitivity study provides a new starting point to further investigate the stability issue for digital systems. More efforts are expected to be taken to develop new tractable stability measure based on this framework and to apply the measure into the optimal FWL structure design to explore structures with higher degree of stability robustness.

It is hoped that the research results obtained in this thesis will help to establish a basis for further investigations and future work towards more robust and efficient implementation of digital systems against the FWL effects.

Appendix A

A Brief Description of GA to Solve Optimization Problems

Genetic algorithms (GA) attempt to mimic (in an over-simplified manner) the natural evolution of species in order to solve difficult optimization problems. The book [98] by D. E. Goldberg provides an elaborate framework for the readers to obtain a comprehensive understanding of the genetic algorithms in search, optimization and machine learning. We just present in this appendix a brief description about how to apply GA to solve some optimization problems in this thesis. For those who are interested in more details, please refer to the book above.

The flowchart in Figure A.1 shows the procedure for GA to solve the optimal structure problems defined in Chapters 4 to 6. These problems are summarized as below.

- Type I: (4.16) and (4.26), which are to maximize the stability robustness μ_2 with respect to $\{\gamma_k\}$ and $\{\Delta_k\}$

- Type II: (5.16) and (5.29), which are to maximize the stability robustness μ_2 with respect to $\{\gamma_k\}$, $\{\Delta_k\}$ and $\{\bar{\eta}_m\}$
- Type III: (6.27) and (6.28), which are to minimize the roundoff noise gain G with respect to $\{\gamma_k\}$ and $\{\bar{\eta}_m\}$.

Denote θ as a solution candidate for one of the above problems

$$\theta \triangleq \begin{cases} \begin{bmatrix} \gamma_1 & \cdots & \gamma_K & \Delta_1 & \cdots & \Delta_K \end{bmatrix}, & \text{for Type I} \\ \begin{bmatrix} \gamma_1 & \cdots & \gamma_K & \Delta_1 & \cdots & \Delta_K & \bar{\eta}_1 & \cdots & \bar{\eta}_p \end{bmatrix}, & \text{for Type II} \\ \begin{bmatrix} \gamma_1 & \cdots & \gamma_K & \bar{\eta}_1 & \cdots & \bar{\eta}_p \end{bmatrix}, & \text{for Type III} \end{cases}$$

and $F(\theta)$ as the corresponding objective function

$$F(\theta) \triangleq \begin{cases} \mu_2, & \text{for Type I and II} \\ G, & \text{for Type III.} \end{cases}$$

The objective of GA is to act upon a set (or population) of solution candidates (or individuals), and cause the population to evolve towards the desired (optimum) solution that optimizes the corresponding objective function (maximize μ_2 or minimize G).

According to Figure A.1, the first step of GA is to generate an initial population, which can be either random or pre-specified. The individuals in the population are usually represented as binary strings. Let B_γ , B_Δ and $B_{\bar{\eta}}$ be the number of bits used to represent γ_k , Δ_k and $\bar{\eta}_m$, respectively, and M be the total number of solution candidates. The initial population can be represented

with a matrix of the dimension $\mathcal{R}^{M \times L}$, in which each row string represents a solution θ , and L is the total bit length of θ

$$L = \begin{cases} K(B_\gamma + B_\Delta), & \text{for Type I} \\ K(B_\gamma + B_\Delta) + pB_{\bar{\eta}}, & \text{for Type II} \\ KB_\gamma + pB_{\bar{\eta}}, & \text{for Type III.} \end{cases}$$

In Step 2, we first decode each solution candidate under the conditions given in (2.10) and (5.24), and then evaluate the corresponding objective function $F(\theta)$ for every candidate solution θ in the population. The objective function determines the strength of the candidate's ability to survive in the environment (fitness) and hence is used to select the candidates with a high level of fitness for reproduction. The optimum value for objective function and the best solution candidate are recorded for the current generation.

After Step 2, the evolutionary computation operators, such as reproduction, crossover, and mutation are applied from Step 3 to Step 5 to obtain the next generation of population. Reproduction operator copies solution strings from the current generation into a mating pool taking the fitness of solution candidates into consideration. Then the crossover operator is applied to the selected mating pool to create offsprings between the parent strings. The offsprings have the possibility of further perturbation by mutation. Usually, random operations on some of the strings are performed with a low probability. The mutation will introduce a degree of diversity to the population, prevent a premature convergence and help to sample unexplored regions of the search space. Through the evolutionary operations from Steps 3 to 5, the next generation of population can be obtained. Then Step 6 evaluates the objective function of each candidate in

the new population and yields the corresponding best solution candidate.

The operations from Step 3 to Step 7 make the solutions of optimizing the objective function propagate in an increasing number. These operations are repeated until the termination conditions in Step 7 are satisfied. For example, the GA can be terminated after a specified number of generations or when the fitness of the optimal solutions has no visible improvements over several generations.

Figure A.1: Flowchart of the GA to solve optimization problems

Author's Publications

1. Jinxin Hao and Gang Li, "An Efficient Controller Structure with Minimum Roundoff Noise Gain," *Accepted by Automatica*, Jun. 2006.
2. Jinxin Hao, Gang Li, and Chunru Wan, "Two Classes of Efficient Digital Controller Structures with Stability Consideration," *IEEE Transactions on Automatic Control*, vol. 51, pp. 164-170, Jan. 2006.
3. Jinxin Hao and Gang Li, "Polynomial Operator Based Sparse Controller Structures with Stability Consideration," *IEE Proceedings Control Theory and Applications*, vol. 152, pp. 521-530, Sept. 2005.
4. Jinxin Hao and Gang Li, "Improved ρ -realization for Digital Controller Implementation," *Electronics Letters*, vol. 40, pp. 976-978, Aug. 2004.
5. Jinxin Hao and Gang Li, "Optimal ρ -realization for Digital Controller Implementation with Stability Consideration," *Electronics Letters*, vol. 40, pp. 226-227, Feb. 2004.
6. Jinxin Hao and Gang Li, "An Efficient Structure for Finite Precision Implementation of Digital Systems," in *Proceedings of the Fifth International Conference on Information, Communications and Signal Processing (ICICS)*, Bangkok, Thailand, Dec. 2005.
7. Jinxin Hao, Gang Li, and Jun Wu, "Pole Deviation Analysis for Digital Systems Based on Second Order Perturbation Theory," in *Proceedings of the IEEE Symposium on Circuits and Systems (ISCAS)*, Kobe, Japan, May 2005.
8. Jinxin Hao and Gang Li, "Polynomial Operator-Based Digital Controller Structures of High Stability Performance and Computation Efficiency," in *Proceedings of the Ninth International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Kunming, China, Dec. 2004.

9. Jinxin Hao and Gang Li, "An Improved Stability Measure for Digital Filter Implementation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, Apr. 2003.
10. Gang Li, Zixue Zhao, and Jinxin Hao, "A Generalized Direct-Form II Transposed Structure for IIR Filter Implementation with Minimal Round-off Noise Gain," in *Proceedings of the IEEE Symposium on Circuits and Systems (ISCAS)*, Bangkok, Thailand, May 2003.

Bibliography

- [1] L. R. Rabiner and B. Gold, *Theory and Applications of Digital Signal Processing*. Prentice-Hall, 1975.
- [2] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing - Principles, Algorithms, and Applications*. Prentice-Hall, 3rd ed., 1996.
- [3] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck, *Discrete-Time Signal Processing*. Prentice-Hall, NJ, 1999.
- [4] S. Boyd and C. H. Barratt, *Linear Controller Design: Limits of Performance*. Prentice-Hall, Englewood Cliffs, NJ, 1991.
- [5] R. E. Skelton, T. Iwasaki, and K. M. Grigoriadis, *A Unified Algebraic Approach to Linear Control Design*. Taylor & Francis Ltd, London, 1998.
- [6] C. T. Mullis and R. A. Roberts, *Digital Signal Processing*. Addison-Wesley, 1987.
- [7] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. Communication and Control Engineering Series. London, U.K.: Springer-Verlag, 1993.
- [8] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. 23, pp. 551-562, Sept. 1976.
- [9] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 273-281, Aug. 1977.
- [10] T. Lin, M. Kawamata, and T. Higuchi, "A unified study on the roundoff noise in 2-D state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 724-730, Jul. 1986.

- [11] W. S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital-filter structures with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 965-973, Oct. 1986.
- [12] L. M. Smith and B. W. Bomar, "An algorithm for constrained roundoff noise minimization in digital filters with application to two-dimensional filters," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1359-1368, Nov. 1988.
- [13] T. Hinamoto, T. Hamanaka, and S. Maekawa, "A generalized study on the synthesis of 2-D state-space digital filters with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1037-1042, Aug. 1988.
- [14] T. Thong and B. Liu, "Error spectrum shaping in narrowband recursive digital filters," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 25, pp. 200-203, Apr. 1977.
- [15] T. L. Chang and S. A. White, "An error cancellation digital filter structure and its distributed-arithmetic implementation," *IEEE Trans. Circuits Syst.*, vol. 28, pp. 339-342, Apr. 1981.
- [16] D. C. Munson and D. Liu, "Narrowband recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. 28, pp. 160-163, Feb. 1981.
- [17] W. E. Higgins and D. C. Munson, "Noise reduction strategies for digital filters: error spectrum shaping versus the optimal linear state-space formulation," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 30, pp. 963-973, Dec. 1982.
- [18] M. Renfors, "Roundoff noise in error-feedback state-space filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 619-622, Apr. 1983.
- [19] W. E. Higgins and D. C. Munson, "Optimal and suboptimal error spectrum shaping for cascade-form digital filters," *IEEE Trans. Circuits Syst.*, vol. 31, pp. 429-437, May 1984.
- [20] P. P. Vaidyanathan, "On error-spectrum shaping in state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. 32, pp. 88-92, Jan. 1985.
- [21] T. I. Laakso and I. O. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Trans. Signal Processing*, vol. 40, pp. 1096-1107, May 1992.

- [22] T. Hinamoto, H. Ohnishi, and W. S. Lu, "Roundoff noise minimization of state-space digital filters using separate and joint error feedback/coordinate transformation optimizations," *IEEE Trans. Circuits Syst. I*, vol. 50, pp. 23-33, Jan. 2003.
- [23] P. Agathoklis and A. Kanellakis, "Roundoff noise 2-D digital filters using residue feedback," in *Proc. IEEE SoutheastCon 'Creative Technology Transfer - A Global Affair'*, pp. 12-15, Apr. 1994.
- [24] T. Hinamoto, S. Karino, and N. Kuroda, "2-D state-space digital filters with error spectrum shaping," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 766-769, May 1996.
- [25] T. Hinamoto, N. Kuroda, and T. Kuma, "Error feedback for noise reduction in 2-D digital filters with quadrantly symmetric or antisymmetric coefficients," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, pp. 2461-2464, Jun. 1997.
- [26] T. Hinamoto, S. Karino, N. Kuroda, and T. Kuma, "Error spectrum shaping in two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 46, pp. 1203-1215, Oct. 1999.
- [27] T. Hinamoto, K. Higashi, and W. S. Lu, "Separate/joint optimization of error feedback and coordinate transformation for roundoff noise minimization in two-dimensional state-space digital filters," *IEEE Trans. Signal Processing*, vol. 51, pp. 2436-2445, Sept. 2003.
- [28] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 31, pp. 884-888, Oct. 1984.
- [29] L. Thiele, "Design of sensitivity and roundoff noise optimal state-space discrete systems," *Int. J. Circuits Theory Appl.*, vol. 12, pp. 39-46, 1984.
- [30] L. Thiele, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 502-510, May 1986.
- [31] W. S. Gray, *A Geometric Approach to the Parametric Sensitivity of Dynamical Systems*. Ph. D. dissertation, Georgia Institute of Technology, Georgia, USA, 1989.
- [32] W. J. Lutz and S. L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1114-1122, Sept. 1988.

- [33] W. Y. Yan and J. B. Moore, "On L_2 -sensitivity minimization of linear state-space systems," *IEEE Trans. Circuits Syst. I*, vol. 39, pp. 641-648, Aug. 1992.
- [34] C. Xiao, "Improved L_2 -sensitivity for state-space digital system," *IEEE Trans. Signal Processing*, vol. 45, pp. 837-840, Apr. 1997.
- [35] G. Li, "Two dimensional system optimal realizations with L_2 -sensitivity minimization," *IEEE Trans. Signal Processing*, vol. 46, pp. 809-813, Mar. 1998.
- [36] T. Hinamoto and T. Takao, "Synthesis of 2-D state-space filter structures with low frequency-weighted sensitivity," *IEEE Trans. Circuits Syst. II*, vol. 39, pp. 646-651, Sept. 1992.
- [37] T. Hinamoto, Y. Zempo, Y. Nishina, and W. S. Lu, "An analytical approach for the synthesis of two-dimensional state-space filter structure with minimum weighted sensitivity," *IEEE Trans. Circuits and Syst. I*, vol. 46, pp. 1172-1183, Oct. 1999.
- [38] T. Hinamoto, S. Yokoyama, T. Inoue, W. Zeng, and W. Lu, "Analysis and minimization of L_2 -sensitivity for linear systems and two-dimensional state-space filters using general controllability and observability gramians," *IEEE Trans. Circuits Syst. I*, vol. 49, pp. 1279-1289, Sept. 2002.
- [39] T. Hinamoto and Y. Sugie, " L_2 -sensitivity analysis and minimization of 2-D separable-denominator state-space digital filters," *IEEE Trans. Signal Processing*, vol. 50, pp. 3107-3114, Dec. 2002.
- [40] P. E. Mantey, "Eigenvalue sensitivity and state variable selection," *IEEE Trans. Automat. Contr.*, vol. 13, pp. 263-269, Jun. 1968.
- [41] R. E. Skelton and D. A. Wagie, "Minimal root sensitivity in linear systems," *J. Guidance Contr.*, vol. 7, pp. 570-574, Sept.-Oct. 1984.
- [42] D. Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using residue feedback," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1210-1220, Oct. 1986.
- [43] G. Li, "On pole and zero sensitivity of linear systems," *IEEE Trans. Circuits Syst. I*, vol. 44, pp. 583-590, Jul. 1997.

- [44] S. Chen, J. Wu, R. H. Istepanian, and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 2149-2153, Nov. 1999.
- [45] J. Wu, S. Chen, G. Li, and J. Chu, "Optimal finite-precision state-estimate feedback controller realizations of discrete-time systems," *IEEE Trans. Automat. Contr.*, vol. 45, pp. 1550-1554, Aug. 2000.
- [46] G. Li and M. Gevers, "Optimal finite-precision implementation of a state-estimate feedback controller," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 1487-1499, Dec. 1990.
- [47] D. Williamson and K. Kadiman, "Optimal finite wordlength linear quadratic regulation," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1218-1228, 1989.
- [48] K. Liu, R. Skelton, and K. Grigoriadis, "Optimal controllers for finite wordlength implementation," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1294-1304, 1992.
- [49] G. Li, J. Wu, S. Chen, and K. Y. Zhao, "Optimum structures of digital controllers in sampled-data systems: a roundoff noise analysis," *IEE Proc. Control Theory Appl.*, vol. 149, pp. 247-255, May 2002.
- [50] A. G. Madievski, B. D. O. Anderson, and M. Gevers, "Optimum realizations of sampled-data controllers for FWL sensitivity minimization," *Automatica*, vol. 31, pp. 367-379, Mar. 1995.
- [51] R. H. Istepanian and J. F. Whidborne, *Digital Controller Implementation and Fragility: A Modern Perspective*. Advances in Industrial Control Series. London, U.K.: Springer-Verlag, 2001.
- [52] I. J. Fialho and T. T. Georgiou, "On stability and performance of sampled-data systems subject to wordlength constraint," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 2476-2481, Dec. 1994.
- [53] I. J. Fialho and T. T. Georgiou, "Optimal finite wordlength digital controller realizations," in *Proc. American Control Conf.*, San Diego, USA, pp. 4326-4327, Jun. 1999.
- [54] G. Li, "On the structure of digital controllers with finite word length consideration," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 689-693, May 1998.

- [55] R. H. Istepanian, G. Li, J. Wu, and J. Chu, "Analysis of sensitivity measures of finite-precision digital controller structures with closed-loop stability bounds," *IEE Proc. Control Theory Appl.*, vol. 145, pp. 472-478, May 1998.
- [56] J. Wu, S. Chen, G. Li, R. H. Istepanian, and J. Chu, "An improved closed-loop stability related measure for finite-precision digital controller realizations," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 1162-1166, Jul. 2001.
- [57] J. Wu, S. Chen, J. F. Whidborne, and J. Chu, "A unified closed-loop stability measure for finite-precision digital controller realizations implemented in different representation schemes," *IEEE Trans. Automat. Contr.*, vol. 48, pp. 816-822, May 2003.
- [58] W. S. Yu and H. J. Ko, "Improved eigenvalue sensitivity for finite-precision digital controller realizations via orthogonal Hermitian transform," *IEE Proc. Control Theory Appl.*, vol. 50, pp. 365-375, Apr. 2003.
- [59] B. W. Bomar and J. C. Hung, "Minimum roundoff noise digital filters with some power-of-two coefficients," *IEEE Trans. Circuits Syst.*, vol. 31, pp. 833-840, 1984.
- [60] G. Amit and U. Shaked, "Small roundoff realization of fixed-point digital filters and controllers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 880-891, Jun. 1988.
- [61] M. Iwatsuki, M. Kawamata, and T. Higuchi, "Statistical sensitivity and minimum sensitivity structures with fewer coefficients in discrete-time linear systems," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 72-80, Jan. 1989.
- [62] G. Li, B. D. O. Anderson, M. Gevers, and J. E. Perkins, "Optimal FWL design of state-space digital systems with weighted sensitivity minimization and sparseness consideration," *IEEE Trans. Circuits Syst.*, vol. 39, pp. 365-377, May 1992.
- [63] G. Li, M. Gevers, and Y. X. Sun, "Performance analysis of a new structure for digital filter implementation," *IEEE Trans. Circuits Syst. I*, vol. 47, pp. 474-482, Apr. 2000.
- [64] J. Kauraniemi, T. I. Laakso, I. Harttimo, and S. J. Ovaska, "Roundoff noise minimization in a direct form delta operator structure," in *Proc. IEEE Int. Acoust., Speech, Signal Processing*, Atlanta, GA, May, 1996.

- [65] J. Kauraniemi and T. I. Laakso, "Roundoff noise analysis of modified delta operator direct form structures," in *Proc. IEEE Int. Symp. Circuits Syst.*, 1997.
- [66] J. Kauraniemi, T. I. Laakso, I. Harttimo, and S. J. Ovaska, "Delta operator realizations of direct-form IIR filters," *IEEE Trans. Circuits Syst. II*, vol. 45, pp. 41-45, Jan. 1998.
- [67] N. Wong and T. S. Ng, "Roundoff noise minimization in a modified direct form delta operator IIR structure," *IEEE Trans. Circuits Syst. II*, vol. 47, pp. 1533-1536, Dec. 2000.
- [68] N. Wong and T. S. Ng, "Improved roundoff noise performance in a direct-form IIR filter using a modified delta operator," *IEEE Trans. Circuits Syst.*, vol. 2, pp. 773-776, May 2001.
- [69] N. Wong and T. S. Ng, "A generalized direct-form delta operator-based IIR filter with minimum noise gain and sensitivity," *IEEE Trans. Circuit Syst. II*, vol. 48, pp. 425-431, Apr. 2001.
- [70] V. Peterka, "Control of uncertain processes: applied theory and algorithms," *Kybernetika*, vol. 22, pp. 1-102, 1986.
- [71] R. H. Middleton and G. C. Goodwin, *Digital Control and Estimation: A Unified Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1990.
- [72] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Signal Processing*, vol. 41, pp. 629-637, Feb. 1993.
- [73] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parametrizations," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 803-807, May 1993.
- [74] S. Chen, J. Wu, R. H. Istepanian, J. Chu, and J. F. Widborne, "Optimizing stability bounds of finite-precision controller structures for sampled-data systems in the delta-operator domain," *IEE Proc. Control Theory Appl.*, vol. 146, pp. 517-526, Jun. 1999.
- [75] S. Chen, R. H. Istepanian, J. Wu, and J. Chu, "Comparative study on optimizing closed-loop stability bounds of finite-precision controller structures with shift and delta operators," *Systems and Control Letters*, vol. 40, pp. 153-163, 2000.

- [76] S. R. Parker and S. F. Hess, "Limit-cycle oscillations in digital filters," *IEEE Trans. Circuits Theory*, vol. 19, pp. 687-697, Nov. 1971.
- [77] M. Buttner, "A novel approach to eliminate limit cycles in digital filters with minimum increase in the quantization noise," in *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 291-294, Apr. 1976.
- [78] P. S. R. Diniz and A. Antoniou, "More economical state-space digital filter structures which are free of constant-input limit cycles," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 807-815, Aug. 1986.
- [79] B. D. Green and L. E. Turner, "New limit cycle bounds for digital filters," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 365-374, Apr. 1988.
- [80] P. H. Bauer, "Limit cycle bounds for floating-point implementations of second-order recursive digital filters," *IEEE Trans. Circuits Syst. II*, vol. 40, pp. 493-501, Aug. 1993.
- [81] B. W. Bomar, "Low-roundoff-noise limit-cycle-free implementation of recursive transfer functions on a fixed-point digital signal processor," *IEEE Trans. Industr. Electron.*, vol. 41, pp. 70-78, Feb. 1994.
- [82] A. N. Willson, Jr., "Limit cycle due to adder overflow in digital filters," *IEEE Trans. Circuits Theory*, vol. 19, pp. 342-346, Jul. 1972.
- [83] W. T. Mills, C. T. Mullis, and R. A. Roberts, "Digital filter realizations without overflow oscillations," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 26, pp. 334-338, Aug. 1978.
- [84] J. H. F. Ritzerfield, "A condition for the overflow stability of second-order digital filters that is satisfied by all scaled state-space structures using saturation," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1049-1057, Aug. 1989.
- [85] B. W. Bomar, "Finite Wordlength Effects," in *The Circuits and Filters Handbook* (W. K. Chen, ed.), pp. 2739-2757, Boca Raton, FL: CRC Press, 2nd ed., 2003.
- [86] R. H. Istepanian, J. Wu, and J. F. Whidborne, "Controller realizations of a teleoperated dual-wrist assembly system with finite word length considerations," *IEEE Trans Control Syst. Technology*, vol. 9, pp. 624-628, Apr. 2001.
- [87] T. Kailath, *Linear Systems*. Prentice-Hall, Englewood Cliffs, 1980.

-
- [88] J. F. Whidborne and D. W. Gu, "Optimal finite-precision controller and filter implementation using floating-point arithmetic," in *Proc. 15th IFAC World Congr.*, Barcelona, Spain, Jul. 2002.
- [89] L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, P. M. Young, and J. C. Doyle, "A formula for computation of the real stability radius," *Automatica*, vol. 31, pp. 879-890, Jun. 1995.
- [90] G. W. Stewart, *Introduction to Matrix Computations*. Academic Press, 1973.
- [91] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. Oxford, U.K.: Clarendon, 1965.
- [92] J. Demmel, "Nearest defective matrices and the geometry of ill-conditioning," in *Reliable Numerical Computation*, pp. 35-56, edited by M. G. Cox and S. Hammarling, Clarendon Press, Oxford, 1990.
- [93] D. F. Delchamps, "Stabilizing a linear system with quantized state feedback," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 916-924, Aug. 1990.
- [94] R. Bellman, *Introduction to Matrix Analysis*. 2nd ed., New York: McGraw-Hill, 1970.
- [95] *Optimization Toolbox for Use with MATLAB*. ver. 2.2, The MathWorks Inc., 2002.
- [96] T. Chen and B. A. Francis, "Input-output stability of sampled-data systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 50-58, Jan. 1991.
- [97] S. Chen and B. L. Luk, "Adaptive simulated annealing for optimization in signal processing applications," *Signal Processing*, vol. 79, pp. 117-128, 1999.
- [98] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.
- [99] J. H. Holland, *Adaptation in Natural and Artificial System*. Ann Arbor, University of Michigan Press, 1975.
- [100] D. Suckley, "Genetic algorithms in the design of FIR filters," *Circuits, Devices and Systems, IEE Proceedings G*, vol. 138, pp. 234-238, Apr. 1991.

-
- [101] A. Roberts and G. Wade, "A structured GA for FIR filter design," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Chelmsford, UK, vol. 1, pp. 16/1-16/8, Nov. 1993.
 - [102] P. B. Wilson and M. D. Macleod, "Low implementation cost IIR digital filter design using genetic algorithms," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Chelmsford, UK, vol. 1, pp. 4/1-4/8, Nov. 1993.
 - [103] M. S. White and S. J. Flockton, "A comparative study of natural algorithms for adaptive IIR filtering," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Chelmsford, UK, vol. 1, pp. 22/1-22/8, Nov. 1993.
 - [104] T. Wicks and S. Lawson, "Genetic algorithm design of wave digital filters with a restricted coefficient set," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Chelmsford, UK, vol. 1, pp. 17/1-17/7, Nov. 1993.
 - [105] J. F. Whidborne and R. H. Istepanian, "Optimal finite-precision PID controller structures using genetic algorithm," in *Proc. 14th IFAC World Congr.*, Beijing, China, pp. 181-186, 1999.
 - [106] J. F. Whidborne and R. H. Istepanian, "Genetic algorithm approach to designing finite-precision controller structures," *IEE Proc. Control Theory Appl.*, vol. 148, pp. 377-382, Sept. 2001.
 - [107] M. D. Greenberg, *Advanced engineering mathematics*. Prentice-Hall, 1998.