DSC SUBMISSION TO THE UNIVERSITY OF SOUTHAMPTON

Intelligent Nonlinear Learning Machines

by
Dr. S. Chen
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
sqc@ecs.soton.ac.uk
http://www.ecs.soton.ac.uk/∼sqc/

Acknowledgments

I am indebted to many people, who have enhanced my understanding of the subject of this submission during my 18-year career. First of all, I would like to thank Professor Peter D. Roberts of the City University, my PhD supervisor, who taught me how to do research and how to write research papers. My particular gratitude is due to Professor Stephen A. Billings of the University of Sheffield, who introduced me to the important topic of nonlinear system identification and has been inspirational in shaping my views concerning the various aspects of intelligent nonlinear systems. Professor Peter M. Grant, Professor Bernard Mulgrew and Professor Steve McLaughlin of the University of Edinburgh have enhanced my understanding regarding adaptive systems and signal processing. I would like to thank them for the enlightenment gained during my stay at Edinburgh. My special thanks are due to Professor Chris J. Harris and Professor Lajos Hanzo. Without their encouragements, support and help, this submission would not have been completed. I would also like to express my gratitude to the School of Electronics and Computer Science for creating a scholarly environment I have enjoyed over the past six years.

I am particularly grateful to many of my academic colleagues and students for their support as well as valuable contributions. Special thanks are due to all my co-authors cited in the submission, particularly to Dr. Xia Hong of the University of Reading and Professor Jun Wu of the Zhejiang University, China.

Finally, I would like to dedicate this submission to my beloved parents as a modest token of my appreciation of their unabating support and to my son, Alan, for making the life sparkling and for helping me to appreciate any intellectual gift one was born with.

Sheng Chen
School of ECS
University of Southampton

BRIEF CURRICULUM VITAE OF DR. S. CHEN

OFFICE ADDRESS

School of Electronics and Computer Science
University of Southampton
Highfield, Southampton SO17 1BJ, UK
Tel.: 023 8059 6660, Fax: 023 8059 4508
E-mail: sqc@ecs.soton.uk.ac, http://www.ecs.soton.ac.uk/∼sqc/

EDUCATION AND QUALIFICATIONS

January 1978 – January 1982: East China Petroleum Institute, BSc in Control Engineering, First Class Honors

October 1983 – October 1986: City University, London, PhD in Control Engineering

WORK EXPERIENCE AND EMPLOYMENT

1 October 1986 – 30 April 1989: Postdoctoral research associate, Department of Control Engineering, University of Sheffield, UK

1 May 1989 – 31 October 1993: Research fellow, Department of Electrical Engineering, University of Edinburgh, UK

1 November 1993 – 31 August 1996: Senior lecturer, Department of Electrical and Electronic Engineering, University of Portsmouth, UK

1 September 1996 – 31 August 1999: Reader, Department of Electrical and Electronic Engineering, University of Portsmouth, UK

1 September 1999 – present: Reader, School of Electronics and Computer Science, University of Southampton, UK

AWARDS AND DISTINCTIONS

British Council Scholarship for overseas PhD students, 1984

IEE Marconi Premium awarded for the paper “Orthogonal least squares algorithm for training multi-output radial basis function networks”, 1994

Included in the list of the most highly cited researchers within the engineering category in the database of the world’s most highly cited researchers, compiled by Institute for Scientific Information (ISI) of the USA,

2004

The best paper award for the paper “Orthogonal least square with boosting for regression”, presented at 5th International Conference on Intelligent Data Engineering and Automated Learning, 2004

INTERNATIONAL SCHOLARSHIP

I have served as PhD examiner for the Universities of Reading, Edinburgh and Leeds, UK; the laboratoire des signaux et systèmes, CNRS-Supélec, France.

I have served as MPhil examiner for Nanyang Technological University, Singapore; City University of Hong Kong, Hong Kong, China.

I have served as IPC member for the IEEE International Conference on Mechatronics and Machine Vision in Practice, 2001 to 2005; TPC member for the IEEE International Conference on Communications, Advanced Wireless Communications Systems Symposium, 2002; TPC member for the International Conference on Wireless Networks, 2002 to 2004; TPC member for the International Conference on Intelligent Data Engineering and Automated Learning, 2004, 2005; and IPC member for the International Conference on Natural Computation, 2005.

I have been a regular reviewer for ten IEEE Transactions, three IEE Proceedings, and many other major international journals; as well as a regular reviewer for some major IEEE International Conferences, including ICC, GLOBECOM, CDC and ACC, and many other major international conferences.

I have served as international advisory committee member for the following Journal: Advanced Electronic Communications – Research & Education; Executive committee member of the Chinese Automation and Computer Society in UK.

RESEARCH

During my 18-year career **I have authored and co-authored 12 book chapters, 115 journal papers and 91 conference papers. More than half of my research papers have been published after I joined the University of Southampton in 1999**. Many of my publications are IEEE publications and hence can be found in Portable Data Format (PDF) on the IEEE Xplore database. I have over 50 journal papers in the IEEE Xplore database. My full publications can also be found at the Publications Database of the School of ECS:

`http://eprints.ecs.soton.ac.uk/`

and the journal papers relevant to the subject of this submission are enclosed in printed format.

My research is wide-ranging. Apart from conferences, I published in the following nine different IEEE Transactions: Transactions on Automatic Control; Circuits and Systems for Video Technology; Communications; Neural Networks; Signal Processing; Systems, Man and Cybernetics Part A; Systems, Man and Cybernetics Part B; Vehicular Technology; Wireless Communications; plus IEEE Signal Processing Letters. I also published in 3 IEE journals and many other international journals.

Many of my journal papers are highly cited. I have 6 journal papers, each having an ISI citation index of over 100, and my classical paper "Orthogonal least squares learning algorithm for radial basis function networks," published in *the IEEE Transactions on Neural Networks* in 1991, received an ISI citation index of over 430 by the end of November 2004. In February 2004, my name was included in the list of the most highly cited researchers within the engineering category in the database of the world's most highly cited researchers, compiled by ISI.

THE SUBMISSION

The various topics covered by this submission will be organised around the eight broad subject areas covered in my peer-reviewed journal papers **[1]-[103]**, listed in the reference section of this submission. As indicated in the title, the main theme behind this submission is the research of **intelligent nonlinear learning** systems.

The papers constituting this submission were co-authored by the following students, research assistants and academic colleagues.

Supervised Undergraduate and PhD Students: N.N. Ahmad, M.Y. Alias, L.C. Anderson, R.J. Backhouse, E.S. Chng, T.B. Cook, S. Galt, Z. He, H.B. Jamaluddin, W. Luo, A.K. Samingan, J. Wigger, A. Wolfgang, Y. Wu, Z.R. Yang.

Supervised Visitors and Research Assistants: K. Alkadhimi, P.E. An, J. Chu, M.J. Korenberg, G. Li, S. Urwin-Wright, J. Wu, K.Y. Zhao.

Academic Co-authors: S.A. Billings, M. Brown, C.F.N. Cowan, G.J. Gibson, P.M. Grant, S. Gunn, L. Hanzo, C.J. Harris, X. Hong, R.H. Istepannian, Y. Liu, B. Luk, S. McLaughlin, B. Mulgrew, D. Sander, P.M. Sharkey, J.F. Whidborne.

The work presented in this submission has not been submitted for a degree in the University of Southampton or any other university before.

Contents

DSC SUBMISSION TO THE UNIVERSITY OF SOUTHAMPTON

Intelligent Nonlinear Learning Machines

by

Dr. S. Chen

1 The Motivation and Background

Learning is fundamental to all walks of science and engineering. Fig. 1 depicts a generic learning model or machine. Such a learning machine takes an input provided by the environment and produces an output. The purpose of the learning algorithm is to adjust the model so that the model output matches a reference as close as possible in some statistic sense. In so-called supervised learning, the environment also provides the reference in the form of a desired output or target for the model. By contrast, in unsupervised learning, the environment does not provide a desired output for the learning machine. The learning machine must somehow construct a reference itself, often based on some statistical information regarding the underlying data generating mechanism. The classical learning machine deals with situations, where the model is linear and the learning rule is also "linear". For a linear model, the model output is linear with respect to both the model inputs and model parameters. The linear learning machine typically adopts a quadratic cost function of the modelling error for adjusting the model parameters, and the gradient of this cost function of a linear model is proportional to the product of the error signal and model input. This is the basic characteristic of linear learning that defines a linear learning machine. Linear learning has been studied widely and issues concerning a linear learning system are well understood.

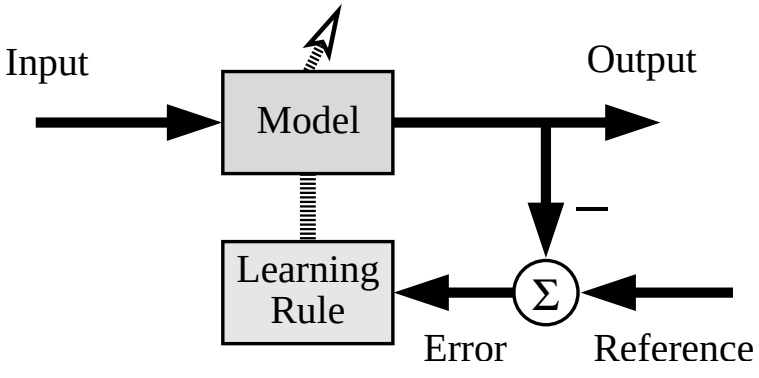

Figure 1: Schematic of Generic Learning Machine.

Most practical systems are however nonlinear to some extent, and hence nonlinear models or learning rules are required to achieve an adequate performance. This submission deals with nonlinear learning machines. A nonlinear learning machine is defined by either or both of the following characteristics:

- The model output is nonlinear with respect to the model input.
- The learning rule that is used to adjust the model parameters is nonlinear.

That is, a nonlinear model and/or a nonlinear learning rule constitute a nonlinear learning system. Note that a nonlinear model by definition must be nonlinear in the model input but it can be either linear in the model parameters or nonlinear in the model parameters. A purely linear-in-the-parameters nonlinear model has an obvious advantage, since linear learning methods can be adopted, but it may require a much larger model dimension, compared to a nonlinear-in-the-parameters representation. Interestingly, a nonlinear learning machine may consist of a linear model with a nonlinear learning rule. A nonlinear learning rule accrues, for example, by adopting a more complicated and relevant cost function than the quadratic function of the model error. Essentially, for a nonlinear learning machine, the gradient of the cost function with respect to the model parameters is no longer proportional to the product of the error signal and the model input. This has some serious implications for the learning process.

For most applications, a nonlinear learning machine offers significant performance improvements over a linear one, as a nonlinear approach is capable of approximating or modelling the underlying data generating mechanism much better. Compared to the linear learning approach, however, the nonlinear learning approach presents some major difficulties. The main challenges are:

1. **The curse of dimensionality**. A nonlinear model can easily become excessively large. A huge model not only has little practical use but also leads to over-fitting. An over-fitted model simply fits into the noise that is present in the training data. This prevents the model from discovering the underlying data generating mechanism and leads to poor generalisation performance.
2. **The complexity of learning process**. A nonlinear learning process is often associated with high computational requirements, and typically nonlinear learning problems are numerically ill-posed. The slow convergence rate of a nonlinear learning process is also a major consideration. Moreover, care must be exercised to avoid a nonlinear learning system for being trapped at some bad local solution.

To overcome or alleviate these difficulties, a successful nonlinear learning machine must be an *intelligent* one.

First of all, a nonlinear learning machine must be intelligent and hence capable of determining the correct model structure and arriving at a parsimonious representation. The objective of modelling from data is not that a model can explain the training data well. Rather, a model should possesses good generalisation performance. That is, it should capture the underlying dynamics that generate the training data instead of simply fitting into the training data. A model should also be easy to interpret and extract knowledge from, since a model is used to aid the understanding of the underlying data generating mechanism. All these depend crucially on the ability of the learning process to derive an appropriate *sparse* model.

A nonlinear learning machine must be intelligent and hence capable of handling computation efficiently, being robust to numerical ill-conditioning and escaping from bad local solution traps. In today's data rich environment, efficiency of a learning machine is a crucial factor for success. In many practical applications, such as those found in communication systems, fast convergence speed of a learning machine is essential

to meet time constraints. The cost functions associated with nonlinear learning are often multi-modal and typically highly complex. To be successful, a learning algorithm must be robust and able to find a global optimal solution or at least avoid bad local solutions.

Proposing a range of intelligent nonlinear learning machines to meet the above-mentioned challenges and to successfully apply them in various key practical applications has been the motivation of my research and constitutes the focus of this submission. The novel contributions of my research efforts spanning the past decade or so are outlined in this submission with reference to the journal papers **[1]-[103]**. The submission is organised in eight broad subject areas, and all these research areas can be unified within the nonlinear learning machine framework of Fig. 1. The eight topics covered in this submission may be grouped into the four parts.

Part 1 of the submission is composed of three chapters, each covering an important topic of machine learning. Specifically, Chapter 2 contains my novel contributions to data regression modelling, Chapter 3 outlines my work on classification, and Chapter 4 summarises my recent research in probability density function estimation. Regression and classification are supervised learning problems, while probability density function estimation is an unsupervised one.

Part 2 focuses specifically on the applications of intelligent nonlinear learning to communication systems, and consists of three chapters. Chapter 5 presents my original contributions to the minimum bit error rate (MBER) design of linear receivers, while Chapter 6 summarises my novel research on nonlinear receiver design. These two topics belong to the class of supervised learning, as training is assumed. My work on blind equalisation, which can be classified as unsupervised learning, is summarised in Chapter 7.

Chapter 8 forms Part 3 of the submission, which contains my recent contributions to finite-precision digital controller design. This problem is an interesting and challenging nonlinear learning problem, requiring intelligent approaches to tackling it. Referring to Fig. 1, for this problem, the *model* is a closed-loop control system consisting of a linear plant model and a linear controller, the *reference* is the designed closed-loop performance if the controller is implemented in infinite precision or at a sufficiently high precision, and the *output* is the actual closed-loop performance given an actual finite-precision controller. Learning or optimisation aims to find some optimal finite-precision realizations of the controller, which minimises some measure of the difference between the desired closed-loop performance and the actual closed-loop performance (as well as maintaining the closed-loop stability margin).

Evolutionary computation methods have found wide applications in intelligent nonlinear learning, owing to their ability to find globally optimal solutions. Part 4 of the submission consists of a single chapter, Chapter 9, which collects my research on the applications of evolutionary computation and optimisation methods to various nonlinear learning problems.

2 Regression

Having a good generalisation capability and a sparse representation are two key requirements in creating an intelligent learning machine. These two considerations have been the main motivations for my research on learning machines. The journal papers **[1]-[18]** summarise my recent new results in the area of intelligent

regression modelling. Most of these advances were achieved after I joined the University of Southampton in 1999, and these new contributions provide significant enhancements to my previous research in this area.

Since the late 1980s, I have been working on nonlinear regression modelling, and before joining the University of Southampton I developed a range of tools or algorithms **[19]-[39]**. The most successful one of these nonlinear modelling toolkits has been the orthogonal least squares (OLS) algorithm for forward selection regression **[26],[34]**. The basic principle of the OLS algorithm is simple but remarkably effective. From a large pool of the candidate regressors, the algorithm selects significant and relevant regressors one by one to form the final model, and the selection procedure is terminated when some prescribed modelling accuracy is achieved, yielding a parsimonious subset model that contains only significant regressors. This selection is carried out by incrementally minimising the mean square error over the training data set. Mathematically, the model regressors form a set of bases that span the model space. By a simple and efficient orthogonalisation, these original model bases are transferred into the orthogonal bases. Each orthogonal base corresponds to an original model regressor. This allows the subset model selection to be carried out efficiently in the new orthogonal space. Since its derivation, the OLS algorithm has become very popular among nonlinear data modelling practitioners, owing to its ability to efficiently construct parsimonious models that generalise well. It has become a standard nonlinear modelling toolkit and finds wide applications in many diverse fields of science and engineering. The paper **[26]** has received an ISI citation index 432 and the paper **[34]** has received an ISI citation index 171 by the end of November, 2004.

My recent works **[1]-[18]** involve enhancing the OLS algorithm with a range of advanced techniques for aiding intelligent modelling. The main novel contributions are summarized as follows.

1. *Combining the OLS algorithm with regularisation techniques*. By introducing regularisation in the orthogonal modelling space and choosing the regularisation penalty as the l_2-norm of the model weights in the orthogonal space, the computational efficiency of the OLS forward selection procedure is preserved. The combined approach offers an enhanced generalisation performance with sparser representations. Furthermore, regularisation also helps in deciding when to terminate the model selection procedure. An alternative regularisation method called the basis pursuit, which employs the regularisation penalty as the l_1-norm of the model weights, has also been combined with the OLS algorithm.

2. *Combining the OLS algorithm with optimal experimental design criteria*. An optimal experimental design aims to optimise both the parameter efficiency and model robustness. By combining the OLS algorithm with an optimal experimental design, such as the D-optimality design, an enhanced model construction algorithm is obtained. With an appropriate choice of the composite cost function, the model construction becomes automatic, and the user does not have to specify any additional criterion to terminate the model selection procedure. Regularisation can also be incorporated naturally with this combined algorithm to further enforce the model sparsity.

3. *Model construction by directly optimising the generalisation capability*. By adopting the leave-one-out test score, which is a measure of the model generalisation capability, within the orthogonal forward selection framework, the resultant algorithm efficiently constructs a sparse model by incrementally optimising the model's generalisation capability. This is the first time that a model generalisation criterion is directly utilised in **selecting** the model regressors, rather than the usual use of helping to

decide the model size. Coupled with the advanced regularisation techniques, this approach results in a state-of-the-art toolkit for sparse regression modelling.

4. *Intelligent knowledge discovery and rule extraction from data*. The OLS decomposition has been extended to the neurofuzzy rule-based model construction. Specifically, a locally regularised OLS algorithm, combined with a D-optimality experimental design for the subspace based rule selection, has been developed for the fuzzy rule regularisation and subspace information extraction. This provides a powerful construction algorithm for extracting the relevant and significant rules from a large dictionary and offers a state-of-the-art technique for discovering knowledge from data.

Below are my new contributions in regression given in detail with reference to the corresponding journal papers.

2.1 Orthogonal Least Squares Combined with Regularisation

Regularisation is one of the principal techniques for improving the generalisation properties of a learning machine. For the large class of linear-in-the-weights nonlinear models, such as the kernel-based model or more generally the radial basis function (RBF) network, a useful regularisation technique is based on an l_2-norm penalty cost of the model weights, as this can be naturally incorporated into the classic quadratic least squares cost function associated with the learning problem. From the viewpoint of the Bayesian learning framework, this choice of regularisation is equivalent to choosing a Gaussian distribution as a prior for the model weights. Classically, regularisation is applied to large full-size models for the purpose of alleviating over-fitting and hence improving the model's generalisation properties. The beauty of the Bayesian learning is to "let the data speak" – the Bayesian procedure learns not only the model parameters but also the regularisation parameters or equivalently the hyperparameters of the prior distribution. The latter is usually carried out via an iterative evidence procedure. Updating the regularisation parameters is based on the inverse of the combined cost function Hessian matrix, which is non-trivial and often becomes numerically ill-conditioned. Moreover, for a complex large model, some of the eigenvalues of this Hessian matrix may even become numerically negative, causing severe learning difficulties.

The paper **[1]** is the first known journal publication to introduce regularisation within the framework of the OLS forward subset model construction. It points out that the parsimonious principle alone is not entirely immune to over-fitting and, therefore, regularisation is also useful for the small model construction. The main contribution of the journal paper **[1]**, however, is the introduction of regularisation in the orthogonal model space, rather than in the original model space. Because of the one-to-one relationship between the original model regressors and the orthogonal model bases, this type of regularisation is exactly equivalent to the original one defined in the original model space, but it offers significant computational advantages. Firstly, the efficient OLS subset selection procedure remains unaffected by the introduction of regularisation. Secondly and most importantly, the Hessian matrix of the composite cost function becomes diagonal and is always well-conditioned (as the OLS algorithm has a built-in mechanism to automatically guarantee well-conditioning). Thus, updating the regularisation parameters is computationally trivial. Coupling of the OLS subset selection procedure and this regularisation technique enhances both, yielding a computationally efficient algorithm for constructing the sparse models that generalise well.

Many nonlinear models, such as the RBF network or kernel-based model, exploit the advantages of linear learning. This is achieved however only after some nonlinear parameters of the model, such as the kernel width, have been fixed or learnt by some other means, typically based on a cross validation. The choice of the RBF width for example has a critical influence on the generalisation capability of a RBF network. It is also plausible that using a different kernel width will require a different (optimal) regularisation parameter. The journal paper **[2]** proposed a novel model construction algorithm by combining a genetic algorithm (GA) optimisation with the regularised OLS learning. This novel approach can be viewed as forming a two-layered learning hierarchy. At the lower layer, with the given RBF width and regularisation parameter, the regularised OLS algorithm constructs a parsimonious subset model from the training data set. The upper layer involves choosing the optimal RBF width and regularisation parameter jointly by optimising the model's generalisation capability using an additional validation data set. As the model's generalisation capability is a highly complex function of both the RBF width and regularisation parameter, the GA is used to arrive at a globally optimal solution. This construction approach attains a significantly enhanced performance at the cost of some additional computational complexity involving the GA optimisation.

In many applications of regularisation, such as those considered in **[1],[2]**, a single common regularisation parameter is employed for every model weight. This corresponds to choosing a same prior distribution for every model weight. For many learning machines, such as kernel models, each regressor is modelled on a data point. Each data (regressor) has different importance or relevance, and this should be reflected in its prior distribution. This is the motivation behind using the multiple-regularisers or local regularisation, in which each model weight is associated with an individual regularisation parameter. Local regularisation has a remarkable property of enforcing model sparsity. During the iterative optimisation, many of the regularisation parameters are driven to very large values and this in turn forces the corresponding model weights to take very small near-zero values. Thus, the model regressors associated with these near-zero weights can be pruned out from the final model. My novel research on this advanced regularisation technique is summarised in the journal paper **[3]**. In the submission **[3]**, a locally regularised OLS algorithm is developed for constructing sparse multi-output regression models. This algorithm combines the advantages of both the OLS model selection, which has the ability to select only the significant regressors to explain the training data, and the local regularisation, which enforces the sparsity of the models. The end result is an efficient construction algorithm that is capable of producing parsimonious regression models with excellent generalisation performance. As local regularisation is introduced in the orthogonal model space, the efficiency of the subset selection procedure is ensured. Moreover, it has been shown in **[3]** that the decision concerning when to terminate the model selection procedure is greatly assisted by the local regularisation.

2.2 Orthogonal Least Squares Combined with Optimal Experimental Design

Optimal experimental designs have been used to construct smooth model response surfaces based on the appropriate setting of the experimental variables under well-controlled experimental conditions. In an optimal experimental design, the model's adequacy is evaluated according to a specific design criterion that is a statistical measure of the goodness of experimental design by virtue of design efficiency and experimental effort. For regression models, in particular, the model's adequacy is measured as a function of the eigenvalues of the design matrix, since it is known that the eigenvalues of the design matrix are linked to the covariance matrix of the least squares parameter estimate. There are a variety of optimal experimental

design criteria based on different aspects of experimental design. The D-optimality criterion is the most effective in optimising the parameter efficiency and model robustness via the maximisation of the determinant of the design matrix. The traditional nonlinear model structure determination based on optimal experimental designs is however inefficient and computationally prohibitive, incurring the curse of dimensionality. The journal papers **[4]-[6]** of the submission summarise my novel research on applying optimal experimental designs to aid efficient parsimonious regression modelling.

To appreciate the benefits of incorporating an optimal experimental design in the subset model construction, it is informative to point out that there are two key questions to be answered for any model construction process, namely which regressors to select and when to terminate the selection. The OLS algorithm is very effective in selecting significant regressors based on the training performance. The number of regressors that the model should have cannot be quantified based on the same training performance, since typically this training performance is continuously improved as more model terms are added, which eventually leads to over-fitting or a poor generalisation performance. Although the local regularisation can help to alleviate this difficulty, ideally the modeller requires some additional criterion to decide when to stop the selection procedure and ultimately to generate a parsimonious model that generalises well. Let us consider the usual information based criteria. An information based criterion is typically a composite cost function consisting of the mean square error over the training data set and some model complexity penalty term. Such an information criterion is at least locally convex, and there exists an "optimal" model size at which the composite cost function reaches a minimum value. While the information based criteria provide the information concerning when to terminate the model selection, they do not tell the modeller which model terms should be selected. That is, the use of an information based criterion in forward regression only affects the termination point of the model selection, but does not penalise the regressor that would cause poor model performance (e.g. too large a variance of the parameter estimate or ill-posedness of the regression matrix), if it is selected. Optimal experimental design criteria offer better solutions as they are directly linked to the model's efficiency and the robustness of parameter estimate.

The work of **[4]-[6]** employed a composite cost function consisting of the training mean square error and the negative log function of the determinant of the design matrix. The regressor selection is thus based on incrementally minimising the training mean square error and simultaneously maximising the determinant of the corresponding design matrix. Due to the orthogonal decomposition, the design matrix is diagonal, and this ensures that the resultant OLS forward selection procedure is as computationally efficient as the original OLS algorithm. The composite cost function combining the training performance and the D-optimality criterion is at least locally convex, and there exists an optimal model size at which it achieves a minimum value. Thus, the cost function used in the regressor selection also provides the decision regarding when to stop the selection. The modeller only has to specify a weighting for the D-optimality cost in the composite cost function and the entire model construction procedure becomes automatic. The value of this weighting does not critically influence the model selection procedure and it can be chosen with ease from a wide range of values. Furthermore, this approach can be naturally combined with the local regularisation. While the D-optimality design improves the model's efficiency and parameter robustness, the local regularisation enforces the model's sparsity and avoids over-fitting. The coupling effects of these two complimentary methods further enhance each other, yielding an efficient yet simple algorithm for constructing sparse regression models with excellent generalisation performance, especially under highly noisy learning conditions.

2.3 Orthogonal Least Squares with Leave-One-Out Test Score and Local Regularisation

The ultimate goodness of a model is characterised by its generalisation performance, yet the achievable training performance has been the main driving force in model construction, namely in deciding which regressors to select. Cross validation is an approach commonly employed to ensure an adequate model's generalisation capability. In a typical cross validation application, the model is constructed based on a training data set and the constructed model is evaluated using a separate validation data set. The performance over the validation set can be seen as an indicator of the model's generalisation performance. The optimal model can thus be determined as the one that minimises this generalisation mean square error over the validation set. It should be obvious that in this approach the model's generalisation capability is only used as a measure of the model complexity, but is not involved in the selection of the model regressors. This underlying "problem" will remain, so long as the basic criterion for the model construction procedure is the training performance. Even in the locally regularised OLS algorithm combined with the D-optimality experimental design, which takes into account the model's generalisation ability in selecting regressors, the main component in the model construction criterion is the training performance. Arguably, a better and more natural approach is directly using a measure of the model's generalisation capability in the model selection procedure rather than only using it as a measure of the model complexity. This is the main motivation behind the work presented in **[7],[8]**.

A commonly used cross validation technique is the delete-one cross validation. The idea is illustrated by the problem of choosing the best model from the K candidates using an N-sample training data set. For each candidate, each data point in the training set is sequentially set aside in turn, the model is estimated using the remaining $N-1$ data points, and the prediction error is derived using the only data point that was removed from training. The leave-one-out (LOO) test score of the model is the mean square error averaged over all the N prediction errors. To select the best model from the K candidates, the same modelling procedure is applied to each of the K candidates, and the model candidate with the minimum LOO test statistic is selected. For the linear-in-the-parameters models, there is an elegant way of generating the LOO test score, without actually sequentially splitting the training data set and repeatedly estimating the associated models. However, the traditional nonlinear model structure determination based on the delete-one cross validation is computationally prohibitive. The journal papers **[7],[8]** in the submission developed an efficient way for sparse regression modelling using the LOO test score. The method proposed in **[7],[8]** uses the locally regularised OLS algorithm as the basic toolkit for model construction, except that the regressor selection is based on incrementally minimising the LOO test score, rather than the original regularised training mean square error. Significant advantages of this novel model construction algorithm are:

- The model construction is based on directly optimising the model's generalisation capability, without resorting to use a separate validation data set. Moreover, owing to the orthogonal decomposition, the computation of the LOO test statistic becomes efficient. This ensures the computational efficiency of the model construction procedure.
- The LOO test score is at least locally convex. Thus the criterion used in selecting regressors also provides the information regarding when to stop the selection. The model construction procedure becomes fully automatic, and the user does not require to specify any algorithmic parameter for this construction procedure.

In summary, combining the LOO test score with the locally regularised OLS learning yields a truly automatic and efficient algorithm capable of producing very sparse models with excellent generalisation performance, using a single training data set only. The comparative study given in **[7]** clearly demonstrates that this method compares favourably with the other existing state-of-the-art sparse regression modelling techniques.

2.4 Orthogonal Least Squares Combined with Basis Pursuit and D-optimality

The regularisation adopted in Sections 2.1 to 2.3 is based on the l_2-norm of the model weights. This penalty function can be "merged" naturally into the main quadratic cost function, and the learning problem remains to be a quadratic optimisation problem, which admits efficient solutions. Moreover, efficient methods exist to optimise the regularisation parameters. An alternative regularisation technique is known as the basis pursuit or least-angle regression, which is based on the l_1-norm of the model weights. A Bayesian interpretation for the basis pursuit method is that it adopts an exponential prior for the model parameters (compare this with a Gaussian prior for the l_2-norm regularisation). The basis pursuit technique is not particularly computationally efficient, because it effectively changes a quadratic optimisation problem having a simple solution into a more sophisticated problem for which generally complex non-quadratic optimisation is required. Furthermore, optimising the regularisation parameters becomes a highly complex problem for which no simple solution exists. An important advantage of the basis pursuit is that it may result in a much sparser model by forcing more parameters to zero than the model derived from the l_2-norm regularisation. The l_2-norm method is good at producing small but non-zero parameter values. Normally, the basis pursuit method is applied to train the large-size full model by forcing many weights to be zero and hence pruning out the corresponding model terms to derive a sparse solution. This process is generally associated with a high computational complexity.

The paper **[9]** developed an efficient model construction method that simultaneously optimises the model's approximation capability, sparsity and robustness by incorporating the basis pursuit regularisation into the combined OLS algorithm and D-optimality design. The cost function for the subset model selection includes a D-optimality criterion that maximises the determinant of the design matrix of the subset model to ensure the model robustness and to allow the model selection procedure to be automatically terminated at a sparse solution. The model weights derived in each forward regression step are initially estimated via the OLS algorithm, and they are then tuned with the aid of a gradient-descent learning algorithm based on the basis pursuit that minimises the l_1-norm of the model's weight vector. Since both the parameter tuning procedure based on the basis pursuit and the regressor selection based on the D-optimality design are integrated with the orthogonal forward regression (OFR), the inherent computational efficiency associated with the original OLS algorithm is maintained. A simple heuristic method of choosing the basis pursuit regularisation parameters is also proposed. This OFR algorithm using both the basis pursuit and D-optimality design provides a powerful additional tool in the toolkit of intelligent nonlinear modelling.

2.5 Efficient Backward Elimination Methods

Two basic approaches of constructing parsimonious regression models are the forward selection and backward elimination. Forward selection starts with a subset model containing no regressor and selects signif-

icant regressors from a large set of candidates one by one to add to the subset model. This approach is computationally more efficient and in particular the forward regression based on the OLS algorithm has proved to be very successful in parsimonious modelling. In contrast, a completely backward elimination procedure eliminates regressors from a full model that contains many regressors, which is a more difficult and computationally complex task, particularly when the full model is very large. Integrating forward selection and backward elimination offers a better model construction method than a purely forward selection one, at the cost of an increased complexity. Alternatively, backward elimination can be employed as a postprocessing procedure, and this can be used to form the hybrid approaches to prune the models that are identified via other approaches. A model constructed via forward regression can gain extra sparsity using backward elimination. This is because forward selection is basically a greedy algorithm, and a regressor selected in a forward manner may become insignificant at a later stage and can then be removed. Adopting backward elimination as a postprocessing tool is computationally affordable.

Conventional backward elimination removes model regressors one at a time based on the least deterioration in the model fit, i.e. based on the training performance. It is highly desired to incorporate some measure of model generalisation capability into this backward elimination procedure. The journal paper **[10]** derived three efficient backward elimination methods: **1)** backward elimination based on the A-optimality design, **2)** backward elimination based on the D-optimality design, and **3)** backward elimination based on the basis pursuit. The A-optimality experimental design aims to minimise the covariance of the parameter estimate, which is proportional to the trace of the inverse of the design matrix. The similarity of the three methods is that each is based on the balance between the model fit and one of the three alternative generalisation measures. Such a balance also ensures that the backward elimination processes can be automatically terminated. The A-optimality and D-optimality based pruning processes require some orthogonalisation operation between the pruned model and the deleted regressor. The pruning formula of the l_1-norm basis pursuit is shown to be extremely simple, without the need for an orthogonalisation process.

2.6 Robust and Efficient Neurofuzzy Rule Base Knowledge Extraction and Estimation

Discovering knowledge from data is an ultimate objective of data modelling. A neurofuzzy network has an inherent model transparency that helps users to understand the system's behaviours, to oversee the critical system operating regions, and/or to extract the physical laws or relationships that underpin the system's operation. Based on a Takagi-Sugeno fuzzy rule inference and model representation, a neurofuzzy model can be functionally expressed as an operating point dependent fuzzy model with a local linear description that directly lends itself to conventional estimation and control synthesis. The model output is decomposed into a convex combination of the outputs of the individual rules, and the basis function can be interpreted as a fuzzy membership function of the individual rules. This property is critically desirable in the context of problems requiring insight into the underlying phenomena, i.e. into the interpretation of the internal system behaviour and/or knowledge (rule) representation of the underlying process. The curse of dimensionality however is a main obstacle in the transparent neurofuzzy rule base knowledge extraction and estimation. For a high-dimensional system, a massive number of the potential rules can be generated, making it impossible for a human to comprehend the resultant rule set. Consequently, the major purpose of neurofuzzy model construction algorithms is to select a parsimonious model structure having a good generalisation capability with a transparent interpretable representation.

The paper **[11]** developed a novel robust neurofuzzy construction algorithm for modelling observed data sets in the form of a set of fuzzy rules. Based on the Takagi-Sugeno inference mechanism, a one to one mapping between a fuzzy rule base and a model matrix feature subspace is established. This link enables the rule based knowledge to be extracted from the matrix subspace to enhance the model transparency. In order to achieve the best possible model robustness and sparsity, an extended Gram-Schmidt orthogonal subspace decomposition has been introduced and combined with the two effective and complementary robust modelling approaches of the regularisation and the D-optimality experimental design. The model rule bases are decomposed into the orthogonal subspaces (compare this with the orthogonal bases in the conventional orthogonal decomposition), so as to enhance the model transparency with the capability of interpreting the derived rule base energy level and hence enabling the selection of the significant and relevant subspaces or rule bases. With an appropriately chosen weighting for the D-optimality cost function, the entire neurofuzzy model construction procedure becomes automatic. Thus, with the important extension from an orthogonal decomposition to an orthogonal subspace decomposition, the major contribution of **[11]** is to extend the locally regularised OLS algorithm with the D-optimality design for the conventional data modelling to the fuzzy rule regularisation and subspace based information extraction. This neurofuzzy model construction algorithm provides a powerful tool for extracting the relevant and significant rules from a large rule set and offers a state-of-the-art technique for discovering knowledge from data.

2.7 Nonlinear System Modelling and Prediction

This section collects some of my recent publications related to three different applications of nonlinear data regression. Specifically, Subsection 2.7.1 deals with nonlinear dynamic system identification and prediction **[12]-[14]**, a novel postprocessing technique designed for image compression **[15],[16]** is presented in Subsection 2.7.2, and Subsection 2.7.3 summarises an application of neural networks to the terrain prediction for employment by legged robots **[17],[18]**.

2.7.1 Dynamic System Identification and Prediction

I am one of the first researchers to apply the radial basis function (RBF) network to nonlinear dynamic system identification and to demonstrate its advantages over other types of artificial neural networks, such as the multilayer perceptron, as summarised in my early publications **[25],[26],[28],[29]**. By placing the RBF centres on the training input data points, the RBF network becomes a linear-in-the-weights model and the OLS algorithm provides an effective means of identifying a sparse RBF network model. Alternatively, the RBF centres can be obtained using a clustering algorithm and the linear weights of the RBF network can then be estimated using the usual least squares solution. Traditionally, the κ-means clustering algorithm is used to obtain the RBF centres. This clustering algorithm can only achieve a local optimal solution, which depends on the initial locations of the cluster centres. A consequence of this local optimality is that some initial centres may become stuck in regions of the input domain with few or no input patterns, and never move to where they are needed. This wastes resources and may result in an unnecessarily large RBF network. The work **[12]** overcomes this drawback by adopting an enhanced clustering algorithm. By using a cluster variation-weighted measure, this enhanced clustering algorithm guarantees to converge to an optimal cluster configuration, regardless of the initial centre locations. Coupled with a cross validation, the method

of **[12]** offers a simple but effective means of identifying nonlinear dynamic systems using a RBF model.

The RBF network has enjoyed considerable success in applications to nonlinear system identification and time-series prediction. Most of the successful results, however, were obtained when the RBF network is applied to stationary systems or signals. The performance of the RBF model applied to non-stationary systems and signals is less satisfactory. This is because the standard RBF network, like many feedforward neural network models, does not characterise the associated temporal variability sufficiently well. Since the real-world systems and signals are often not only highly nonlinear but also highly non-stationary, it is desirable to develop the models which can handle the signals that exhibit both of these characteristics. This is what motivated the work in **[13]** to develop a novel gradient RBF (GRBF) network for nonlinear and non-stationary time series prediction. For the non-stationary time series involving variations of local mean and trend, the series can be made stationary by applying a suitable differential operation on the signal. This is the principle behind the linear auto-regressive integrated moving average model used to predict non-stationary signals. By incorporating a similar mechanism into the RBF network, the resultant model structure has a better predictive performance for non-stationary homogeneous time series. In addition to using the first-order difference of the signal as the input vector of the network, the response of each RBF node is also modified with a first-order local model term which can be interpreted as a local one-step predictor. The OLS algorithm is modified accordingly to provide an efficient construction algorithm for identifying the parsimonious GRBF network models having an excellent predictive performance for nonlinear and non-stationary time series.

The book chapter contribution **[14]** provides a comprehensive treatment of methods and techniques designed for nonlinear dynamic system identification. A general principle of system identification is that the model should not be more complex than required for capturing the underlying system dynamics. This concept, known as the parsimonious principle, is particularly relevant in nonlinear model building, because the size of a nonlinear model can become, to a great extent, excessively large. An overly complicated model may simply conform to the noise in the training data, resulting in over-fitting. An over-fitted model does not capture the underlying system structure sufficiently well and can, in fact, perform poorly on new data. This consideration is the main motivation and central theme of my research reported in **[14]**. Conventional nonlinear models as well as various neural network models are reviewed. A substantive array of system identification techniques is discussed in the context of both the nonlinear-in-the-parameters and linear-in-the-parameters models, with emphasis on determining the model structure. Model validation methods are also provided, and the concept of (nonlinear) identifiability as well as the novel idea of local model fitting are introduced. The work **[14]** offers a comprehensive reference for researchers and practitioners engaged in nonlinear dynamic system identification.

2.7.2 Postprocessing Designed for Image Compression

Image coding is a technique involved for reducing the bandwidth and storage capacity requirements. However, the family of lossy image-coding methods causes quantisation distortions in the reconstructed images. In practice, there is always a tradeoff between the affordable coding bit rate and the coded image quality. Generally, increasing the coding bit rate improves the quality of the reconstructed image, but the coding bit rate is limited by the channel bandwidth or by the storage capacity. As an attractive alternative, image

postenhancement may be used for improving the quality of the reconstructed image, without increasing the coding bit rate. Existing postprocessing methods can be divided into two categories: those employing filtering to smooth the blocking artifacts in reconstructed images and those formulating postprocessing as an image-recovery problem. Since filtering also causes unwanted over-smoothing on image edges, the class of filtering methods is not appropriate for applications, which require genuinely good image quality with minimum distortions. The second class of methods rely heavily on the accuracy of *a priori* image models used and on the optimisation algorithms adopted. These methods often involve adaptive filtering to avoid excessive smoothing. In addition, existing postprocessing methods are specifically designed for block-based coding methods using fixed coding block sizes, such as those in transform coding (TC) and vector quantisation (VC), where blocking artifacts constitute serious sources of distortions. These adaptive filtering techniques are unable to improve the quality of predictive coding, where blocking artifacts do not exist and blurred edges are the main coding distortions. These methods are also impractical to use for quadtee (QT) coding, which has variable block sizes.

The motivation of my related research **[15],[16]** was to create a postprocessing technique, which is able to correct the coding distortions of all major coding schemes, and the journal papers **[15],[16]** developed a generic postprocessing technique. The key of this postprocessing method is the ability to recover the quantisation distortion image, defined as the difference between the original and decoded images. It can be shown that the main coding impairments are due to edge distortions, including blurred edges and spurious edges. Spurious edges are caused by blocking artifacts. This suggests that the basic task of postprocessing is to correct these edge distortions. The novel distortion-recovery model developed in **[15],[16]** consists of a visual feature extractor to extract edge information from the decoded image, and a mapping to map the visual features of the decoded image onto the quantisation distortion image. Specifically, visually important edge features are computed as multi-scale first-order derivatives of the decoded image. Interestingly, this gradient extractor imitates certain characteristics of visual cortex. As the exact relationship between the gradient features of the decoded image and the quantisation distortion image is unknown, an artificial neural network, referred to as the neural network visual model (NNVM), is trained to learn this relationship. The advantages of the proposed postprocessing technique have been demonstrated in a range of experiments involving four coding systems, namely TC, VC, QT coding and predictive coding schemes. The experimental results obtained confirm that the NNVM achieves significant improvements in the quality of the reconstructed images, in terms of both the objective distortion measure used and subjective viewing assessment. More specifically, the experiments were conducted on eight test images in conjunction with the aforementioned four coding schemes, each having two different coding bit rates. The average coding peak signal to noise ratio (PSNR) over these experimental conditions was approximately 30 dB without postprocessing. It is well-known that the higher the PSNR is the more difficult for a postenhancement technique to achieve postprocessing gain. The experimental conditions were designed specifically to create high PSNR conditions, so as to test the proposed postprocessing method under "unfavourable" situations. Even under such a high average PSNR condition, the proposed postprocessing technique achieved an average PSNR gain of over 1 dB[1], which is much better than other existing postprocessing methods can achieve.

[1]Due to a limited length, not all the experimental results could be presented in the journal papers **[15],[16]**. More detailed experimental results are given in the PhD thesis "A Generic Postprocessing Technique for Image Coding Applications," by my PhD student, Zhongmin He, in 1999.

2.7.3 Terrain Prediction for Legged Robots

Roughly half of the land surface of the earth is inaccessible to conventional wheeled or tracked vehicles. Legged robots present significant advantages over wheeled or tracked mechanisms because of their ability to move in very rough and unstructured terrains and to step over obstacles. However, without efficient walking strategies these advantages cannot be realized. The walking gait algorithm is a decision based process regarding which leg should be lifted or placed to provide time-space coordination of the motion of the various legs of the robot. It is the most crucial process in the control of legged robot motion. A terrain mapping system, which can accurately predict unknown terrain, is capable of aiding the walking gait algorithm to provide smooth and efficient walking motion of legged robots in unstructured environments. The work reported in **[17],[18]** applied an artificial neural network as terrain predictor to aid the navigation of legged robots across rough terrains. In the experiments involving an eight-leg robot called Robug IV roaming in a range of terrain environments, this neural network terrain predictor was demonstrated to provide significantly improved prediction accuracy over an existing terrain mapping system previously implemented on Robug IV.

3 Classification

Classification is another basic learning problem in machine learning. It has many similarities with regression, but also has some important differences. Notably, the desired output in a classification problem only takes a finite number of discrete values. For example, in two-class classification problems, the desired output or class label is binary. The construction of a classifier is often linked to probability distribution estimation. This is because, if the conditional probability density functions of the classes are known, the optimal Bayesian classifier can readily be formulated. For regression, the least squares or mean square error is the principal criterion in deriving appropriate learning mechanisms. The ultimate objective of classification is to maintain a high classification accuracy or a low error rate, and this should be the basic criterion for deriving learning rules in classification applications. My research in the field of classification is both research and application oriented, and the four journal papers **[40]-[43]**, which provide a summary of my recent contributions to this field, were selected for inclusion in the submission. My novel contributions in the research field of classification can be summarised as follows.

1. *Deriving a robust maximum likelihood learning method for heteroscedastic probabilistic neural networks by incorporating a statistical bootstrap technique known as the Jack-knife into the expectation-maximisation (EM) algorithm.*

2. *Developing an efficient sparse kernel classifier construction technique based on an orthogonal forward selection (OFS) procedure that chooses significant and relevant kernel terms by incrementally maximising the Fisher ratio of class separability measure.*

3. *Developing a stochastic gradient based sample-by-sample learning technique, called the least error rate algorithm, for real-time adaptation of a classifier, that directly minimises the (approximate) error rate of the classifier.*

These novel contributions in classification are now described below in more detail with reference to the corresponding journal papers.

3.1 Robust Training of Heteroscedastic Probabilistic Neural Networks

The key to applying the optimal Bayes strategy for pattern classification is the estimation of the conditional probability density functions (PDFs) of the classes, and a classical PDF estimate is the Parzen window estimator. The probabilistic neural networks (PNNs) implement the Parzen window estimator using mixtures of Gaussian kernels. Traditionally, a common variance is used for every Gaussian basis function, and this is termed as the homoscedastic PNNs. The expectation-maximisation (EM) algorithm is typically applied to provide an optimal maximum likelihood (ML) training algorithm for PNNs. If each Gaussian basis function is allowed to have a different variance, a much more parsimonious PNN can be used to adequately approximate the underlying conditional PDFs of the training data but the associated learning problem is much more difficult. Indeed, when extending the EM training algorithm to this class of heteroscedastic PNNs, often numerical difficulties occur. The journal paper **[40]** analysed the root of this numerical problem, and incorporated a robust statistical technique, referred to as the Jack-knife, into the EM algorithm to derive a robust and effective ML training method for heteroscedastic PNNs.

The Jack-knife constitutes a simple yet effective statistical bootstrap method that is widely used in robust statistical analysis for bias reduction and interval estimation. Basically, the Jack-knife procedure partitions a sample space into several subsets and observes the influence of each subset on the estimation process. The Jack-knife estimator is very effective in reducing the bias of the estimate. Another desired property of the Jack-knife technique is that it can remove the effect of a few outliers within a data space, giving rise to a robust estimate. Note that the effect of a few outliers is the main source that may cause the EM algorithm to collapse when training heteroscedastic PNNs, as analysed mathematically as well as demonstrated numerically in **[40]**. Thus, this numerical difficulty can be overcome by incorporating the Jack-knife technique in the EM algorithm, which represents the main contribution of **[40]**. The resultant robust ML training algorithm for heteroscedastic PNNs is then applied in **[40]** to the real data involving success or failure prediction of UK private construction companies, and the results obtained clearly demonstrate the effectiveness of this robust ML learning method.

3.2 Kernel Classifier Construction with Fisher Ratio of Class Separability Measure

Just as in regression application, having a good generalisation capability in terms of test error rate and a parsimonious representation are the two key requirements of kernel classifier construction. Support vector machine (SVM) based methods are often regarded as the state-of-the-art techniques for kernel classifier design. The SVM method can be viewed to start with a large classifier model (each training input data corresponding to a candidate kernel term), and learning involves the removal of the insignificant kernel terms or equivalently to retain the significant support vectors. The SVM method is known to have excellent generalisation performance. However, the ability of the SVM method to produce sparse models is overstated, and many practical applications have demonstrated that kernel classifiers constructed by the SVM methods may not be sufficiently sparse. As a design alternative, the relevance vector machine (RVM) method has

the ability to produce much sparser classifiers with similarly good generalisation performance as that of the SVM method. The RVM method however is computationally much more demanding than the SVM method and the resultant learning problem is often inherently ill-posed. Forward selection, which selects significant kernel terms one by one to form a sparse classifier, has clear computational advantages. In my early work **[26]**, by formulating the classifier construction as regression modelling, the orthogonal least squares (OLS) algorithm was applied to select a parsimonious radial basis function (RBF) classifier in an efficient orthogonal forward selection (OFS) procedure. The least squares cost used by the OLS algorithm for selecting kernel terms is, however, not the most relevant one in the context of the classification objective. What is really needed here is to find a suitable criterion for selecting significant kernel terms, which is directly linked to the classification objective.

The degree of class separation is directly linked to the test error rate, and the Fisher ratio class separability measure thus offers a suitable criterion for selecting significant kernel terms. This was first suggested in:

K.Z. Mao, "RBF neural network center selection based on Fisher ratio class separability measure," *IEEE Trans. Neural Networks*, Vol.13, pp.1211-1217, Sept. 2002.

My contribution **[41],[42]** involves refining the OFS procedure based on the Fisher ratio class separability measure in a generic classifier design algorithm, applying it to some typical problems and demonstrating its computational advantages over the other state-of-the-art kernel-based classifier design methods. By formulating the kernel classifier design as regression modelling, orthogonal decomposition can be applied to orthogonalise the "regression matrix", and the Fisher ratio class separability measure can be efficiently computed for each orthogonal base candidate. At each stage of the OFS procedure, a candidate base, namely a candidate kernel term, is selected if it produces the maximum value of the Fisher ratio among the candidates that have not yet been selected. The orthogonalisation process and the computation of the kernel weights remain identical to those employed by the original OLS algorithm used in regression. This OFS procedure is inherently well-conditioned and efficient. The applications reported in **[41],[42]** show that it has the same excellent generalisation performance and a similar level of sparsity as the RVM method, but it is computationally much simpler than the RVM method. Thus, this OFS algorithm based on the Fisher ratio class separability measure provides a viable alternative for constructing truly sparse kernel classifiers that generalise well.

3.3 Stochastic Least Error Rate Training for Classifiers

As is mentioned previously, the mean square error (MSE) criterion is not a generally suitable cost function for classifier design. For off-line or block-data based learning, such as those reported in Sections 3.1 and 3.2, the need for adopting a more relevant criterion to train classifiers has been widely recognised. Given the underlying pattern space, that is, the information available to a classifier, the maximum a *posteriori* probability or Bayes classifier provides the true optimal performance. We should refer to the definition of the minimum error rate (MER) as the achievable error rate for a classifier with an additional constraint of a given structure (e.g. a RBF classifier with a given number of hidden nodes). The basic question is then whether it is possible to achieve this MER and how close it is to the true optimal performance of the Bayes classifier.

Kernel-based classifier construction algorithms, such as the support vector machine (SVM) method and the orthogonal forward selection (OFS) algorithm with the Fisher ratio class separability measure, have been shown to be capable of approximating the optimal Bayesian performance. All these training algorithms are however block-data based and cannot be implemented in a truly sample-by-sample adaptive manner. In many applications involving nonlinear (e.g. neural network) classifiers, adaptive training or sample-by-sample adaptation is required to meet real-time computational constraints. Application examples include neural network equalizers and multiuser detectors in communication systems. Typically, the adaptive training of neural network classifiers is usually carried out using some stochastic gradient algorithm, such as the well-known least mean square (LMS) algorithm, based on the minimum MSE (MMSE) criterion. A somewhat undesirable situation exists that, on one hand, the performance of a classifier is evaluated using the probability of error while, on the other hand, a different MSE criterion is used during the learning stage.

The journal paper **[43]** developed a truly sample-by-sample adaptive training method based on directly minimising the classifier's approximate error rate for two-class classification problems, and applied the resultant stochastic least error rate (LER) algorithm to nonlinear equalisers for binary signalling. The derivation of this LER algorithm is based on a reasonable assumption that the underlying pattern space is a finite-state machine contaminated by additive Gaussian white noise. Thus, the distribution of the pattern space is a multi-variate Gaussian mixture. By virtually linearising the output of a nonlinear classifier in the vicinity of the states of the pattern space, the probability density function (PDF) of the classifier's output can be accurately approximated as a mixture of Gaussians. Given the PDF, the probability of error of a classifier can be explicitly formulated, and the MER classifier is defined as the one that minimises this probability of error, which can be obtained for example using a gradient based optimisation method initialising the search from a suitably chosen initial solution. Since the states of the pattern space are generally unknown, the PDF of the classifier's output is also unknown. However, this PDF can be estimated using the Parzen window estimate based on a block of training data. The Parzen window based method estimates an unknown PDF by placing a symmetric unimodal kernel function on each data sample and combining these kernels to form a PDF estimate. This method is known to produce a remarkably accurate PDF estimate. For the case of sample-by-sample adaptation, a single sample based PDF "estimate" is used and, conceptually, from this one-sample PDF estimate, a one-sample or instantaneous "error probability" is derived. Using the instantaneous gradient of this error probability naturally gives rise to a stochastic gradient algorithm referred to as the LER algorithm. The results involving nonlinear equalisation reported in **[43]** have demonstrated that a small RBF network trained by the LER algorithm can closely approximate the optimal Bayesian performance, and the work in **[43]** also confirms that the same RBF network trained by the LMS algorithm converges well in terms of the MSE, but may produce a poor error rate performance.

4 Probability Density Function Estimation

Estimation of the PDF from a representative sample drawn from the underlying density is a problem of fundamental importance in both machine learning and many other fields of engineering. However, PDF estimation is a hard and often ill-posed learning problem. It is also an unsupervised one, as the learning process relies only on observations or samples drawn from the underlying density to be estimated and the learning machine does not have the corresponding "desired outputs". There exist two basic approaches to density

estimation, namely parametric and nonparametric methods. In the parametric approach, the unknown density is assumed to belong to a parametric set satisfying certain conditions and density estimation becomes a problem of typically estimating these parameters using the maximum likelihood principle. The nonparametric approach does not have the limitation of assuming that the unknown density belongs to a particular form from a parametric set. A well-known nonparametric density estimation technique is the classical Parzen window based PDF estimate, which is remarkably simple and accurate. The particular problem associated with the Parzen window method however is the high computational cost required for calculating the PDF value of an observation which has a complexity proportional to the sample size, as the Parzen window based PDF estimate employs the full data sample set in defining a density estimate for subsequent observations. In today's data rich environment, this can be a serious problem in practical applications.

Recently, the SVM method has been proposed as a promising tool for sparse kernel density estimation. The motivation of the SVM density estimation technique comes from the claim that the SVM method can effectively perform function approximations in high-dimensional spaces from finite data sets with sparse representations. Currently, the machine learning community is actively engaged in the investigation of the SVM density estimation method and a range of other related techniques, such as the data reduction method reported in:

M. Girolami and C. He, "Probability density estimation from optimally condensed data samples," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.25, No.10, pp.1253–1264, 2003.

These SVM-type methods can be interpreted as search techniques starting from the full data sample set with each data sample used as a candidate kernel centre, and learning is used to gradually discard (zero out) many of the kernel weights, leaving only the most significant non-zero weights for the support vectors. This leads to a sparse representation of the kernel density estimate. My novel research in sparse density estimation **[44]** adopts an alternative forward selection approach. The motivation of adopting this alternative approach arises from my research in regression and classification. In the context of regression and classification, it is known that a properly designed forward selection algorithm can provide sparser representations than the SVM method with equally good generalisation performance, as it was shown in Chapters 2 and 3 of this submission.

4.1 A Novel Sparse Density Construction Method

My recent journal paper **[44]** presents an efficient construction algorithm for obtaining sparse kernel density estimates based on a regression approach that directly optimises the model's generalisation capability. To demonstrate how the kernel density estimation, which is an unsupervised learning problem, can be reformulated as "supervised" regression modelling, it is worth emphasising that the main objective of density estimation is that the cumulative distribution function (CDF) corresponding to the density estimate should realise or approximate the true CDF of the underlying density. The true CDF of the unknown density is unknown. However, this CDF can be accurately approximated from an empirical CDF computed solely from the observation samples. Hence, by substituting the empirical CDF in the place of the true CDF as the "desired response" for the density estimation, the problem becomes a supervised regression modelling one. All the sparse regression techniques presented in Chapter 2 may then readily be applied to construct parsi-

monious representations of kernel density estimates achieving an excellent test performance. As discussed in Section 2.3, the orthogonal least squares (OLS) with leave-one-out (LOO) test statistic and local regularisation is particularly effective in obtaining very sparse models with excellent generalisation capability. This construction algorithm was therefore chosen in **[44]** for sparse kernel density construction.

The work in **[44]** represents a significant contribution to the important topic of PDF estimation. The advantages of this novel density construction method can be summarised as follows:

- The proposed algorithm incrementally minimises the leave-one-out test score, which is a measure of generalisation capability. Thus, the resultant density construction technique is based on directly optimising the model's generalisation performance. A local regularisation method is naturally incorporated into the density construction process to enhance the achievable sparsity.
- This method is simple to implement and computationally efficient, owning to the application of an orthogonal forward regression. Moreover, the density construction process is entirely automatic, and except for the kernel width the algorithm contains no parameters that require tuning. This is in contrast to the SVM based methods, which have to carefully choose some of the algorithmic parameters to ensure a successful density construction. The method formulates density estimation as a regression, where many data modelling practitioners feel more at home.

The experimental results reported in **[44]** have demonstrated the ability of this novel algorithm to effectively construct a very sparse kernel density estimate having a comparable accuracy to that of the full sample set based optimised Parzen window estimate. The results also show that the proposed algorithm compares favourably with the SVM based methods in terms of both the achievable test accuracy and sparsity, when constructing kernel density estimates.

5 Minimum Bit Error Rate Linear Receivers for Communication Systems

One of the prime tasks of receivers in wireless communication systems is to overcome or alleviate the distortion effects of bandlimited channels and/or multiple access interference, so as to recover or detect the transmitted data symbols as reliably as possible. Linear receivers are the most widely used receiver structures, due to their advantages of low complexity, well-established adaptive procedures and convergence properties. A linear receiver can loosely be defined as the one that linearly combines the received signals, and we will allow the definition to include the class of linear-combiner aided decision feedback equalisers (DFEs). A DFE is inherently a nonlinear structure, as it feeds back the detected symbols. However, we can view the decision feedback as a space translation, which translates the original observation space into a new reduced-dimensional space. In this translated observation space, the DFE becomes a linear equaliser. My journal papers **[45],[46]** gave a detailed analysis of this space-translation property. In this chapter and in Chapter 6, we will assume that training is available for the adaptive receiver structures. That is, we consider supervised learning. In communication systems, this is achieved by organising the transmissions into a frame structure, and each frame includes a short pre-determined training symbol sequence that is not part of the transmitted data and it is known to both the transmitter and receiver. A synchronised receiver can generate this training sequence locally and uses it as the desired response to adapt the receiver parameters.

This *trained* receiver strategy is in contrast to the *blind* receiver strategy covered in Chapter 7, where the receiver does not have access to a training sequence and must adapt the receiver parameters based on the received signal only, using unsupervised learning.

Traditionally, the Wiener or minimum mean square error (MMSE) solution is regarded as the optimal solution for the generic linear receiver. This thinking has its roots in adaptive filtering, and a range of well-documented adaptive filtering techniques have found applications in adaptive communication receiver design. In particular, a stochastic gradient based learning algorithm, the least mean square (LMS) algorithm, offers a truly adaptive low-complexity means of realising the MMSE linear receiver solution. The ultimate objective of a communication receiver however is to minimise the error probability in detecting the transmitted data symbols or the bit error rate (BER). Thus, a linear receiver can be considered as a linear classifier with certain real-time computational constraints. My publications **[45]-[56]** are instrumental in rediscovering the minimum BER (MBER) principle and revitalising active research in developing a rich set of adaptive MBER techniques for a variety of communication systems. My novel contributions are summarised as follows:

- Derived the generic linear MBER design that is applicable for all the major state-of-the-art communication systems, and demonstrate its superior performance in comparison to the MMSE solution.
- Developed a stochastic adaptive algorithm for approximating the exact linear MBER solution, which has a similarly low complexity as the LMS algorithm, and therefore makes the MBER design practically applicable.

My detailed work in this novel research area is outlined below with reference to the relevant journal papers.

5.1 Minimum Bit Error Rate Linear Equalisation

Equalisation plays a vital role in combating distortion and interference in communication links and high-density data storage systems. A linear equaliser may be viewed as the inverse of the channel having a linear distortion. This leads to the zero-forcing solution that completely eliminates the intersymbol interference (ISI). The zero-forcing equaliser however can cause serious noise enhancement, resulting in severe degradation in detection accuracy. The MMSE solution is traditionally viewed as offering a trade-off between the two conflicting requirements of eliminating ISI and not enhancing noise, and the linear MMSE design is the most widely used receiver design in practice. My journal papers **[45],[46]** showed that the MMSE equalisation solution does not necessarily achieve the MBER performance and is therefore sub-optimal. The MBER equalisation solution designed for binary signalling, namely for binary phase shift keying (BPSK), is derived in **[45]**, and the extension to minimum symbol error rate (MSER) equalisation is given in **[46]** for the general multilevel modulation scheme.

5.1.1 Asymptotic Minimum Bit Error Rate Equaliser

Unlike for the MMSE design, there exists no closed-form solution for the MBER equalisation and hence nonlinear iterative optimisation is required to obtain a numerical MBER solution in general. This is par-

ticularly inconvenient in theoretical analysis of the MBER design and in performance comparison of the MMSE and MBER equalisers. This lack of theoretical results is also partly to blame for the slow adoption of the MBER approach within the communications research community. One way of overcoming this difficulty is to study the properties of the MBER solution under the asymptotic condition of large signal to noise ratio (SNR), which may provide some theoretical insights into the MBER solution. The work reported in **[47]** analysed the asymptotic MBER solution for large SNRs and discovered that asymptotically the linear support vector machine (SVM) classification solution is the MBER equalisation solution for the BPSK modulation scheme. The significance of this result is that the linear SVM equaliser solution is much easier to obtain, involving only simple quadratic optimisation, and this provides an efficient means of obtaining an approximate MBER equalisation design without involving costly nonlinear optimisation. Moreover, geometric interpretation of the maximum margin property of the SVM solution enabled the work in **[47]** to provide insights into why the MMSE and MBER equalisers behave very differently in the asymptotic case. Thus, for the first time, it shows in a simple and easy-to-understand geometric interpretation, why the MMSE solution is in general sub-optimal. The publication **[48]** further generalises the results to the case of multilevel modulation schemes and provides a simple and efficient means of realising the generic linear MSER equaliser.

5.1.2 Stochastic Least Bit Error Rate Equaliser

In practice, equalisers must operate adaptively under the most harsh real-time computational constraints. An important reason for the popularity of the MMSE design is that it facilitates the use of standard adaptive filtering techniques, such as the least mean square (LMS) algorithm. At each sampling instant, the LMS algorithm adapts the filter parameters along the direction of the negative stochastic gradient of the squared instantaneous error between the filter desired output and actual output. Thus, the LMS algorithm has a very low computational complexity that meets real-time operational constraints of communication systems. It is well-known from adaptive filtering theory that the simple stochastic learning strategy of the LMS algorithm is guaranteed to converge to the MMSE solution with an appropriate choice of the learning step size. Without developing similar low-complexity adaptive algorithms for realising or implementing the MBER solution, it is unlikely that the MBER equaliser design will find its way into practical communication systems. The journal paper **[49]** derived a novel adaptive MBER equaliser, which has a similarly low complexity as the adaptive LMS equaliser. The novelty of the work in **[49]** is the recognition that the key to realising an adaptive MBER equaliser is the estimation of the probability density function (PDF) of the equaliser output. Given this PDF, the BER of the equaliser becomes known. By definition, the MBER solution is the equaliser tap vector that minimises this BER. The PDF of the equaliser output is generally unknown. However, it can be approximated very accurately using a block of training data based on Parzen window or kernel density estimation. This leads to a block-data adaptive MBER solution. By considering a single-sample PDF "estimate", conceptually, an instantaneous "BER" is obtained. Minimising this instantaneous error probability naturally leads to the stochastic gradient adaptive MBER equaliser. This new stochastic learning strategy based on the MBER principle was first developed in **[49]** and later has been referred to as the least bit error rate (LBER) algorithm.

My journal paper **[50]** developed this concept of adaptive MBER equalisation further into the general theory of adaptive MBER filtering for communication systems. It points out that the MMSE approach is

optimal only if the filter output is Gaussian distributed. Since the filter output of a communication receiver is generally a mixed sum of Gaussian distributions and therefore non-Gaussian, the MMSE solution is inherently non-optimal. By contrast, the adaptive MBER filtering effectively exploits the non-Gaussian distribution of the filter output and, consequently, it can provide significant performance gains in terms of a smaller BER in comparison to the MMSE approach. The work **[50]** also highlights an interesting analogy between the traditional adaptive filtering approach based on the MMSE criterion and the proposed novel adaptive MBER filtering approach. The second-order statistics required to find the MMSE solution can be estimated using a block of samples and, by considering a single-sample estimate, a stochastic gradient MMSE algorithm, namely the LMS algorithm, is derived. The PDF required to determine the MBER solution can be approximated with a kernel density estimate based on a block of samples and, by considering a single-sample density estimate, a stochastic gradient adaptive MBER algorithm, namely the LBER algorithm, is formulated.

The journal paper **[51]** extended the result of adaptive MBER equalisation designed for BPSK modulation and developed an approach based on direct minimisation of the symbol error rate (SER) for communication systems employing a multilevel pulse-amplitude modulation scheme. The Parzen window technique is again adopted for approximating the PDF of the equaliser's output, and this naturally leads to a block-data based adaptive minimum SER (MSER) algorithm, which iteratively minimises the estimated SER of the equaliser by adjusting the equaliser's weights using the conjugate gradient optimisation method. It is shown that this block-data based adaptive MSER algorithm converges rapidly and the length of the data block required for achieving an accurate approximation of the MSER solution is reasonably small. Sample-by-sample adaptive implementation of the MSER equalisation solution is then considered and a stochastic gradient adaptive MSER algorithm, referred to as the least symbol error rate (LSER), is derived. This LSER algorithm has a low computational complexity, comparable to that of the simple LMS algorithm. Simulation results suggest that the LSER algorithm converges reasonably fast with appropriate initialisations.

5.2 Minimum Bit Error Rate CDMA Multiuser Detection

Code-division multiple-access (CDMA) constitutes an attractive multiuser scheme that allows users to transmit at the same carrier frequency in an uncoordinated manner. However, this creates multiuser interference (MUI) which, if not controlled, can seriously degrade the quality of reception. Mutually orthogonal spreading codes for different users can provide an inherent immunity to MUI in the case of synchronous systems. Unfortunately, multipath interference is often encountered in CDMA systems and will reduce this inherent immunity to MUI by destroying the orthogonality of the spreading codes. A multiuser detector (MUD) is required at the receiver to combat the MUI in order to achieve an adequate performance. In a CDMA system, the objective of the receiver is to detect the transmitted information bits of one (at mobile-end) or many (at base station) users. The first case, usually referred to as the downlink, is particular challenging, as a mobile handset has limited computational power. For such applications, a linear detector structure is most widely used, and the MMSE design is often regarded as the state-of-the-art technique for CDMA multiuser detection. This popularity is again due to the fact that the MMSE multiuser detector can readily be implemented efficiently using standard adaptive filtering techniques. The ultimate performance criterion of a MUD is its BER, just as in the case of single-user channel equalisation, and the MMSE MUD is generally sub-optimal in this context.

The publications **[52]-[54]** developed a novel adaptive MBER multiuser detector for downlink applications. The proposed MBER approach can effectively exploit the non-Gaussian nature of the MUI. The adaptive MBER MUD derived in **[52]-[54]** has been shown to be much more robust to the near-far effect than the MMSE based MUD, and is particularly effective in combating the adverse communication conditions, where interfering users have a high power. This is because, in the situation where a few strong interfering users exist, the non-Gaussian distribution is evident, and the MMSE solution can become considerably inferior in comparison to the MBER solution. Technically, the key to the development of this MBER MUD is again that of deriving the probability distribution of a linear detector's output. The PDF of the linear detector's output is non-Gaussian, consisting of a mixed sum of Gaussian distributions, although the number of Gaussian components is much larger compared to the single-user channel equalisation, owing to the channel-induced intersymbol interference (ISI) as well as the MUI. By accurately approximating this PDF of the linear detector's output based on Parzen window estimation, a block-data adaptive MBER CDMA detector is readily formulated. This further leads to an adaptive LBER CDMA detector that has a similarly low complexity as the LMS based CDMA detector. The publications **[52],[54]** concentrate on adaptive MBER multiuser detection for CDMA systems employing BPSK signalling, while the work in **[53]** extends the results to CDMA systems using quadrature phase shift keying (QPSK) modulation. These works have clearly demonstrated the efficiency of the MBER filtering technique in interference-limited communication systems.

5.3 Multiantenna Aided Minimum Bit Error Rate Multiuser Detection

In an effort to further increase the achievable system capacity, antenna arrays can be employed for supporting multiple users in a space division multiple access (SDMA) communications scenario. To gain insight into the multiuser-supporting capability of an SDMA system, it is useful to draw some comparisons with CDMA multiuser systems. In a CDMA system, each user is separated by a unique user-specific spreading code. In contrast, an SDMA system differentiates each user by the associated unique user-specific channel impulse response (CIR) encountered at the receiver antennas. In a simplistic but conceptually appealing interpretation, one could argue that the unique user-specific CIR plays the role of a user-specific CDMA signature. In this analogy the CIR-signatures are not orthogonal to each other, but this is not a serious limitation, because even orthogonal spreading codes become non-orthogonal upon convolution by the CIR. However, owing to the non-orthogonal nature of the CIRs, an effective multiuser receiver is required for separating the users in an SDMA system. We consider the multiple receiver-antenna aided receiver structure based on the temporal reference approach, i.e. having training. The most popular design is constituted by the MMSE solution. However, as demonstrated in the previous section for CDMA systems, minimising the mean square error does not guarantee that the BER of the system is also minimised. The publications **[50],[55],[56]** have developed an adaptive multiantenna aided MBER multiuser detection technique.

The journal paper **[50]** considered a multiantenna system where the antennas have close spacing, typically half the wavelength, and hence the CIRs encountered at the receiver antennas are correlated. Adaptive MBER beamforming is developed in **[50]** as an example of the general adaptive MBER filtering technique for interference-limited communication systems. It is demonstrated that MBER beamforming has significant advantages over MMSE beamforming in a so-called overloaded system, where the number of the users supported is more than the number of the receiver antennas. For an overloaded system, the BER performance

of the MMSE design may break down, owing to the lack of sufficient degrees of freedom to "null" all the interfering users. The MBER design in contrast is often capable of achieving the desired linearly separable property and therefore guaranteeing an adequate BER performance. The publication **[55]** derived an MBER multiuser detector designed for SDMA aided orthogonal frequency division multiplexing (OFDM) systems. Coupled with an adaptive estimation of the frequency domain channel transfer function (FDCHTF) matrix, this novel MBER multiuser detector constitutes a state-of-the-art adaptive technique for SDMA aided OFDM communication systems. The work in **[56]** investigated genetic algorithm (GA) assisted error probability optimisation for adaptive beamforming. The motivation of this work is that the BER cost function is generally highly complex, and gradient-based optimisation often requires careful initialisation and tuning of the algorithmic parameters in order to ensure the convergence to a global minimum. To circumvent these implementation challenges, the GA offers an attractive means for the block-data based direct minimisation of the BER of the beamformer. One of my recent major research efforts has been focused on the development of novel adaptive receiver structures for SDMA communication systems, including space-time equalisation techniques for interference-limited systems. At the time of writing this submission, significant advances have been made and the results have been submitted for journal publications[2].

6 Symbol-Decision Based Nonlinear Receivers for Communication Systems

In Chapter 5, linear receiver structures were considered. A linear receiver benefits from a low computational complexity but also exhibits an inherently limited capability. Specifically, a linear receiver employs a linear classifier and hence can only form a linear decision (classification) boundary. Thus, linear receivers only perform adequately, when the different classes of the underlying noise-free signal states can be separated by a linear decision boundary. This is referred to as a linearly separable scenario. The output phasor constellation of a severely dispersive and interference limited communication system may not always be linearly separable. Even for linearly separable phasor constellations, the optimal decision boundary is in general nonlinear. Hence, a nonlinear receiver structure may offer significant performance gains over a linear one, although usually at the cost of an increased complexity. There are two basic strategies for nonlinear receiver structures, namely sequence detection that estimates the entire transmitted symbol sequence as a whole and symbol based detection that makes decisions on a symbol-by-symbol basis. The optimal sequence estimation technique is the maximum *a posteriori* probability sequence estimation method, which provides the truly optimal performance. In practice, maximum likelihood sequence estimation (MLSE) provides a performance almost indistinguishable from that of the optimal maximum *a posteriori* probability sequence estimation. The high implementation complexity and deferring decisions associated with the sequence estimation strategy are however often unacceptable in many practical communication systems. An alternative detection strategy is to make symbol-by-symbol decisions based on the information provided by a finite-memory receiver. We note that the linear receiver discussed in the previous chapter is based on this finite-memory symbol-decision based structure. A symbol-decision based nonlinear receiver can thus be viewed as a nonlinear classifier subject to real-time detection constraints. The optimal symbol-decision

[2] S. Chen, N.N. Ahmad and L. Hanzo, "Adaptive minimum bit error rate beamforming," *IEEE Trans. Wireless Communications*, to appear, 2005.

S. Chen and L. Hanzo, "Space-time decision feedback equalization assisted multiuser detection for multiple antenna aided SDMA systems," submitted to *IEEE Trans. Wireless Communications*, 2005.

based method is known to be the Bayesian symbol-decision solution. My research in the field of nonlinear receivers is mainly concentrated on the symbol-decision based structure for practical purposes.

The publications **[57]-[76]** in the submission cover various issues in nonlinear receiver design, ranging from novel receiver structures, adaptive algorithms and establishing their connection with neural network based classifiers to efficient performance evaluation and implementations. My novel contributions in this important research area can be summarised as follows:

1. *Interpret the optimal Bayesian symbol-decision solution in the context of channel equalisation and derive the optimal Bayesian decision feedback equaliser (DFE). Explain the linkage between the Bayesian equaliser and the radial basis function (RBF) neural network, as well as develop a range of efficient adaptive algorithms for the adaptive implementation of the Bayesian equalisation solution.*

2. *Conduct a comprehensive simulation study involving realistic mobile fading communication environments to evaluate the adaptive Bayesian symbol-decision DFE, and conclusively demonstrate that it has the ability to outperform the adaptive MLSE. This is particularly interesting because the Bayesian DFE is known to be inferior to the optimal MLSE for transmission over time-invariant channels.*

3. *Develop efficient importance sampling simulation techniques for evaluating the symbol error rate (SER) of the optimal Bayesian DFE, and derive a novel method for the reduced-complexity implementation of the Bayesian equalisation solution.*

4. *Extend the concept of Bayesian symbol-decision based nonlinear equalisation into the generic Bayesian symbol-decision multiuser receiver for interference limited communication systems, including CDMA and multiantenna aided SDMA systems.*

My detailed research works in nonlinear receiver design are given below with reference to the publications **[57]-[76]**.

6.1 Single-User Nonlinear Channel Equalisation

By viewing channel equalisation as a nonlinear classification problem, the journal paper **[57]** derived the optimal symbol-decision equalisation solution for binary phase shift keying (BPSK) modulation and defined the terminology *Bayesian equaliser*. My work **[57]** also points out that the Bayesian equaliser and the RBF network have an identical structure and therefore establishes the inherent linkage with the research involving neural network based equalisers. Two strategies were proposed for the adaptive implementation of the Bayesian equaliser. The first technique, known as the indirect method, identifies the channel first and then calculates the equaliser parameters, namely the RBF centres or channel state vectors, based on the estimated channel impulse response (CIR). The second method directly estimates the RBF centres using a supervised clustering algorithm. The journal paper **[58]** extended the work of **[57]** to the general multilevel modulation scheme and proposed a novel adaptive Bayesian DFE. A drawback of the nonlinear Bayesian solution is its high computational complexity due to the large number of channel states associated with the underlying conditional probability distributions, particularly in the context of multilevel signalling. The work **[58]** has shown that the decision feedback can significantly reduce the number of states associated with the Bayesian

solution and therefore substantially reduces the computational complexity. Moreover, it is demonstrated that the decision feedback makes the underlying classification problem much simpler. This clearly explains why the Bayesian DFE outperforms the Bayesian equaliser using no decision feedback.

In general, communication systems are modelled by considering complex-valued channels in conjunction with complex-valued modulation schemes, such as quadrature phase shift keying (QPSK) or quadrature amplitude modulation (QAM). The journal paper **[59]** proposed a novel complex-valued RBF network architecture and developed two learning algorithms for finding the RBF centres. The complex-valued orthogonal least squares (OLS) algorithm, an extension of the conventional real-valued one, is a batch-based learning algorithm capable of constructing an adequate network structure, while a complex-valued version of the hybrid clustering and least squares algorithm offers the benefits of real-time adaptation capability. The proposed complex-valued RBF network has a single-hidden-layer based architecture. The inputs and outputs of the RBF network are both complex-valued. Each node in the hidden layer has a real radially symmetric response spread around the complex-valued node centre vector. The output layer contains a set of complex-valued linear combiners associated with complex-valued connection weights. This RBF network can be viewed as a nonlinear mapping from the complex-valued multi-dimensional input space onto the complex-valued multi-dimensional output space, and it provides a powerful tool for nonlinear signal processing involving complex-valued signals. In many applications, the real-valued response of a hidden node in this complex-valued RBF network has an intrinsic physical interpretation – it actually realises some conditional density function of the underlying data generating mechanism, and this is further exploited in the companion journal paper **[60]**, which developed both the Bayesian equaliser using no decision feedback and the Bayesian DFE for complex-valued channel equalisation involving QPSK modulation. The work **[60]** has shown that the optimal Bayesian equaliser designed for complex-valued channels and modulation schemes is structurally equivalent to the complex-valued RBF network introduced in **[59]** and exploited this intimate connection to develop fast training algorithms for implementing the Bayesian equaliser based on the complex-valued RBF network. A novel strategy of utilising decision feedback was employed to improve the equaliser performance as well as to reduce its computational complexity.

In real-life mobile communication systems, typically fast frequency-selective fading environments are encountered, and it is critical that an equaliser is able to adapt its taps rapidly to changing channel conditions. The main objective of my journal paper **[61]** is to carefully investigate the performance of the adaptive Bayesian DFE under time-dispersive mobile fading channel environments. For stationary channels, it is well-known that the performance of the adaptive Bayesian DFE is inferior to that of the adaptive MLSE. This often leads to the question of why "inventing" the Bayesian symbol-decision equalisers. The work **[61]** also aims to address this issue by comparing both the adaptive Bayesian DFE and the adaptive MLSE under the same dispersive fading conditions commonly found in real-life mobile communication systems. A computer simulator was created to provide realistic dispersive and fast fading channel conditions[3], and an extensive simulation study was carried out to evaluate the adaptive Bayesian DFE, with the conventional (linear) adaptive DFE and the adaptive MLSE used as two benchmarkers. The conclusions of this simulation study are as follows. In terms of implementation and computational complexity, the adaptive Bayesian DFE is slightly more complex than the conventional DFE, but is of lower complexity than the adaptive MLSE. In terms of its symbol detection error rate, the adaptive Bayesian DFE dramatically

[3] 4-QAM was used and the system design parameters were comparable to the GSM specifications

outperforms the conventional DFE. Moreover, for severely fading multipath channels, the adaptive MLSE exhibits a significant degradation in comparison to the optimal performance and becomes inferior to the adaptive Bayesian DFE. This observation is not entirely surprising. For time-varying channels, the MLSE accumulates tracking errors in its channel estimate, owing to its sequence detection nature, and this causes a serious performance degradation. Moreover, employing decision-directed adaptation during the actual data transmission is essential for transmission over rapid time-varying channels. But the long decision delay associated with the MLSE makes decision-directed adaptation ineffective. By contrast, the Bayesian DFE makes decision on a symbol-by-symbol basis and therefore does not accumulate tacking errors in its channel estimation. Furthermore, due to its short fixed decision delay, the Bayesian DFE suffers less from the time delay in channel estimation and hence can exploit decision-directed adaptation more effectively. In summary, the Bayesian DFE has both implementation and performance advantages over the theoretically optimal MLSE for transmission over frequency-selective fast fading mobile channels.

The publication **[62]** is a survey of linear and nonlinear channel equalisation techniques. It unifies various previous works using neural networks under the framework of the Bayesian equalisation solution. By adopting a Bayesian approach, it is explained why the nonlinear classification capability of neural networks is desired for the equalisation process. Furthermore, this geometric interpretation of the equalisation process allows us to choose the most appropriate network structure for equalisation applications. The RBF network is shown to realise precisely the Bayesian equalisation solution. A further advantage of the Bayesian approach is that decision feedback can be utilised to reduce the associated processing complexity. The adaptive Bayesian equaliser based on the RBF network offers significant performance improvements over the traditional adaptive linear scheme at the cost of a small increase in computational complexity. Typically, several dB signal to noise ratio improvements can be obtained by the former at the error probability of 10^{-4}. Moreover, the required training time of the adaptive Bayesian equaliser is comparable to that of the adaptive linear scheme. Although the MLSE achieves the best theoretical performance when the channel is known, it suffers from a serious performance degradation in a highly nonstationary propagation environment. In contrast, the adaptive Bayesian equaliser is robust and may outperform the adaptive MLSE in this practical situation. My more recent journal paper **[63]** exploits the phasor classification based nature of the equalisation process and points out that the recent advances in the state-of-the-art sparse kernel classification algorithms adopted from machine learning offer viable design alternatives for implementing batch-based adaptive Bayesian equalisers.

6.2 Nonlinear Receivers in the Presence of Co-channel Interference

In frequency division multiple access/time division multiple access (FDMA/TDMA) cellular mobile systems, co-channel interference is the dominant limiting factor influencing the system's design and performance. The journal papers **[64],[65]** investigated adaptive equalisation in the presence of both intersymbol interference (ISI) as well as co-channel interference (CCI). In the work **[64]**, the RBF network was designed to realise a sophisticated nonlinear adaptive equaliser capable of operating under poor signal to interference ratio (SIR) and signal to noise ratio (SNR) conditions. A two-stage learning strategy was derived by exploiting the nature of the data structure and this enabled the RBF network to adaptively implement the optimal Bayesian symbol-decision based equaliser. At the first stage of the learning, a supervised clustering algorithm was employed to model the effects of the channel ISI. This learning stage is extremely simple and

robust, and it is capable of producing a good approximation to the optimal Bayesian solution. At the second stage of the learning, the network structure was expanded and an unsupervised clustering algorithm was incorporated into the learning procedure so that the network became capable of modelling the effects of the CCI, closely approximating the full optimal symbol-decision equalisation solution. The number of kernels used by the full Bayesian equaliser in the presence of both the channel ISI and CCI was equal to the product of the numbers of channel and co-channel states. This may become excessive for certain practical systems. The study **[64]** has shown that even by using the first-stage learning only, which considers modelling the channel states only, the RBF equaliser can effectively exploit the difference between a non-Gaussian interference signal and the Gaussian noise to the benefit of the achievable equalisation performance.

The high computational complexity associated with the full Bayesian solution in the presence of both ISI and CCI can be significantly reduced by utilising decision feedback. Moreover, the Bayesian DFE often considerably outperforms the Bayesian equaliser. The journal paper **[65]** derived an adaptive Bayesian DFE which incorporates CCI compensation. By exploiting the structure of the CCI, this adaptive Bayesian DFE is able to distinguish an interfering signal from white noise and utilises this information to improve its performance. The adaptive implementation of this scheme includes identifying the channel model using the least mean square algorithm and estimating the co-channel states. It is shown that only dominant or significant co-channel states have to be estimated, and this further reduces the associated computational complexity. Moreover, to speed up the learning, a novel scheme was proposed which estimates the scalar co-channel states and then maps them into the vector states required by the Bayesian DFE. Further interesting results have been obtained in comparing the performance of the adaptive Bayesian DFE and the adaptive MLSE in the presence of strong CCI. The adaptive Bayesian DFE employing the proposed simple scheme to compensate for the CCI was shown to outperform the adaptive MLSE that treats the CCI as an additional coloured noise source.

6.3 CDMA Nonlinear Multiuser Detection

For interference limited communication systems, such as CDMA systems, the underlying channel output phasor constellation may become linearly inseparable. A linear multiuser detector (MUD) cannot achieve adequate performance and hence a nonlinear MUD is preferred in such a situation. The optimal symbol-decision based MUD is the Bayesian MUD, which is computationally expensive to realise as it requires the knowledge of both the desired user's states as well as the interfering users' states. One way of adaptively implementing the Bayesian MUD is to estimate the desired user's channel impulse response (CIR) as well as the interfering users' CIRs. In the downlink case, however, the receiver only has access to the desired user's training sequence and hence it is difficult to estimate the interfering users' CIRs unless for example the blind subspace based eigenvalue decomposition aided solutions are employed. The employment of decision feedback, although normally very effective in terms of complexity reduction, has little use in reducing the computational load associated with downlink detection applications. A possible strategy is to estimate the interfering users' states using unsupervised clustering. Such an unsupervised clustering process is typically complex and has a limited multiuser resolution, particularly when the number of interfering users is large. Alternatively, a RBF MUD can be constructed by using only the desired user's states, with some degradation with respect to the optimal Bayesian performance.

To circumvent the difficulties in the adaptive implementation of the Bayesian MUD in the context of downlink applications, the journal paper **[66]** proposed a batch-based RBF MUD, which can be constructed by the support vector machine (SVM) algorithm based only on a block of the received signals and corresponding desired user's training symbols. The SVM classification is a general machine learning approach that is capable of constructing a sparse RBF or kernel classifier from a large training data set, while having an excellent generalisation capability. The simulation results provided in **[66]** demonstrated that the SVM MUD can closely approximate the Bayesian single-user performance. My experience with the SVM MUD indicates however that the size of the detector may not be always small enough for practical purposes. Other kernel classification algorithms, such as the relevance vector machine (RVM), can alternatively be used to construct a sparser RBF MUD centre set with similarly good generalisation performance to that of the SVM method. In particular, the orthogonal forward selection (OFS) algorithm based on the so-called Fisher ratio class separability measure, which was described in Section 3.2, offers further computational advantages in constructing a sparse RBF MUD centre set. This will be demonstrated in the next section, where the employment of multiuser detection is considered for multiantenna aided SDMA communication systems.

6.4 Multiantenna Aided Space-Time Nonlinear Receivers

Nonlinear multiuser detection was considered in **[41],[42]** in the context of multiple-receive-antenna assisted SDMA communication systems. A beamforming structure was employed in these studies, namely, antenna elements having a typical spacing of half a wavelength. In the context of SDMA, the spatial separation in terms of the angle of arrival between the desired signal and the closest interfering signal dominates the achievable system performance and hence the system's user capacity. When this angular separation is below a certain threshold, linear beamforming ultimately fails, since the signals transmitted by the individual users become linearly inseparable. In conjunction with nonlinear spatial processing the achievable system capacity can be significantly increased, since an adequate performance can be maintained even in case of a low angular separation, compared to linear beamforming. The journal papers **[41],[42]** investigated an adaptive implementation of the beamforming based Bayesian MUD using the RBF network structure. The OFS algorithm based on the Fisher ratio class separability measure was employed to construct a parsimonious RBF detector from a block of training data, which can closely approximate the optimal Bayesian performance. This novel block-data based construction algorithm has the same excellent generalisation capability as other state-of-the-art kernel modelling methods, such as the SVM and RVM, but additionally offers considerable computational advantages over the SVM and RVM techniques.

The publication **[67]** considered the generic SDMA communication system using multiple receive antennas, where each user and each receiver antenna is characterised by a user-antenna specific channel impulse response (CIR), and developed a RBF network assisted space-time equalisation scheme for dispersive fading channels. The uplink scenario was investigated and, therefore, decision feedback was employed to improve the MUD's performance as well as to reduce the complexity of the Bayesian MUD. By estimating the CIRs associated with every user, the optimal Bayesian space-time decision feedback equaliser (ST-DFE) was constructed. Simulation was carried out in conjunction with dispersive CIRs faded at the normalised Doppler frequency of 0.0005. A Kalman filter was employed for channel estimation, which predicts the CIR coefficients needed in the feedforward section and estimates the channel for the feedback section. The simulation results obtained clearly demonstrated that the RBF network assisted ST-DFE receiver structure outperforms

the conventional linear ST-DFE receiver structure and it is less sensitive to both error propagation and channel estimation errors.

6.5 Importance Sampling Simulation for Performance Evaluation

For the class of equalisers based on a symbol-by-symbol decision with decision feedback, the maximum *a posteriori* probability equaliser with decision feedback or the Bayesian DFE is known to provide the best performance. The complexity of this optimal Bayesian solution, however, increases exponentially with the CIR length and with the size of the modulated symbol constellation. Furthermore, due to its complicated structure, the performance analysis of the Bayesian DFE is usually based on conventional Monte Carlo simulation, which is computationally costly even for high-BER, low signal to noise ratio (SNR) conditions. To obtain a reliable BER estimate, at least 100 errors should occur during a simulation in order to maintain a sufficiently low estimation error variance. Thus, for a BER level of 10^{-6}, at least 10^8 data samples are needed. Investigating the Bayesian DFE, when a BER performance better than 10^{-6} is required, is computationally very demanding using a conventional Monte Carlo simulation.

Importance sampling (IS) refers to a simulation technique, which aims to reduce the variance of the error rate estimator. By reducing the variance of the error rate estimator, IS can achieve a given precision from shorter simulation runs, compared with a conventional Monte Carlo simulation. The basic idea behind IS is that certain values of the input random variables in a simulation have more impact on the error probability being estimated than others. If these "important" values are emphasized by sampling more frequently, the estimator variance can be reduced. The fundamental issue in IS simulation is then the choice of the biased distribution, which encourages the exploration of the important regions of the input variables. One of the most effective IS techniques is the mean translation approach, where the distribution is moved toward the erroneous decision region. This usually corresponds to shifting the density to a decision boundary. It is highly desired that a chosen IS technique is asymptotically efficient (AE). Loosely speaking, an AE estimator requires a number of simulation trials, which grows lower than exponentially, as the error ratio tends to zero. Thus, when AE estimators are available, it is realistic to attempt extremely low error probability simulations.

The application of a mean-translation based IS technique to practical simulation systems is by no means a straightforward and easy task. Iltis developed a randomized bias technique for the IS simulation of Bayesian equalisers operating without a decision feedback in:

R.A. Iltis, "A randomized bias technique for the importance sampling simulation of Bayesian equalizers," *IEEE Trans. Communications*, Vol.43, No.2/3/4, pp.1107–1115, 1995.

This IS simulation method however can only guarantee asymptotic efficiency for certain channels. The novel contribution of my journal papers **[72],[73]** is the derivation of a mean-translation based IS technique for the performance evaluation of the Bayesian DFE, which achieves *asymptotic efficiency*. The publication **[72]** considered the IS simulation of the Bayesian DFE using binary signalling. By viewing decision feedback as a geometric translation of the channel output constellation space into a reduced dimensional space by removing the ambiguity associated with the symbols that were already decided upon, the Bayesian

DFE is "converted" to the Bayesian equaliser having no feedback in the translated space. Hence, we can achieve the desired property that the two subsets associated with the opposite-class channel output states are always linearly separable. A design procedure was developed, which determines the set of hyperplanes that form the asymptotic Bayesian decision boundary and constructs the convex regions associated with individual states by intersecting hyperplanes that are reachable from the states concerned[4]. This provides the appropriate bias vectors for the simulation density to ensure asymptotic efficiency of the IS simulation. The journal paper **[73]** further extended these results to general multi-level modulation schemes. Based on a geometric translation of the multiple subsets of noise-free channel output states, the asymptotic Bayesian decision boundary separating any two neighbouring signal classes can be deduced. Furthermore, by exploiting the symmetric nature of the distribution within each subset of channel states, the symbol error rate of the Bayesian DFE applied to generic multi-level constellation was shown to be a scaled function of the equivalent "binary" Bayesian DFE's BER evaluated for any two neighbouring signal subsets. These two properties enable the extension of the IS simulation technique originally derived for the binary Bayesian DFE to the general multi-level constellation case.

6.6 Efficient Reduced-Complexity Implementation of Optimal Nonlinear Receivers

The optimal maximum *a posteriori* probability symbol-decision solution for the generic nonlinear receiver structure offers significant performance gains over the simple linear receiver structure, although this is achieved at the cost of a considerable increase in computational complexity. In single-user channel equalisation applications, for example, the complexity of the Bayesian solution increases exponentially with the channel impulse response (CIR) length and the number of bits/symbol. For interference limited communication systems, the computational requirements associated with the optimal Bayesian solution may become prohibitive. The significance of my publications **[74]-[76]** is that they derived an efficient reduced-complexity implementation of the optimal Bayesian equalisation solution, which has a computational complexity equal to the sum of *a few* linear equalisers' complexity. Geometrically, the linear equaliser or conventional DFE partitions the observation space with the aid of a hyperplane using a linear discriminant function, which requires only a few multiplications and additions. The root of the high complexity associated with the Bayesian equaliser or DFE lies in the fact that the Bayesian solution partitions the observation space using a hypersurface and the optimal nonlinear discriminant function required to determine this hypersurface is expensive to compute. It can be shown that asymptotically, as the SNR tends to infinity, the Bayesian decision hypersurface becomes piecewise linear and is made up of a set of hyperplanes. In practice, at large rather than infinite SNRs, the performance difference between using the Bayesian decision boundary and a piecewise-linear approximation becomes negligible. This asymptotic property was exploited for example in the derivation of an importance sampling simulation technique developed for the performance evaluation of the Bayesian solution, presented in the previous section. It also motivated my research on using multiple hyperplanes to partition the signal space **[74]-[76]**.

My journal paper **[74]** considered the case of the Bayesian DFE in conjunction with binary signalling. The design procedure is simple yet effective. It was shown that each component hyperplane, which forms part of the asymptotic Bayesian decision boundary, is determined by a single pair of the closest opposite-

[4] A hyperplane is reachable from a state if a bias vector can be constructed which shifts the state to the hyperplane, see **[72]**

class channel output states, which were hence termed a *dominant* state pair. The number of the dominant state pairs or component hyperplanes is determined purely by the CIR. Thus, the design procedure involves automatically finding the set of these dominant state pairs and constructing a separating hyperplane for each pair. The resultant equaliser consists of a set of linear discriminant functions and a Boolean logic function. This simple equaliser structure asymptotically realises the optimal Bayesian solution. In practice, for small to moderate SNRs, the performance degradation in comparison to using the full Bayesian solution is often negligible. The attainable complexity reduction capability can be demonstrated by the numerical example given in **[74]**, where the CIR consists of 4 taps and a binary signalling scheme is employed. The full Bayesian DFE detector requires 80 multiplications, 127 additions and 16 $\exp(\bullet)$ function evaluations to make a symbol decision. By contrast, the multiple-hyperplane design procedure identifies seven hyperplanes that form the asymptotic Bayesian decision boundary for this example. The resultant multiple-hyperplane-based detector requires 28 multiplications and 28 additions to make a symbol decision, which is seven times of the complexity required by the conventional (linear) DFE but substantially lower than that of the full Bayesian design.

The journal paper **[75]** extended the design to the Bayesian DFE contrived for the general multi-level modulation scheme. Two properties, the *shifting* property and *symmetric* property, facilitate the extension of the design procedure from the binary Bayesian DFE to the generic case. First of all, decision feedback ensures that the underlying equalisation problem is linearly separable. The shifting property manifests the following fact. Any two neighbouring subsets of channel states, corresponding to the two neighbouring symbol points in the symbol constellation, are connected in the observation space by the "reverse" channel tap vector[5]. The symmetric property implies that the phasors of each subclass of channel states are distributed symmetrically around the subclass centre. Thus, the design procedure can be based on any two neighbouring subclasses. Furthermore, the reduction in detector complexity achieved with the aid of this signal space partitioning approach is more significant for high-order multi-level modulation schemes, compared to the binary one. In **[75]**, a numerical example is used to demonstrate this complexity reduction capability, where the channel has 3 taps and a 4-ary modulation scheme is employed. For this particular example, the design procedure constructs a five-hyperplane based detector. The full Bayesian DFE requires 380 additions, 256 multiplications and 64 $\exp(\bullet)$ evaluations to detect a symbol. The multiple-hyperplane detector, however, needs only 25 additions and 15 multiplications to make a decision, which is less than 6% of the complexity required by the full Bayesian DFE.

7 Blind Equalisation

To cope with unknown and time-varying channel conditions, often a supervised learning strategy is adopted. During the training period, a predefined training symbol sequence is sent by the transmitter. A (synchronised) receiver generates this training sequence locally and uses it as the reference to estimate the CIR and/or adapt the receiver equaliser. Training "wastes" the precious system bandwidth and is to be avoided if possible in order to improve the effective system throughput. Furthermore, in some communication systems, such as multi-point communication networks, using regular training is impossible. In such situations, an unsupervised learning strategy, commonly known as *blind* equalisation, must be adopted. In blind adaptation

[5]Let the tap vector be $\mathbf{a} = [a_0\ a_1 \cdots a_{n-1}]^T$. Then, the reverse tap vector is defined by $\mathbf{a}_{\rm rev} = [a_{n-1} \cdots a_1\ a_0]^T$

the receiver must estimate the CIR and/or adapt the equaliser purely based on the received signal samples, with the aid of some knowledge of statistics regarding the transmitted data symbols. A blind equaliser must rely on some form of higher-order statistics, since the second-order statistics cannot resolve the ambiguity of the channel phase response. Blind equalisation techniques can be conveniently divided into three families.

The *first family* of blind adaptive algorithms constructs a transversal or finite-duration impulse response (FIR) filter, i.e. a linear equaliser, which aims for removing the effects of the CIR. These blind equalisers, commonly referred to as Bussgang algorithms, typically adjust the equaliser coefficients recursively by optimising some nonconvex criterion using a stochastic gradient algorithm. A Bussgang-type blind equaliser typically has a low computational complexity, but suffers from the drawback of slow convergence. The *second family* of blind equalisation algorithms identifies a channel model using techniques based on higher order cumulants or the equivalent higher order spectra. Once the CIR has been obtained, it can be employed to design an equaliser. Because the process of channel estimation is separated from channel equalisation, a variety of existing equalisation schemes can be utilised. This second class of blind equalisers, although very general and powerful, requires a large number of received data samples and extensive computations to estimate the higher order cumulants. The *third family* of blind equalisation schemes adopts the principle of joint maximum likelihood channel estimation and data detection. This class of blind equalisers is thus capable of offering optimal performance, but it is computationally expensive if not prohibitive. A major advantage of this third family is that relatively few signal samples are required to achieve adequate equalisation.

The publications **[77]-[82]** represent my research in this important field of blind equalisation, and my contributions have spread across the whole spectrum of blind equalisation problems. Specifically, the journal papers **[77]-[79]** provide my novel contributions to the Bussgang class of blind equalisers, while the journal paper **[80]** summarises my work in the higher-order cumulant based family of blind equalisation schemes. Finally, the publications **[81],[82]** offer my generic results focussed on the third class of blind equalisation techniques using the joint maximum likelihood estimation of the CIR and data. These contributions are given below in more detail.

7.1 Low-Complexity Blind Equalisation

Blind equalisation improves the system's effective bandwidth efficiency by avoiding the use of a training sequence. Furthermore, using regular training for multi-point communication systems is infeasible and blind equalisers provide a practical means of combating the detrimental effects of the channel intersymbol interference (ISI) in such systems. For communication systems employing high bandwidth-efficiency quadrature amplitude modulation (QAM) signalling, the constant modulus algorithm (CMA) based FIR equaliser is by far the most popular blind equalisation scheme. It imposes modest computational requirements and hence readily meets real-time computational constraints. The CMA is also robust to imperfect carrier recovery. A particular problem of the CMA, however, is that it only achieves a moderate level of mean square error (MSE) after convergence, which may not be sufficiently low for the system to obtain an adequate BER performance. A possible solution is to switch to a decision directed (DD) adaptation, which should be able to minimise the residual CMA steady state MSE. However, in order for such a reconfiguration to be successful, the CMA's steady state MSE should be sufficiently low. In practice, such a low level of MSE may not always be achievable by the CMA. The convergence speed of the CMA is also typically slow.

My journal paper **[77]** proposed a novel low-complexity blind adaptive scheme for high-order M-QAM communication systems. The proposed blind algorithm is based on the idea of maximising the "*a posteriori*" probability density function (PDF) of the equaliser output as a function of the equaliser weights. To accomplish a fast as well as reliable convergence and to keep the complexity to a minimum, a multi-stage procedure is adopted. At the first stage, a 4-cluster PDF model is adopted as though the data constellation is an equivalent 4-QAM one. The aim of this stage is to classify the equaliser outputs correctly into one of the four quadrants in the complex plane with a high probability. At the second stage, a 16-cluster PDF model is used and it is divided into 4 sub-sets, one for each quadrant. If the equaliser output appears in a particular quadrant, the corresponding 4-cluster sub-model is used to adapt the equaliser weights. After the stage two, the complex plane is divided into 16 square regions, each containing a 4-cluster sub-model. The procedure is continuing until after the $\log_2 \sqrt{M}$-th stage, when the correct data constellation is restored. The computational complexity of this multi-stage blind equaliser, quantified in terms of the number of multiplications and additions per weight update, is only slightly more than that of the simple CMA-based blind equaliser. Simulation results presented in **[77]** show that this blind equaliser has a much faster convergence speed and better steady-state equalisation performance than the CMA. However, a drawback of this blind equaliser is that its adaptive process requires $\log_2 \sqrt{M}$ switching operations and each stage of adaptation needs a different set of algorithmic parameters.

In an effort to circumvent the difficulties in deciding whether the CMA equaliser can be switched to and when it can be switched to a DD adaptation, De Castro and co-workers proposed an interesting solution in:

F.C.C. De Castro, M.C.F. De Castro and D.S. Arantes, "Concurrent blind deconvolution for channel equalization," in *Proc. ICC'2001* (Helsinki, Finland), June 11-15, 2001, Vol.2, pp.366–371.

Rather than switching to a DD adaptation after the CMA has converged, De Castro and co-workers proposed to operate a DD equaliser concurrently with a CMA equaliser. The weight adaptation of the DD equaliser follows that of the CMA equaliser, and the DD adjustment only takes place if the CMA is seen to achieve a successful adjustment with a high probability. At a cost of slightly more than doubling the complexity of the simple CMA, this concurrent CMA and DD equaliser is shown to obtain a dramatic improvement in equalisation performance over the CMA. My journal paper **[78]** provided a comparative study of this combined CMA and DD equaliser and my multi-stage blind equaliser discussed in the previous paragraph, in the fractionally-spaced (i.e. sampling faster than symbol rate) equalisation context. The conclusions obtained in this comparative study are as follows. The multi-stage blind equaliser is computationally simpler, has a faster convergence rate and achieves better steady-state equalisation performance than the combined CMA and DD blind equaliser. An advantage of the combined CMA and DD blind equaliser is that it has a fewer number of algorithmic parameters and is much easier to tune than the multi-stage blind equaliser.

My journal paper **[79]** proposed a novel combined CMA and soft decision directed (SDD) blind equaliser, which offers a state-of-the-art technique for low-complexity blind equalisation of high-order QAM schemes. In the SDD sub-equaliser, a local 4-cluster PDF model is used to approximate the PDF of the equaliser output. The operations of the CMA and SDD sub-equalisers occur in a truly concurrent manner, and there is no need to compute an intermediate output after the CMA adaptation, as the combined CMA and DD scheme does, to decide whether weight-adaptation of the DD equaliser can take place. Rather than being committed to a single hard decision as the DD scheme does, alternative decisions are also considered in a local region

that includes this hard decision, and each tentative decision is weighted by an exponential term, which is a function of the distance between the equaliser's soft output and the associated tentative decision. This soft decision nature enables a simultaneous update of both the constituent equalisers without serious error propagation, while simplifying the scheme's operation. This also has an effect of improving the achievable convergence speed. This SDD scheme in fact corresponds to the last stage of the multi-stage blind equaliser discussed previously. This combined CMA and SDD scheme is computationally simpler than the concurrent CMA and DD one, and just like the latter, the combined CMA and SDD blind equaliser is easy to tune. Extensive simulation results have shown that both the blind equalisers have the same steady-state equalisation performance but the combined CMA and SDD equaliser has a faster convergence than the latter.

7.2 Optimal Blind Channel Identification with Higher-Order Cumulant Fitting

For the family of blind equalisers based on higher-order cumulant fitting techniques, usually a two-stage strategy is adopted, which first identifies a channel model using higher-order cumulant (HOC) fitting algorithms and then employs the estimated channel model to design an equaliser. The key stage of this approach is its ability to obtain an accurate channel estimate. Once the CIR is available, a variety of existing equaliser design methods can be employed, ranging from a simple linear inverse filter to a sophisticated maximum likelihood sequence estimator, depending on the required balance tradeoff between the achievable performance and the complexity imposed. The typical cost functions of HOC fitting are, however, multimodal, and conventional gradient optimisation techniques may converge to locally optimal solutions, unless a good initial value of the channel parameters is provided, which is not always possible. The journal paper **[80]** in the submission is one of the first papers in the literature to propose the use of a genetic algorithm (GA) for optimal blind channel identification in conjunction with HOC fitting. More specifically, in the work **[80]**, a computationally efficient micro GA implementation is adopted to optimise the fourth-order cumulant-based cost function.

Some observations can readily be drawn from the study of **[80]**. In the simulation, the proposed micro GA-based scheme always approaches the globally optimal channel estimate and the optimisation process converges rapidly. Compared to other existing methods of HOC fitting, the proposed technique appears to be more accurate and robust. This is demonstrated by the associated very small standard deviations of the estimated CIR taps over different runs with randomly chosen channel populations and for a wide range of channel conditions. For other existing methods of HOC fitting, it is a common phenomenon that the estimation accuracy reduces and the estimation variance increases, as the SNR decreases. For our GA-based method, the simulation results suggest that, at least for the channel conditions simulated, the SNR has little effect on the achievable convergence rate and estimation accuracy. The number of cost function evaluations required for the GA scheme to converge to a globally optimal solution was found to be typically a few thousand in our simulations. This compares favourably to other existing schemes. If some of the CIR taps have zero values, our method is capable of identifying them as the non-existing channel taps by providing their estimated values at least an order smaller in magnitude than the dominant CIR taps. This is particularly important in practice, when an over-length channel model is used in order to avoid time-consuming and difficult model-order selection process.

7.3 Blind Joint Maximum Likelihood Channel Estimation and Data Detection

In blind equalisation, both the CIR and the transmitted data sequence are unknown. In theory, their optimal estimates can be obtained via the joint maximum likelihood (ML) optimisation over the CIR and the data. The computational requirement of such a joint optimisation procedure is, however, often prohibitively large. In practice, often suboptimal strategies or approximations are adopted. In conjunction with a known CIR, an optimal ML solution for data detection is the maximum likelihood sequence estimation (MLSE), which can be carried out using the Viterbi algorithm (VA). Given the data sequence, an optimal ML solution applicable to channel estimation is constituted by the classic least squares channel estimate. A straightforward way of carrying out joint estimation when both the CIR and the data sequence are unknown is to employ a batch iterative process between data decoding and CIR estimation. The blind trellis search technique designed for joint channel and data estimation may be viewed as a recursive process, where an "enlarged" VA retains several surviving sequences and associated CIR estimates for each state of the trellis. Another interesting batch processing technique designed for joint ML estimation is the quantised channel algorithm. It maintains a family of candidate CIR estimates with discrete (quantised) parameters. Each CIR model is used by the VA to tentatively decode the data, and the algorithm selects the most likely quantised CIR and data combination. My journal paper **[81]** proposed a novel scheme for joint ML channel and data estimation using a GA. This paper is one of the first research papers, which deal with the application of GA-based optimisation to joint ML detection problems. The significance of the work **[81]** is that it demonstrates the effectiveness of the GA in obtaining the optimal or near optimal solution for a joint ML estimation problem, which otherwise would be computationally prohibitive for a conventional optimisation technique to solve.

The proposed GA-based scheme designed for blind joint ML channel and data estimation is remarkably simple. The algorithm developed in **[81]** adopts a two-layer strategy for the joint optimisation over the CIR and the data by combining the GA with the VA. At the top layer, an efficient version of the GA known as the micro GA searches through the CIR parameter space to optimise the ML criterion. The bottom layer consists of a number of VA units, one for each member of the CIR population. Each VA unit decodes the data based on the given CIR model and feeds back the corresponding likelihood metric value to the GA. The GA evolves the CIR population to find the globally optimal solution. The conclusions drawn from the study of **[81]** are as follows. The proposed GA-based scheme converges consistently to a globally optimal solution with a small estimation variance. Compared to other existing batch-type methods used for joint ML channel and data estimation, the GA-based scheme is more robust and accurate as well as computationally more efficient in terms of the total number of required VA evaluations. A significant advantage of this GA-based scheme is that it requires a small number of received data samples, ranging from a few dozens to hundred, to achieve accurate blind equalisation. The algorithm is particularly effective in combating the *curse of dimensionality* associated with high-order modulation constellations.

Among the three families of blind equalisation methods, the third family based on the principle of joint ML CIR estimation and data sequence detection is the computationally most expensive one. The scheme presented in the publication **[82]** has a "low complexity" in comparison to other blind algorithms of the third class. To achieve low complexity, this algorithm performs a joint channel estimation and symbol detection, rather than an optimal data sequence detection. At each symbol instant, each of the possible feedforward symbol sequences is used to produce a least mean square CIR estimate. Each of these "conditional" channel

estimates is then employed to design a Bayesian DFE for symbol detection, and each Bayesian DFE provides a tentative symbol decision. The final symbol decision is chosen from the set of tentative symbol decisions as the one that is the most likely. This final detected symbol is fed back to the equaliser feedback section and is used to update an "unconditional" channel estimate. This blind equaliser is *fast*, as it is computationally much simpler than other blind algorithms based on joint channel and data estimation. The drawback of this scheme is that it does not offer the optimal joint ML solution for blind equalisation.

8 Optimising Controller Integrity for Finite-Precision Implementations

The current controller design methodology often assumes that the controller is implemented exactly, even though in reality a control law can only be realised in finite precision. The justification of this assumption is usually on the ground that the plant uncertainty is the most significant source of uncertainty in the control system. Increasingly, however, researchers have realised that the controller uncertainty has significant influence on the performance of the control system. A stable control system may achieve a lower than predicted performance or may even become unstable, when the control law is implemented with a finite-precision device due to the *finite word length* (FWL) effects. This is highlighted in the context of the so-called fragility puzzles, where some high-performance robust optimal controllers are seen to become fragile if the controller design does not take into account the finite precision related implementation uncertainty. Ironically, these controllers have been designed to tolerate uncertainty in the plant, and yet small perturbations of the controller parameters may render the designed closed-loop system unstable. The fragility issues are strongly related to and interconnected with the FWL controller implementation issues. Although the number of controller implementations using high-precision floating-point processors is increasing due to their reduced price and increasing availability, for reasons of cost, simplicity, speed, memory space and power consumption, the use of fixed-point processors is still more desirable for many industrial and consumer applications, particularly for mass market applications in the automotive and consumer electronics sectors. Furthermore, due to their reliability and well-understood properties, predominantly fixed-point processors are used in safety-critical systems. In conjunction with a fixed-point processor, however, the detrimental FWL effects are markedly increased due to a reduced precision. The problem can become serious when a high sampling rate and a high-order controller are used.

Therefore, a major concern when applying advanced control design methods to practical systems is the *reliability* and *achievability* of the designed controller performance. Care must be exercised in implementing robust optimal control laws, as an inappropriate realisation may result in degraded performance of the control system. This may even lead to doubts about the effectiveness of the designed control strategy and about the benefits of advanced control techniques in general. Although there is an increasing awareness about the detrimental FWL effects, traditionally these have rarely been accounted for in the design or implementation of advanced control strategies. The publications **[83]-[98]** in the submission represent my contributions to the research of integrated controller design under both plant and controller uncertainties. More specifically, the contributions presented in **[83]-[98]** are significant because they:

1. Provide a comprehensive understanding of the FWL effects on the achievable closed-loop stability and performance, and develop a universal approach to optimising the robustness of closed-loop stability

and controller performance in a finite-precision implementation.

2. Develop a unified methodology for combining optimal FWL implementation requirements as well as robust optimal controller designs, and derive an integrated design procedure for optimising the controller's integrity under both model uncertainty and controller implementation uncertainty.

Generally speaking, there are two types of FWL errors in the digital controller. The first one is the perturbation of controller parameters implemented with FWL precision and the second one is the rounding errors that occur in arithmetic operations. Typically, the effects of these two types of errors are investigated separately for reasons of mathematical tractability. The effects of the first type of FWL errors are typically investigated with the aid of some closed-loop stability robustness related measures. The second type of FWL errors is usually quantified in terms of the roundoff noise gain. My research in finite-precision digital controller design covers both of these two critical aspects of FWL implementations. Below my novel contributions are detailed with reference to the corresponding journal papers.

8.1 Closed-Loop Stability Margin Maximisation for Fixed-Point Implementation

A designed controller must first of all ensure closed-loop stability. In a finite-precision implementation, however, the actually implemented controller parameters will differ from the designed values, due to the associated FWL effects. If this controller parameter perturbation moves some of the closed-loop eigenvalues across the stability boundary, the closed loop becomes unstable. The problem is particularly serious in fixed-point implementations having short word lengths. However, losing closed-loop stability must be prevented, which constitutes the most crucial issue in finite-precision controller design. It is known that the FWL effects inflicted on the closed-loop stability depend on the controller realisation structure. This property can be utilised to select an appropriate controller realisation in order to improve the robustness of closed-loop stability under controller parameter perturbations. The journal papers **[83]-[93]** concentrate on solving the generic problem of optimal controller realisation with maximum closed-loop stability margin for fixed-point implementation.

In practice, it is vital to know when the FWL error will cause closed-loop instability. In other words, given a specific controller realisation, we would like to know how robust the closed-loop stability will be with respect to the controller's parameter perturbation. This ultimately means that we would like to know the largest open "hypersphere" in the controller perturbation space, within which closed-loop remains stable. The size or radius of this "hypersphere" provides a true measure of the closed-loop stability robustness. From this measure, the minimum word length required to ensure closed-loop stability for the given controller realisation can be estimated. The optimal controller realisation in theory is defined as the one that maximises this measure. However, how to calculate the value of this true closed-loop stability measure is still an open problem. To circumvent this difficulty, a practical approach is to adopt approximation and derive some computationally tractable "lower bound" of this true measure. In the classical work by Gevers and Li:

M. Gevers and G. Li, *Parametrisation in Control, Estimation and Filtering Problems: Accuracy Aspects*. Springer Verlag: London, 1993

an l_2-norm FWL stability measure was formulated based on the closed-loop pole sensitivity analysis. This l_2-norm FWL stability measure is computationally tractable and can be regarded as a lower bound of the true stability measure in some sense. In the work by Gevers and Li, a practical design framework for the optimal controller realisation problem has been developed based on this stability robustness measure.

8.1.1 Optimal Finite-Precision Controller Design Based on l_1-Norm Stability Measure

The main contribution of the works **[83]-[87]** is to formulate an l_1-norm stability measure and to develop a generic optimal FWL controller design using this pole-sensitivity based l_1-norm stability measure. This l_1-norm stability measure is computationally tractable and can be shown to be a more "accurate" approximation to the true stability measure than the l_2-norm based measure. Thus, it provides a better criterion for optimising the closed-loop stability bound than the classical l_2-norm measure. Based on this new l_1-norm stability measure, optimal FWL controller realisation problems have been solved for various controller structures, including the PID controller and output feedback controller **[83]-[85]**, the state-estimate feedback controller **[86]**, and the reduced-order observer-based controller **[87]**. My further specific contributions are:

1. An efficient global optimisation procedure was developed based on the adaptive simulated annealing algorithm to solve the generic optimal FWL controller realisation problem **[83]-[86]**. The cost function of a generic optimal FWL controller realisation problem is non-smooth as well as non-convex, and hence traditional optimisation methods, such as the simplex search algorithm, can only find a locally optimal solution. By adopting an efficient global optimisation strategy, "global" optimal controller realisations can be obtained, which have a better FWL closed-loop stability margin.

2. The effects of sampling rate on the FWL closed-loop stability margin was extensively investigated, and it has been shown that canonical controller forms, such as the controllable canonical realisation, may have a poor FWL closed-loop stability margin and would require a prohibitively large word length to guarantee stability, particularly in the fast sampling case **[83],[85]**. This reinforces the need to search for an optimal controller realisation for FWL fixed-point implementations.

8.1.2 An Improved l_1-Norm Stability Measure for FWL Controller Design

In all the works based on pole sensitivity analysis prior to my publication **[88]**, the closed-loop eigenvalue sensitivity was considered. However, it is well-known that the stability of a linear system depends only on the moduli of its eigenvalues. An eigenvalue is specified by its modulus and argument. Thus all the previously proposed stability measures include the unnecessary eigenvalue arguments, and this redundancy is the main source that imposes an unnecessarily conservative constraint on estimating the true stability bound. By considering the sensitivity of the closed-loop eigenvalue moduli, the journal paper **[88]** proposed a new improved l_1-norm stability measure. This new stability measure is computationally tractable and provides a more accurate approximation of the unknown true stability measure, i.e. it is a lower bound closer to this true stability measure, than other existing measures based on the pole-sensitivity approach. This improved stability measure yields a more accurate estimate of the minimum word length required to guarantee closed-loop stability for a fixed-point controller realisation, than any other existing pole-sensitivity

measure. Based on this improved measure, the optimal controller realisation problem was formulated and solved in **[88]** for the generic controller structure that includes the output-feedback controller, the full-order observer-based controller and the reduced-order observer-based controller.

My journal paper **[89]** compared the two alternative approaches based on the pole sensitivity and complex stability radius measures, which were both designed for optimising the closed-loop stability robustness of fixed-point digital controllers. It was claimed by some researchers that the complex stability radius measure was better (less conservative) than the measures based on pole sensitivity analysis. The work **[89]** showed that this claim is no longer valid for the pole sensitivity approach that adopts the improved l_1-norm stability measure. The conclusions of the comparative study **[89]** are as follows.

- Both the pole sensitivity and complex stability radius approaches involve some approximations in estimating the true stability measure. Therefore, they are both conservative measures. As the assumptions stipulated for these two different lower bounds of the intractable true stability measure are different, it is difficult to say which measure is less conservative in estimating the true minimum word length. It is generally dependent on the specific scenario. In particular, the corresponding optimal controller realisations obtained by solving the two related optimisation problems are generally different. For the output-feedback controller structure, experience shows that the two approaches are often compatible in that the two optimal controller realisations usually have similarly good FWL characteristics in the context of fixed-point implementations.

- An important advantage of the complex stability radius measure is that the corresponding optimisation problem can be formulated as a linear matrix inequality (LMI) problem, and this LMI problem is easier to solve than the nonlinear optimisation problem associated with the pole sensitivity approach. The latter can have many solutions. The pole sensitivity approach however is applicable to the general controller structure that includes output-feedback and observer-based controllers and that is parameterised either by shift or delta operators. The approach based on the complex stability radius measure in its present form can only be applied to output-feedback controllers, and it is not apparent how to generalise to observer-based controllers or controllers in the delta operator domain.

8.1.3 Optimal Finite-Precision Controller Realisations in Delta-Operator Domain

A digital controller structure is typically described and realised with the shift operator. However, a controller can alternatively be described and realised with the delta operator. Two major advantages are claimed for the use of the delta-operator realisation: a theoretically unified formulation of both continuous-time and discrete-time systems, and better numerical properties in FWL implementations, particularly at high sampling rates. However, for fixed-point implementations having a short word length, the advantages of the delta form are limited due to the low dynamic range of the fixed-point arithmetic. Previous research was mainly concentrated on improving the efficiency of delta-operator parameterisation by modifying the delta operator form. My works **[90]-[92]** adopted a more fundamental approach to optimising the closed-loop stability robustness of digital controllers parameterised in the delta operator with respect to FWL errors. By finding optimal delta-operator controller realisations for fixed-point implementations, all the potential advantages of delta-operator parameterisation can be achieved. It turns out that the pole-sensitivity approach

used for analysing FWL digital controllers in the shift operator domain, which was reported in the previous two subsections, can be extended to the study of delta-operator based controller realisations.

The journal paper **[90]** derived an l_1-norm stability measure applicable to delta-operator based controller realisations for the output-feedback controller structure. The optimal controller realisation problem in the delta operator domain was formulated based on this stability measure. Because the optimisation criterion for this optimal realisation problem is non-smooth as well as non-convex, an efficient non-gradient based global optimisation method based on the adaptive simulated annealing was employed in **[90]** to search for an optimal controller realisation. The results obtained in the study of **[90]** confirm that the optimal FWL realisations of the delta-operator based controller have better closed-loop stability margins than those of the shift-operator based controller, especially under fast sampling conditions. The work **[91]** further extended the delta-operator parameterisation to the general controller structure that includes both output-feedback and observer-based controllers. A unified formulation was adopted to include both the shift and delta operator parameterisations and to analyse the underlying relationship between these two controller parameterisations. Under this unified formulation with the l_1-norm stability measure as the optimisation criterion, the generic optimal controller realisation problem was derived and solved in **[91]**, and the results derived in this study demonstrate that one can always obtain an optimal delta controller realisation, which has a better closed-loop stability margin than the optimal shift realisation in FWL fixed-point implementation.

8.1.4 Sparse Realisations Based on the Improved l_1-Norm Stability Measure

Determining the best controller structure for a fixed-point implementation is far from trivial. The main difficulties lie in the nonlinear and non-smooth nature of the problem, as well as in the multi-objective requirements. Current research has focused on assessing optimal controller realisations in terms of their closed-loop stability. This is natural, as the closed-loop stability is the most critical issue. However, other closed-loop controller performance measures can also be important and should be taken into account. Moreover, in real-time applications where maintaining computational efficiency is crucial, a sparse controller structure is highly desired. A controller is said to have a sparse structure if it contains many trivial elements of 0, 1 or -1. Apart from computational advantages, these trivial controller coefficients do not cause any error in a fixed-point implementation. Having a sparse structure is particularly important in the context of real-time applications of high-order controllers. Canonical controller forms are usually very sparse but may have poor closed-loop stability margins. The optimal controller realisation obtained by maximising some stability measure is on the other hand generally non-sparse. Finding sparse controller realisations with good closed-loop stability characteristics in FWL implementations poses a complex multi-objective problem.

Previous work has derived the design procedure for obtaining sparse controller realisations based on the l_2-norm stability measure. The original contribution of my journal paper **[93]** is that of proposing a new practical design procedure for obtaining sparse suboptimal controller realisations based on the improved l_1-norm stability measure. In the study **[93]**, a new improved FWL closed-loop stability measure was proposed, which takes into account the number of trivial elements in a controller realisation. The true optimal realisation that maximises this measure will possess an optimal trade-off between robustness to FWL errors and a sparse structure. This new stability measure is computationally tractable but it is not a continuous function of the controller parameters. Thus, the true optimal sparse realisation that maximises this new sta-

bility measure is difficult to obtain. However, the improved l_1-norm stability measure is a close lower bound of this new measure, and we know how to optimise the former. This leads to a practical iterative design procedure to search for a suboptimal solution with a sparse structure. Specifically, the controller realisation that maximises the improved l_1-norm stability measure is first obtained, but the resultant "optimal" realisation is non-sparse. A stepwise algorithm is then applied to make this controller realisation sparse without overly sacrificing the achievable FWL stability robustness. The proposed method has some advantages over the existing methods used for obtaining sparse solutions: it is more accurate in estimating the robustness of the FWL closed-loop stability, while the computational complexity is considerably reduced.

8.2 Optimal Controller Realisations for Floating-Point Implementation

With decreasing in price and increasing in availability, the use of floating-point processors in controller implementations has increased dramatically. Optimising controller integrity for floating-point implementations has become an important practical topic. Floating-point representations have quite different characteristics from fixed-point representations. The dynamic range of a floating-point representation is determined by its exponent part. Overflow or underflow occurs when the number of bits for the exponent part is insufficient. Under the condition that the number of exponent bits is sufficient, the achievable accuracy or precision is determined by the mantissa part. Note that the perturbation resulting from using a finite-precision floating-point arithmetic is multiplicative, unlike the additive perturbation resulting from a finite-precision fixed-point arithmetic. The characteristics of the floating-point representation are highly complex and more difficult to analyse than those of a fixed-point format. The publications **[94],[95]** represent my research in the field of optimal finite-precision controller design used for floating-point implementations. The most significant contribution of my journal paper **[94]** is:

- Pointing out for the first time that optimal FWL floating-point controller designs have to consider both the mantissa and exponent parts of the representation format.

All the previous research works prior to my publication **[94]** only considered the stability robustness with respect to the mantissa part, as the robustness of closed-loop stability depends only on the mantissa part. This would however imply an unlimited exponent word length. Maximising some stability measure will result in a minimum mantissa word length that can still guarantee closed-loop stability. However, in doing so, the dynamic range of the controller realisation might have been increased, leading to an increase in the required exponent word length. In practice, the total word length is finite and fixed, and sufficient number of bits must first be allocated to the exponent part of the representation, while the remaining bits are then used for the mantissa part. Therefore, a better approach is to maximise the robustness of closed-loop stability with respect to both the mantissa and exponent parts, i.e. minimising the total word length required to ensure closed-loop stability. Note that the discussion here equally applies to fixed-point representations.

Specifically, in the work **[94]**, a novel exponent measure was first derived to quantify the dynamic range of a floating-point controller realisation. This exponent measure defines the minimum exponent word length required for the controller realisation. A closed-loop stability measure was then developed to quantify the robustness of closed-loop stability with respect to the precision of floating-point representation. This mantissa measure was derived by considering the sensitivities of the closed-loop eigenvalue moduli, in a

way similar to the derivation of the improved l_1-norm stability measure for fixed-point implementation. The FWL properties of a floating-point controller realisation can be characterised by a composite measure consisting of both the exponent and mantissa measures. Based on this novel composite stability measure, the optimal FWL controller realisation problem was formulated and solved in **[94]**, where it was demonstrated that the proposed design procedure yields computationally efficient controller realisations with enhanced FWL closed-loop stability performance. In particular, the total word length required to guarantee closed-loop stability is minimised.

An additional contribution of **[94]** is to provide a comparison between two alternative approaches to FWL controller design. Most works dedicated to FWL controller design adopt an indirect strategy, which relies on the following property. A control law can be implemented with different realisations, and these different realisations are all equivalent if they are implemented at an infinite precision. However, different controller realisations possess different degrees of robustness to FWL errors. The control law is assumed to be given by some controller design methods, which may not take into account any FWL considerations, and the FWL design has to select optimal realisations for the given control law by optimising some FWL criteria. Most of my works belong to the category of this indirect approach. An alternative but better approach is to explicitly incorporate the FWL effects into the controller design process. This direct strategy should be a preferred approach, since it does not impose specific assumptions on the controller. However, except for a few special cases, applying the direct approach to various controller design methods is still an open problem. But this difficulty does not exist in the context of the indirect strategy where the controller synthesis and controller realisation constitute two separate steps. Various existing controller design methods can be used to attain a transfer function or an initial realisation of the controller, which can then be optimised to satisfy the FWL implementation requirements.

8.3 Unified Framework for Finite-Precision Controller Design

In practice, a controller is realised by a digital processor of finite bit length in a particular representation format, namely, fixed-point, floating-point or block-floating-point format. My journal paper **[96]** developed a unified framework for finite-precision controller design in different representation schemes. In a generic arithmetic scheme, the total word length is divided into a sign bit, a word length characterising the dynamic range, and another word length representing the resolution or precision. Specifically, in a fixed-point representation, the dynamic range is defined by the integer part of the representation and the precision by its fractional part. For a floating-point format, the exponent part specifies the dynamic range of the representation and its mantissa part defines the accuracy, while in a block-floating-point format, the block exponent and the block mantissa define the dynamic range and precision of the representation, respectively. It is also known that the perturbation imposed by the FWL fixed-point representation is additive, while it is multiplicative for the FWL floating-point representation, and neither multiplicative nor additive for the block-floating-point representation. By unifying the three FWL controller designs obeying the above-mentioned three different formats within a common framework, the contributions of my publication **[96]** are summarised as follows.

A unified FWL closed-loop stability measure was derived which is applicable to all the three above-mentioned arithmetic schemes. Unlike most of the existing works which only take into account the precision of a representation scheme under the assumption of an unlimited dynamic range, both the dynamic range and

the precision of an arithmetic scheme are considered in this new unified measure. To facilitate the design of optimal finite-precision controller realisations, a computationally tractable FWL closed-loop stability measure was then introduced and the method of computing the value of this measure for a given controller realisation was given. For each arithmetic scheme, the optimal controller realisation is defined as the solution that maximises the corresponding stability measure, and a numerical optimisation approach is adopted to solve the resultant optimal realisation problem. Simulation results have confirmed that the optimal floating-point controller realisation implemented in floating-point format is the best in terms of robustness to FWL errors. The results have also shown that with the aid of dividing the controller coefficients into blocks, block-floating-point realisations can have better robustness to FWL errors than fixed-point ones but choosing an appropriate division of blocks is difficult in practice. These results agree with the common understanding of the number representation formats. It is well known that the fixed-point format is the best in terms of hardware cost, arithmetic operation simplicity and execution speed, while the floating-point format is the worst in terms of these aspects. The proposed design procedure provides the designer with useful quantitative information regarding finite precision computational properties, namely robustness to FWL errors and the estimated minimum word length required for guaranteeing closed-loop stability. This allows the designer to choose an optimal controller realisation in an appropriate representation scheme to achieve the best possible computational efficiency and closed-loop controller performance.

8.4 Roundoff Noise Minimisation for Fixed-Point Implementation

The rounding errors in arithmetic operations will cause the true closed-loop performance to deviate from the designed performance. In the journal paper **[97]**, the effect of roundoff noise in a digital controller was analysed in the context of the generic sampled-data system and the optimal digital controller structure problem was formulated and solved for the shift-operator parameterised controller with roundoff noise consideration. The contribution of the paper **[97]** are threefold. The first one is to give a thorough analysis of the effect of roundoff noise in the digital controller on the output of the sampled-data system. Based on this analysis, a new measure, called the averaged roundoff noise gain, was proposed. This novel measure, unlike the existing ones, was derived for the generic hybrid system rather than its discrete-time counterpart and hence can take the inter-sample behaviour into account. The second contribution is to derive an efficient method of evaluating this measure by fast sampling plant, which can avoid the numerical problems involved in direct computation of the newly defined measure. The exact expression for the covariance matrix of the controller state vector was also derived in order to scale the realisations with the l_2 norm and hence to prevent the signals in the controller from overflow. It was shown that the proposed new measure is controller realisation dependent, and the third contribution of this paper is to present an analytical solution to the optimal controller structure problem, which identifies those controller realisations that minimise the averaged roundoff noise gain subject to the l_2-scaling constraint. Both theoretical analysis and simulation results show that the optimal controller realisations obtained with the aid of the proposed approach are superior to those obtained using the traditional roundoff-noise analysis based on a digital control system.

The optimal shift-operator based controller realisation that minimises the averaged roundoff noise gain subject to the l_2-scaling constraint is fully parameterised. In practice, a sparse controller realisation is preferred. It is also highly desired to analyse the effect of roundoff noise in a delta-operator parameterised controller. The significance of my journal paper **[98]** is to propose a generic framework of designing sparse

controller realisations with small roundoff noise based on a polynomial-operator parameterisation of the controller. This polynomial-operator realisation is a generalisation of the direct forms in the shift and delta operators. Specifically, in **[98]**, a new implementation model of a state-space controller realisation was proposed, where each coefficient matrix of the realisation is separated into a trivial part, which only contains elements from $\{-1, 0, 1\}$ and hence causes no rounding error, as well as a non-trivial part. Secondly, based on this proposed model, we analysed the output deviation of the closed-loop system due to the roundoff noise in the digital controller. An analytical expression was obtained for the roundoff noise gain. The problem of identifying the optimal realisations in the shift operator was already solved in **[97]**. The third contribution, which is the most interesting one, was to derive a new sparse controller realisation based on the polynomial-operator parameterisation. This polynomial-operator parameterisation provides the designer with more degrees of freedom to reduce the roundoff noise than the shift or delta operator parameterisation. The problem of finding optimal polynomial operators can be solved with the aid of a simple exhaustive search. In the design example given in **[98]**, the roundoff noise gain produced by the proposed sparse realisation in conjunction with the optimal polynomial operators is much smaller than that of the optimal fully parameterised realisation in the shift operator. Moreover, the proposed sparse realization yields a roundoff noise which is only half of that produced by the l_2-scaled sparse canonical controller realisation in the delta operator. It also has a simpler implementation than this delta-operator based realisation.

9 Evolutionary Computation and Optimisation

Many nonlinear machine learning applications involve nonlinear optimisation problems with non-smooth and/or non-convex optimisation criteria. It is therefore not surprising that evolutionary computation and optimisation methods have found wide applications in nonlinear learning. This chapter collects several my applications of using evolutionary computation and optimisation methods to solve complex nonlinear learning problems [80],[81],[83]-[85],[99]-[103]. More specifically, the family of global optimisation algorithms, such as genetic algorithms (GAs) and adaptive simulated annealing (ASA), have been adopted to optimise nonlinear learning machines in a variety of applications. The GA and ASA belong to a class of so-called guided random search methods. The underlying mechanisms used for guiding the optimisation search process are, however, very different for the two methods. The class of GAs is population based, and evolves a solution population according to the principles of the evolution of species in nature. It is by far the best known and most widely applied global optimisation scheme in machine learning and engineering applications. The ASA by contrast evolves a single solution in the parameter space with the aid of certain guiding principles that imitate the random behaviour of molecules during the annealing process. Unlike the conventional simulated annealing, the ASA adopts an important mechanism called the reannealing scheme, which not only speeds up the search process but also makes the optimisation process robust to the requirement of different problems. One of my significant contributions [100]-[103] was to help popularising the ASA by demonstrating that it offers a viable alternative to the GA in various applications.

9.1 Establish Efficient Walking Gaits for Legged Robots Using Genetic Algorithms

Legged robots present significant advantages over wheeled or tracked mechanisms due to their ability to move in very rough and unstructured terrains and to step over obstacles. However, without efficient walking

strategies these advantages cannot be realised. Robots that stumble and bump into obstacles are classic examples of robots with inefficient walking strategies – with each terrain contact there is a loss of energy and accelerated wear and tear on the vehicle, resulting in a reduction of the expected lifespan. In the design and development of a legged robot, many factors have to be considered. As a consequence, creating a legged robot that can efficiently and autonomously negotiate a wide range of terrains is a challenging task. When designing a legged robot it is often useful to consider biological walking systems, which tend to be much more versatile and seem to be more effective and elegant, in order to emulate these or similar mechanisms in the design. Indeed, many researchers working in the area of legged robotics have traditionally looked towards the natural world for inspiration and solutions, reasoning that these evolutionary solutions are appropriate and effective because they have passed the hard tests for survival over time and generations. The journal paper **[99]** reported the research conducted in using a method of "natural evolution emulation" to improve the mechanisms used for controlling the stepping sequence of a legged robot.

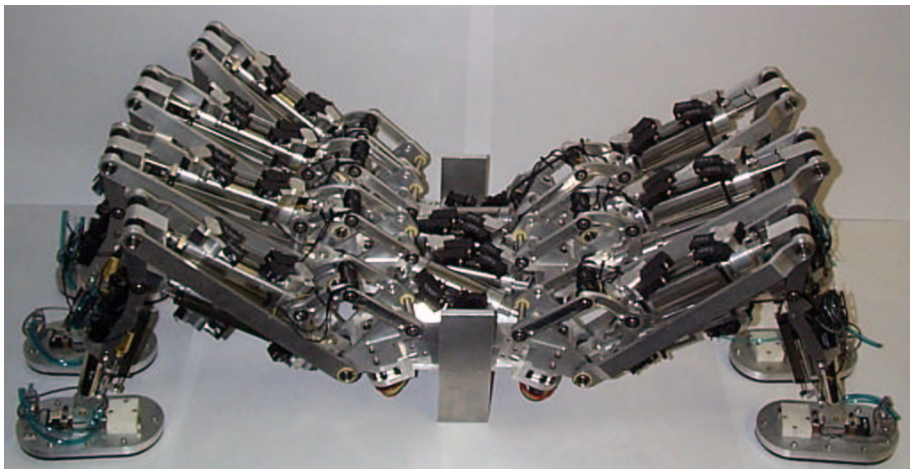

Figure 2: Robug IV. A fully extended robot leg has an approximate length of 1 m.

In **[99]**, it was shown that by employing GA-based optimisation within the walking system of a legged robot, favourable gait behaviour can be achieved. By analysing these GA-derived solutions it is then possible to learn ideal behaviour patterns and gain valuable knowledge for updating the control mechanisms or, alternatively, for improving future robot designs. The research work was conducted on an 8-legged robot called Robug IV. A picture of Robug IV is shown in Figure 2. Two sets of experimental results reported in **[99]** are worth mentioning. The first test was for the robot walking over flat terrain in normal operating conditions, that is, walking with a laterally (sideways) directed body velocity. The second test was identical to the first test except that one leg was made inoperative to simulate a mechanical breakage, thus rendering the leg useless. In the first test, the walking gait derived by the GA for the fully operational robot has the characteristics of the walking behaviour of the ghost crab over flat terrain. It is thus interesting to see that the GA can "emulate" natural evolution to arrive at a similar solution. A more practical application of the GA-based design tool is to a situation in which there is no simple way of copying nature, and this was the purpose of the second test. Research data concerning the walking behaviour of crabs with a leg removed is very limited and thus in this situation there is no comprehensive data of a "natural solution" as such. In the second test, it was observed that the GA's evolution successfully adopted the walking gait to avoid potentially very hazardous states of unstable or critical-stability situations.

9.2 Evolutionary Computation for Global Optimisation Applications

The optimisation criteria associated with many nonlinear machine learning problems are typically multimodal and/or non-smooth. Conventional gradient-based algorithms are ineffective in these applications due to the problem of potentially encountering local minima or the difficulty in calculating the associated gradients. Optimisation methods that require no gradient calculation and can achieve a globally optimal solution offer considerable advantages in solving these difficult optimisation problems. The GA and ASA, which solve an optimisation problem by relying only on the cost function value, are known to be capable of finding globally optimal solutions. In my research on nonlinear learning machines, I have frequently used the GA and ASA to solve difficult multimodal and/or non-smooth optimisation problems. Four typical examples are summarised here. The GA is of course well-known to researchers and practicians. In my publications **[100],[101]**, I presented the ASA in a form that is easy to understand and to program. A contribution of my works is therefore to make this potentially powerful global optimisation technique accessible to a wide audience in the engineering community.

9.2.1 Optimising Higher-Order Cumulant Criterion for Blind Channel Identification

An important class of blind identification techniques designed for non-minimum phase channels is based on the higher order cumulant (HOC) fitting method. HOC cost functions are, however, multimodal, and conventional gradient methods may converge to locally optimal or "wrong" solutions. To overcome the problem of local minima, global optimisation methods are highly advantageous. In my journal paper **[80]**, the family of GAs was applied for optimal blind channel identification with HOC fitting. In the publication **[100]**, the ASA was adopted to optimise an HOC fitting cost function used for blind channel identification. The results reported in **[100]** show that the ASA is capable of finding a globally optimal solution for blind channel identification with the aid of HOC fitting. The purpose of the work **[100]** was to demonstrate to the research community that, unlike the conventional simulated annealing, the ASA achieves much faster convergence in the HOC fitting process. It was observed that the ASA achieved a similar convergence speed as the GA during the HOC fitting process. Thus, it offers a viable alternative to GAs in this class of important multimodal optimisation problems.

9.2.2 Optimising Joint Maximum Likelihood Criterion for Blind Equalisation

The joint maximum likelihood (ML) channel estimation and data detection, discussed in Section 7.3, is a highly complex nonlinear optimisation problem. Simply performing an iteration between the least squares estimate for the channel impulse response (CIR) assuming a certain data sequence and the maximum likelihood sequence estimate for the data sequence while assuming a specific CIR will generally be suboptimal. Moreover, the success or failure of such a strategy heavily depends on the initial conditions. The power of the GA in solving this highly complex joint optimisation problem was demonstrated in my journal paper **[81]**. The GA as a global optimiser is capable of efficiently and accurately searching through a fraction of the channel parameter space and yet approaching the ML solution. In the extensive simulation performed in **[81]**, the combined GA used for optimal channel estimate and the Viterbi algorithm (VA) employed for ML data detection always converged to the true joint ML solution found for blind equalisation. My journal

paper **[101]** further investigated an alternative global optimiser, called the ASA, again in the context of this novel blind equalisation application. The results of **[101]** show that the combined ASA used for optimal channel estimate and the VA employed for ML data detection achieves a similar convergence speed as the GA based scheme. The ASA based scheme was shown to be slightly more accurate than the GA aided one. This can be explained by the quantisation effects of the binary encoding scheme used by the GA. In this type of optimisation applications, the ASA has an added advantage as it requires less effort to program, compared to the GA. A motivation of **[101]** was to introduce the ASA algorithm to the signal processing community, which was less familiar with this global optimizer. In **[101]**, the mechanisms of the ASA as well as its detailed implementation issues were presented in a clear and easy-to-understand manner.

9.2.3 Digital IIR Filter Design

Adaptive infinite-impulse-response (IIR) filtering has been an active area of research for many years, and many properties of IIR filters are well known. A major concern in IIR filtering applications is that the cost function of IIR filters is generally multi-modal with respect to the filter coefficients, and the usual gradient-based algorithm can easily be trapped at local minima. In order to arrive at a global minimum solution, global optimisation techniques are needed. Global optimisation methods, however, require extensive computations and are usually batch-type algorithms, as the cost function employed must be evaluated for a block of data. In contrast, gradient-based learning can be recursively implemented to update the filter coefficients as each new data sample is acquired. Despite of these drawbacks, applying global optimisation methods to IIR filter design is attractive, since in many practical applications a global optimal solution can be much better than local optimal ones. When considering global optimisation methods contrived for digital IIR filter design, the GA seems to have attracted considerable attention, and most of the publications in this area are concerned with the applications of the GA. My journal papers **[101],[102]** are the first publications in the literature to apply the ASA algorithm for digital IIR filter design. The results reported in **[101],[102]** demonstrate that the efficiency of the ASA appears to be similar to that of the GA when used for digital IIR filter design, expressed in terms of the total number of cost function evaluations required to achieve a global optimal solution. This suggests that the ASA offers a viable alternative to digital IIR filter design. Adopting a moving window strategy, a batch-recursive version of the ASA was proposed for adaptive applications in **[102]**. This constitutes a major contribution to the research and application of the ASA algorithm.

9.2.4 Optimising Stability Bounds for Finite-Precision Controller Realisations

In order to optimise controller integrity for finite word length (FWL) fixed-point implementations, a particular approach discussed in Section 8.1 was based on closed-loop pole sensitivity analysis **[103]**. In this pole sensitivity analysis approach, optimal controller realisations were chosen by maximising various measures of the closed-loop stability margin. The closed-loop stability margin maximisation problem is evidently non-smooth **[83]-[93]**, since the derivatives of its associated optimisation criteria may have discontinuities. Thus, the optimisation must be based on a direct search using cost function values without the aid of cost function derivatives. Conventional optimisation techniques, such as the Rosenbrock and simplex algorithms, solve an optimisation problem by relying exclusively on cost function values and, therefore, can be adopted for solving optimal FWL controller realisation problems. These conventional optimisation methods are

however only local optimisers. The cost function associated with an optimal FWL controller realisation problem can be severely non-convex. Thus, the results obtained by using a conventional optimisation algorithm heavily depend on the initial conditions. It has been observed that the optimisation process sometimes converges to unsatisfactory local solutions. In my research of finite-precision digital controller designs, I have opted for using the ASA algorithm to solve the highly non-convex and non-smooth optimisation problems associated with closed-loop stability margin maximisation **[84]-[86],[88],[90]-[92],[103]**. The ASA is a global optimiser that uses the cost function value only in performing optimisation search. It is efficient in terms of its convergence speed and is easy to program. Thus, the ASA algorithm is ideal for solving the optimal FWL controller realisation problem. Experience has confirmed that the ASA is able to provide significantly better results than those typically obtained by conventional optimisation techniques.

10 Conclusions and Future Research

In this DSc submission my research conducted at the University of Southampton during the past six years plus some of my significant contributions before 1999 were summarised. The theme of the submission is **intelligent nonlinear learning machines**. From the more than 200 research papers that I co-authored during my 18-year career I have opted for using 103 journal papers and book contributions **[1]-[103]**, which were best aligned with the main theme of the submission. Over half of the 103 research papers cited in the submission were published after I joined the University of Southampton. My full list of publications is also attached.

As mentioned in Chapter 1, the eight topics, covered in Chapters 2 to 9 of the submission, can be grouped into four broad subjects. The related conclusions and some future research directions are now briefly summarised for each of these four subjects.

10.1 Computational Intelligence and Machine Learning

The learning algorithms presented in Chapters 2 to 4 are significant because they represents state-of-the-art techniques for modelling from data. These nonlinear learning machines are *intelligent*, because they are capable of constructing sparse representations with excellent generalisation ability in a computationally efficient way. An interesting future research objective that may be used for further improving these nonlinear learning machines is to develop a generic "kernel hunting" technique. More specifically, rather than employing a fixed common kernel variance for every kernel regressor and positioning the kernel regressors on the training input data points, as most of the kernel modelling methods do, kernel hunting aims to position kernel regressors at the most appropriate positions and to shape each kernel regressor by adjusting its individual kernel covariance matrix. Such an approach can be expected to produce sparser models with improved modelling capability and generalisation performance. To maintain computational efficiency, the well tested forward selection strategy coupled with orthogonalisation can be incorporated into this kernel hunting approach. During the writing of this submission, work has already begun along this direction, and some interesting initial results have been obtained, which have been submitted for publication:

S. Chen, X.X. Wang and D.J. Brown, "Sparse incremental regression modelling using correlation criterion

with boosting search," *IEEE Signal Processing Letters*, to appear, 2005.

S. Chen, X.X. Wang, X. Hong and C.J. Harris, "Kernel classifier construction using orthogonal forward selection and boosting with Fisher ratio class separability measure,"submitted to *IEEE Trans. Neural Networks*, 2004.

10.2 Adaptive Receiver Design For Future Communications Systems

During the past five years, the minimum bit error rate (MBER) principle has evolved from merely an alternative to the standard minimum mean square error (MMSE) design for single-input single-output (SISO) channel equalisation to become a universal adaptive linear receiver design method for various advanced communication systems, such as multiuser detection in code-division multiple-access (CDMA) and multiple antenna aided space-division multiple-access (SDMA) systems. My research results reported in Chapter 5 of the submission clearly demonstrate that the MBER design offers significant system performance gains over the MMSE benchmarker in generic applications. It has also been shown that the adaptive MBER receiver is robust to adverse channel conditions. Moreover, the concept of MBER design is having an impact on our understanding of the fundamental system capacity. For example, for SDMA systems a practical rule-of-thumb is that to support N users one needs at least N receive antennas. My research however has demonstrated that the MBER receiver is capable of supporting more users than the number of antennas. Further research will couple the adaptive MBER receiver with channel coding to form a powerful iterative or turbo receiver scheme. Another possible future research direction is to unify our adaptive MBER receiver with the MBER transmitter design currently pursued by other researchers.

My research on adaptive nonlinear receiver design, reported in Chapter 6 of the submission, has clearly demonstrated that such an adaptive nonlinear receiver is capable of meeting the stringent requirements of the third generation and future communication systems. The key in practical adaptive nonlinear receiver design is the computational affordability. Simply "expanding" an optimal nonlinear receiver design to multiuser detection can easily become computationally intractable as the number of users increases. An interesting solution, which is currently researched in the Communications Research Group at Southampton, is to apply iterative or turbo space-time nonlinear equalisation principle to form **affordable** and (near) **optimal** adaptive nonlinear multiuser detection for the generic SDMA system. The front end of the receiver consists of a set of parallel interference cancellers (PICs). Each PIC is followed by a single-user adaptive radial basis function (RBF) detector equipped with a single-user channel estimator. The detected user bits after channel decoding are re-modulated and fed back to both the RBF detectors and PICs.

10.3 Finite-Precision Digital Controller Design

My significant contribution to finite-precision digital controller design has been to present a unified framework for optimising controller integrity in finite word length (FWL) implementations. Currently, a generic optimal FWL controller realisation is obtained through numerical optimisation. For practical purpose, it is highly desired that a solution can be obtained via a closed-form design rather than a numerical one. In fact, an engineer probably would favour a suboptimal closed-form solution to an optimal solution that has to be solved via some difficult numerical optimisation. Future research will focus on deriving easy-to-compute,

preferably closed-form designs for optimal FWL implementation. During the writing of this submission, some interesting initial results have been obtained. For fixed-point implementation and for the pole sensitivity analysis based on the l_2-norm stability measure, we have derived the optimal closed-form controller realisation set, and the results have been submitted for publication:

- J. Wu, S. Chen, G. Li and J. Chu, "A search algorithm for a class of optimal finite-precision controller realization problems with saddle points," submitted to *SIAM J. Control and Optimization*, 2004.

Since the optimal controller realisations that maximise the given stability measure form a closed-form set, the redundancy in this set can be exploited to further minimise the dynamic range of the controller realisation, to design sparse realisations and to minimise the roundoff noise gain of the controller realisation, without reducing the closed-loop stability robustness.

10.4 Evolutionary Computation and Global Optimisation

In my research on nonlinear learning machines, I have found that evolutionary computation methods constitute valuable tool kits. Specifically, I have used genetic algorithms (GAs) in generic learning machine designs, and applied the global optimisation methods, such as the GA and adaptive simulated annealing (ASA), in optimising key learning parameters as well as in solving the optimisation problems associated with some machine learning applications. Using evolutionary computation methods in machine learning will remain a focus of my research. I have always been fascinated by the topic of optimisation in engineering. During the writing of this DSc submission, I have developed a simple yet very efficient guided random search algorithm for global optimisation applications, and the results have been submitted for publication:

- S. Chen, X.X. Wang and C.J. Harris, "Experiments with repeating weighted boosting search for optimization in signal processing applications," *IEEE Trans. Systems, Man and Cybernetics, Part B*, to appear, 2005.

References

[1] S. Chen, E.S. Chng and K. Alkadhimi, "Regularised orthogonal least squares algorithm for constructing radial basis function networks," *Int. J. Control*, Vol.64, No.5, pp.829–837, 1996.

[2] S. Chen, Y. Wu and B.L. Luk, "Combined genetic algorithm optimisation and regularised orthogonal least squares learning for radial basis function networks," *IEEE Trans. Neural Networks*, Vol.10, No.5, pp.1239–1243, 1999.

[3] S. Chen, "Multi-output regression using a locally regularised orthogonal least square algorithm," *IEE Proc. Vision, Image and Signal Processing*, Vol.149, No.4, pp.185–195, 2002.

[4] S. Chen, X. Hong and C.J. Harris, "Sparse kernel regression modelling using combined locally regularized orthogonal least squares and D-optimality experimental design," *IEEE Trans. Automatic Control*, Vol.48, No.6, pp.1029–1036, 2003.

[5] S. Chen, X. Hong and C.J. Harris, "Sparse multi-output radial basis function network construction using combined locally regularised orthogonal least square and D-Optimality experimental design," *IEE Proc. Control Theory and Applications*, Vol.150, No.2, pp.139–146, 2003.

[6] X. Hong, C.J. Harris, S. Chen and P.M. Sharkey, "Robust nonlinear model identification methods using forward regression," *IEEE Trans. Systems, Man and Cybernetics, Part A*, Vol.33, No.4, pp.514–523, 2003.

[7] S. Chen, X. Hong, C.J. Harris and P.M. Sharkey, "Sparse modeling using orthogonal forward regression with PRESS statistic and regularization," *IEEE Trans. Systems, Man and Cybernetics, Part B*, Vol.34, No.2, pp.898–911, 2004.

[8] X. Hong, S. Chen and P.M. Sharkey, "Automatic kernel regression modelling using combined leave-one-out test score and regularised orthogonal least squares," *Int. J. Neural Systems*, Vol.14, No.1, pp.27–37, 2004.

[9] X. Hong, M. Brown, S. Chen and C.J. Harris, "Sparse model identification using orthogonal forward regression with basis pursuit and D-optimality," *IEE Proc. Control Theory and Applications*, Vol.151, No.4, pp.491–498, 2004.

[10] X. Hong, C.J. Harris, M. Brown and S. Chen, "Backward elimination methods for associative memory network pruning," *Int. J. Hybrid Intelligent Systems*, Vol.1, No.2, pp.90–98, 2004.

[11] X. Hong, C.J. Harris and S. Chen, "Robust neurofuzzy rule base knowledge extraction and estimation using subspace decomposition combined with regularization and D-optimality," *IEEE Trans. Systems, Man and Cybernetics, Part B*, Vol.34, No.1, pp.598–608, 2004.

[12] S. Chen, "Nonlinear time series modelling and prediction using Gaussian RBF networks with enhanced clustering and RLS learning," *Electronics Letters*, Vol.31, No.2, pp.117–118, 1995.

[13] E.S. Chng, S. Chen and B. Mulgrew, "Gradient radial basis function network for nonlinear and non-stationary time series prediction," *IEEE Trans. Neural Networks*, Vol.7, No.1, pp.190–194, 1996.

[14] S.A. Billings and S. Chen, "The determination of multivariable nonlinear models for dynamic systems," in: C.T. Leondes, Ed. *Control and Dynamic Systems*, Volume 7 of *Neural Network Systems Techniques and Applications*, San Diego: Academic Press, 1998, pp.231–278.

[15] S. Chen, Z. He and P.M. Grant, "Artificial neural network visual model for image coder enhancement," *Neurocomputing*, Vol.30, No.1-4, pp.339–346, 2000.

[16] S. Chen, Z. He and B.L. Luk, "A generic postprocessing technique for image compression," *IEEE Trans. Circuits and Systems for Video Technology*, Vol.11, No.4, pp.546–553, 2001.

[17] S. Urwin-Wright, D. Sanders and S. Chen, "Terrain prediction for an eight legged robot," *J. Robotic Systems*, Vol.19, No.2, pp.91–98, 2002.

[18] S. Urwin-Wright, D. Sanders and S. Chen, "Predicting terrain contours using a feed-forward neural network," *Engineering Applications of Artificial Intelligence*, Vol.16, No.5/6, pp.465-472, 2003.

[19] E.S. Chng, S. Chen and B. Mulgrew, "Efficient computational schemes for the orthogonal least squares algorithm," *IEEE Trans. Signal Processing*, Vol.43, No.1, pp.373–376, 1995.

[20] S. Chen and J. Wigger, "Fast orthogonal least squares algorithm for efficient subset model selection," *IEEE Trans. Signal Processing*, Vol.43, No.7, pp.1713–1715, 1995.

[21] P.E. An, M. Brown, S. Chen and C.J. Harris, "Comparative aspects of neural network algorithms for on-line modelling of dynamic processes," *Proc. I. MECH. E., Pt.I, J. Systems and Control Eng.*, Vol.207, pp.223–241, 1993.

[22] S. Chen and S.A. Billings, "Neural networks for non-linear dynamic system modelling and identification," *Int. J. Control*, Vol.56, No.2, pp.319–346, 1992.

[23] S. Chen, P.M. Grant and C.F.N. Cowan, "Orthogonal least squares algorithm for training multi-output radial basis function networks," *IEE Proc. Part F*, Vol.139, No.6, pp.378–384, 1992.

[24] S.A. Billings, H.B. Jamaluddin and S. Chen, "Properties of neural networks with applications to modelling nonlinear dynamical systems," *Int. J. Control*, Vol.55, No.1, pp.193–224, 1992.

[25] S. Chen, S.A. Billings and P.M. Grant, "Recursive hybrid algorithm for non-linear system identification using radial basis function networks," *Int. J. Control*, Vol.55, No.5, pp.1051–1070, 1992.

[26] S. Chen, C.F.N. Cowan and P.M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks*, Vol.2, No.2, pp.302–309, 1991.

[27] S.A. Billings, H.B. Jamaluddin and S. Chen, "A comparison of the backpropagation and recursive prediction error algorithms for training neural networks," *Mechanical Systems and Signal Processing*, Vol.5, No.3, pp.233–255, 1991.

[28] S. Chen, S.A. Billings, C.F.N. Cowan and P.M. Grant, "Non-linear systems identification using radial basis functions," *Int. J. Systems Sci.*, Vol.21, No.12, pp.2513–2539, 1990.

[29] S. Chen, S.A. Billings, C.F.N. Cowan and P.M. Grant, "Practical identification of NARMAX models using radial basis functions," *Int. J. Control*, Vol.52, No.6, pp.1327–1350, 1990.

[30] S. Chen, C.F.N. Cowan, S.A. Billings and P.M. Grant, "Parallel recursive prediction error algorithm for training layered neural networks," *Int. J. Control*, Vol.51, No.6, pp.1215–1228, 1990.

[31] S. Chen, S.A. Billings and P.M. Grant, "Non-linear systems identification using neural networks," *Int. J. Control*, Vol.51, No.6, pp.1191–1214, 1990.

[32] S. Chen and S.A. Billings, "Modelling and analysis of non-linear time series," *Int. J. Control*, Vol.50, No.6, pp.2151–2171, 1989.

[33] S.A. Billings and S. Chen, "Extended model set, global data and threshold model identification of severely non-linear systems," *Int. J. Control*, Vol.50, No.5, pp.1897–1923, 1989.

[34] S. Chen, S.A. Billings and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, Vol.50, No.5, pp.1873–1896, 1989.

[35] S.A. Billings, S. Chen and M.J. Korenberg, "Identification of MIMO non-linear systems using a forward-regression orthogonal estimator," *Int. J. Control*, Vol.49, No.6, pp.2157–2189, 1989.

[36] S. Chen and S.A. Billings, "Representation of non-linear systems: the NARMAX model," *Int. J. Control*, Vol.49, No.3, pp.1013–1032, 1989.

[37] S. Chen and S.A. Billings, "Recursive prediction error estimator for non-linear models," *Int. J. Control*, Vol.49, No.2, pp.569–594, 1989.

[38] S.A. Billings, S. Chen and R.J. Backhouse, "The identification of linear and non-linear models of a turbocharged automotive diesel engine," *Mechanical Systems and Signal Processing*, Vol.3, No.2, pp.123–142, 1989.

[39] S.A. Billings and S. Chen, "Identification of non-linear rational systems using a prediction-error estimation algorithm," *Int. J. Systems Sci.*, Vol.20, No.3, pp.467–494, 1989.

[40] Z.R. Yang and S. Chen, "Robust maximum likelihood training of heteroscedastic probabilistic neural networks," *Neural Networks*, Vol.11, No.4, pp.739–747, 1998.

[41] S. Chen, L. Hanzo and A. Wolfgang, "Kernel-based nonlinear beamforming construction using orthogonal forward selection with Fisher ratio class separability measure," *IEEE Signal Processing Letters*, Vol.11, No.5, pp.478–481, 2004.

[42] S. Chen, L. Hanzo and A. Wolfgang, "Nonlinear multiantenna detection methods," *EURASIP J. Applied Signal Processing*, Vol.2004, No.9, pp.1225–1237, 2004.

[43] S. Chen, B. Mulgrew and L. Hanzo, "Least bit error rate adaptive nonlinear equalizers for binary signalling," *IEE Proc. Communications*, Vol.150, No.1, pp.29–36, 2003.

[44] S. Chen, X. Hong and C.J. Harris, "Sparse kernel density construction using orthogonal forward regression with leave-one-out test score and local regularization," *IEEE Trans. Systems, Man and Cybernetics, Part B*, Vol.34, No.4, pp.1708–1717, 2004.

[45] S. Chen, B. Mulgrew, E.S. Chng and G. Gibson, "Space translation properties and the minimum-BER linear-combiner DFE," *IEE Proc. Communications*, Vol.145, No.5, pp.316–322, 1998.

[46] S. Chen and B. Mulgrew, "The minimum-SER linear-combiner decision feedback equaliser," *IEE Proc. Communications*, Vol.146, No.6, pp.347–353, 1999.

[47] S. Chen, S. Gunn and C.J. Harris, "Decision feedback equalizer design using support vector machines," *IEE Proc. Vision, Image and Signal Processing*, Vol.147, No.3, pp.213–219, 2000.

[48] S. Chen, A.K. Samingan and L. Hanzo, "Optimal decision feedback equalizer for M-PAM signals using a support vector machine solution," *Electronics Letters*, Vol.36, No.20, pp.1742–1744, 2000.

[49] B. Mulgrew and S. Chen, "Adaptive minimum-BER decision feedback equalisers for binary signalling," *Signal Processing*, Vol.81, No.7, pp.1479–1489, 2001.

[50] S. Chen, "Adaptive minimum bit-error-rate filtering," *IEE Proc. Vision, Image and Signal Processing*, Special Issue on Non-linear and Non-Gaussian Signal Processing, Vol.151, No.1, pp.76–85, 2004.

[51] S. Chen, L. Hanzo and B. Mulgrew, "Adaptive minimum symbol-error-rate decision feedback equalization for multi-level pulse-amplitude modulation," *IEEE Trans. Signal Processing*, Vol.52, No.7, pp.2092–2101, 2004.

[52] S. Chen, A.K. Samingan, B. Mulgrew and L. Hanzo, "Adaptive minimum-BER linear multiuser detection for DS-CDMA signals in multipath channels," *IEEE Trans. Signal Processing*, Vol.49, No.6, pp.1240–1247, 2001.

[53] A.K. Samingan, S. Chen and L. Hanzo, "Adaptive minimum-BER linear multiuser detection for CDMA signals in multipath channels with 4-QAM constellation," *Electronics Letters*, Vol.37, No.11, pp.721–723, 2001.

[54] S. Chen, "Least bit error rate adaptive multiuser detection," in: L.P. Wang, Ed. *Soft Computing in Communications*, Berlin: Springer Verlag, 2003, pp.389–408.

[55] M.Y. Alias, A.K. Samingan, S. Chen and L. Hanzo, "Multiple antenna aided OFDM employing minimum bit error rate multiuser detection," *Electronics Letters*, Vol.39, No.24, pp.1769–1770, 2003.

[56] A. Wolfgang, N.N. Ahmad, S. Chen and L. Hanzo, "Genetic algorithm assisted error probability optimisation for beamforming," *Electronics Letters*, Vol.40, No.5, pp.320–322, 2004.

[57] S. Chen, B. Mulgrew and P.M. Grant, "A clustering technique for digital communications channel equalisation using radial basis function networks," *IEEE Trans. Neural Networks*, Vol.4, No.4, pp.570–579, 1993.

[58] S. Chen, B. Mulgrew and S. McLaughlin, "Adaptive Bayesian equaliser with decision feedback," *IEEE Trans. Signal Processing*, Vol.41, No.9, pp.2918–2927, 1993.

[59] S. Chen, S. McLaughlin and B. Mulgrew, "Complex-valued radial basis function network, Part I: network architecture and learning algorithms," *Signal Processing*, Vol.35, No.1, pp.19–31, 1994.

[60] S. Chen, S. McLaughlin and B. Mulgrew, "Complex-valued radial basis function network, Part II: application to digital communications channel equalisation," *Signal Processing*, Vol.36, pp.175–188, 1994.

[61] S. Chen, S. McLaughlin, B. Mulgrew and P.M. Grant, "Adaptive Bayesian decision feedback equaliser for dispersive mobile radio channels," *IEEE Trans. Communications*, Vol.43, No.5, pp.1937–1946, 1995.

[62] S. Chen, "Adaptive equalisation using neural networks," in: A.F. Murray, Ed. *Applications of Neural Networks*, Dordrecht: Kluwer, 1995, pp.241-265.

[63] S. Chen, S.R. Gunn and C.J. Harris, "The relevance vector machine technique for channel equalization application," *IEEE Trans. Neural Networks*, Vol.12, No.6, pp.1529–1532, 2001.

[64] S. Chen and B. Mulgrew, "Overcoming co-channel interference using an adaptive radial basis function equaliser," *Signal Processing*, Vol.28, No.1, pp.91–107, 1992.

[65] S. Chen, S. McLaughlin, B. Mulgrew and P.M. Grant, "Bayesian decision feedback equaliser for overcoming co-channel interference," *IEE Proc. Communications*, Vol.143, No.4, pp.219–225, 1996.

[66] S. Chen, A.K. Samingan and L. Hanzo, "Support vector machine multiuser receiver for DS-CDMA signals in multipath channels," *IEEE Trans. Neural Networks*, Vol.12, No.3, pp.604-611, 2001.

[67] A. Wolfgang, S. Chen and L. Hanzo, "Radial basis function network assisted space-time equalization for dispersive fading environments," *Electronics Letters*, Vol.40, No.16, pp.1006–1007, 2004.

[68] P.M. Grant, S. Chen and B. Mulgrew, "Nonlinear adaptive filter performance in typical applications," *A. Rev. Control*, Vol.20, No.0, pp.107–118, 1996.

[69] S. Chen, G.J. Gibson, C.F.N. Cowan and P.M. Grant, "Reconstruction of binary signals using an adaptive radial-basis-function equalizer," *Signal Processing*, Vol.22, No.1, pp.77–93, 1991.

[70] S. Chen, G.J. Gibson and C.F.N. Cowan, "Adaptive channel equalization using a polynomial-perceptron structure," *IEE Proc. Part I*, Vol.137, No.5, pp.257–264, 1990.

[71] S. Chen, G.J. Gibson, C.F.N. Cowan and P.M. Grant, "Adaptive equalization of finite non-linear channels using multilayer perceptrons," *Signal Processing*, Vol.20, No.2, pp.107–119, 1990.

[72] S. Chen, "Importance sampling simulation for evaluating the lower-bound BER of the Bayesian DFE," *IEEE Trans. Communications*, Vol.50, No.2, pp.179–182, 2002.

[73] S. Chen, "Importance sampling simulation for evaluating lower-bound symbol error rate of the Bayesian DFE with multi-level signalling schemes," *IEEE Trans. Signal Processing*, Vol.50, No.5, pp.1229–1236, 2002.

[74] S. Chen, B. Mulgrew and L. Hanzo, "Asymptotic Bayesian decision feedback equalizer using a set of hyperplanes," *IEEE Trans. Signal Processing*, Vol.48, No.12, pp.3493–3500, 2000.

[75] S. Chen, L. Hanzo and B. Mulgrew, "Decision feedback equalization using multiple-hyperplane partitioning for detecting ISI-corrupted M-ary PAM signals," *IEEE Trans. Communications*, Vol.49, No.5, pp.760–764, 2001.

[76] S. Chen and L. Hanzo, "Importance sampling simulation and multiple-hyperplane realization of the Bayesian decision feedback equaliser," in: J.G. McWhirter and I.K. Proudler, Eds. *Mathematics in Signal Processing V*, Oxford, UK: Oxford University Press, 2002, pp.157–167.

[77] S. Chen, S. McLaughlin, P.M. Grant and B. Mulgrew, "Multi-stage blind clustering equaliser," *IEEE Trans. Communications*, Vol.43, No.3, pp.701–705, 1995.

[78] S. Chen, T.B. Cook and L.C. Anderson, "A comparative study of two blind FIR equalizers," *Digital Signal Processing*, Vol.14, No.1, pp.18–36, 2004.

[79] S. Chen, "Low complexity concurrent constant modulus algorithm and soft decision directed scheme for blind equalisation," *IEE Proceedings – Vision, Image and Signal Processing*, Vol.150, No.5, pp.312–320, 2003.

[80] S. Chen, Y. Wu and S. McLaughlin, "Genetic algorithm optimisation for blind channel identification with higher-order cumulant fitting," *IEEE Trans. Evolutionary Computation*, Vol.1, No.4, pp.259–265, 1997.

[81] S. Chen and Y. Wu, "Maximum likelihood joint channel and data estimation using genetic algorithms," *IEEE Trans. Signal Processing*, Vol.46, No.5, pp.1469–1473, 1998.

[82] S. Chen, S. McLaughlin, P.M. Grant and B. Mulgrew, "Fast blind equalisation based on a Bayesian decision feedback equaliser" *Electronics Letters*, Vol.29, No.10, pp.891–893, 1993.

[83] J. Wu, R.H. Istepanian and S. Chen, "Stability issues of finite precision controller structures for sampled data systems," *Int. J. Control*, Vol.72, No.15, pp.1331–1342, 1999.

[84] S. Chen, J. Wu, R.H. Istepanian and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures," *IEEE Trans. Automatic Control*, Vol.44, No.11, pp.2149–2153, 1999.

[85] R.H. Istepanian, S. Chen, J. Wu and J.F. Whidborne, "Optimal finite-precision controller realization of sample data systems," *Int. J. Systems Science*, Vol.31, No.4, pp.429–438, 2000.

[86] J. Wu, S. Chen, G. Li and J. Chu, "Optimal finite-precision state-estimate feedback controller realization of discrete-time systems," *IEEE Trans. Automatic Control*, Vol.45, No.8, pp.1550–1554, 2000.

[87] J. Wu, S. Chen, G. Li and J. Chu, "Finite world length implementation for digital reduced order observer based controllers," *Developments in Chemical Engineering and Mineral Processing*, Vol.9, No.1/2, pp.41–48, 2001.

[88] 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. Automatic Control*, Vol.46, No.7, pp.1162–1166, 2001.

[89] S. Chen, J. Wu and G. Li, "Two approaches based on pole sensitivity and stability radius measures for finite precision digital controller realizations," *Systems and Control Letters*, Vol.45, No.4, pp.321–329, 2002.

[90] S. Chen, J. Wu, R.H. Istepanian, J. Chu and J.F. Whidborne, "Optimizing stability bounds of finite-precision controller structures for sampled-data systems in the delta operator domain," *IEE Proc. Control Theory and Applications*, Vol.146, No.6, pp.517–526, 1999.

[91] J. Wu, S. Chen, G. Li, R.H. Istepanian and J. Chu, "Shift and delta operator realizations for digital controllers with finite-word-length considerations," *IEE Proc. Control Theory and Applications*, Vol.147, No.6, pp.664–672, 2000.

[92] 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, No.3, pp.153–163, 2000.

[93] J. Wu, S. Chen, G. Li and J. Chu, "Constructing sparse realizations of finite-precision digital controllers based on a closed-loop stability related measure," *IEE Proc. Control Theory and Applications*, Vol.150, No.1, pp.61–68, 2003.

[94] J. Wu, S. Chen, J.F. Whidborne and J. Chu, "Optimal realizations of floating-point implemented digital controllers with finite word length considerations," *Int. J. Control*, Vol.77, No.5, pp.427–440, 2004.

[95] J. Wu and S. Chen, "Stable controller coefficient perturbation in floating point implementation," in: V.D. Blondel and A. Megretsik, Eds. *Unsolved Problems in Mathematical Systems and Control Theory*, Princeton University Press, 2004, pp280–283.

[96] 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. Automatic Control*, Vol.48, No.5, pp.816–822, 2003.

[97] 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 and Applications*, Vol.149, No.3, pp.247–255, 2002.

[98] G. Li, J. Wu and S. Chen, "Sparse controller realization with small roundoff noise," *IEE Proc. Control Theory and Applications*, Vol.151, No.2, pp.246–251, 2004.

[99] B.L. Luk, S. Galt and S. Chen, "Using genetic algorithms to establish efficient walking gaits for an eight-legged robot," *Int. J. Systems Science*, Vol.32, No.6, pp.703–713, 2001.

[100] S. Chen, B.L. Luk and Y. Liu, "Application of adaptive simulated annealing to blind channel identification with HOC fitting," *Electronics Letters*, Vol.34, No.3, pp.234–235, 1998.

[101] S. Chen and B.L. Luk, "Adaptive simulated annealing for optimization in signal processing applications," *Signal Processing*, Vol.79, No.1, pp.117–128, 1999.

[102] S. Chen, R.H. Istepanian and B.L. Luk, "Digital IIR filter design using adaptive simulated annealing," *Digital Signal Processing*, Vol.11, No.3, pp.241–251, 2001.

[103] S. Chen and J. Wu, "The determination of optimal finite precision controller realizations using a global optimization strategy: a pole sensitivity approach," chapter 6 in: R.S.H. Istepanian and J.F. Whidborne, Eds. *Digital Controller Implementation and Fragility: A Modern Perspective*, London: Springer Verlag, 2001, pp.87–104.