

PSU LIBRARIES ANNEX



2152 04 04 004



ADVANCES IN
NEURAL
INFORMATION
PROCESSING
SYSTEMS 13

EDITED BY
TODD K. LEEN,
THOMAS G. DIETTERICH,
AND VOLKER TRESP

GOOGLE EXHIBIT 1016
Google v. K.Mizra
IPR2026-00304



EDITORS
LEEN
DIETTERICH
TRESP

ADVANCES IN
NEURAL INFORMATION
PROCESSING SYSTEMS

13

Engin.
QP363
.3
.A43
v.13
2000

The MIT Press
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142
<http://mitpress.mit.edu>



LEEAH 0-262-12241-3

ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 13

ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 13

Proceedings of the 2000 Conference

edited by

Todd K. Leen, Thomas G. Dietterich and Volker Tresp

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

© 2001 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording or information storage and retrieval) without permission in writing from the publisher.

This book was printed and bound in the United States of America.

ISSN: 1049-5258

ISBN: 0-262-12241-3

Cambridge, Massachusetts
London, England

Contents

Preface	xv
NIPS Committees	xvii
Reviewers	xix

Part I Cognitive Science

<i>Who Does What? A Novel Algorithm to Determine Function Localization,</i> Ranit Aharonov-Barki, Isaac Meilijson and Eytan Ruppin	3
<i>A Productive, Systematic Framework for the Representation of Visual Structure,</i> Shimon Edelman and Nathan Intrator	10
<i>The Interplay of Symbolic and Subsymbolic Processes in Anagram Problem Solving,</i> David B. Grimes and Michael C. Mozer	17
<i>Hippocampally-Dependent Consolidation in a Hierarchical Model of Neocortex,</i> Szabolcs Káli and Peter Dayan	24
<i>Position Variance, Recurrence and Perceptual Learning,</i> Zhaoping Li and Peter Dayan	31
<i>The Use of MDL to Select among Computational Models of Cognition,</i> In J. Myung, Mark A. Pitt, Shaobo Zhang and Vijay Balasubramanian	38
<i>Active Inference in Concept Learning,</i> Jonathan D. Nelson and Javier R. Movellan	45
<i>The Early Word Catches the Weights,</i> Mark A. Smith, Garrison W. Cottrell and Karen L. Anderson	52
<i>Structure Learning in Human Causal Induction,</i> Joshua B. Tenenbaum and Thomas L. Griffiths	59
<i>Adaptive Object Representation with Hierarchically-Distributed Memory Sites,</i> Bosco S. Tjan	66

Part II Neuroscience

<i>What Can a Single Neuron Compute?,</i> Blaise Agüera y Arcas, Adrienne L. Fairhall and William Bialek	75
<i>Dendritic Compartmentalization Could Underlie Competition and Attentional Biasing of Simultaneous Visual Stimuli,</i> Kevin A. Archie and Bartlett W. Mel	82
<i>Place Cells and Spatial Navigation Based on 2D Visual Feature Extraction, Path Integration, and Reinforcement Learning,</i> Angelo Arleo, Fabrizio Smeraldi, Stéphane Hug and Wulfram Gerstner	89
<i>Modelling Spatial Recall, Mental Imagery and Neglect,</i> Suzanna Becker and Neil Burgess	96

<i>Stability and Noise in Biochemical Switches</i> , William Bialek	103
<i>Temporally Dependent Plasticity: An Information Theoretic Account</i> , Gal Chechik and Naftali Tishby	110
<i>A New Model of Spatial Representation in Multimodal Brain Areas</i> , Sophie Deneve, Jean-Rene Duhamel and Alexandre Pouget	117
<i>Multiple Timescales of Adaptation in a Neural Code</i> , Adrienne L. Fairhall, Geoffrey D. Lewen, William Bialek and Robert R. de Ruyter van Steveninck	124
<i>Dopamine Bonuses</i> , Sham Kakade and Peter Dayan	131
<i>Finding the Key to a Synapse</i> , Thomas Natschl�ger and Wolfgang Maass	138
<i>Processing of Time Series by Neural Circuits with Biologically Realistic Synaptic Dynamics</i> , Thomas Natschl�ger, Wolfgang Maass, Eduardo D. Sontag and Anthony Zador . . .	145
<i>Spike-Timing-Dependent Learning for Oscillatory Networks</i> , Silvia Scarpetta, Zhaoping Li and John Hertz	152
<i>Universality and Individuality in a Neural Code</i> , Elad Schneidman, Naama Brenner, Naftali Tishby, Robert R. de Ruyter van Steveninck and William Bialek	159
<i>Natural Sound Statistics and Divisive Normalization in the Auditory System</i> , Odelia Schwartz and Eero P. Simoncelli	166
<i>Development of Hybrid Systems: Interfacing a Silicon Neuron to a Leech Heart Interneuron</i> , Mario F. Simoni, Gennady S. Cymbalyuk, Michael Q. Sorensen, Ronald L. Calabrese and Stephen P. DeWeerth	173
<i>Whence Sparseness?</i> , Carl van Vreeswijk	180

Part III Theory

<i>Efficient Learning of Linear Perceptrons</i> , Shai Ben-David and Hans Ulrich Simon .	189
<i>Algorithmic Stability and Generalization Performance</i> , Olivier Bousquet and Andr� Elisseeff	196
<i>Competition and Arbors in Ocular Dominance</i> , Peter Dayan	203
<i>From Margin to Sparsity</i> , Thore Graepel, Ralf Herbrich and Robert C. Williamson .	210
<i>Permitted and Forbidden Sets in Symmetric Threshold-Linear Networks</i> , Richard H.R. Hahnloser and H. Sebastian Seung	217
<i>A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work</i> , Ralf Herbrich and Thore Graepel	224
<i>On Reversing Jensen's Inequality</i> , Tony Jebara and Alex Pentland	231
<i>Second Order Approximations for Probability Models</i> , Hilbert J. Kappen and Wim Wiegerinck	238

<i>Some New Bounds on the Generalization Error of Combined Classifiers,</i> Vladimir Koltchinskii, Dmitriy Panchenko and Fernando Lozano	245
<i>Sparsity of Data Representation of Optimal Kernel Machine and Leave-one-out Estimator,</i> Adam Kowalczyk	252
<i>Foundations for a Circuit Complexity Theory of Sensory Processing,</i> Robert A. Legenstein and Wolfgang Maass	259
<i>A Tighter Bound for Graphical Models,</i> Martijn A. R. Leisink and Hilbert J. Kappen	266
<i>Learning Curves for Gaussian Processes Regression: A Framework for Good Approximations,</i> Dörthe Malzahn and Manfred Oppel	273
<i>Weak Learners and Improved Rates of Convergence in Boosting,</i> Shie Mannor and Ron Meir	280
<i>Learning Continuous Distributions: Simulations With Field Theoretic Priors,</i> Ilya Nemenman and William Bialek	287
<i>Occam's Razor,</i> Carl Edward Rasmussen and Zoubin Ghahramani	294
<i>The Kernel Trick for Distances,</i> Bernhard Schölkopf	301
<i>Regularization with Dot-Product Kernels,</i> Alex J. Smola, Zoltán L. Óvári and Robert C. Williamson	308
<i>Analysis of Bit Error Probability of Direct-Sequence CDMA Multiuser Demodulators,</i> Toshiyuki Tanaka	315
<i>Error-correcting Codes on a Bethe-like Lattice,</i> Renato Vicente, David Saad and Yoshiyuki Kabashima	322
<i>Algebraic Information Geometry for Learning Machines with Singularities,</i> Sumio Watanabe	329
<i>Computing with Finite and Infinite Networks,</i> Ole Winther	336
<i>Stagewise Processing in Error-correcting Codes and Image Restoration,</i> K. Y. Michael Wong and Hidetoshi Nishimori	343
<i>Learning Winner-take-all Competition Between Groups of Neurons in Lateral Inhibitory Networks,</i> Xiaohui Xie, Richard H.R. Hahnloser and H. Sebastian Seung	350
<i>Convergence of Large Margin Separable Linear Classification,</i> Tong Zhang	357

Part IV Algorithms and Architecture

<i>A Support Vector Method for Clustering,</i> Asa Ben-Hur, David Horn, Hava T. Siegelmann and Vladimir Vapnik	367
<i>A Variational Mean-Field Theory for Sigmoidal Belief Networks,</i> Chiranjib Bhattacharyya and S. Sathya Keerthi	374
<i>Direct Classification with Indirect Data,</i> Timothy X Brown	381
<i>Model Complexity, Goodness of Fit and Diminishing Returns,</i> Igor V. Cadez and Padhraic Smyth	388

<i>A Linear Programming Approach to Novelty Detection,</i> Colin Campbell and Kristin P. Bennett	395
<i>Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping,</i> Rich Caruana, Steve Lawrence and Lee Giles	402
<i>Incremental and Decremental Support Vector Machine Learning,</i> Gert Cauwenberghs and Tomaso Poggio	409
<i>Vicinal Risk Minimization,</i> Olivier Chapelle, Jason Weston, Léon Bottou and Vladimir Vapnik	416
<i>Gaussianization,</i> Scott Saobing Chen and Ramesh A. Gopinath	423
<i>The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity,</i> David Cohn and Thomas Hofmann	430
<i>Improved Output Coding for Classification Using Continuous Relaxation,</i> Koby Crammer and Yoram Singer	437
<i>Sparse Representation for Gaussian Process Models,</i> Lehel Csató and Manfred Opper	444
<i>Explaining Away in Weight Space,</i> Peter Dayan and Sham Kakade	451
<i>An Adaptive Metric Machine for Pattern Classification,</i> Carlotta Domeniconi, Jing Peng and Dimitrios Gunopulos	458
<i>High-temperature Expansions for Learning Models of Nonnegative Data,</i> Oliver B. Downs	465
<i>Incorporating Second-Order Functional Knowledge for Better Option Pricing,</i> Charles Dugas, Yoshua Bengio, François Bédille, Claude Nadeau and René Garcia	472
<i>Discovering Hidden Variables: A Structure-Based Approach,</i> Gal Elidan, Noam Lotner, Nir Friedman and Daphne Koller	479
<i>Accumulator Networks: Suitors of Local Probability Propagation,</i> Brendan J. Frey and Anitha Kannan	486
<i>Sequentially Fitting "Inclusive" Trees for Inference in Noisy-OR Networks,</i> Brendan J. Frey, Relu Patrascu, Tommi S. Jaakkola and Jodi Moran	493
<i>A New Approximate Maximal Margin Classification Algorithm,</i> Claudio Gentile	500
<i>Propagation Algorithms for Variational Bayesian Learning,</i> Zoubin Ghahramani and Matthew J. Beal	507
<i>The Kernel Gibbs Sampler,</i> Thore Graepel and Ralf Herbrich	514
<i>'N-Body' Problems in Statistical Learning,</i> Alexander G. Gray and Andrew W. Moore	521
<i>Large Scale Bayes Point Machines,</i> Ralf Herbrich and Thore Graepel	528
<i>Beyond Maximum Likelihood and Density Estimation: A Sample-Based Criterion for Unsupervised Learning of Complex Models,</i> Sepp Hochreiter and Michael C. Mozer	535

<i>Ensemble Learning and Linear Response Theory for ICA,</i> Pedro A.d.F.R. Højen-Sørensen, Ole Winther and Lars Kai Hansen	542
<i>Generalizable Singular Value Decomposition for Ill-posed Datasets,</i> Ulrik Kjems, Lars Kai Hansen and Stephen C. Strother	549
<i>Algorithms for Non-negative Matrix Factorization,</i> Daniel D. Lee and H. Sebastian Seung	556
<i>Text Classification using String Kernels,</i> Huma Lodhi, John Shawe-Taylor, Nello Cristianini and Chris Watkins	563
<i>Constrained Independent Component Analysis,</i> Wei Lu and Jagath C. Rajapakse	570
<i>Active Support Vector Machine Classification,</i> Olvi L. Mangasarian and David R. Musicant	577
<i>The Unscented Particle Filter,</i> Rudolph van der Merwe, Arnaud Doucet, Nando de Freitas and Eric Wan	584
<i>A Mathematical Programming Approach to the Kernel Fisher Algorithm,</i> Sebastian Mika, Gunnar Rätsch and Klaus-Robert Müller	591
<i>Automatic Choice of Dimensionality for PCA,</i> Thomas P. Minka	598
<i>On Iterative Krylov-Dogleg Trust-Region Steps for Solving Neural Networks</i> <i>Nonlinear Least Squares Problems,</i> Eiji Mizutani and James W. Demmel	605
<i>An Information Maximization Approach to Overcomplete and Recurrent</i> <i>Representations,</i> Oren Shriki, Haim Sompolinsky and Daniel D. Lee	612
<i>Sparse Greedy Gaussian Process Regression,</i> Alex J. Smola and Peter Bartlett	619
<i>Kernel Expansions with Unlabeled Examples,</i> Martin Szummer and Tommi S. Jaakkola	626
<i>Sparse Kernel Principal Component Analysis,</i> Michael E. Tipping	633
<i>Data Clustering by Markovian Relaxation and the Information Bottleneck</i> <i>Method,</i> Naftali Tishby and Noam Slonim	640
<i>Active Learning for Parameter Estimation in Bayesian Networks,</i> Simon Tong and Daphne Koller	647
<i>Mixtures of Gaussian Processes,</i> Volker Tresp	654
<i>Tree-Based Modeling and Estimation of Gaussian Processes on Graphs with</i> <i>Cycles,</i> Martin J. Wainwright, Erik B. Sudderth and Alan S. Willsky	661
<i>Feature Selection for SVMs,</i> Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio and Vladimir Vapnik	668
<i>On a Connection between Kernel PCA and Metric Multidimensional Scaling,</i> Christopher K. I. Williams	675
<i>Using the Nyström Method to Speed Up Kernel Machines,</i> Christopher K. I. Williams and Matthias Seeger	682
<i>Generalized Belief Propagation,</i> Jonathan S. Yedidia, William T. Freeman and Yair Weiss	689

<i>A Gradient-Based Boosting Algorithm for Regression Problems,</i> Richard S. Zemel and Toniann Pitassi	696
<i>Regularized Winnow Methods,</i> Tong Zhang	703

Part V Implementation

<i>A Silicon Primitive for Competitive Learning,</i> David Hsu, Miguel Figueroa and Chris Diorio	713
<i>Smart Vision Chip Fabricated Using Three Dimensional Integration Technology,</i> Hiroyuki Kurino, M. Nakagawa, Kang Wook Lee, Tomonori Nakamura, Yuusuke Yamada, Ki Tae Park and Mitsumasa Koyanagi	720
<i>Homeostasis in a Silicon Integrate and Fire Neuron,</i> Shih-Chii Liu and Bradley A. Minch	727
<i>Fast Training of Support Vector Classifiers,</i> Fernando Pérez-Cruz, Pedro L. Alarcón-Diana, Angel Navia-Vázquez and Antonio Artés-Rodríguez	734
<i>Four-legged Walking Gait Control Using a Neuromorphic Chip Interfaced to a Support Vector Learning Algorithm,</i> Susanne Still, Bernhard Schölkopf, Klaus Hepp and Rodney J. Douglas	741

Part VI Speech, Handwriting and Signal Processing

<i>New Approaches Towards Robust and Adaptive Speech Recognition</i> (invited paper), Hervé Bourlard, Samy Bengio and Katrin Weber	751
<i>Speech Denoising and Dereverberation Using Probabilistic Models,</i> Hagai Attias, John C. Platt, Alex Acero and Li Deng	758
<i>Combining ICA and Top-Down Attention for Robust Speech Recognition,</i> Un-Min Bae and Soo-Young Lee	765
<i>Learning Joint Statistical Models for Audio-Visual Fusion and Segregation,</i> John W. Fisher III, Trevor Darrell, William T. Freeman and Paul Viola	772
<i>Factored Semi-Tied Covariance Matrices,</i> Mark J.F. Gales	779
<i>Higher-Order Statistical Properties Arising from the Non-Stationarity of Natural Signals,</i> Lucas Parra, Clay Spence and Paul Sajda	786
<i>One Microphone Source Separation,</i> Sam T. Roweis	793
<i>Minimum Bayes Error Feature Selection for Continuous Speech Recognition,</i> George Saon and Mukund Padmanabhan	800
<i>Periodic Component Analysis: An Eigenvalue Method for Representing Periodic Structure in Speech,</i> Lawrence K. Saul and Jont B. Allen	807
<i>FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks,</i> Malcolm Slaney and Michele Covell	814

<i>Noise Suppression Based on Neurophysiologically-motivated SNR Estimation for Robust Speech Recognition,</i> Jürgen Tchorz, Michael Kleinschmidt and Birger Kollmeier	821
--	-----

Part VII Visual Processing

<i>Shape Context: A New Descriptor for Shape Matching and Object Recognition,</i> Serge Belongie, Jitendra Malik and Jan Puzicha	831
<i>Emergence of Movement Sensitive Neurons' Properties by Learning a Sparse Code for Natural Moving Images,</i> Rafal Bogacz, Malcolm W. Brown and Christophe Giraud-Carrier	838
<i>The Manhattan World Assumption: Regularities in Scene Statistics which Enable Bayesian Inference,</i> James M. Coughlan and Allen L. Yuille	845
<i>Feature Correspondence: A Markov Chain Monte Carlo Approach,</i> Frank Dellaert, Steven M. Seitz, Sebastian Thrun and Charles Thorpe	852
<i>Keeping Flexible Active Contours on Track using Metropolis Updates,</i> Trausti T. Kristjansson and Brendan J. Frey	859
<i>Color Opponency Constitutes a Sparse Representation for the Chromatic Structure of Natural Scenes,</i> Te-Won Lee, Thomas Wachtler and Terrence J. Sejnowski	866
<i>Learning Segmentation by Random Walks,</i> Marina Meilă and Jianbo Shi	873
<i>Partially Observable SDE Models for Image Sequence Recognition Tasks,</i> Javier R. Movellan, Paul Mineiro and Ruth J. Williams	880
<i>Learning Sparse Image Codes using a Wavelet Pyramid Architecture,</i> Bruno A. Olshausen, Phil Sallee and Michael S. Lewicki	887
<i>Learning and Tracking Cyclic Human Motion,</i> Dirk Ormoneit, Hedvig Sidenbladh, Michael J. Black and Trevor Hastie	894
<i>Redundancy and Dimensionality Reduction in Sparse-Distributed Representations of Natural Objects in Terms of Their Local Features,</i> Penio S. Penev	901
<i>Rate-coded Restricted Boltzmann Machines for Face Recognition,</i> Yee Whye Teh and Geoffrey E. Hinton	908
<i>Divisive and Subtractive Mask Effects: Linking Psychophysics and Biophysics,</i> Barbara Zenger and Christof Koch	915

Part VIII Applications

<i>From Mixtures of Mixtures to Adaptive Transform Coding,</i> Cynthia Archer and Todd K. Leen	925
<i>A Neural Probabilistic Language Model,</i> Yoshua Bengio, Réjean Ducharme and Pascal Vincent	932

<i>A Comparison of Image Processing Techniques for Visual Speech Recognition Applications</i> , Michael S. Gray, Terrence J. Sejnowski and Javier R. Movellan . . .	939
<i>Support Vector Novelty Detection Applied to Jet Engine Vibration Spectra</i> , Paul Hayton, Bernhard Schölkopf, Lionel Tarassenko and Paul Anuzis	946
<i>Recognizing Hand-written Digits Using Hierarchical Products of Experts</i> , Guy Mayraz and Geoffrey E. Hinton	953
<i>Sex with Support Vector Machines</i> , Baback Moghaddam and Ming-Hsuan Yang . .	960
<i>Probabilistic Semantic Video Indexing</i> , Milind R. Naphade, Igor Kozintsev and Thomas Huang	967
<i>Interactive Parts Model: An Application to Recognition of On-line Cursive Script</i> , Predrag Neskovic, Philip C. Davis and Leon N. Cooper	974
<i>Learning Switching Linear Models of Human Motion</i> , Vladimir Pavlović, James M. Rehg and John MacCormick	981
<i>Bayes Networks on Ice: Robotic Search for Antarctic Meteorites</i> , Liam Pedersen, Dimi Apostolopoulos and Red Whittaker	988
<i>The Use of Classifiers in Sequential Inference</i> , Vasin Punyakanok and Dan Roth . .	995
<i>Machine Learning for Video-Based Rendering</i> , Arno Schödl and Irfan Essa	1002
<i>Bayesian Video Shot Segmentation</i> , Nuno Vasconcelos and Andrew Lippman . . .	1009

Part IX Control, Navigation and Planning

<i>Programmable Reinforcement Learning Agents</i> , David Andre and Stuart J. Russell .	1019
<i>Exact Solutions to Time-Dependent MDPs</i> , Justin A. Boyan and Michael L. Littman	1026
<i>Decomposition of Reinforcement Learning for Admission Control of Self-Similar Call Arrival Processes</i> , Jakob Carlström	1033
<i>Reinforcement Learning with Function Approximation Converges to a Region</i> , Geoffrey J. Gordon	1040
<i>Hierarchical Memory-Based Reinforcement Learning</i> , Natalia Hernandez-Gardiol and Sridhar Mahadevan	1047
<i>Automated State Abstraction for Options using the U-Tree Algorithm</i> , Anders Jonsson and Andrew G. Barto	1054
<i>Robust Reinforcement Learning</i> , Jun Morimoto and Kenji Doya	1061
<i>Kernel-Based Reinforcement Learning in Average-Cost Problems: An Application to Optimal Portfolio Choice</i> , Dirk Ormoneit and Peter Glynn	1068
<i>Using Free Energies to Represent Q-values in a Multiagent Reinforcement Learning Task</i> , Brian Sallans and Geoffrey E. Hinton	1075
<i>Balancing Multiple Sources of Reward in Reinforcement Learning</i> , Christian R. Shelton	1082

APRICODD: Approximate Policy Construction Using Decision Diagrams, Robert St-Aubin, Jesse Hoey and Craig Boutilier	1089
Index of Authors	1097
Keyword Index	1101

The following symbols that appear in the table of contents indicate the type of contribution: **P** for a paper, **L** for a letter, **S** for a short survey, **R** for a review, **A** for an abstract, **B** for a book review, **C** for a conference report, **D** for a dissertation, **E** for an editorial, **F** for a feature, **G** for a general interest, **H** for a historical survey, **I** for an invited paper, **J** for a journal, **K** for a keynote, **L** for a letter, **M** for a memorial, **N** for a note, **O** for an obituary, **P** for a paper, **Q** for a questionnaire, **R** for a review, **S** for a short survey, **T** for a technical report, **U** for an unpublished work, **V** for a video, **W** for a workshop, **X** for an exhibition, **Y** for a yearbook, **Z** for a zine.

The conference of maintaining high quality articles with frequent delivery is a well known enough fact, and a well known problem. The 1992 papers presented at the conference were chosen by a panel of 14 members, each of whom was carefully reviewed by three or four referees. The referees were given more than two months to review the papers. Panel members were given a list of 14 papers to review, and the panel members were given a list of 14 papers to review.

The conference of maintaining high quality articles with frequent delivery is a well known enough fact, and a well known problem. The 1992 papers presented at the conference were chosen by a panel of 14 members, each of whom was carefully reviewed by three or four referees. The referees were given more than two months to review the papers. Panel members were given a list of 14 papers to review, and the panel members were given a list of 14 papers to review.

The conference of maintaining high quality articles with frequent delivery is a well known enough fact, and a well known problem. The 1992 papers presented at the conference were chosen by a panel of 14 members, each of whom was carefully reviewed by three or four referees. The referees were given more than two months to review the papers. Panel members were given a list of 14 papers to review, and the panel members were given a list of 14 papers to review.

The conference of maintaining high quality articles with frequent delivery is a well known enough fact, and a well known problem. The 1992 papers presented at the conference were chosen by a panel of 14 members, each of whom was carefully reviewed by three or four referees. The referees were given more than two months to review the papers. Panel members were given a list of 14 papers to review, and the panel members were given a list of 14 papers to review.

Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping

Rich Caruana
CALD, CMU
5000 Forbes Ave.
Pittsburgh, PA 15213
caruana@cs.cmu.edu

Steve Lawrence
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
lawrence@research.nj.nec.com

Lee Giles
Information Sciences
Penn State University
University Park, PA 16801
giles@ist.psu.edu

Abstract

The conventional wisdom is that backprop nets with excess hidden units generalize poorly. We show that nets with excess capacity generalize well when trained with backprop and early stopping. Experiments suggest two reasons for this: 1) Overfitting can vary significantly in different regions of the model. Excess capacity allows better fit to regions of high non-linearity, and backprop often avoids overfitting the regions of low non-linearity. 2) Regardless of size, nets learn task subcomponents in similar sequence. Big nets pass through stages similar to those learned by smaller nets. Early stopping can stop training the large net when it generalizes comparably to a smaller net. We also show that conjugate gradient can yield worse generalization because it overfits regions of low non-linearity when learning to fit regions of high non-linearity.

1 Introduction

It is commonly believed that large multi-layer perceptrons (MLPs) generalize poorly: nets with too much capacity overfit the training data. Restricting net capacity prevents overfitting because the net has insufficient capacity to learn models that are too complex. This belief is consistent with a VC-dimension analysis of net capacity vs. generalization: the more free parameters in the net the larger the VC-dimension of the hypothesis space, and the less likely the training sample is large enough to select a (nearly) correct hypothesis [2].

Once it became feasible to train large nets on real problems, a number of MLP users noted that the overfitting they expected from nets with excess capacity did not occur. Large nets appeared to generalize as well as smaller nets — sometimes better. The earliest report of this that we are aware of is Martin and Pittman in 1991: “*We find only marginal and inconsistent indications that constraining net capacity improves generalization*” [7].

We present empirical results showing that MLPs with excess capacity often do not overfit. On the contrary, we observe that large nets often generalize better than small nets of sufficient capacity. Backprop appears to use excess capacity to better fit regions of high non-linearity, while still fitting regions of low non-linearity smoothly. (This desirable behavior can disappear if a fast training algorithm such as conjugate gradient is used instead of backprop.) Nets with excess capacity trained with backprop appear first to learn models similar to models learned by smaller nets. If early stopping is used, training of the large net can be halted when the large net’s model is similar to models learned by smaller nets.

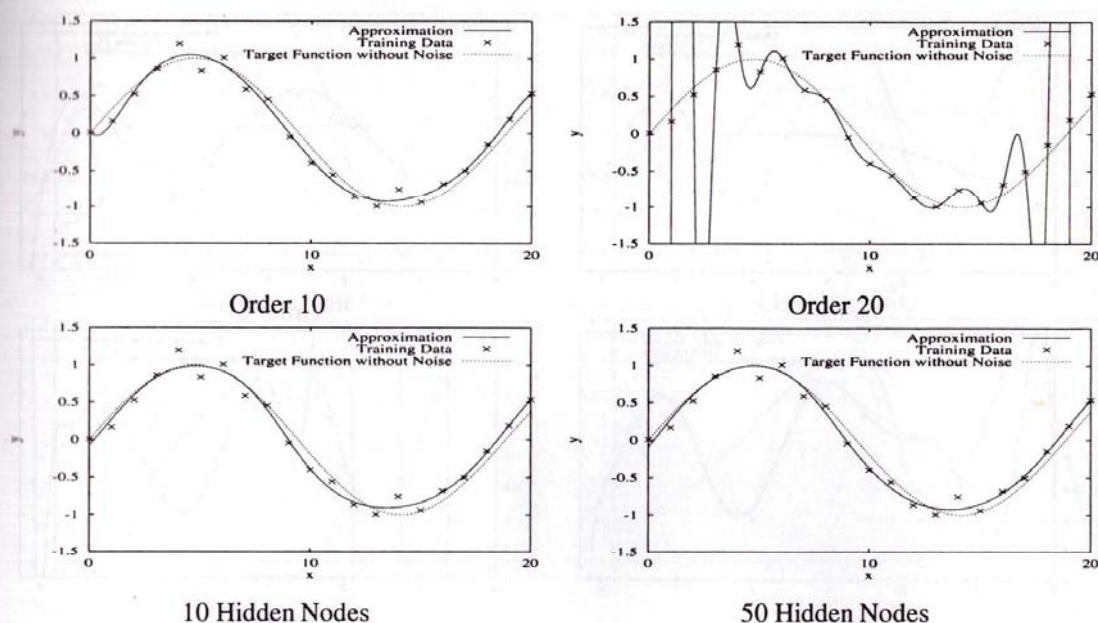


Figure 1: Top: Polynomial fit to data from $y = \sin(x/3) + \nu$. Order 20 overfits. Bottom: Small and large MLPs fit to same data. The large MLP does not overfit significantly more than the small MLP.

2 Overfitting

Much has been written about overfitting and the bias/variance tradeoff in neural nets and other machine learning models [2, 12, 4, 8, 5, 13, 6]. The top of Figure 1 illustrates polynomial overfitting. We created a training dataset by evaluating $y = \sin(x/3) + \nu$ at $0, 1, 2, \dots, 20$ where ν is a uniformly distributed random variable between -0.25 and 0.25 . We fit polynomial models with orders 2-20 to the data. Underfitting occurs with order 2. The fit is good with order 10. As the order (and number of parameters) increases, however, significant overfitting (poor generalization) occurs. At order 20, the polynomial fits the training data well, but interpolates poorly.

The bottom of Figure 1 shows MLPs fit to the data. We used a single hidden layer MLP, backpropagation (BP), and 100,000 stochastic updates. The learning rate was reduced linearly to zero from an initial rate of 0.5 (reducing the learning rate improves convergence, and linear reduction performs similarly to other schedules [3]). This schedule and number of updates trains the MLPs to completion. (We examine early stopping in Section 4.) As with polynomials, the smallest net with one hidden unit (HU) (4 weights) underfits the data. The fit is good with two HU (7 weights). Unlike polynomials, however, networks with 10 HU (31 weights) and 50 HU (151 weights) also yield good models. MLPs with seven times as many parameters as data points trained with BP do not significantly overfit this data. The experiments in Section 4 confirm that this bias of BP-trained MLPs towards smooth models is not limited to the simple 2-D problem used here.

3 Local Overfitting

Regularization methods such as weight decay typically assume that overfitting is a global phenomenon. But overfitting can vary significantly in different regions of a model. Figure 2 shows polynomial fits for data generated from the following equation:

$$y = \begin{cases} -\cos(x) + \nu & 0 \leq x < \pi \\ \cos(3(x - \pi)) + \nu & \pi \leq x \leq 2\pi \end{cases} \quad (\text{Equation 1})$$

Five equally spaced points were generated in the first region, and 15 in the second region, so that the two regions have different data densities and different underlying functions. Overfitting is different in the two regions. In Figure 2 the order 6 model fits the left region

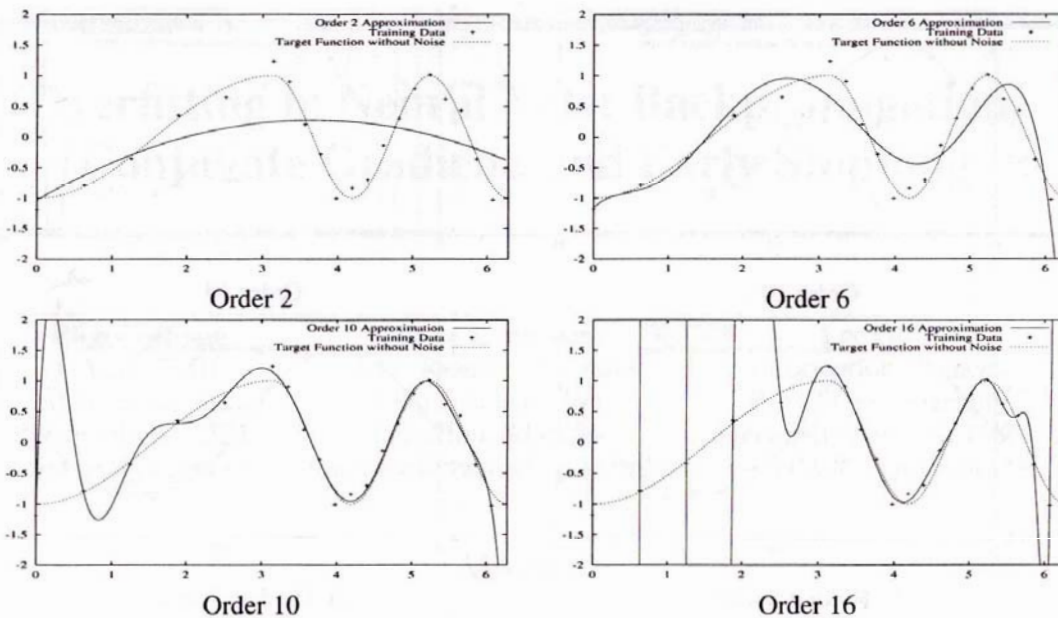


Figure 2: Polynomial approximation of data from Equation 1 as the order of the model is increased from 2 to 16. The overfitting behavior differs in the left and right hand regions.

well, but larger models overfit it. The order 6 model underfits the region on the right, and the order 10 model fits it better. No model performs well on both regions. Figure 3 shows MLPs trained on the same data (20,000 batch updates, learning rate linearly reduced to zero starting at 0.5). Small nets underfit. Larger nets, however, fit the entire function well *without significant overfitting in the left region*.

The ability of MLPs to fit both regions of low and high non-linearity well (without overfitting) depends on the training algorithm. Conjugate gradient (CG) is the most popular second order method. CG results in lower *training* error for this problem, but overfits significantly. Figure 4 shows results for 10 trials for BP and CG. Large BP nets generalize better on this problem — even the optimal size CG net is prone to overfitting. The degree of overfitting varies in different regions. When the net is large enough to fit the region of high non-linearity, overfitting is often seen in the region of low non-linearity.

4 Generalization, Network Capacity, and Early Stopping

The results in Sections 2 and 3 suggest that BP nets are less prone to overfitting than expected. But MLPs can and do overfit. This section examines overfitting vs. net size on seven problems: NETtalk [10], 7 and 12 bit parity, an inverse kinematic model for a robot arm (thanks to Sebastian Thrun for the simulator), Base 1 and Base 2: two sonar modeling problems using data collected from a robot wondering hallways at CMU, and vision data used to learn to steer an autonomous car [9]. These problems exhibit a variety of characteristics. Some are Boolean. Others are continuous. Some have noise. Others are noise-free. Some have many inputs or outputs. Others have few inputs or outputs.

4.1 Results

For each problem we used small training sets (100–1000 points, depending on the problem) so that overfitting was possible. We trained fully connected feedforward MLPs with one hidden layer whose size varied from 2 to 800 HU (about 500–100,000 parameters). All the nets were trained with BP using stochastic updates, learning rate 0.1, and momentum 0.9.

We used early stopping for regularization because it doesn't interfere with backprop's ability to control capacity locally. Early stopping combined with backprop is so effective that very large nets can be trained without significant overfitting. Section 4.2 explains why.

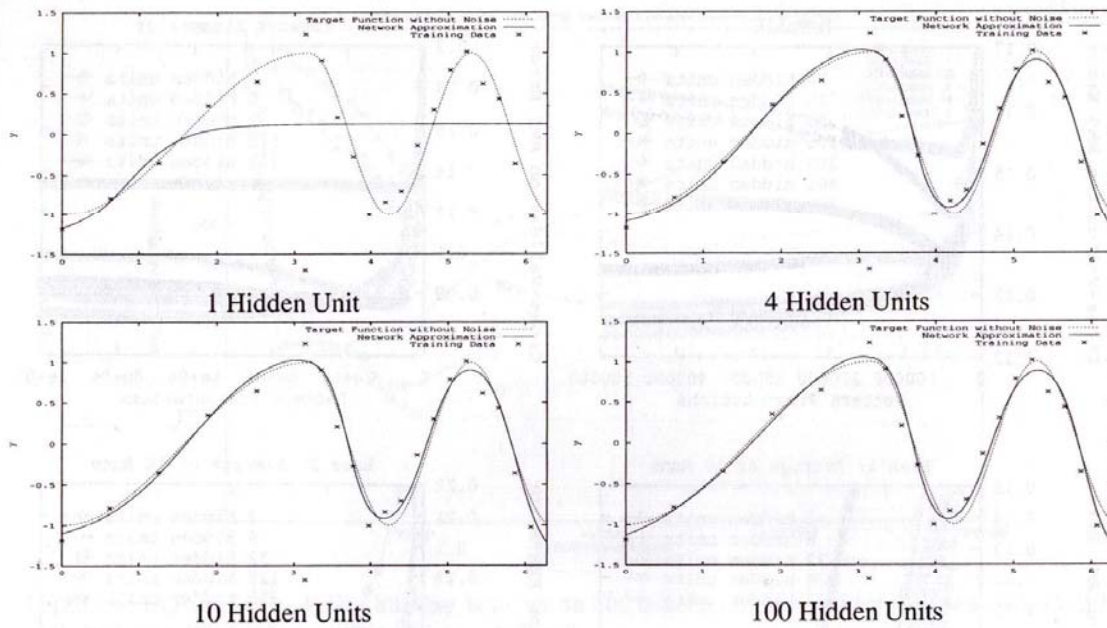


Figure 3: MLP approximation using backpropagation (BP) training of data from Equation 1 as the number of hidden units is increased. No significant overfitting can be seen.

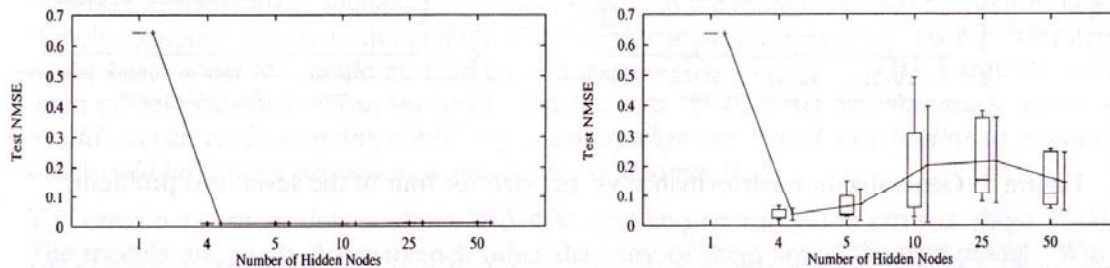


Figure 4: Test Normalized Mean Squared Error for MLPs trained with BP (left) and CG (right). Results are shown with both box-whiskers plots and the mean plus and minus one standard deviation.

Figure 5 shows generalization curves for four of the problems. Examining the results for all seven problems, we observe that on only three (Base 1, Base 2, and ALVINN), do nets that are too large yield worse generalization than smaller networks, but the loss is surprisingly small. Many trials were required before statistical tests confirmed that the differences between the optimal size net and the largest net were significant. Moreover, the results suggest that generalization is hurt more by using a net that is a little too small than by using one that is far too large, i.e., it is better to make nets too large than too small.

For most tasks and net sizes, we trained well beyond the point where generalization performance peaked. Because we had complete generalization curves, we noticed something unexpected. On some tasks, small nets overtrained considerably. The NETtalk graph in Figure 5 is a good example. Regularization (e.g., early stopping) is critical for nets of all sizes — not just ones that are *too big*. Nets with restricted capacity can overtrain.

4.2 Why Excess Capacity Does Not Hurt

BP nets initialized with small weights can develop large weights only after the number of updates is large. Thus BP nets consider hypotheses with small weights before hypotheses with large weights. Nets with large weights have more representational power, so simple hypotheses are explored before complex hypotheses.

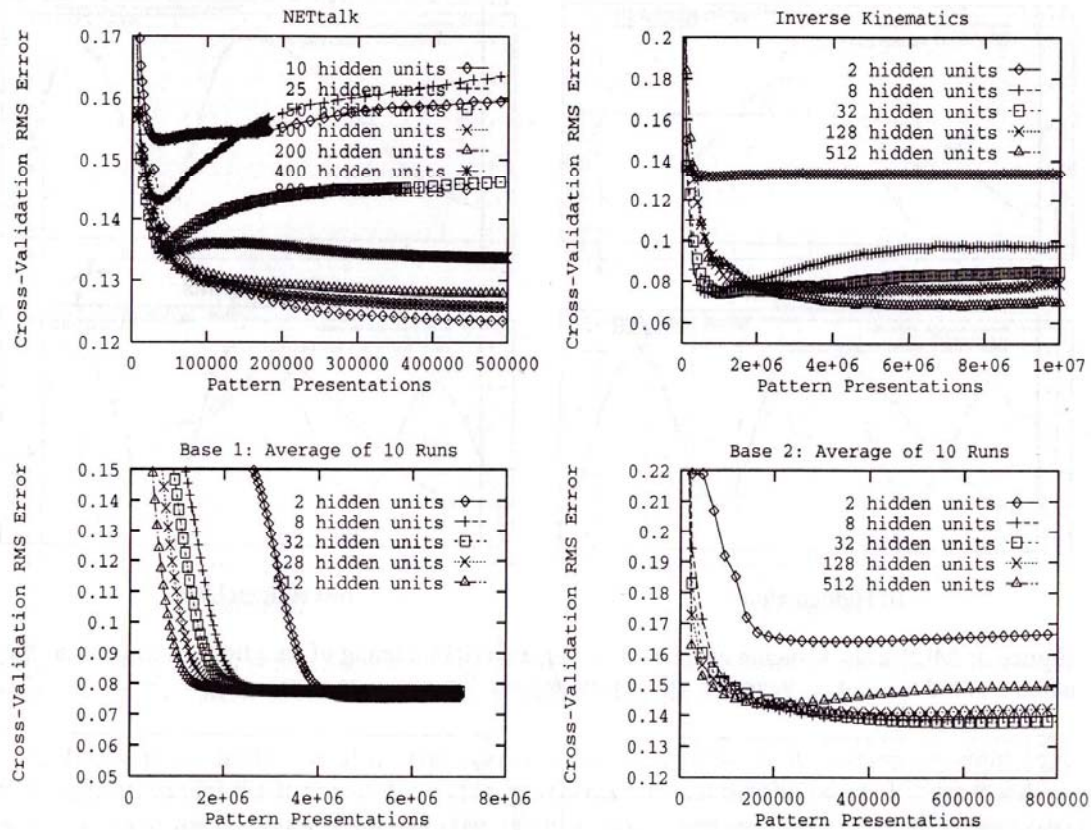


Figure 5: Generalization performance vs. net size for four of the seven test problems.

We analyzed *what* nets of different size learn *while* they are trained. We compared the input/output behavior of nets at different stages of learning on large samples of test patterns. We compare the input/output behavior of two nets by computing the squared error between the predictions made by the two nets. If two nets make the same predictions for all test cases, they have learned the same model (even though each model is represented differently), and the squared error between the two models is zero. If two nets make different predictions for test cases, they have learned different models, and the squared error between them is large. This is not the error the models make predicting the true labels, but the difference between predictions made by two different models. Two models can have poor generalization (large error on true labels), but have near zero error compared to each other if they are similar models. But two models with good generalization (low error on true labels) must have low error compared to each other.

The first graph in Figure 5 shows learning curves for nets with 10, 25, 50, 100, 200, and 400 HU trained on NETtalk. For each size, we saved the net from the epoch that generalized best on a large test set. This gives us the best model of each size found by backprop. We then trained a BP net with 800 HU, and after each epoch compared this net's model with the best models saved for nets of 10–400 HU. This lets us compare the sequence of models learned by the 800 HU net to the best models learned by smaller nets.

Figure 6 shows this comparison. The horizontal axis is the number of backprop passes applied to the 800 HU net. The vertical axis is the error between the 800 HU net model and the best model for each smaller net. The 800 HU net starts off distant from the good smaller models, then becomes similar to the good models, and then diverges from them. This is expected. What is interesting is that the 800 HU net first becomes closest to the best

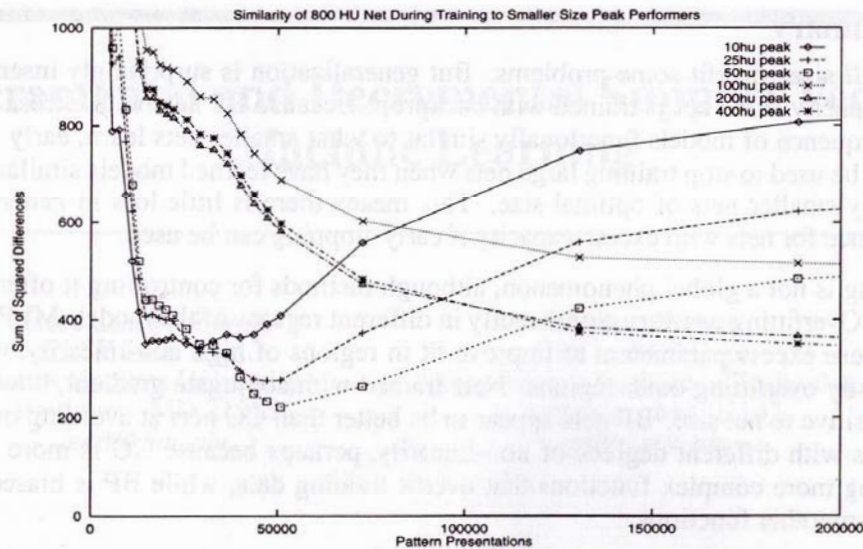


Figure 6: I/O similarity during training between an 800 hidden unit net and smaller nets (10, 25, 50, 100, 200, and 400 hidden units) trained on NETtalk.

10 HU net, then closest to the 25 HU net, then closest to the 50 HU net, etc. As it is trained, the 800 HU net learns a sequence of models similar to the models learned by smaller nets. If early stopping is used, training of the 800 HU net can be stopped when it behaves similar to the best model that could be learned with nets of 10, 25, 50, ... HU. Large BP nets learn models similar to those learned by smaller nets. *If a BP net with too much capacity would overfit, early stopping could stop training when the model was similar to a model that would have been learned by a smaller net of optimal size.*

The error between models is about 200–400, yet the generalization error is about 1600. The models are much closer to each other than any of them are to the true model. With early stopping, what counts is the closest approach of each model to the target function, not where models end up late in training. With early stopping there is little disadvantage to using models that are too large because their learning trajectories are similar to those followed by smaller nets of more optimal size.

5 Related Work

Our results show that models learned by backprop are biased towards “smooth” solutions. As nets with excess capacity are trained, they first explore smoother models similar to the models smaller nets would have learned. Weigend [11] performed an experiment that showed BP nets learn a problem’s eigenvectors in sequence, learning the 1st eigenvector first, then the 2nd, etc. His result complements our analysis of what nets of different sizes learn: if large nets learn an eigenvector sequence similar to smaller nets, then the models learned by the large net will pass through intermediate stages similar to what is learned by small nets (but iff nets of different sizes learn the eigenvectors equally well, which is an assumption we do not need to make.)

Theoretical work by [1] supports our results. Bartlett notes: “the VC-bounds seem loose; neural nets often perform successfully with training sets that are considerably smaller than the number of weights.” Bartlett shows (for classification) that the number of training samples only needs to grow according to A^{2l} (ignoring log factors) to avoid overfitting, where A is a bound on the total weight magnitudes and l is the number of layers in the network. This result suggests that a net with smaller weights will generalize better than a similar net with large weights. Examining the weights from BP and CG nets shows that BP training typically results in smaller weights.

6 Summary

Nets of *all* sizes overfit some problems. But generalization is surprisingly insensitive to excess capacity if the net is trained with backprop. Because BP nets with excess capacity learn a sequence of models functionally similar to what smaller nets learn, early stopping can often be used to stop training large nets when they have learned models similar to those learned by smaller nets of optimal size. This means there is little loss in generalization performance for nets with excess capacity if early stopping can be used.

Overfitting is not a global phenomenon, although methods for controlling it often assume that it is. Overfitting can vary significantly in different regions of the model. MLPs trained with BP use excess parameters to improve fit in regions of high non-linearity, while not significantly overfitting other regions. Nets trained with conjugate gradient, however, are more sensitive to net size. BP nets appear to be better than CG nets at avoiding overfitting in regions with different degrees of non-linearity, perhaps because CG is more effective at learning more complex functions that overfit training data, while BP is biased toward learning smoother functions.

References

- [1] Peter L. Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems*, volume 9, page 134. The MIT Press, 1997.
- [2] E.B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1(1):151–160, 1989.
- [3] C. Darken and J.E. Moody. Note on learning rate schedules for stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 3, pages 832–838. Morgan Kaufmann, 1991.
- [4] S. Geman et al. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.
- [5] A. Krogh and J.A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, volume 4, pages 950–957. Morgan Kaufmann, 1992.
- [6] Y. Le Cun, J.S. Denker, and S.A. Solla. Optimal Brain Damage. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 598–605, San Mateo, 1990. (Denver 1989), Morgan Kaufmann.
- [7] G.L. Martin and J.A. Pittman. Recognizing hand-printed letters and digits using backpropagation learning. *Neural Computation*, 3:258–267, 1991.
- [8] J.E. Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems*, volume 4, pages 847–854. Morgan Kaufmann, 1992.
- [9] D.A. Pomerleau. Alvin: An autonomous land vehicle in a neural network. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1, pages 305–313, San Mateo, 1989. (Denver 1988), Morgan Kaufmann.
- [10] T. Sejnowski and C. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168, 1987.
- [11] A. Weigend. On overfitting and the effective number of hidden units. In *Proceedings of the 1993 Connectionist Models Summer School*, pages 335–342. Lawrence Erlbaum Associates, 1993.
- [12] A.S. Weigend, D.E. Rumelhart, and B.A. Huberman. Generalization by weight-elimination with application to forecasting. In *Advances in Neural Information Processing Systems*, volume 3, pages 875–882. Morgan Kaufmann, 1991.
- [13] D. Wolpert. On bias plus variance. *Neural Computation*, 9(6):1211–1243, 1997.