

RECURRENT NEURAL NETWORKS FOR PREDICTION

**LEARNING ALGORITHMS,
ARCHITECTURES AND STABILITY**

DANILO P. MANDIC | JONATHON A. CHAMBERS

Contents

Preface	xv
1 Introduction	1
1.1 Some Important Dates in the History of Connectionism	2
1.2 The Structure of Neural Networks	2
1.3 Perspective	4
1.4 Neural Networks for Prediction: Perspective	5
1.5 Structure of the Book	6
1.6 Readership	8
2 Fundamentals	9
2.1 Perspective	9
2.1.1 Chapter Summary	9
2.2 Adaptive Systems	9
2.2.1 Configurations of Adaptive Systems Used in Signal Processing	10
2.2.2 Blind Adaptive Techniques	12
2.3 Gradient-Based Learning Algorithms	12
2.4 A General Class of Learning Algorithms	14
2.4.1 Quasi-Newton Learning Algorithm	15
2.5 A Step-by-Step Derivation of the Least Mean Square (LMS) Algorithm	15
2.5.1 The Wiener Filter	16
2.5.2 Further Perspective on the Least Mean Square (LMS) Algorithm	17
2.6 On Gradient Descent for Nonlinear Structures	18
2.6.1 Extension to a General Neural Network	19
2.7 On Some Important Notions From Learning Theory	19
2.7.1 Relationship Between the Error and the Error Function	19
2.7.2 The Objective Function	20
2.7.3 Types of Learning with Respect to the Training Set and Objective Function	20
2.7.4 Deterministic, Stochastic and Adaptive Learning	21
2.7.5 Constructive Learning	21

2.7.6	Transformation of Input Data, Learning and Dimensionality	22
2.8	Learning Strategies	24
2.9	General Framework for the Training of Recurrent Networks by Gradient-Descent-Based Algorithms	24
2.9.1	Adaptive Versus Nonadaptive Training	24
2.9.2	Performance Criterion, Cost Function, Training Function	25
2.9.3	Recursive Versus Nonrecursive Algorithms	25
2.9.4	Iterative Versus Noniterative Algorithms	25
2.9.5	Supervised Versus Unsupervised Algorithms	25
2.9.6	Pattern Versus Batch Learning	26
2.10	Modularity Within Neural Networks	26
2.11	Summary	29
3	Network Architectures for Prediction	31
3.1	Perspective	31
3.2	Introduction	31
3.3	Overview	32
3.4	Prediction	33
3.5	Building Blocks	35
3.6	Linear Filters	37
3.7	Nonlinear Predictors	39
3.8	Feedforward Neural Networks: Memory Aspects	41
3.9	Recurrent Neural Networks: Local and Global Feedback	43
3.10	State-Space Representation and Canonical Form	44
3.11	Summary	45
4	Activation Functions Used in Neural Networks	47
4.1	Perspective	47
4.2	Introduction	47
4.3	Overview	51
4.4	Neural Networks and Universal Approximation	51
4.5	Other Activation Functions	54
4.6	Implementation Driven Choice of Activation Functions	57
4.7	MLP versus RBF Networks	60
4.8	Complex Activation Functions	60
4.9	Complex Valued Neural Networks as Modular Groups of Compositions of Möbius Transformations	65
4.9.1	Möbius Transformation	65
4.9.2	Activation Functions and Möbius Transformations	65
4.9.3	Existence and Uniqueness of Fixed Points in a Complex Neural Network via Theory of Modular Groups	67
4.10	Summary	68

5	Recurrent Neural Networks Architectures	69
5.1	Perspective	69
5.2	Introduction	69
5.3	Overview	72
5.4	Basic Modes of Modelling	72
5.4.1	Parametric versus Nonparametric Modelling	72
5.4.2	White, Grey and Black Box Modelling	73
5.5	NARMAX Models and Embedding Dimension	74
5.6	How Dynamically Rich are Nonlinear Neural Models?	75
5.6.1	Feedforward versus Recurrent Networks for Nonlinear Modelling	76
5.7	Wiener and Hammerstein Models and Dynamical Neural Networks	77
5.7.1	Overview of Block-Stochastic Models	77
5.7.2	Connection Between Block-Stochastic Models and Neural Networks	78
5.8	Recurrent Neural Network Architectures	81
5.9	Hybrid Neural Network Architectures	84
5.10	Nonlinear ARMA Models and Recurrent Networks	86
5.11	Summary	89
6	Neural Networks as Nonlinear Adaptive Filters	91
6.1	Perspective	91
6.2	Introduction	91
6.3	Overview	92
6.4	Neural Networks and Polynomial Filters	92
6.5	Neural Networks and Nonlinear Adaptive Filters	95
6.6	Training Algorithms for Recurrent Neural Networks	101
6.7	Learning Strategies for a Neural Predictor/Identifier	101
6.7.1	Learning Strategies for a Neural Adaptive Recursive Filter	103
6.7.2	Equation Error Formulation	104
6.7.3	Output Error Formulation	104
6.8	Filter Coefficient Adaptation for IIR Filters	105
6.8.1	Equation Error Coefficient Adaptation	107
6.9	Weight Adaptation for Recurrent Neural Networks	107
6.9.1	Teacher Forcing Learning for a Recurrent Perceptron	108
6.9.2	Training Process for a NARMA Neural Predictor	109
6.10	The Problem of Vanishing Gradients in Training of Recurrent Neural Networks	109
6.11	Learning Strategies in Different Engineering Communities	111
6.12	Learning Algorithms and the Bias/Variance Dilemma	111
6.13	Recursive and Iterative Gradient Estimation Techniques	113
6.14	Exploiting Redundancy in Neural Network Design	113
6.15	Summary	114

7	Stability Issues in RNN Architectures	115
7.1	Perspective	115
7.2	Introduction	115
7.3	Overview	118
7.4	A Fixed Point Interpretation of Convergence in Networks with a Sigmoid Nonlinearity	118
7.4.1	Some Properties of the Logistic Function	118
7.4.2	Logistic Function, Rate of Convergence and Fixed Point Theory	121
7.5	Convergence of Nonlinear Relaxation Equations Realised Through a Recurrent Perceptron	124
7.6	Relaxation in Nonlinear Systems Realised by an RNN	127
7.7	The Iterative Approach and Nesting	130
7.8	Upper Bounds for GAS Relaxation within FCRNNs	133
7.9	Summary	133
8	Data-Reusing Adaptive Learning Algorithms	135
8.1	Perspective	135
8.2	Introduction	135
8.2.1	Towards an <i>A Posteriori</i> Nonlinear Predictor	136
8.2.2	Note on the Computational Complexity	137
8.2.3	Chapter Summary	138
8.3	A Class of Simple <i>A Posteriori</i> Algorithms	138
8.3.1	The Case of a Recurrent Neural Filter	140
8.3.2	The Case of a General Recurrent Neural Network	141
8.3.3	Example for the Logistic Activation Function	141
8.4	An Iterated Data-Reusing Learning Algorithm	142
8.4.1	The Case of a Recurrent Predictor	143
8.5	Convergence of the <i>A Posteriori</i> Approach	143
8.6	<i>A Posteriori</i> Error Gradient Descent Algorithm	144
8.6.1	<i>A Posteriori</i> Error Gradient Algorithm for Recurrent Neural Networks	146
8.7	Experimental Results	146
8.8	Summary	147
9	A Class of Normalised Algorithms for Online Training of Recurrent Neural Networks	149
9.1	Perspective	149
9.2	Introduction	149
9.3	Overview	150
9.4	Derivation of the Normalised Adaptive Learning Rate for a Simple Feedforward Nonlinear Filter	151
9.5	A Normalised Algorithm for Online Adaptation of Recurrent Neural Networks	156
9.6	Summary	160

10	Convergence of Online Learning Algorithms in Neural Networks	161
10.1	Perspective	161
10.2	Introduction	161
10.3	Overview	164
10.4	Convergence Analysis of Online Gradient Descent Algorithms for Recurrent Neural Adaptive Filters	164
10.5	Mean-Squared and Steady-State Mean-Squared Error Convergence	167
10.5.1	Convergence in the Mean Square	168
10.5.2	Steady-State Mean-Squared Error	169
10.6	Summary	169
11	Some Practical Considerations of Predictability and Learning Algorithms for Various Signals	171
11.1	Perspective	171
11.2	Introduction	171
11.2.1	Detecting Nonlinearity in Signals	173
11.3	Overview	174
11.4	Measuring the Quality of Prediction and Detecting Nonlinearity within a Signal	174
11.4.1	Deterministic Versus Stochastic Plots	175
11.4.2	Variance Analysis of Delay Vectors	175
11.4.3	Dynamical Properties of NO ₂ Air Pollutant Time Series	176
11.5	Experiments on Heart Rate Variability	181
11.5.1	Experimental Results	181
11.6	Prediction of the Lorenz Chaotic Series	195
11.7	Bifurcations in Recurrent Neural Networks	197
11.8	Summary	198
12	Exploiting Inherent Relationships Between Parameters in Recurrent Neural Networks	199
12.1	Perspective	199
12.2	Introduction	199
12.3	Overview	204
12.4	Static and Dynamic Equivalence of Two Topologically Identical RNNs	205
12.4.1	Static Equivalence of Two Isomorphic RNNs	205
12.4.2	Dynamic Equivalence of Two Isomorphic RNNs	206
12.5	Extension to a General RTRL Trained RNN	208
12.6	Extension to Other Commonly Used Activation Functions	209
12.7	Extension to Other Commonly Used Learning Algorithms for Recurrent Neural Networks	209
12.7.1	Relationships Between β and η for the Backpropagation Through Time Algorithm	210
12.7.2	Results for the Recurrent Backpropagation Algorithm	211

12.7.3	Results for Algorithms with a Momentum Term	211
12.8	Simulation Results	212
12.9	Summary of Relationships Between β and η for General Recurrent Neural Networks	213
12.10	Relationship Between η and β for Modular Neural Networks: Perspective	214
12.11	Static Equivalence Between an Arbitrary and a Referent Modular Neural Network	214
12.12	Dynamic Equivalence Between an Arbitrary and a Referent Modular Network	215
12.12.1	Dynamic Equivalence for a GD Learning Algorithm	216
12.12.2	Dynamic Equivalence Between Modular Recurrent Neural Networks for the ERLS Learning Algorithm	217
12.12.3	Equivalence Between an Arbitrary and the Referent PRNN	218
12.13	Note on the β - η - W Relationships and Contractivity	218
12.14	Summary	219
Appendix A The \mathcal{O} Notation and Vector and Matrix Differentiation		221
A.1	The \mathcal{O} Notation	221
A.2	Vector and Matrix Differentiation	221
Appendix B Concepts from the Approximation Theory		223
Appendix C Complex Sigmoid Activation Functions, Holomorphic Mappings and Modular Groups		227
C.1	Complex Sigmoid Activation Functions	227
C.1.1	Modular Groups	228
Appendix D Learning Algorithms for RNNs		231
D.1	The RTRL Algorithm	231
D.1.1	Teacher Forcing Modification of the RTRL Algorithm	234
D.2	Gradient Descent Learning Algorithm for the PRNN	234
D.3	The ERLS Algorithm	236
Appendix E Terminology Used in the Field of Neural Networks		239
Appendix F On the <i>A Posteriori</i> Approach in Science and Engineering		241
F.1	History of <i>A Posteriori</i> Techniques	241
F.2	The Usage of <i>A Posteriori</i>	242
F.2.1	<i>A Posteriori</i> Techniques in the RNN Framework	242

F.2.2	The Geometric Interpretation of <i>A Posteriori</i> Error Learning	243
-------	--	-----

Appendix G	Contraction Mapping Theorems	245
-------------------	-------------------------------------	-----

G.1	Fixed Points and Contraction Mapping Theorems	245
G.1.1	Contraction Mapping Theorem in \mathbb{R}	245
G.1.2	Contraction Mapping Theorem in \mathbb{R}^N	246
G.2	Lipschitz Continuity and Contraction Mapping	246
G.3	Historical Perspective	247

Appendix H	Linear GAS Relaxation	251
-------------------	------------------------------	-----

H.1	Relaxation in Linear Systems	251
H.1.1	Stability Result for $\sum_{i=1}^m a_i = 1$	253
H.2	Examples	253

Appendix I	The Main Notions in Stability Theory	263
-------------------	---	-----

Appendix J	Deseasonalising Time Series	265
-------------------	------------------------------------	-----

References	267
-------------------	-----

Index	281
--------------	-----