Hebb Learning of Features based on their Information Content
Ferdinand Peper, Hideki Noda
This paper investigates the stationary points of a Hebb learning rule with a sigmoid nonlinearity in it. We show mathematically that when the input has a low information content, as measured by the input's variance, this learning rule suppresses learning, that is, forces the weight vector to converge to the zero vector. When the information content exceeds a certain value, the rule will automatically begin to learn a feature in the input. Our analysis suggests that under certain conditions it is the first principal component that is learned. The weight vector length remains bounded, provided the variance of the input is finite. Simulations confirm the theoretical results derived.
****

A Variational Principle for Model-based Morphing
Lawrence Saul, Michael Jordan
Given a multidimensional data set and a model of its density, we consider how to define the optimal interpolation between two points. This is done by assigning a cost to each path through space, based on two competing goals-one to interpolate through regions of high density, the other to minimize arc length. From this path functional, we derive the Euler-Lagrange equations for extremal motionj given two points, the desired interpolation is found by solv(cid:173) ing a boundary value problem. We show that this interpolation can be done efficiently, in high dimensions, for Gaussian, Dirichlet, and mixture models.
****

ARTEX: A Self-organizing Architecture for Classifying Image Regions
Stephen Grossberg, James Williamson
A self-organizing architecture is developed for image region classi(cid:173) fication. The system consists of a preprocessor that utilizes multi(cid:173) scale filtering, competition, cooperation, and diffusion to compute a vector of image boundary and surface properties, notably texture and brightness properties. This vector inputs to a system that incrementally learns noisy multidimensional mappings and their probabilities. The architecture is applied to difficult real-world image classification problems, including classification of synthet(cid:173) ic aperture radar and natural texture images, and outperforms a recent state-of-the-art system at classifying natural textures.
****

The CONDENSATION Algorithm - Conditional Density Propagation and Applications to Visual Tracking
Andrew Blake, Michael Isard
The power of sampling methods in Bayesian reconstruction of noisy signals is well known. The extension of sampling to temporal prob(cid:173) lems is discussed. Efficacy of sampling over time is demonstrated with visual tracking.
****

A Silicon Model of Amplitude Modulation Detection in the Auditory Brainstem
André van Schaik, Eric Fragnière, Eric Vittoz
Detectim of the periodicity of amplitude modulatim is a major step in the determinatim of the pitch of a SOODd. In this article we will present a silicm model that uses synchrroicity of spiking neurms to extract the fundamental frequency of a SOODd. It is based m the observatim that the so called 'Choppers' in the mammalian Cochlear Nucleus synchrmize well for certain rates of amplitude modulatim, depending m the cell's intrinsic chopping frequency. Our silicm model uses three different circuits, i.e., an artificial cochlea, an Inner Hair Cell circuit, and a spiking neuron circuit
****

Adaptive On-line Learning in Changing Environments
Noboru Murata, Klaus-Robert Müller, Andreas Ziehe, Shun-ichi Amari
An adaptive on-line algorithm extending the learning of learning idea is propos

ed and theoretically motivated. Relying only on gra(cid:173)dient flow information it can be applied to learning continuous functions or distributions, even when no explicit loss function is gi(cid:173)ven and the Hessian is not available. Its efficiency is demonstrated for a non-stationary blind separation task of acoustic signals.

****

Minimizing Statistical Bias with Queries

David Cohn

I describe a querying criterion that attempts to minimize the error of a learner by minimizing its estimated squared bias. I describe experiments with locally-weighted regression on two simple prob(cid:173)lems, and observe that this "bias-only" approach outperforms the more common "variance-only" exploration approach, even in the presence of noise.

****

A Micropower Analog VLSI HMM State Decoder for Wordspotting

John Lazzaro, John Wawrzynek, Richard P. Lippmann

We describe the implementation of a hidden Markov model state decoding system, a component for a wordspotting speech recogni(cid:173)tion system. The key specification for this state decoder design is microwatt power dissipation; this requirement led to a continuous(cid:173)time, analog circuit implementation. We characterize the operation of a 10-word (81 state) state decoder test chip.

****

Dual Kalman Filtering Methods for Nonlinear Prediction, Smoothing and Estimation

Eric Wan, Alex Nelson

Prediction, estimation, and smoothing are fundamental to signal processing. To perform these interrelated tasks given noisy data, we form a time series model of the process that generates the data. Taking noise in the system explicitly into account, maximum(cid:173)likelihood and Kalman frameworks are discussed which involve the dual process of estimating both the model parameters and the un(cid:173)derlying state of the system. We review several established meth(cid:173)ods in the linear case, and propose severa! extensions utilizing dual Kalman filters (DKF) and forward-backward (FB) filters that are applicable to neural networks. Methods are compared on several simulations of noisy time series. We also include an example of nonlinear noise reduction in speech.

****

Source Separation and Density Estimation by Faithful Equivariant SOM

Juan Lin, Jack Cowan, David Grier

We couple the tasks of source separation and density estimation by extracting the local geometrical structure of distributions ob(cid:173)tained from mixtures of statistically independent sources. Our modifications of the self-organizing map (SOM) algorithm results in purely digital learning rules which perform non-parametric his(cid:173)togram density estimation. The non-parametric nature of the sep(cid:173)aration allows for source separation of non-linear mixtures. An anisotropic coupling is introduced into our SOM with the role of aligning the network locally with the independent component con(cid:173)tours. This approach provides an exact verification condition for source separation with no prior on the source distributions.

****

Continuous Sigmoidal Belief Networks Trained using Slice Sampling

Brendan J. Frey

Real-valued random hidden variables can be useful for modelling latent structure that explains correlations among observed vari(cid:173)ables. I propose a simple unit that adds zero-mean Gaussian noise to its input before passing it through a sigmoidal squashing func(cid:173)tion. Such units can produce a variety of useful behaviors, ranging from deterministic to binary stochastic to continuous stochastic. I show how "slice sampling" can be used for inference and learning in top-down networks of these units and demonstrate learning on two simple problems.

****

Compositionality, MDL Priors, and Object Recognition

Elie Bienenstock, Stuart Geman, Daniel Potter

Images are ambiguous at each of many levels of a contextual hi(cid:173) erarchy. Nevertheless, the high-level interpretation of most scenes is una mbiguous, as evidenced by the superior performance of hu(cid:173) mans. Thi s observation argues for global vision models, such as de(cid:173) formable temp lates. Unfortunately, such models are computation(cid:173) ally intractable f or unconstrained problems. We propose a composi(cid:173) tional model in which primitives are recursively composed, subject to syntactic restrictions, to for m tree-structured objects and object groupings. Ambiguity is propagated up t he hierarchy in the form of multiple interpretations, which are later resolved by a Bayesian, equivalently minimum-description-Iength, cost functional.

****

A Mean Field Algorithm for Bayes Learning in Large Feed-forward Neural Networks

Manfred Opper, Ole Winther

We present an algorithm which is expected to realise Bayes optimal predictions in large feed-forward networks. It is based on mean field methods developed w ithin statistical mechanics of disordered sys(cid:173) tems. We give a deri vation for the single layer perceptron and show that the algorithm also pr ovides a leave-one-out cross-validation test of the predictions. Simulatio ns show excellent agreement with theoretical results of statistical mechanic s.

****

Ordered Classes and Incomplete Examples in Classification

Mark Mathieson

The classes in classification tasks often have a natural ordering, and the trai ning and testing examples are often incomplete. We propose a non(cid:173) line ar ordinal model for classification into ordered classes. Predictive, simula tion-based approaches are used to learn from past and classify fu(cid:173) ture incomplete examples. These techniques are illustrated by making prognoses for patients who have suffered severe head injuries.

****

Unification of Information Maximization and Minimization

Ryotaro Kamimura

In the present paper, we propose a method to unify information maximi zation and minimization in hidden units. The information maximization and m inimization are performed on two different lev(cid:173) els: collective and ind ividual level. Thus, two kinds of information: collective and individual in formation are defined. By maximizing collective information and by mini mizing individual information, simple networks can be generated in term s of the number of con(cid:173) nections and the number of hidden units. Obtained networks are expected to give better generalization and improved interpretation of internal representations. This method was applied to the i nfer(cid:173) ence of the maximum onset principle of an artificial language . In this problem, it was shown that the individual information min(c id:173) imization is not contradictory to the collective information max (cid:173) imization. In addition, experimental results confirmed improved generalization performance, because over-training can significantly be suppre ssed.

****

Multi-Task Learning for Stock Selection

Joumana Ghosn, Yoshua Bengio

Artificial Neural Networks can be used to predict future returns of st ocks in order to take financial decisions . Should one build a separate network for each stock or share the same network for all the stocks? In th is paper we also explore other alternatives, in which some layers are shared and others are not shared. When the prediction of future retu rns for different stocks are viewed as different tasks, sharing some p arameters across stocks is a form of multi-task learning. In a series

of experiments with Canadian stocks, we obtain yearly returns that are more than 14% above various benchmarks.

****

Cholinergic Modulation Preserves Spike Timing Under Physiologically Realistic Fluctuating Input

Akaysha Tang, Andreas Bartels, Terrence J. Sejnowski

Neuromodulation can change not only the mean firing rate of a neuron, but also its pattern of firing . Therefore, a reliable neu(cid:173) ral coding scheme, whether a rate coding or a spike time based coding, must be robust in a dynamic neuromodulatory environ(cid:173) ment. The common observation that cholinergic modulation leads to a reduction in spike frequency adaptation implies a modifica(cid:173) tion of spike timing, which would make a neural code based on precise spike timing difficult to maintain. In this paper, the effects of cholinergic modulation were studied to test the hypothesis that precise spike timing can serve as a reliable neural code. Using the whole cell patch-clamp technique in rat neocortical slice prepara(cid:173) tion and compartmental modeling techniques, we show that cholin(cid:173) ergic modulation, surprisingly, preserved spike timing in response to a fluctuating inputs that resembles in vivo conditions. This re(cid:173) sult suggests that in vivo spike timing may be much more resistant to changes in neuromodulator concentrations than previous physi(cid:173) ological studies have implied.

****

Analytical Mean Squared Error Curves in Temporal Difference Learning

Satinder Singh, Peter Dayan

We have calculated analytical expressions for how the bias and variance of the estimators provided by various temporal difference value estimation algorithms change with offline updates over trials in absorbing Markov chains using lookup table representations. We illustrate classes of learning curve behavior in various chains, and show the manner in which TD is sensitive to the choice of its step(cid:173) size and eligibility trace parameters.

****

Dynamics of Training

Siegfried Bös, Manfred Opper

A new method to calculate the full training process of a neural net(cid:173) work is introduced. No sophisticated methods like the replica trick are used. The results are directly related to the actual number of training steps. Some results are presented here, like the maximal learning rate, an exact description of early stopping, and the neces(cid:173) sary number of training steps. Further problems can be addressed with this approach.

****

A Constructive RBF Network for Writer Adaptation

John Platt, Nada Matic

This paper discusses a fairly general adaptation algorithm which augments a standard neural network to increase its recognition ac(cid:173) curacy for a specific user. The basis for the algorithm is that the output of a neural network is characteristic of the input, even when the output is incorrect. We exploit this characteristic output by using an Output Adaptation Module (OAM) which maps this out(cid:173) put into the correct user-dependent confidence vector. The OAM is a simplified Resource Allocating Network which constructs ra(cid:173) dial basis functions on-line. We applied the OAM to construct a writer-adaptive character recognition system for on-line hand(cid:173) printed characters. The OAM decreases the word error rate on a test set by an average of 45%, while creating only 3 to 25 basis functions for each writer in the test set.

****

MLP Can Provably Generalize Much Better than VC-bounds Indicate

Adam Kowalczyk, Herman Ferrá

Results of a study of the worst case learning curves for a partic(cid:173) ular class of probability distribution on input space to MLP with hard

threshold hidden units are presented. It is shown in partic(cid:173) ular, that in the thermodynamic limit for scaling by the number of connections to the first hidden layer, although the true learning curve behaves as ~ a-I for a ~ 1, its VC-dimension based bound is trivial (= 1) and its VC-entropy bound is trivial for a ::; 6.2. It is also shown that bounds following the true learning curve can be derived from a formalism based on the density of error patterns.

****

3D Object Recognition: A Model of View-Tuned Neurons

Emanuela Bricolo, Tomaso Poggio, Nikos K. Logothetis

In 1990 Poggio and Edelman proposed a view-based model of ob(cid:173) ject recognition that accounts for several psychophysical properties of certain recognition tasks. The model predicted the existence of view-tuned and view-invariant units, that were later found by Lo(cid:173) gothetis et al. (Logothetis et al., 1995) in IT cortex of monkeys trained with views of specific paperclip objects. The model, how(cid:173) ever, does not specify the inputs to the view-tuned units and their internal organization. In this paper we propose a model of these view-tuned units that is consistent with physiological data from single cell responses.

****

Learning Bayesian Belief Networks with Neural Network Estimators

Stefano Monti, Gregory Cooper

In this paper we propose a method for learning Bayesian belief networks from data. The method uses artificial neural networks as probability estimators, thus avoiding the need for making prior assumptions on the nature of the probability distributions govern(cid:173) ing the relationships among the participating variables. This new method has the potential for being applied to domains containing both discrete and continuous variables arbitrarily distributed. We compare the learning performance of this new method with the performance of the method proposed by Cooper and Herskovits in [7]. The experimental results show that, although the learning scheme based on the use of ANN estimators is slower, the learning accuracy of the two methods is comparable. Category: Algorithms and Architectures.

****

Approximate Solutions to Optimal Stopping Problems

John Tsitsiklis, Benjamin Van Roy

We propose and analyze an algorithm that approximates solutions to the problem of optimal stopping in a discounted irreducible ape(cid:173) riodic Markov chain. The scheme involves the use of linear com(cid:173) binations of fixed basis functions to approximate a Q-function. The weights of the linear combination are incrementally updated through an iterative process similar to Q-Iearning, involving sim(cid:173) ulation of the underlying Markov chain. Due to space limitations, we only provide an overview of a proof of convergence (with prob(cid:173) ability 1) and bounds on the approximation error. This is the first theoretical result that establishes the soundness of a Q-Iearning(cid:173) like algorithm when combined with arbitrary linear function ap(cid:173) proximators to solve a sequential decision problem. Though this paper focuses on the case of finite state spaces, the results extend naturally to continuous and unbounded state spaces, which are ad(cid:173) dressed in a forthcoming full-length paper.

****

An Orientation Selective Neural Network for Pattern Identification in Particle Detectors

Halina Abramowicz, David Horn, Ury Naftaly, Carmit Sahar-Pikielny

We present an algorithm for identifying linear patterns on a two(cid:173) dimensional lattice based on the concept of an orientation selective cell, a concept borrowed from neurobiology of vision. Construct(cid:173) ing a multi-layered neural network with fixed architecture which implements orientation selectivity, we define output elements cor(cid:173) responding to different orientations, which allow us to make a se(cid:173) lection deci

sion. The algorithm takes into account the granularity of the lattice as well as the presence of noise and inefficiencies. The method is applied to a sample of data collected with the ZEUS detector at HERA in order to identify cosmic muons that leave a linear pattern of signals in the segmented calorimeter. A two dimensional representation of the relevant part of the detector is used. The algorithm performs very well. Given its architecture, this system becomes a good candidate for fast pattern recognition in parallel processing devices.
****

Early Brain Damage
Volker Tresp, Ralph Neuneier, Hans-Georg Zimmermann
Optimal Brain Damage (OBD) is a method for reducing the num(cid:173)ber of weights in a neural network. OBD estimates the increase in cost function if weights are pruned and is a valid approximation if the learning algorithm has converged into a local minimum. On the other hand it is often desirable to terminate the learning pro(cid:173)cess before a local minimum is reached (early stopping). In this paper we show that OBD estimates the increase in cost function incorrectly if the network is not in a local minimum. We also show how OBD can be extended such that it can be used in connec(cid:173)tion with early stopping. We call this new approach Early Brain Damage, EBD. EBD also allows to revive already pruned weights. We demonstrate the improvements achieved by EBD using three publicly available data sets.
****

Text-Based Information Retrieval Using Exponentiated Gradient Descent
Ron Papka, James Callan, Andrew Barto
The following investigates the use of single-neuron learning algo(cid:173)rithms to improve the performance of text-retrieval systems that accept natural-language queries. A retrieval process is explained that transforms the natural-language query into the query syntax of a real retrieval system: the initial query is expanded using statis(cid:173)tical and learning techniques and is then used for document ranking and binary classification. The results of experiments suggest that Kivinen and Warmuth's Exponentiated Gradient Descent learning algorithm works significantly better than previous approaches.
****

An Analog Implementation of the Constant Average Statistics Constraint For Sensor Calibration
John Harris, Yu-Ming Chiang
We use the constant statistics constraint to calibrate an array of sensors that contains gain and offset variations. This algorithm has been mapped to analog hardware and designed and fabricated with a 2um CMOS technology. Measured results from the chip show that the system achieves invariance to gain and offset variations of the input signal.
****

A Neural Model of Visual Contour Integration
Zhaoping Li
We introduce a neurobiologically plausible model of contour inte(cid:173)gration from visual inputs of individual oriented edges. The model is composed of interacting excitatory neurons and inhibitory in(cid:173)terneurons, receives visual inputs via oriented receptive fields (RFs) like those in VI. The RF centers are distributed in space. At each location, a finite number of cells tuned to orientations spanning 1800 compose a model hypercolumn. Cortical interactions modify neural activities produced by visual inputs, selectively amplifying activities for edge elements belonging to smooth input contours. El(cid:173)ements within one contour produce synchronized neural activities. We show analytically and empirically that contour enhancement and neural synchrony increase with contour length, smoothness and closure, as observed experimentally. This model gives testable predictions, and in addition, introduces a feedback mechanism al(cid:173)lowing higher visual centers to enhance, suppress, and segment contours.

****

Unsupervised Learning by Convex and Conic Coding
Daniel Lee, H. Sebastian Seung
Unsupervised learning algorithms based on convex and conic en(cid:173) coders are proposed. The encoders find the closest convex or conic combination of basis vectors to the input. The learning algorithms produce basis vectors t hat minimize the reconstruction error of the encoders. The convex algorithm d evelops locally linear models of the input, while the conic algorithm discovers features. Both al(cid:173) gorithms are used to model handwritten d igits and compared with vector quantization and principal component analysis. The neural network implementations involve feedback connections that project a reconstruction back to the input layer.

****

Reconstructing Stimulus Velocity from Neuronal Responses in Area MT
Wyeth Bair, James Cavanaugh, J. Movshon
We employed a white-noise velocity signal to study the dynamics of the resp onse of single neurons in the cortical area MT to visual motion. Responses wer e quantified using reverse correlation, opti(cid:173) mal linear reconstruction filters, and reconstruction signal-to-noise ratio (SNR). The SNR and lower bound estimates of information rate were lower than we expected. Ninety percent of the informa(cid:173) tion was transmitted below 18 Hz, and the highe st lower bound on bit rate was 12 bits/so A simulated opponent motion energ y sub(cid:173) unit with Poisson spike statistics was able to out-perform the M T neurons. The temporal integration window, measured from the re(cid:173) verse correlation half-width, ranged from 30-90 ms. The window was narrower when a stimulus moved faster, but did not change when temporal frequen cy was held constant.

****

Multidimensional Triangulation and Interpolation for Reinforcement Learning
Scott Davies
Dynamic Programming, Q-Iearning and other discrete Markov Decision Proce ss solvers can be -applied to continuous d-dimensional state-spaces by quant izing the state space into an array of boxes. This is often problematic above two dimensions: a coarse quantization can lead to poor policies, and fine quantization is too expensive. Possible solutions are variable-resolution dis cretization, or function approximation by neural nets. A third option, w hich has been little studied in the reinforcement learning literature, is interpolation on a coarse grid. In this paper we study interpolatio n tech(cid:173) niques that can result in vast improvements in the onlin e behavior of the resulting control systems: multilinear interpolation, a nd an interpolation algorithm based on an interesting regular triangula tion of d-dimensional space. We adapt these interpolators under three r einforcement learning paradigms: (i) offline value iteration with a known model, (ii) Q-Iearning, and (iii) online value iteration with a previ ously unknown model learned from data. We describe empirical results, and the resulting implications for practical learning of continuous non-linear dyn amic control.

****

Viewpoint Invariant Face Recognition using Independent Component Analysis and At tractor Networks
Marian Bartlett, Terrence J. Sejnowski
We have explored two approaches to recogmzmg faces across changes in po se. First, we developed a representation of face images based on independent component analysis (ICA) and compared it to a principal component analysi s (PCA) representation for face recognition. The ICA basis vectors for this data set were more spatially local than the PCA basis vectors and the ICA representa(cid:173) tion had greater invariance to changes in pose. Second, we present a model for the development of viewpoint invariant response s to faces from visual experience in a biological system. The temporal c ontinuity of natural visual experience was incorporated into an attracto

r network model by Hebbian learning following a lowpass temporal filter o n unit activities. When combined with the tem(cid:173) poral filter, a basic Hebbian update rule became a generalization of Griniasty et al. (1993), which associates temporally proximal input patterns into basins of a ttraction. The system acquired rep(cid:173) resentations of faces that were la rgely independent of pose.
****

On the Effect of Analog Noise in Discrete-Time Analog Computations
Wolfgang Maass, Pekka Orponen
We introduce a model for noise-robust analog computations with discrete time that is flexible enough to cover the most important concrete case s, such as computations in noisy analog neural nets and networks of noi sy spiking neurons. We show that the presence of arbitrarily small amounts o f analog noise reduces the power of analog computational models to that of finite automata, and we also prove a new type of upper bound fo r the VC-dimension of computational models with analog noise.
****

Combinations of Weak Classifiers
Chuanyi Ji, Sheng Ma
To obtain classification systems with both good generalization per(cid:173) form ance and efficiency in space and time, we propose a learning method b ased on combinations of weak classifiers, where weak clas(cid:173) sifiers are linear classifiers (perceptrons) which can do a little better than making rando m guesses. A randomized algorithm is proposed to find the weak classifiers. They· are then combined through a ma(cid:173) jority vote. As demonstrated t hrough systematic experiments, the method developed is able to obtain combinat ions of weak classifiers with good generalization performance and a fast training time on a variety of test problems and real applications.
****

Reinforcement Learning for Dynamic Channel Allocation in Cellular Telephone Syst ems
Satinder Singh, Dimitri Bertsekas
In cellular telephone systems, an important problem is to dynami(cid:173) cally allocate the communication resource (channels) so as to max(cid:173) imize ser vice in a stochastic caller environment. This problem is naturally formu lated as a dynamic programming problem and we use a reinforcement learning (RL) method to find dynamic channel allocation policies that are better than p revious heuristic solutions. The policies obtained perform well for a broad v ariety of call traf(cid:173) fic patterns. We present results on a large cellular system with approximately 4949 states.
****

Neural Learning in Structured Parameter Spaces - Natural Riemannian Gradient
Shun-ichi Amari
The parameter space of neural networks has a Riemannian met(cid:173) ric structure. The natural Riemannian gradient should be used instead of t he conventional gradient, since the former denotes the true steepest descen t direction of a loss function in the Riemannian space. The behavior of the s tochastic gradient learning algorithm is much more effective if the natural g radient is used. The present paper studies the information-geometrical struct ure of perceptrons and other networks, and prove that the on-line learn ing method based on the natural gradient is asymptotically as efficien t as the optimal batch algorithm. Adaptive modification of the learning constant is proposed and analyzed in terms of the Riemannian mea(cid:173) sure and is shown to be efficient. The natural gradient is finally appli ed to blind separation of mixtured independent signal sources.
****

Extraction of Temporal Features in the Electrosensory System of Weakly Electric Fish
Fabrizio Gabbiani, Walter Metzner, Ralf Wessel, Christof Koch
The encoding of random time-varying stimuli in single spike trains of electrose

nsory neurons in the weakly electric fish Eigenmannia was investigated using methods of statistical signal processing. At the first stage of the electrosensory system, spike trains were found to encode faithfully the detailed time-course of random stimuli, while at the second stage neurons responded specifically to features in the temporal waveform of the stimulus. Therefore stimulus infor(cid:173) mation is processed at the second stage of the electrosensory system by extracting temporal features from the faithfully preserved image of the environment sampled at the first stage.

****

A Model of Recurrent Interactions in Primary Visual Cortex

Emanuel Todorov, Athanassios Siapas, David Somers

A general feature of the cerebral cortex is its massive intercon(cid:173) nectivity - it has been estimated anatomically [19] that cortical neurons receive upwards of 5,000 synapses, the majority of which originate from other nearby cortical neurons. Numerous experi(cid:173) ments in primary visual cortex (VI) have revealed strongly nonlin(cid:173) ear interactions between stimulus elements which activate classical and non-classical receptive field regions. Recurrent cortical con(cid:173) nections likely contribute substantially to these effects. However, most theories of visual processing have either assumed a feedfor(cid:173) ward processing scheme [7], or have used recurrent interactions to account for isolated effects only [1, 16, 18]. Since nonlinear sys(cid:173) tems cannot in general be taken apart and analyzed in pieces, it is not clear what one learns by building a recurrent model that only accounts for one, or very few phenomena. Here we develop a relatively simple model of recurrent interactions in VI, that re(cid:173) flects major anatomical and physiological features of intracortical connectivity, and simultaneously accounts for a wide range of phe(cid:173) nomena observed physiologically. All phenomena we address are strongly nonlinear, and cannot be explained by linear feedforward models.

****

On a Modification to the Mean Field EM Algorithm in Factorial Learning

A. Dunmur, D. Titterington

A modification is described to the use of mean field approxima(cid:173) tions in the E step of EM algorithms for analysing data from latent structure models, as described by Ghahramani (1995), among oth(cid:173) ers. The modification involves second-order Taylor approximations to expectations computed in the E step. The potential benefits of the method are illustrated using very simple latent profile models.

****

VLSI Implementation of Cortical Visual Motion Detection Using an Analog Neural Computer

Ralph Etienne-Cummings, Jan Van der Spiegel, Naomi Takahashi, Alyssa Apsel, Paul Mueller

Two dimensional image motion detection neural networks have been implemented using a general purpose analog neural computer. The neural circuits perform spatiotemporal feature extraction based on the cortical motion detection model of Adelson and Bergen. The neural computer provides the neurons, synapses and synaptic time-constants required to realize the model in VLSI hardware. Results show that visual motion estimation can be implemented with simple sum-and(cid:173) threshold neural hardware with temporal computational capabilities. The neural circuits compute general 20 visual motion in real-time.

****

Local Bandit Approximation for Optimal Learning Problems

Michael Duff, Andrew Barto

In general, procedures for determining Bayes-optimal adaptive controls for Markov decision processes (MDP's) require a pro(cid:173) hibitive amount of computation-the optimal learning problem is intractable. This paper proposes an approximate approach in which bandit processes are used to model, in a certain "local" sense, a given MDP. Bandit processes constitute an

important subclass of MDP's, and have optimal learning strategies (defined in terms of Gittins indices) that can be computed relatively efficiently. Thus, one scheme for achieving approximately-optimal learning for general MDP's proceeds by taking actions suggested by strategies that are optimal with respect to local bandit models.
(cid:173)
****

Learning Appearance Based Models: Mixtures of Second Moment Experts
Christoph Bregler, Jitendra Malik
This paper describes a new technique for object recognition based on learning appearance models. The image is decomposed into local regions which are described by a new texture representation called "Generalized Second Moments" that are derived from the output of multiscale, multiorientation filter banks. Class-characteristic local texture features and their global composition is learned by a hierarchical mixture of experts architecture (Jordan & Jacobs). The technique is applied to a vehicle database consisting of 5 general car categories (Sedan, Van with back-doors, Van without back-doors, old Sedan, and Volkswagen Bug). This is a difficult problem with considerable in-class variation. The new technique has a 6.5% misclassification rate, compared to eigen-images which give 17.4% misclassification rate, and nearest neighbors which give 15 .7% misclassification rate.
(cid:173)
****

Interpreting Images by Propagating Bayesian Beliefs
Yair Weiss
A central theme of computational vision research has been the realization that reliable estimation of local scene properties requires propagating measurements across the image. Many authors have therefore suggested solving vision problems using architectures of locally connected units updating their activity in parallel. Unfortunately, the convergence of traditional relaxation methods on such architectures has proven to be excruciatingly slow and in general they do not guarantee that the stable point will be a global minimum. In this paper we show that an architecture in which Bayesian Beliefs about image properties are propagated between neighboring units yields convergence times which are several orders of magnitude faster than traditional methods and avoids local minima. In particular our architecture is non-iterative in the sense of Marr [5]: at every time step, the local estimates at a given location are optimal given the information which has already been propagated to that location. We illustrate the algorithm's performance on real images and compare it to several existing methods.
(cid:173)
****

Why did TD-Gammon Work?
Jordan Pollack, Alan Blair
Although TD-Gammon is one of the major successes in machine learning, it has not led to similar impressive breakthroughs in temporal difference learning for other applications or even other games. We were able to replicate some of the success of TD-Gammon, developing a competitive evaluation function on a 4000 parameter feed-forward neural network, without using back-propagation, reinforcement or temporal difference learning methods. Instead we apply simple hill-climbing in a relative fitness environment. These results and further analysis suggest that the surprising success of Tesauro's program had more to do with the co-evolutionary structure of the learning task and the dynamics of the backgammon game itself.
(cid:173)
****

NeuroScale: Novel Topographic Feature Extraction using RBF Networks
David Lowe, Michael Tipping
Dimension-reducing feature extraction neural network techniques which also preserve neighbourhood relationships in data have traditionally been the exclusive domain of Kohonen self organising maps. Recently, we introduced a novel dimension-reducing feature extraction process, which is also topographic, based upon a Radial Basis Function architecture. It has been ob

served that the gener(cid:173) alisation performance of the system is broadly i nsensitive to model order complexity and other smoothing factors such as the kernel widths, contrary to intuition derived from supervised neural net(cid:173) work models. In this paper we provide an effective demonstratio n of this property and give a theoretical justification for the apparent 'self -regularising' behaviour of the 'NEUROSCALE' architecture.

****

Time Series Prediction using Mixtures of Experts

Assaf Zeevi, Ron Meir, Robert Adler

We consider the problem of prediction of stationary time series, using the architecture known as mixtures of experts (MEM). Here we suggest a mi xture which blends several autoregressive models. This study focuses on so me theoretical foundations of the predic(cid:173) tion problem in this co ntext. More precisely, it is demonstrated that this model is a universal approximator, with respect to learn(cid:173) ing the unknown prediction fun ction . This statement is strength(cid:173) ened as upper bounds on the m ean squared error are established. Based on these results it is possible to compare the MEM to other families of models (e.g., neural networks and state dependent mod(cid:173) els). It is shown that a degenerate version of the MEM is in fact equivalent to a neural network, and the number of expe rts in the architecture plays a similar role to the number of hidden u nits in the latter model.

****

MIMIC: Finding Optima by Estimating Probability Densities

Jeremy De Bonet, Charles Isbell, Paul Viola

In many optimization problems, the structure of solutions reflects complex re lationships between the different input parameters. For example, experience may tell us that certain parameters are closely related and should not be explored independently. Similarly, ex(cid:173) perience may establish that a subset of parameters must take on particular values. Any search of t he cost landscape should take advantage of these relationships. We presen t MIMIC, a framework in which we analyze the global structure of the optimi zation land(cid:173) scape. A novel and efficient algorithm for the estim ation of this structure is derived. We use knowledge of this structure to g uide a randomized search through the solution space and, in turn, to re(ci d:173) fine our estimate ofthe structure. Our technique obtains significant sp eed gains over other randomized optimization procedures.

****

Neural Network Models of Chemotaxis in the Nematode Caenorhabditis Elegans

Thomas Ferrée, Ben Marcotte, Shawn Lockery

We train recurrent networks to control chemotaxis in a computer model of the nematode C. elegans. The model presented is based closely on the body me chanics, behavioral analyses, neuroanatomy and neurophysiology of C. elegans, each imposing constraints rel(cid:173) evant for information processing. S imulated worms moving au(cid:173) tonomously in simulated chemical environmen ts display a variety of chemotaxis strategies similar to those of biological worms.

****

GTM: A Principled Alternative to the Self-Organizing Map

Christopher Bishop, Markus Svensén, Christopher Williams

The Self-Organizing Map (SOM) algorithm has been extensively studied and has been applied with considerable success to a wide variety of problems. H owever, the algorithm is derived from heuris(cid:173) tic ideas and this lea ds to a number of significant limitations. In this paper, we consider t he problem of modelling the probabil(cid:173) ity density of data in a space of several dimensions in terms of a smaller number of latent, or hidden, variables. We introduce a novel form of latent variable model, w hich we call the GTM algo(cid:173) rithm (for Generative Topographic Mappin g), which allows general non-linear transformations from latent space to data space, and which is trained using the EM (expectation-maximizatio

n) algo(cid:173) rithm. Our approach overcomes the limitations of the SOM, whi le introducing no significant disadvantages. We demonstrate the per(cid:173) f ormance of the GTM algorithm on simulated data from flow diag(cid:173) nostics for a multi-phase oil pipeline.

****

Support Vector Method for Function Approximation, Regression Estimation and Sign al Processing

Vladimir Vapnik, Steven Golowich, Alex Smola

The Support Vector (SV) method was recently proposed for es(cid:173) ti mating regressions, constructing multidimensional splines, and solving lin ear operator equations [Vapnik, 1995]. In this presenta(cid:173) tion we re port results of applying the SV method to these problems.

****

Smoothing Regularizers for Projective Basis Function Networks

John Moody, Thorsteinn Rögnvaldsson

Smoothing regularizers for radial basis functions have been studied extensively, but no general smoothing regularizers for projective basis junctions (PBFs), s uch as the widely-used sigmoidal PBFs, have heretofore been proposed. We de(cid:173) rive new classes of algebraically-simple mH'-order smoothing regularizers for networks of the form f(W, x) = L7=1 Ujg [x T Vj + Vjol + uo, with general projective basis functions g[.]. These regularizers are:

****

Bangs, Clicks, Snaps, Thuds and Whacks: An Architecture for Acoustic Transient P rocessing

Fernando Pineda, Gert Cauwenberghs, R. Edwards

We propose a neuromorphic architecture for real-time processing of acoustic transients in analog VLSI. We show how judicious normalization of a time-frequency signal allows an elegant and robust implementation of a correlation algorithm. The algorithm uses binary multiplexing instead of analog-analog multiplication. This removes the need for analog storage and analog-multiplication. Simulations show that the resulting algorithm has the same out-of-sample classification performance (-93% correct) as a baseline template-matching algorithm.

****

Salient Contour Extraction by Temporal Binding in a Cortically-based Network

Shih-Cheng Yen, Leif Finkel

It has been suggested that long-range intrinsic connections in striate cortex may play a role in contour extraction (Gilbert et aI., 1996). A nu mber of recent physiological and psychophysical studies have examined t he possible role of long range connections in the modulation of contrast de tection thresholds (Polat and Sagi, 1993,1994; Kapadia et aI., 1995; Kovacs and Julesz, 1994) and various pre-attentive detection tasks (Kovacs and Julesz, 1993; Field et aI., 1993). We have developed a network archite cture based on the anatomical connectivity of striate cortex, as well a s the temporal dynamics of neuronal processing, that is able to reproduc e the observed experimental results. The network has been tested on real images and has applications in terms of identifying salient contours in automatic image processing systems.

****

Learning Decision Theoretic Utilities through Reinforcement Learning

Magnus Stensmo, Terrence J. Sejnowski

Probability models can be used to predict outcomes and compensate for missing d ata, but even a perfect model cannot be used to make decisions unless the utili ty of the outcomes, or preferences between them, are also provided. This arise s in many real-world problems, such as medical di(cid:173) agnosis, where the c ost of the test as well as the expected improvement in the outcome must be co nsidered. Relatively little work has been done on learning the utilities of out comes for optimal decision making. In this paper, we show how temporal-differen ce reinforcement learning (TO(A» can be used to determine decision theoretic ut ilities within the context of a mixture model and apply this new approach to a

problem in medical di(cid:173) agnosis. TO(A) learning of utilities reduces th e number of tests that have to be done to achieve the same level of performance compared with the probability model alone, which results in significant cost s avings and in(cid:173) creased efficiency.

****

Spatial Decorrelation in Orientation Tuned Cortical Cells

Alexander Dimitrov, Jack Cowan

In this paper we propose a model for the lateral connectivity of ori entation-selective cells in the visual cortex based on information(cid:173) theo retic considerations. We study the properties of the input sig(cid:173) nal t o the visual cortex and find new statistical structures which have not been processed in the retino-geniculate pathway. Applying the idea that the s ystem optimizes the representation of incoming signals, we derive the lat eral connectivity that will achieve this for a set of local orientation-sel ective patches, as well as the complete spatial structure of a layer of su ch patches. We compare the results with various physiological measurements.

****

Sequential Tracking in Pricing Financial Options using Model Based and Neural Ne twork Approaches

Mahesan Niranjan

This paper shows how the prices of option contracts traded in finan(cid:173) cia l markets can be tracked sequentially by means of the Extended Kalman Filter al gorithm. I consider call and put option pairs with identical strike price and t ime of maturity as a two output nonlin(cid:173) ear system. The Black-Scholes ap proach popular in Finance liter(cid:173) ature and the Radial Basis Functions ne ural network are used in modelling the nonlinear system generating these observ ations. I show how both these systems may be identified recursively using the EKF algorithm. I present results of simulations on some FTSE 100 Index options data and discuss the implications of viewing the pricing problem in this sequen tial manner.

****

Are Hopfield Networks Faster than Conventional Computers?

Ian Parberry, Hung-Li Tseng

It is shown that conventional computers can be exponentiallx faster than planar Hopfield networks: although there are planar Hopfield networks that take ex ponential time to converge, a stable state of an arbitrary planar Hopfield n etwork can be found by a conventional computer in polynomial time. The theory of 'P.cS-completeness gives strong evidence that such a separation is unlikely for nonpla(cid:173) nar Hopfield networks, and it is demonstra ted that this is also the case for several restricted classes of nonpl anar Hopfield networks, including those who interconnection graphs are t he class of bipar(cid:173) tite graphs, graphs of degree 3, the dual of the k night's graph, the 8-neighbor mesh, the hypercube , the butterfly, the cube -connected cycles, and the shuffle-exchange graph.

****

Learning from Demonstration

Stefan Schaal

By now it is widely accepted that learning a task from scratch, i.e., without any prior knowledge, is a daunting undertaking. Humans, however, r arely at(cid:173) tempt to learn from scratch. They extract initial bias es as well as strategies how to approach a learning problem from instruc tions and/or demonstrations of other humans. For learning control, this p aper investigates how learning from demonstration can be applied in th e context of reinforcement learning. We consider priming the Q-function, the value function, the policy, and the model of the task dynamics as pos sible areas where demonstrations can speed up learning. In general nonlinea r learning problems, only model-based rein(cid:173) forcement learning shows significant speed-up after a demonstration, while in the special case of l inear quadratic regulator (LQR) problems, all methods profit from the demonstration. In an implementation of pole balancing on a complex anth

ropomorphic robot arm, we demonstrate that, when facing the complexities of real signal processing, model-based reinforcement learning offers the most robustness for LQR problems. Using the suggested methods, the robot learns pole balancing in just a single trial after a 30 second long demonstration of the human instructor.
****

Clustering Sequences with Hidden Markov Models
Padhraic Smyth
This paper discusses a probabilistic model-based approach to clus(cid:173) tering sequences, using hidden Markov models (HMMs) . The prob(cid:173) lem can be framed as a generalization of the standard mixture model approach to clustering in feature space. Two primary issues are addressed. First, a novel parameter initialization procedure is proposed, and second, the more difficult problem of determining the number of clusters K, from the data, is investigated. Experi(cid:173) mental results indicate that the proposed techniques are useful for revealing hidden cluster structure in data sets of sequences.
****

Promoting Poor Features to Supervisors: Some Inputs Work Better as Outputs
Rich Caruana, Virginia de
In supervised learning there is usually a clear distinction between inputs and outputs - inputs are what you will measure, outputs are what you will predict from those measurements. This paper shows that the distinction between inputs and outputs is not this simple. Some features are more useful as extra outputs than as inputs. By using a feature as an output we get more than just the case values but can. learn a mapping from the other inputs to that feature. For many features this mapping may be more useful than the feature value itself. We present two regression problems and one classification problem where performance improves if features that could have been used as inputs are used as extra outputs instead. This result is surprising since a feature used as an output is not used during testing.
****

Hidden Markov Decision Trees
Michael Jordan, Zoubin Ghahramani, Lawrence Saul
We study a time series model that can be viewed as a decision tree with Markov temporal structure. The model is intractable for exact calculations, thus we utilize variational approximations. We consider three different distributions for the approximation: one in which the Markov calculations are performed exactly and the layers of the decision tree are decoupled, one in which the decision tree calculations are performed exactly and the time steps of the Markov chain are decoupled, and one in which a Viterbi-like assumption is made to pick out a single most likely state sequence. We present simulation results for artificial data and the Bach chorales.
****

Representing Face Images for Emotion Classification
Curtis Padgett, Garrison Cottrell
We compare the generalization performance of three distinct rep(cid:173) resentation schemes for facial emotions using a single classification strategy (neural network). The face images presented to the clas(cid:173) sifiers are represented as: full face projections of the dataset onto their eigenvectors (eigenfaces); a similar projection constrained to eye and mouth areas (eigenfeatures); and finally a projection of the eye and mouth areas onto the eigenvectors obtained from 32x32 random image patches from the dataset. The latter system achieves 86% generalization on novel face images (individuals the networks were not trained on) drawn from a database in which human sub(cid:173) jects consistently identify a single emotion for the face .
****

Separating Style and Content

Joshua Tenenbaum, William Freeman
We seek to analyze and manipulate two factors, which we call style and conte nt, underlying a set of observations. We fit training data with bilinear mod els which explicitly represent the two-factor struc(cid:173) ture. These models can adapt easily during testing to new styles or content, allowing us to so lve three general tasks: extrapolation of a new style to unobserved content; classification of content observed in a new style; and translation of new content observed in a new style. For classification, we embed bilinear mo dels in a probabilistic framework, Separable Mixture Models (SMMsj, which generalizes earlier work on factorial mixture models [7, 3]. Significa nt per(cid:173) formance improvement on a benchmark speech dataset shows the benefits of our approach.
****

Contour Organisation with the EM Algorithm
José Leite, Edwin Hancock
This paper describes how the early visual process of contour organ(cid:173) isat ion can be realised using the EM algorithm. The underlying computation al representation is based on fine spline coverings. Ac(cid:173) cording to our EM approach the adjustment of spline parameters draws on an iterat ive weighted least-squares fitting process. The expectation step of our EM procedure computes the likelihood of the data using a mixture model de fined over the set of spline cover(cid:173) ings. These splines are limited in their spatial extent using Gaus(cid:173) sian windowing functions. The max imisation of the likelihood leads to a set of linear equations in the spline pa rameters which solve the weighted least squares problem. We evaluate the tech nique on the localisation of road structures in aerial infra-red images.
****

Combining Neural Network Regression Estimates with Regularized Linear Weights
Christopher Merz, Michael Pazzani
When combining a set of learned models to form an improved es(cid:173) ti mator, the issue of redundancy or multicollinearity in the set of models must be addressed. A progression of existing approaches and their limit ations with respect to the redundancy is discussed. A new approach, PCR , based on principal components regres(cid:173) sion is proposed to addr ess these limitations. An evaluation of the new approach on a collection of domains reveals that: 1) PCR was the most robust combination method as the redundancy of the learned models increased, 2) redundancy could be handl ed without eliminating any of the learned models, and 3) the principal compo(cid:173) nents of the learned models provided a continuum of "regularized" weigh ts from which PCR * could choose.
****

Triangulation by Continuous Embedding
Marina Meila, Michael Jordan
When triangulating a belief network we aim to obtain a junction tree of minimum state space. According to (Rose, 1970), searching for the optimal t riangulation can be cast as a search over all the permutations of the graph's vertices. Our approach is to embed the discrete set of permutati ons in a convex continuous domain D. By suitably extending the cost func tion over D and solving the continous nonlinear optimization task we hope to obtain a good triangulation with respect to the aformentioned cost. This paper presents two ways of embedding the triangulation probl em into continuous domain and shows that they perform well compared to the best known heuristic.
****

Bayesian Model Comparison by Monte Carlo Chaining
David Barber, Christopher Bishop
The techniques of Bayesian inference have been applied with great success to many problems in neural computing including evaluation of regression functions , determination of error bars on predictions, and the treatment of hyper-par ameters. However, the problem of model comparison is a much more challenging

one for which current techniques have significant limitations. In this paper we show how an extended form of Markov chain Monte Carlo, called chaining, is able to provide effective estimates of the relative probabilities of different models. We present results from the robot arm problem and compare them with the corresponding results obtained using the standard Gaussian approximation framework.

****

Practical Confidence and Prediction Intervals

Tom Heskes

We propose a new method to compute prediction intervals. Especially for small data sets the width of a prediction interval does not only depend on the variance of the target distribution, but also on the accuracy of our estimator of the mean of the target, i.e., on the width of the confidence interval. The confidence interval follows from the variation in an ensemble of neural networks, each of them trained and stopped on bootstrap replicates of the original data set. A second improvement is the use of the residuals on validation patterns instead of on training patterns for estimation of the variance of the target distribution. As illustrated on a synthetic example, our method is better than existing methods with regard to extrapolation and interpolation in data regimes with a limited amount of data, and yields prediction intervals which actual confidence levels are closer to the desired confidence levels.

****

Consistent Classification, Firm and Soft

Yoram Baram

A classifier is called consistent with respect to a given set of class-labeled points if it correctly classifies the set. We consider classifiers defined by unions of local separators and propose algorithms for consistent classifier reduction. The expected complexities of the proposed algorithms are derived along with the expected classifier sizes. In particular, the proposed approach yields a consistent reduction of the nearest neighbor classifier, which performs "firm" classification, assigning each new object to a class, regardless of the data structure. The proposed reduction method suggests a notion of "soft" classification, allowing for indecision with respect to objects which are insufficiently or ambiguously supported by the data. The performances of the proposed classifiers in predicting stock behavior are compared to that achieved by the nearest neighbor method.

****

Neural Models for Part-Whole Hierarchies

Maximilian Riesenhuber, Peter Dayan

We present a connectionist method for representing images that explicitly addresses their hierarchical nature. It blends data from neuroscience about whole-object viewpoint sensitive cells in inferotemporal cortex[8] and attentional basis-field modulation in V4[3] with ideas about hierarchical descriptions based on microfeatures.[5,11] The resulting model makes critical use of bottom-up and top-down pathways for analysis and synthesis.[6] We illustrate the model with a simple example of representing information about faces.

****

Bayesian Unsupervised Learning of Higher Order Structure

Michael Lewicki, Terrence J. Sejnowski

Multilayer architectures such as those used in Bayesian belief networks and Helmholtz machines provide a powerful framework for representing and learning higher order statistical relations among inputs. Because exact probability calculations with these models are often intractable, there is much interest in finding approximate algorithms. We present an algorithm that efficiently discovers higher order structure using EM and Gibbs sampling. The model can be interpreted as a stochastic recurrent network in which ambiguity in lower-level states is resolved through f

eedback from higher levels. We demonstrate the performance of the algorithm on bench(cid:173) mark problems.
****

An Architectural Mechanism for Direction-tuned Cortical Simple Cells: The Role of Mutual Inhibition
Silvio Sabatini, Fabio Solari, Giacomo Bisio
A linear architectural model of cortical simple cells is presented. The model evidences how mutual inhibition, occurring through synaptic coupling functions asymmetrically distributed in space, can be a possible basis for a wide variety of spatio-temporal simple cell response properties, including direction selectivity and velocity tuning. While spatial asymmetries are included explicitly in the structure of the inhibitory interconnections, temporal asymmetries originate from the specific mutual inhibition scheme considered. Extensive simulations supporting the model are reported.
****

Complex-Cell Responses Derived from Center-Surround Inputs: The Surprising Power of Intradendritic Computation
Bartlett Mel, Daniel Ruderman, Kevin Archie
Biophysical modeling studies have previously shown that cortical pyramidal cells driven by strong NMDA-type synaptic currents and/or containing dendritic voltage-dependent Ca++ or Na+ chan(cid:173) nels, respond more strongly when synapses are activated in several spatially clustered groups of optimal size-in comparison to the same number of synapses activated diffusely about the dendritic arbor [8]- The nonlinear intradendritic interactions giving rise to this "cluster sensitivity" property are akin to a layer of virtual non(cid:173) linear "hidden units" in the dendrites, with implications for the cel(cid:173) lular basis of learning and memory [7, 6], and for certain classes of nonlinear sensory processing [8]- In the present study, we show that a single neuron, with access only to excitatory inputs from unori(cid:173) ented ON- and OFF-center cells in the LGN, exhibits the principal nonlinear response properties of a "complex" cell in primary visual cortex, namely orientation tuning coupled with translation invari(cid:173) ance and contrast insensitivity_ We conjecture that this type of intradendritic processing could explain how complex cell responses can persist in the absence of oriented simple cell input [13]-
****

Training Algorithms for Hidden Markov Models using Entropy Based Distance Functions
Yoram Singer, Manfred K. K. Warmuth
We present new algorithms for parameter estimation of HMMs. By adapting a framework used for supervised learning, we construct iterative algorithms that maximize the likelihood of the observations while also attempting to stay "close" to the current estimated parameters. We use a bound on the relative entropy between the two HMMs as a distance mea(cid:173) sure between them. The result is new iterative training algorithms which are similar to the EM (Baum-Welch) algorithm for training HMMs. The proposed algorithms are composed of a step similar to the expectation step of Baum-Welch and a new update of the parameters which replaces the maximization (re-estimation) step. The algorithm takes only negligi(cid:173) bly more time per iteration and an approximated version uses the same expectation step as Baum-Welch. We evaluate experimentally the new algorithms on synthetic and natural speech pronunciation data. For sparse models, i.e. models with relatively small number of non-zero parameters, the proposed algorithms require significantly fewer iterations.
****

A Constructive Learning Algorithm for Discriminant Tangent Models
Diego Sona, Alessandro Sperduti, Antonina Starita
(HSS) developed an algo(cid:173)
****

Effective Training of a Neural Network Character Classifier for Word Recognition
Larry Yaeger, Richard Lyon, Brandyn Webb

We have combined an artificial neural network (ANN) character classifie r with context-driven search over character segmentation, word segmentation, a nd word recognition hypotheses to provide robust recognition of hand-pri nted English text in new models of Apple Computer's Newton MessagePad. We present some innovations in the training and use of ANNs al; character cla ssifiers for word recognition, including normalized output error, frequency ba lancing, error emphasis, negative training, and stroke warping. A recurring th eme of reducing a priori biases emerges and is discussed.

****

Second-order Learning Algorithm with Squared Penalty Term

Kazumi Saito, Ryohei Nakano

This paper compares three penalty terms with respect to the effi(cid:173) cien cy of supervised learning, by using first- and second-order learn(cid:173) ing a lgorithms. Our experiments showed that for a reasonably ade(cid:173) quate pen alty factor, the combination of the squared penalty term and the second-order learning algorithm drastically improves the convergence performance more than 20 times over the other com(cid:173) binations, at the same time bringi ng about a better generalization performance.

****

Multi-effect Decompositions for Financial Data Modeling

Lizhong Wu, John Moody

High frequency foreign exchange data can be decomposed into three comp onents: the inventory effect component, the surprise infonnation (news) com ponent and the regular infonnation component. The presence of the inventory eff ect and news can make analysis of trends due to the diffusion of infonnation (regular information component) difficult. We propose a neural-net-based, indepe ndent component analysis to sep(cid:173) arate high frequency foreign exchange d ata into these three components. Our empirical results show that our proposed m ulti-effect decomposition can reveal the intrinsic price behavior.

****

Microscopic Equations in Rough Energy Landscape for Neural Networks

K. Y. Michael Wong

We consider the microscopic equations for learning problems in neural networks. The aligning fields of an example are obtained from the cavi ty fields, which are the fields if that example were absent in the learning process. In a rough energy landscape, we assume that the den sity of the local minima obey an exponential distribution, yielding macros copic properties agreeing with the first step replica symmetry breaking solutio n. Iterating the microscopic equations provide a learning algorithm, whic h results in a higher stability than conventional algorithms.

****

Dynamic Features for Visual Speechreading: A Systematic Comparison

Michael Gray, Javier Movellan, Terrence J. Sejnowski

Humans use visual as well as auditory speech signals to recognize spoken w ords. A variety of systems have been investigated for per(cid:173) forming thi s task. The main purpose of this research was to sys(cid:173) tematically compare the performance of a range of dynamic visual features on a speechrea ding task. We have found that normal(cid:173) ization of images to elimi nate variation due to translation, scale, and planar rotation yielded su bstantial improvements in general(cid:173) ization performance regardless of the visual representation used. In addition, the dynamic information in th e difference between suc(cid:173) cessive frames yielded better performance than optical-flow based approaches, and compression by local low-pass filteri ng worked sur(cid:173) prisingly better than global principal components analysi s (PCA). These results are examined and possible explanations are explored.

****

The Effect of Correlated Input Data on the Dynamics of Learning

Søren Halkjær, Ole Winther

The convergence properties of the gradient descent algorithm in the case of t he linear perceptron may be obtained from the response function. We d

erive a general expression for the response function and apply it to the case of data with simple input correlations. It is found that correlations severely may slow down learning. This explains the success of PCA as a method for reducing training time. Motivated by this finding we furthermore propose to transform the input data by removing the mean across input variables as well as examples to decrease correlations. Numerical findings for a medical classification problem are in fine agreement with the theoretical results.

****

Learning Temporally Persistent Hierarchical Representations

Suzanna Becker

A biologically motivated model of cortical self-organization is pro(cid:173)posed. Context is combined with bottom-up information via a maximum likelihood cost function. Clusters of one or more units are modulated by a common contextual gating Signal; they thereby organize themselves into mutually supportive predictors of abstract contextual features. The model was tested in its ability to discover viewpoint-invariant classes on a set of real image sequences of cen(cid:173)tered, gradually rotating faces. It performed considerably better than supervised back-propagation at generalizing to novel views from a small number of training examples.

****

Exploiting Model Uncertainty Estimates for Safe Dynamic Control Learning

Jeff Schneider

Model learning combined with dynamic programming has been shown to be effective for learning control of continuous state dynamic systems. The simplest method assumes the learned model is correct and applies dynamic programming to it, but many approximators provide uncertainty estimates on the fit. How can they be exploited? This paper addresses the case where the system must be prevented from having catastrophic failures dur(cid:173)ing learning.

We propose a new algorithm adapted from the dual control literature and use Bayesian locally weighted regression models with dy(cid:173)namic programming. A common reinforcement learning assumption is that aggressive exploration should be encouraged. This paper addresses the con(cid:173)verse case in which the system has to reign in exploration. The algorithm is illustrated on a 4 dimensional simulated control problem.

****

Spectroscopic Detection of Cervical Pre-Cancer through Radial Basis Function Networks

Kagan Tumer, Nirmala Ramanujam, Rebecca Richards-Kortum, Joydeep Ghosh

The mortality related to cervical cancer can be substantially re(cid:173)duced through early detection and treatment. However, cur(cid:173)rent detection techniques, such as Pap smear and colposcopy, fail to achieve a concurrently high sensitivity and specificity. In vivo fluorescence spectroscopy is a technique which quickly, non(cid:173)invasively and quantitatively probes the biochemical and morpho(cid:173)logical changes that occur in pre-cancerous tissue. RBF ensemble algorithms based on such spectra provide automated, and near real(cid:173)time implementation of pre-cancer detection in the hands of non(cid:173)experts. The results are more reliable, direct and accurate than those achieved by either human experts or multivariate statistical algorithms.

****

Radial Basis Function Networks and Complexity Regularization in Function Learning

Adam Krzyzak, Tamás Linder

In this paper we apply the method of complexity regularization to de(cid:173)rive estimation bounds for nonlinear function estimation using a single hidden layer radial basis function network. Our approach differs from the previous complexity regularization neural network function learning schemes in that we operate with random covering numbers and 11 metric entropy, making it po~sibleto consider much broader families of activa(cid:173)tion functions, namely functions of bo

unded variation. Some constraints previously imposed on the network parameters a re also eliminated this way. The network is trained by means of complexity regul arization in(cid:173) volving empirical risk minimization. Bounds on the expecte d risk in tenns of the sample size are obtained for a large class of loss functi ons. Rates of convergence to the optimal loss are also derived.

****

Adaptively Growing Hierarchical Mixtures of Experts

Jürgen Fritsch, Michael Finke, Alex Waibel

We propose a novel approach to automatically growing and pruning Hierarchical Mixtures of Experts. The constructive algorithm pro(cid:173) posed here ena bles large hierarchies consisting of several hundred experts to be train ed effectively. We show that HME's trained by our automatic growing pro cedure yield better generalization per(cid:173) formance than traditional static and balanced hierarchies. Eval(cid:173) uation of the algorithm is performed (1) on vowel classification and (2) within a hybrid vers ion of the JANUS r9] speech recog(cid:173) nition system using a subset of the Switchboard large-vocabulary speaker-independent continuous speech rec ognition database.

****

On-line Policy Improvement using Monte-Carlo Search

Gerald Tesauro, Gregory Galperin

We present a Monte-Carlo simulation algorithm for real-time policy improvement of an adaptive controller. In the Monte-Carlo sim(cid:173) ulation, the long-term expected reward of each possible action is statistically measu red, using the initial policy to make decisions in each step of the simulatio n. The action maximizing the measured expected reward is then taken, resultin g in an improved policy. Our algorithm is easily parallelizable and has been i mplemented on the IBM SP! and SP2 parallel-RISC supercomputers. We have obt ained promising initial results in applying this algo(cid:173) rithm to the domain of backgammon. Results are reported for a wide variety of in itial policies, ranging from a random policy to TD-Gammon, an extremely strong multi-layer neural network. In each case, the Monte-Carlo algorithm g ives a substantial reduction, by as much as a factor of 5 or more, in t he error rate of the base players. The algorithm is also potentially u seful in many other adaptive control applications in which it is possible to simulate the environment.

****

Blind Separation of Delayed and Convolved Sources

Te-Won Lee, Anthony Bell, Russell Lambert

We address the difficult problem of separating multiple speakers with mu ltiple microphones in a real room. We combine the work of Torkkola and Am ari, Cichocki and Yang, to give Natural Gra(cid:173) dient information maxi misation rules for recurrent (IIR) networks, blindly adjusting delays, sepa rating and deconvolving mixed sig(cid:173) nals. While they work well o n simulated data, these rules fail in real rooms which usually involv e non-minimum phase transfer functions, not-invertible using stable IIR filt ers. An approach that sidesteps this problem is to perform infomax on a feedfo rward archi(cid:173) tecture in the frequency domain (Lambert 1996). We demon strate real-room separation of two natural signals using this approach.

****

A New Approach to Hybrid HMM/ANN Speech Recognition using Mutual Information Neu ral Networks

Gerhard Rigoll, Christoph Neukirchen

This paper presents a new approach to speech recognition with hybrid HMM/ANN t echnology. While the standard approach to hybrid HMMI ANN systems is b ased on the use of neural networks as posterior probability estimators, the new approach is based on the use of mutual information neural networks tra ined with a special learning algorithm in order to maximize the mutual inf ormation between the input classes of the network and its resulting sequence of firing output neurons during training. It is shown in this paper that

such a neural network is an optimal neural vector quantizer for a discrete hidden Markov model system trained on Maximum Likelihood principles. One of the main advantages of this approach is the fact, that such neural networks can be easily combined with HMM's of any complexity with context-dependent capabilities. It is shown that the resulting hybrid system achieves very high recognition rates, which are now already on the same level as the best conventional HMM systems with continuous parameters, and the capabilities of the mutual information neural networks are not yet entirely exploited.

****

Competition Among Networks Improves Committee Performance

Paul Munro, Bambang Parmanto

The separation of generalization error into two types, bias and variance (Geman, Bienenstock, Doursat, 1992), leads to the notion of error reduction by averaging over a "committee" of classifiers (Perrone, 1993). Committee perfonnance decreases with both the average error of the constituent classifiers and increases with the degree to which the misclassifications are correlated across the committee. Here, a method for reducing correlations is introduced, that uses a winner-take-all procedure similar to competitive learning to drive the individual networks to different minima in weight space with respect to the training set, such that correlations in generalization perfonnance will be reduced, thereby reducing committee error.

****

Selective Integration: A Model for Disparity Estimation

Michael Gray, Alexandre Pouget, Richard Zemel, Steven Nowlan, Terrence J. Sejnowski

Local disparity information is often sparse and noisy, which creates two conflicting demands when estimating disparity in an image re(cid:173) gion: the need to spatially average to get an accurate estimate, and the problem of not averaging over discontinuities. We have devel(cid:173) oped a network model of disparity estimation based on disparity(cid:173) selective neurons, such as those found in the early stages of process(cid:173) ing in visual cortex. The model can accurately estimate multiple disparities in a region, which may be caused by transparency or oc(cid:173) clusion, in real images and random-dot stereograms. The use of a selection mechanism to selectively integrate reliable local disparity estimates results in superior performance compared to standard back-propagation and cross-correlation approaches. In addition, the representations learned with this selection mechanism are con(cid:173) sistent with recent neurophysiological results of von der Heydt, Zhou, Friedman, and Poggio [8] for cells in cortical visual area V2. Combining multi-scale biologically-plausible image processing with the power of the mixture-of-experts learning algorithm represents a promising approach that yields both high performance and new insights into visual system function.

****

The Neurothermostat: Predictive Optimal Control of Residential Heating Systems

Michael C. Mozer, Lucky Vidmar, Robert Dodier

The Neurothermostat is an adaptive controller that regulates in(cid:173) door air temperature in a residence by switching a furnace on or off. The task is framed as an optimal control problem in which both comfort and energy costs are considered as part of the con(cid:173) trol objective. Because the consequences of control decisions are delayed in time, the Neurothermostat must anticipate heating de(cid:173) mands with predictive models of occupancy patterns and the ther(cid:173) mal response of the house and furnace. Occupancy pattern predic(cid:173) tion is achieved by a hybrid neural net / look-up table. The Neu(cid:173) rothermostat searches, at each discrete time step, for a decision sequence that minimizes the expected cost over a fixed planning horizon. The first decision in this sequence is taken, and this pro(cid:173) cess repeats. Simulations of the Neurothermostat were conducted using artificial occupancy data in which regularity was systemat(cid:173) ically varied, as well as o

ccupancy data from an actual residence. The Neurothermostat is compared ag ainst three conventional poli(cid:173) cies, and achieves reliably lower c osts. This result is robust to the relative weighting of comfort and ene rgy costs and the degree of variability in the occupancy patterns.
****

Softening Discrete Relaxation
Andrew Finch, Richard Wilson, Edwin Hancock
This paper describes a new framework for relational graph match(cid:173) ing. The starting point is a recently reported Bayesian consistency measure which gauges structural differences using Hamming dis(cid:173) tance. The main contributions of the work are threefold. Firstly, we demonstrate how th e discrete components of the cost func(cid:173) tion can be softened. The second contribution is to show how the softened cost function can be used to locate matches using continuous non-linear optimisation. Fin ally, we show how the res(cid:173) ulting graph matching algorithm relates to the standard quadratic assignment problem.
****

A Comparison between Neural Networks and other Statistical Techniques for Modeli ng the Relationship between Tobacco and Alcohol and Cancer
Tony Plate, Pierre Band, Joel Bert, John Grace
Epidemiological data is traditionally analyzed with very simple techniqu es. Flexible models, such as neural networks, have the potential to dis cover unanticipated features in the data. However, to be useful, flexible mo dels must have effective control on overfit(cid:173) ting. This paper reports on a comparative study of the predictive quality of neural networks and other flexible models applied to real and artificial epidemiological data. The resul ts suggest that there are no major unanticipated complex features in the real data, and also demonstrate that MacKay's [1995] Bayesian neural network methodology provides effective control on overfitting while retain(cid:173) ing the ability to discover complex features in the artificial data.
****

An Apobayesian Relative of Winnow
Nick Littlestone, Chris Mesterharm
We study a mistake-driven variant of an on-line Bayesian learn(cid:173) ing algorithm (similar to one studied by Cesa-Bianchi, Helmbold, and Pan izza [CHP96]). This variant only updates its state (learns) on trials in which it makes a mistake. The algorithm makes binary classifications using a linea r-threshold classifier and runs in time lin(cid:173) ear in the number of attr ibutes seen by the learner. We have been able to show, theoretically and in simulations, that this algorithm performs well under assumptions quit e different from those embod(cid:173) ied in the prior of the original Bay esian algorithm. It can handle situations that we do not know how to han dle in linear time with Bayesian algorithms. We expect our techniques to be useful in deriving and analyzing other apobayesian algorithms.
****

LSTM can Solve Hard Long Time Lag Problems
Sepp Hochreiter, Jürgen Schmidhuber
Standard recurrent nets cannot deal with long minimal time lags between relevant signals. Several recent NIPS papers propose alter(cid:173) native methods. We f irst show: problems used to promote various previous algorithms can be solved mo re quickly by random weight guessing than by the proposed algorithms. We then us e LSTM, our own recent algorithm, to solve a hard problem that can neither be qu ickly solved by random search nor by any other recurrent net algorithm we are aw are of.
****

488 Solutions to the XOR Problem
Frans Coetzee, Virginia Stonick
A globally convergent homotopy method is defined that is capable of sequenti ally producing large numbers of stationary points of the multi-layer perceptron mean-squared error surface. Using this al(cid:173) gorithm large subsets o

f the stationary points of two test problems are found. It is shown empiric ally that the MLP neural network appears to have an extreme ratio of saddle points compared to local minima, and that even small neural net work problems have extremely large numbers of solutions.
****

A Mixture of Experts Classifier with Learning Based on Both Labelled and Unlabel led Data
David J. Miller, Hasan Uyar
We address statistical classifier design given a mixed training set co n(cid:173) sisting of a small labelled feature set and a (generally la rger) set of unlabelled features. This situation arises, e.g., for medica l images, where although training features may be plentiful, expensive expertise is re(cid:173) quired to extract their class labels. We propo se a classifier structure and learning algorithm that make effective use of unlabelled data to im(cid:173) prove performance. The learning is based o n maximization of the total data likelihood, i.e. over both the labell ed and unlabelled data sub(cid:173) sets. Two distinct EM learning algorit hms are proposed, differing in the EM formalism applied for unlabelled data. The classifier, based on a joint probability model for features and labels, is a "mixture of experts" structure that is equivalent to the rad ial basis function (RBF) classifier, but unlike RBFs, is amenable to likeli hood-based training. The scope of application for the new method is grea tly extended by the observation that test data, or any new data to classify , is in fact additional, unlabelled data - thus, a combined learning/classif ication operation - much akin to what is done in image segmentation - ca n be invoked whenever there is new data to classify. Experiments with data sets from the UC Irvine database demonstrate that the new learning algorithms and structure achieve substantial performance gains over alter native approaches.
****

Statistical Mechanics of the Mixture of Experts
Kukjin Kang, Jong-Hoon Oh
We study generalization capability of the mixture of experts learn(cid:173) ing from examples generated by another network with the same architecture . When the number of examples is smaller than a crit(cid:173) ical value, th e network shows a symmetric phase where the role of the experts is no t specialized. Upon crossing the critical point, the system undergoes a continuous phase transition to a symme(cid:173) try breaking phase wher e the gating network partitions the input space effectively and each exper t is assigned to an appropriate sub(cid:173) space. We also find that the mi xture of experts with multiple level of hierarchy shows multiple phase transi tions.
****

Predicting Lifetimes in Dynamically Allocated Memory
David Cohn, Satinder Singh
Predictions oflifetimes of dynamically allocated objects can be used to improve time and space efficiency of dynamic memory manage(cid:173) ment in computer p rograms. Barrett and Zorn [1993] used a simple lifetime predictor and demons trated this improvement on a variety of computer programs. In this paper, w e use decision trees to do lifetime prediction on the same programs and show significantly better prediction . Our method also has the advantag e that during training we can use a large number of features and let the de cision tree automatically choose the relevant subset.
****

Reinforcement Learning for Mixed Open-loop and Closed-loop Control
Eric Hansen, Andrew Barto, Shlomo Zilberstein
Closed-loop control relies on sensory feedback that is usually as(cid:173) 1 73) sumed to be free . But if sensing incurs a cost, it may be cost(cid:173) effective to take sequences of actions in open-loop mode. We d e(cid:173) scribe a reinforcement learning algorithm that learns to comb

ine open-loop and closed-loop control when sensing incurs a cost. Al(cid:173) though we assume reliable sensors, use of open-loop control means that actions must sometimes be taken when the current state of the controlled system is uncertain. This is a special case of the hidden-state problem in reinforcement learning, and to cope, our algorithm relies on short-term memory. The main result of the pa(cid:173)per is a rule that significantly limits exploration of possible memory states by pruning memory states for which the estimated value of information is greater than its cost. We prove that this rule allows convergence to an optimal policy.

****

Computing with Infinite Networks

Christopher Williams

For neural networks with a wide class of weight-priors, it can be shown that in the limit of an infinite number of hidden units the prior over functions tends to a Gaussian process. In this paper an(cid:173)alytic forms are derived for the covariance function of the Gaussian processes corresponding to networks with sigmoidal and Gaussian hidden units. This allows predictions to be made efficiently using networks with an infinite number of hidden units, and shows that, somewhat paradoxically, it may be easier to compute with infinite networks than finite ones.

****

Regression with Input-Dependent Noise: A Bayesian Treatment

Christopher Bishop, Cazhaow Quazaz

In most treatments of the regression problem it is assumed that the distribution of target data can be described by a deterministic function of the inputs, together with additive Gaussian noise hav(cid:173)ing constant variance. The use of maximum likelihood to train such models then corresponds to the minimization of a sum-of-squares error function. In many applications a more realistic model would allow the noise variance itself to depend on the input variables. However, the use of maximum likelihood to train such models would give highly biased results. In this paper we show how a Bayesian treatment can allow for an input-dependent variance while over(cid:173)coming the bias of maximum likelihood.

****

The Generalisation Cost of RAMnets

Richard Rohwer, Michal Morciniec

Given unlimited computational resources, it is best to use a crite(cid:173)rion of minimal expected generalisation error to select a model and determine its parameters. However, it may be worthwhile to sac(cid:173)rifice some generalisation performance for higher learning speed. A method for quantifying sub-optimality is set out here, so that this choice can be made intelligently. Furthermore, the method is applicable to a broad class of models, including the ultra-fast memory-based methods such as RAMnets. This brings the added benefit of providing, for the first time, the means to analyse the generalisation properties of such models in a Bayesian framework .

****

Multilayer Neural Networks: One or Two Hidden Layers?

Graham Brightwell, Claire Kenyon, Hélène Paugam-Moisy

We study the number of hidden layers required by a multilayer neu(cid:173)ral network with threshold units to compute a function f from n d to {O, I}. In dimension d = 2, Gibson characterized the functions computable with just one hidden layer, under the assumption that there is no "multiple intersection point" and that f is only defined on a compact set. We consider the restriction of f to the neighbor(cid:173)hood of a multiple intersection point or of infinity, and give neces(cid:173)sary and sufficient conditions for it to be locally computable with one hidden layer. We show that adding these conditions to Gib(cid:173)son's assumptions is not sufficient to ensure global computability with one hidden layer, by exhibiting a new non

-local configuration, the "critical cycle", which implies that f is not computable with one hidden layer.
****

Improving the Accuracy and Speed of Support Vector Machines
Christopher J. C. Burges, Bernhard Schölkopf
Support Vector Learning Machines (SVM) are finding application in patte rn recognition, regression estimation , and operator inver(cid:173) sion fo r ill-posed problems. Against this very general backdrop , any methods f or improving the generalization performance, or for improving the speed i n test phase, of SVMs are of increasing in(cid:173) terest. In this paper we combine two such techniques on a pattern recognition problem. The metho d for improving generalization per(cid:173) formance (the "virtual support ve ctor" method) does so by incor(cid:173) porating known invariances of the p roblem. This method achieves a drop in the error rate on 10,000 NIST tes t digit images of 1.4% to 1.0%. The method for improving the speed (the " reduced set" method) does so by approximating the support vector decision sur(cid:173) face. We apply this method to achieve a factor of fifty speedup in test phase over the virtual support vector machine. The combined approa ch yields a machine which is both 22 times faster than the original mac hine, and which has better generalization performance, achieving 1.1 % error. The virtual support vector method is appli(cid:173) cable to any SVM proble m with known invariances. The reduced set method is applicable to any su pport vector machine.
****

Adaptive Access Control Applied to Ethernet Data
Timothy Brown
This paper presents a method that decides which combinations of traffic can be accepted on a packet data link, so that quality of service (QoS) constraints can be met. The method uses samples of QoS results at dif(cid:173) ferent load conditions to build a neural network decision function. Pre(cid:173) vious si milar approaches to the problem have a significant bias. This bias is l ikely to occur in any real system and results in accepting loads that miss QoS targets by orders of magnitude. Preprocessing the data to either remove the bias or provide a confidence level, the method was applied to sources based on difficult-to-analyze ethernet data traces. With this data, the m ethod produces an accurate access control function that dramatically outperfo rms analytic alternatives. Interestingly, the results depend on throwing aw ay more than 99% of the data.
****

An Adaptive WTA using Floating Gate Technology
W. Kruger, Paul Hasler, Bradley Minch, Christof Koch
We have designed, fabricated, and tested an adaptive Winner(cid:173) Tak e-All (WTA) circuit based upon the classic WTA of Lazzaro, et al [IJ. We have added a time dimension (adaptation) to this circuit to make t he input derivative an important factor in winner selection. To accomplish t his, we have modified the classic WTA circuit by adding floating gate transistors which slowly null their inputs over time. We present a simpli fied analysis and experimen(cid:173) tal data of this adaptive WTA fabricated in a standard CMOS 2f.tm process.
****

Representation and Induction of Finite State Machines using Time-Delay Neural Ne tworks
Daniel Clouse, C. Giles, Bill Horne, Garrison Cottrell
This work investigates the representational and inductive capabili(cid:173) ties of time-delay neural networks (TDNNs) in general, and of two subclasses of TD NN, those with delays only on the inputs (IDNN), and those which include dela ys on hidden units (HDNN) . Both ar(cid:173) chitectures are capable of represe nting the same class of languages, the definite memory machine (DMM) langua ges, but the delays on the hidden units in the HDNN helps it outperform the IDNN on problems composed of repeated features over short time windo

ws.
****

Probabilistic Interpretation of Population Codes
Richard Zemel, Peter Dayan, Alexandre Pouget
We present a theoretical framework for population codes which generaliz es naturally to the important case where the population provides infor mation about a whole probability distribution over an underlying quanti ty rather than just a single value. We use the framework to analyze tw o existing models, and to suggest and evaluate a third model for encoding su ch probability distributions.
****

Analog VLSI Circuits for Attention-Based, Visual Tracking
Timothy Horiuchi, Tonia Morris, Christof Koch, Stephen DeWeerth
A one-dimensional visual tracking chip has been implemented us(cid:173) ing neur omorphic, analog VLSI techniques to model selective visual attention in the con trol of saccadic and smooth pursuit eye move(cid:173) ments. The chip incorporat es focal-plane processing to compute image saliency and a winner-take-all circu it to select a feature for tracking. The target position and direction of motio n are reported as the target moves across the array. We demonstrate its functio n(cid:173) ality in a closed-loop system which performs saccadic and smooth pur suit tracking movements using a one-dimensional mechanical eye.
****

Online Learning from Finite Training Sets: An Analytical Case Study
Peter Sollich, David Barber
We analyse online learning from finite training sets at non(cid:173) in finitesimal learning rates TJ. By an extension of statistical me(cid:173) chanics methods, we obtain exact results for the time-dependent gener alization error of a linear network with a large number of weights N . We find, for example, that for small training sets of size p ~ N, larger learning rates can be used without compromis(cid:173) ing asymptot ic generalization performance or convergence speed. Encouragingly, for o ptimal settings of TJ (and, less importantly, weight decay ,) at given f inal learning time, the generalization per(cid:173) formance of online learning is essentially as good as that of offline learning.
****

Spatiotemporal Coupling and Scaling of Natural Images and Human Visual Sensitivi ties
Dawei Dong
We study the spatiotemporal correlation in natural time-varying images a nd explore the hypothesis that the visual system is con(cid:173) cerned with the optimal coding of visual representation through spatiotemporal decorrelation of the input signal. Based on the measured spatiotemporal power spectrum, the transform needed to decorrelate input signal is derived ana lytically and then compared with the actual processing observed in psychophysic al experiments.
****

Using Curvature Information for Fast Stochastic Search
Genevieve Orr, Todd Leen
We present an algorithm for fast stochastic gradient descent that uses a nonlinear adaptive momentum scheme to optimize the late time convergence rate. The algorithm makes effective use of cur(cid:173) vature information, requi res only O(n) storage and computation, and delivers convergence rates cl ose to the theoretical optimum. We demonstrate the technique on linear and l arge nonlinear back(cid:173) prop networks.
****

ARC-LH: A New Adaptive Resampling Algorithm for Improving ANN Classifiers
Friedrich Leisch, Kurt Hornik
We introduce arc-Ih, a new algorithm for improvement of ANN clas(cid:173) sifie r performance, which measures the importance of patterns by aggregated ne twork output errors. On several artificial benchmark problems, this algorit

hm compares favorably with other resample and combine techniques.
****

Estimating Equivalent Kernels for Neural Networks: A Data Perturbation Approach
A. Burgess

We describe the notion of "equivalent kernels" and suggest that this provides a framework for comparing different classes of regression models, including neural networks and both parametric and non-parametric statistical techniques. Unfortunately, standard techniques break down when faced with models, such as neural networks, in which there is more than one "layer" of adjustable parameters. We propose an algorithm which overcomes this limitation, estimating the equivalent kernels for neural network models using a data perturbation approach. Experimental results indicate that the networks do not use the maximum possible number of degrees of freedom, that these can be controlled using regularisation techniques and that the equivalent kernels learnt by the network vary both in "size" and in "shape" in different regions of the input space.
****

Monotonicity Hints
Joseph Sill, Yaser Abu-Mostafa

A hint is any piece of side information about the target function to be learned. We consider the monotonicity hint, which states that the function to be learned is monotonic in some or all of the input variables. The application of mono tonicity hints is demonstrated on two real-world problems- a credit card application task, and a problem in medical diagnosis. A measure of the monotonicity error of a candidate function is defined and an objective function for the enforcement of monotonicity is derived from Bayesian principles. We report experimental results which show that using monotonicity hints leads to a statistically significant improvement in performance on both problems.
****

A Convergence Proof for the Softassign Quadratic Assignment Algorithm
Anand Rangarajan, Alan L. Yuille, Steven Gold, Eric Mjolsness

The softassign quadratic assignment algorithm has recently emerged as an effective strategy for a variety of optimization prob(cid:173) lems in pattern recognition and combinatorial optimization. While the effectiveness of the algorithm was demonstrated in thousands of simulations, there was no known proof of convergence. Here, we provide a proof of convergence for the most general form of the algorithm.
****

Removing Noise in On-Line Search using Adaptive Batch Sizes
Genevieve Orr

Stochastic (on-line) learning can be faster than batch learning. However, at late times, the learning rate must be annealed to re(cid:173) move the noise present in the stochastic weight updates. In this annealing phase, the convergence rate (in mean square) is at best proportional to l/T where T is the number of input presentations. An alternative is to increase the batch size to remove the noise. In this paper we explore convergence for LMS using 1) small but fixed batch sizes and 2) an adaptive batch size. We show that the best adaptive batch schedule is exponential and has a rate of conver(cid:173) gence which is the same as for annealing, Le., at best proportional to l/T.
****

Size of Multilayer Networks for Exact Learning: Analytic Approach
André Elisseeff, Hélène Paugam-Moisy

This article presents a new result about the size of a multilayer neural network computing real outputs for exact learning of a finite set of real samples. The architecture of the network is feedforward, with one hidden layer and several outputs. Starting from a fixed training set, we consider the network as a function of its weights. We derive, for a wide family of transfer functions, a lower and an upper bound on the number o

f hidden units for exact learning, given the size of the dataset and t
he dimensions of the input and output spaces.
****

Support Vector Regression Machines
Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, Vladimir Va
pnik
A new regression technique based on Vapnik's concept of support vector
s is introduced. We compare support vector regression (SVR) with a com
mittee regression technique (bagging) based on regression trees and ridg
e regression done in feature space. On the basis of these experiments, it
is expected that SVR will have advantages in high dimensionality space
because SVR optimization does not depend on the dimensionality of the input sp
ace.
****

A Hierarchical Model of Visual Rivalry
Peter Dayan
Binocular rivalry is the alternating percept that can result when the two eyes
see different scenes. Recent psychophysical evidence supports an account for
one component of binocular rivalry similar to that for other bistable percepts.
We test the hypothesisl9, 16, 18 that alternation can be generated by com
petition between top(cid:173) down cortical explanations for the inputs, rat
her than by direct competition between the inputs. Recent neurophysiological
ev(cid:173) idence shows that some binocular neurons are modulated with th
e changing percept; others are not, even if they are selective be(cid:173) tween
the stimuli presented to the eyes. We extend our model to a hierarchy to add
ress these effects.
****

Dynamically Adaptable CMOS Winner-Take-All Neural Network
Kunihiko Iizuka, Masayuki Miyamoto, Hirofumi Matsui
The major problem that has prevented practical application of analog neu
ro-LSIs has been poor accuracy due to fluctuating analog device charac
teristics inherent in each device as a result of manufacturing. This p
aper proposes a dynamic control architecture that allows analog silicon neura
l networks to compensate for the fluctuating device characteristics and
adapt to a change in input DC level. We have applied this architec
ture to compensate for input offset voltages of an analog CMOS WTA (Winn
er-Take-AlI) chip that we have fabricated. Experimental data show the effec
tiveness of the architecture.
****

Ensemble Methods for Phoneme Classification
Steve Waterhouse, Gary Cook
This paper investigates a number of ensemble methods for improv(cid:173) ing
the performance of phoneme classification for use in a speech recognition
system. Two ensemble methods are described; boosting and mixtures of experts,
both in isolation and in combination. Re(cid:173) sults are presented on two
speech recognition databases: an isolated word database and a large vocabulary
continuous speech database. These results show that principled ensemble method
s such as boost(cid:173) ing and mixtures provide superior performance to mo
re naive en(cid:173) semble methods such as averaging.
****

Maximum Likelihood Blind Source Separation: A Context-Sensitive Generalization o
f ICA
Barak Pearlmutter, Lucas Parra
In the square linear blind source separation problem, one must find a linear
unmixing operator which can detangle the result Xi(t) of mixing n unknown
independent sources 8i(t) through an unknown n x n mixing matrix A(t) of
causal linear filters: Xi = E j aij * 8 j . We cast the problem as one of
maximum likelihood density estima(cid:173) tion, and in that framework intro
duce an algorithm that searches for independent components using both tempor
al and spatial cues. We call the resulting algorithm "Contextual ICA," after

the (Bell and Sejnowski 1995) Infomax algorithm, which we show to be a special case of cICA. Because cICA can make use of the temporal structure of its input, it is able separate in a number of situations where standard methods cannot, including sources with low kur(cid:173) tosis, colored Gaussian sources, and sources which have Gaussian histograms.

****

Temporal Low-Order Statistics of Natural Sounds

Hagai Attias, Christoph Schreiner

In order to process incoming sounds efficiently, it is advantageous for the auditory system to be adapted to the statistical structure of natural auditory scenes. As a first step in investigating the relation between the system and its inputs, we study low-order statistical properties in several sound ensembles using a filter bank analysis. Focusing on the amplitude and phase in different frequency bands, we find simple parametric descriptions for their distribution and power spectrum that are valid for very different types of sounds. In particular, the amplitude distribution has an exponential tail and its power spectrum exhibits a modified power-law behavior, which is manifested by self-similarity and long-range temporal cor(cid:173) relations. Furthermore, the statistics for different bands within a given ensemble are virtually identical, suggesting translation in(cid:173) variance along the cochlear axis. These results show that natural sounds are highly redundant, and have possible implications to the neural code used by the auditory system.

****

Limitations of Self-organizing Maps for Vector Quantization and Multidimensional Scaling

Arthur Flexer

The limitations of using self-organizing maps (SaM) for either clustering/vector quantization (VQ) or multidimensional scaling (MDS) are being discussed by reviewing recent empirical findings and the relevant theory. SaM 's remaining ability of doing both VQ and MDS at the same time is challenged by a new combined tech(cid:173) nique of online K-means clustering plus Sammon mapping of the cluster centroids. SaM are shown to perform significantly worse in terms of quantization error , in recovering the structure of the clus(cid:173) ters and in preserving the topology in a comprehensive empirical study using a series of multivariate normal clustering problems.

****

A Spike Based Learning Neuron in Analog VLSI

Philipp Häfliger, Misha Mahowald, Lloyd Watts

Many popular learning rules are formulated in terms of continu(cid:173) ous, analog inputs and outputs. Biological systems, however, use action potentials, which are digital-amplitude events that encode analog information in the inter-event interval. Action-potential representations are now being used to advantage in neuromorphic VLSI systems as well. We report on a simple learning rule, based on the Riccati equation described by Kohonen [1], modified for action-potential neuronal outputs. We demonstrate this learning rule in an analog VLSI chip that uses volatile capacitive storage for synaptic weights. We show that our time-dependent learning rule is sufficient to achieve approximate weight normalization and can detect temporal correlations in spike trains.

****

One-unit Learning Rules for Independent Component Analysis

Aapo Hyvärinen, Erkki Oja

Neural one-unit learning rules for the problem of Independent Com(cid:173) ponent Analysis (ICA) and blind source separation are introduced. In these new algorithms, every ICA neuron develops into a sepa(cid:173) rator that finds one of the independent components. The learning rules use very simple constrained Hebbianjanti-Hebbian learning in which decorrelating feedback may be added. To speed up the convergence of these stochastic gradient de

scent rules, a novel com(cid:173) putationally efficient fixed-point algorithm is introduced.

****

Analysis of Temporal-Diffference Learning with Function Approximation
John Tsitsiklis, Benjamin Van Roy

We present new results about the temporal-difference learning al(cid:173) go rithm, as applied to approximating the cost-to-go function of a Markov chain using linear function approximators. The algo(cid:173) rithm we a nalyze performs on-line updating of a parameter vector during a single endles s trajectory of an aperiodic irreducible finite state Markov chain. Results i nclude convergence (with probability 1), a characterization of the limit of convergence, and a bound on the resulting approximation error. In addition t o establishing new and stronger results than those previously available, our analysis is based on a new line of reasoning that provides new intuition about the dynamics of temporal-difference learning. Furthermore, w e discuss the implications of two counter-examples with regards to the Si gnificance of on-line updating and linearly parameterized function approxima tors.

****

Fast Network Pruning and Feature Extraction by using the Unit-OBS Algorithm
Achim Stahlberger, Martin Riedmiller

The algorithm described in this article is based on the OBS algo(cid:1 73) rithm by Hassibi, Stork and Wolff ([1] and [2]). The main disad(ci d:173) vantage of OBS is its high complexity. OBS needs to calculate the in verse Hessian to delete only one weight (thus needing much time to prune a big net) . A better algorithm should use this matrix to remove more than only one weight , because calculating the inverse Hessian takes the m ost time in the OBS algorithm. The algorithm, called Unit- OBS, describe d in this article is a method to overcome this disadvantage. This algorit hm only needs to calculate the inverse Hessian once to remove one whole unit t hus drastically reducing the time to prune big nets. A further advantage of Unit- OBS is that it can be used to do a feature extraction on the input data. This can be helpful on the understanding of unknown pr oblems.

****

Learning with Noise and Regularizers in Multilayer Neural Networks
David Saad, Sara Solla
Sara A. Solla

****

The Learning Dynamcis of a Universal Approximator
Ansgar West, David Saad, Ian Nabney

The learning properties of a universal approximator, a normalized committee m achine with adjustable biases, are studied for on-line back-propagation lea rning. Within a statistical mechanics frame(cid:173) work, numerical stud ies show that this model has features which do not exist in previously s tudied two-layer network models with(cid:173) out adjustable biases, e.g., attr active suboptimal symmetric phases even for realizable cases and noiseless dat a.

****

Gaussian Processes for Bayesian Classification via Hybrid Monte Carlo
David Barber, Christopher Williams

The full Bayesian method for applying neural networks to a pre(cid:173) diction problem is to set up the prior/hyperprior structure for the net and then perform the necessary integrals. However, these inte(cid:173) gral s are not tractable analytically, and Markov Chain Monte Carlo (MCMC) methods are slow, especially if the parameter space is high-dimensional. Using Gaussian processes we can approximate the weight space integral analytical ly, so that only a small number of hyperparameters need be integrated over by MCMC methods. We have applied this idea to classification problems, obtaining ex(cid:173) cellent results on the real-world problems investigate

d so far .
****

Genetic Algorithms and Explicit Search Statistics
Shumeet Baluja
The genetic algorithm (GA) is a heuristic search procedure based on mechanisms abstracted from population genetics. In a previous paper [Baluja & Caruana, 1995], we showed that much simpler algorithms, such as hillcIimbing and Population(cid:173) Based Incremental Learning (PBIL), perform comparably to GAs on an optimiza(cid:173) tion problem custom designed to benefit from the GA's operators. This paper extends these results in two directions. First, in a large-scale empirical comparison of problems that have been reported in GA literature, we show that on many prob(cid:173) lems, simpler algorithms can perform significantly better than GAs. Second, we describe when crossover is useful, and show how it can be incorporated into PBIL.
****

Learning Exact Patterns of Quasi-synchronization among Spiking Neurons from Data on Multi-unit Recordings
Laura Martignon, Kathryn Laskey, Gustavo Deco, Eilon Vaadia
This paper develops arguments for a family of temporal log-linear models to represent spatio-temporal correlations among the spiking events in a group of neurons. The models can represent not just pairwise correlations but also correlations of higher order. Methods are discussed for inferring the existence or absence of correlations and estimating their strength. A frequentist and a Bayesian approach to correlation detection are compared. The frequentist method is based on G 2 statistic with estimates obtained via the Max-Ent principle. In the Bayesian approach a Markov Chain Monte Carlo Model Composition (MC3) algorithm is applied to search over connectivity structures and Laplace's method is used to approximate their posterior probability. Performance of the methods was tested on synthetic data. The methods were applied to experimental data obtained by the fourth author by means of measurements carried out on behaving Rhesus monkeys at the Hadassah Medical School of the Hebrew University. As conjectured, neural connectivity structures need not be neither hierarchical nor decomposable.
****

Efficient Nonlinear Control with Actor-Tutor Architecture
Kenji Doya
A new reinforcement learning architecture for nonlinear control is proposed. A direct feedback controller, or the actor, is trained by a value-gradient based controller, or the tutor. This architecture enables both efficient use of the value function and simple computa(cid:173) tion for real-time implementation. Good performance was verified in multi-dimensional nonlinear control tasks using Gaussian soft(cid:173) max networks.
****

Interpolating Earth-science Data using RBF Networks and Mixtures of Experts
Ernest Wan, Don Bone
We present a mixture of experts (ME) approach to interpolate sparse, spatially correlated earth-science data. Kriging is an interpolation method which uses a global covariation model estimated from the data to take account of the spatial dependence in the data. Based on the close relationship between kriging and the radial basis function (RBF) network (Wan & Bone, 1996), we use a mixture of generalized RBF networks to partition the input space into statistically correlated regions and learn the local covariation model of the data in each region. Applying the ME approach to simulated and real-world data, we show that it is able to achieve good partitioning of the input space, learn the local covariation models and improve generalization.
****

Statistically Efficient Estimations Using Cortical Lateral Connections
Alexandre Pouget, Kechen Zhang
Coarse codes are widely used throughout the brain to encode sen(cid:173) sory and motor variables. Methods designed to interpret these codes,

such as population vector analysis, are either inefficient, i.e., the variance of the estimate is much larger than the smallest possi(cid:173) ble variance, or biologically implausible, like maximum likelihood. Moreover, these methods attempt to compute a scalar or vector estimate of the encoded variable. Neurons are faced with a simi(cid:173) lar estimation problem . They must read out the responses of the presynaptic neurons, but, by contrast, they typically encode the variable with a further population code rather than as a scalar. We show how a non-linear recurrent network can be used to per(cid:173) form these estimation in an optimal way while keeping the estimate in a coarse code format. This work suggests that lateral connec(cid:173) tions in the cortex may be involved in cleaning up uncorrelated noise among neurons representing similar variables.

****

Recursive Algorithms for Approximating Probabilities in Graphical Models

Tommi Jaakkola, Michael Jordan

We develop a recursive node-elimination formalism for efficiently approximating large probabilistic networks. No constraints are set on the network topologies. Yet the formalism can be straightfor(cid:173) wardly integrated with exact methods whenever they are/become applicable. The approximations we use are controlled: they main(cid:173) tain consistently upper and lower bounds on the desired quantities at all times. We show that Boltzmann machines, sigmoid belief networks, or any combination (i.e., chain graphs) can be handled within the same framework. The accuracy of the methods is veri(cid:173) fied experimentally.

****

Clustering via Concave Minimization

Paul Bradley, Olvi Mangasarian, W. Street

The problem of assigning m points in the n-dimensional real space Rn to k clusters is formulated as that of determining k centers in Rn such that the sum of distances of each point to the nearest center is minimized. If a polyhedral distance is used, the problem can be formulated as that of minimizing a piecewise-linear concave function on a polyhedral set which is shown to be equivalent to a bilinear program: minimizing a bilinear function on a polyhe(cid:173) dral set. A fast finite k-Median Algorithm consisting of solving few linear programs in closed form leads to a stationary point of the bilinear program. Computational testing on a number of real(cid:173) world databases was carried out. On the Wisconsin Diagnostic Breast Cancer (WDBC) database, k-Median training set correct(cid:173) ness was comparable to that of the k-Mean Algorithm, however its testing set correctness was better. Additionally, on the Wisconsin Prognostic Breast Cancer (WPBC) database, distinct and clini(cid:173) cally important survival curves were extracted by the k-Median Algorithm, whereas the k-Mean Algorithm failed to obtain such distinct survival curves for the same database.

****

Balancing Between Bagging and Bumping

Tom Heskes

We compare different methods to combine predictions from neu(cid:173) ral networks trained on different bootstrap samples of a regression problem. One of these methods, introduced in [6] and which we here call balancing, is based on the analysis of the ensemble gen(cid:173) eralization error into an ambiguity term and a term incorporating generalization performances of individual networks. We show how to estimate these individual errors from the residuals on valida(cid:173) tion patterns. Weighting factors for the different networks follow from a quadratic programming problem. On a real-world problem concerning the prediction of sales figures and on the well-known Boston housing data set, balancing clearly outperforms other re(cid:173) cently proposed alternatives as bagging [1] and bumping [8].

****

Self-Organizing and Adaptive Algorithms for Generalized Eigen-Decomposition
Chanchal Chatterjee, Vwani Roychowdhury
The paper is developed in two parts where we discuss a new approach to self-organization in a single-layer linear feed-forward network. First, two novel algorithms for self-organization are derived from a two-layer linear hetero-associative network performing a one-of-m classification, and trained with the constrained least-mean-squared classification error criterion. Second, two adaptive algorithms are derived from these self(cid:173) organizing procedures the principal generalized eigenvectors of two correlation matrices from two sequences of random vectors. These novel adaptive algorithms can be implemented in a single-layer linear feed-forward network. We give a rigorous convergence analysis of the adaptive algorithms by using stochastic approximation theory. As an example, we consider a problem of online signal detection in digital mobile communications.
****

Multi-Grid Methods for Reinforcement Learning in Controlled Diffusion Processes
Stephan Pareigis
Reinforcement learning methods for discrete and semi-Markov de(cid:173) cision problems such as Real-Time Dynamic Programming can be generalized for Controlled Diffusion Processes. The optimal control problem reduces to a boundary value problem for a fully nonlinear second-order elliptic differential equation of Hamilton(cid:173) Jacobi-Bellman (HJB-) type. Numerical analysis provides multi(cid:173) grid methods for this kind of equation. In the case of Learning Con(cid:173) trol, however, the systems of equations on the various grid-levels are obtained using observed information (transitions and local cost). To ensure consistency, special attention needs to be directed to(cid:173) ward the type of time and space discretization during the obser(cid:173) vation. An algorithm for multi-grid observation is proposed. The multi-grid algorithm is demonstrated on a simple queuing problem.
****

Noisy Spiking Neurons with Temporal Coding have more Computational Power than Sigmoidal Neurons
Wolfgang Maass
We exhibit a novel way of simulating sigmoidal neural nets by net(cid:173) works of noisy spiking neurons in temporal coding. Furthermore it is shown that networks of noisy spiking neurons with temporal coding have a strictly larger computational power than sigmoidal neural nets with the same number of units.
****

Edges are the 'Independent Components' of Natural Scenes.
Anthony Bell, Terrence J. Sejnowski
Field (1994) has suggested that neurons with line and edge selectivities found in primary visual cortex of cats and monkeys form a sparse, dis(cid:173) tributed representation of natural scenes, and Barlow (1989) has reasoned that such responses should emerge from an unsupervised learning algorithm that attempts to find a factorial code of independent visual features. We show here that non-linear 'infomax', when applied to an ensemble of nat(cid:173) ural scenes, produces sets of visual filters that are localised and oriented. Some of these filters are Gabor-like and resemble those produced by the sparseness-maximisation network of Olshausen & Field (1996). In addition, the outputs of these filters are as independent as possible, since the info(cid:173) max network is able to perform Independent Components Analysis (ICA). We compare the resulting ICA filters and their associated basis functions, with other decorrelating filters produced by Principal Components Analysis (PCA) and zero-phase whitening filters (ZCA). The ICA filters have more sparsely distributed (kurtotic) outputs on natural scenes. They also resem(cid:173) ble the receptive fields of simple cells in visual cortex, which suggests that these neurons form an information-theoretic co-ordinate system for images.

****
For Valid Generalization the Size of the Weights is More Important than the Size of the Network
Peter Bartlett
This paper shows that if a large neural network is used for a pattern classification problem, and the learning algorithm finds a network with small weights that has small squared error on the training patterns, then the generalization performance depends on the size of the weights rather than the number of weights. More specifi(cid:173)cally, consider an i-layer feed-forward network of sigmoid units, in which the sum of the magnitudes of the weights associated with each unit is bounded by A. The misclassification probability con(cid:173)verges to an error estimate (that is closely related to squared error on the training set) at rate O((cA)l(l+1)/2J(log n)jm) ignoring log factors, where m is the number of training patterns, n is the input dimension, and c is a constant. This may explain the gen(cid:173)eralization performance of neural networks, particularly when the number of training examples is considerably smaller than the num(cid:173)ber of weights. It also supports heuristics (such as weight decay and early stopping) that attempt to keep the weights small during training.
****
Neural Network Modeling of Speech and Music Signals
Alex Röbel
Time series prediction is one of the major applications of neural net(cid:173)works. After a short introduction into the basic theoretical foundations we argue that the iterated prediction of a dynamical system may be in(cid:173)terpreted as a model of the system dynamics. By means of RBF neural networks we describe a modeling approach and extend it to be able to model instationary systems. As a practical test for the capabilities of the method we investigate the modeling of musical and speech signals and demonstrate that the model may be used for synthesis of musical and speech signals.
****
Rapid Visual Processing using Spike Asynchrony
Simon Thorpe, Jacques Gautrais
We have investigated the possibility that rapid processing in the visual system could be achieved by using the order of firing in different neurones as a code, rather than more conventional firing rate schemes. Using SPIKENET, a neural net simulator based on integrate-and-fire neurones and in which neurones in the input layer function as analog(cid:173)to-delay converters, we have modeled the initial stages of visual processing. Initial results are extremely promising. Even with activity in retinal output cells limited to one spike per neuron per image (effectively ruling out any form of rate coding), sophisticated processing based on asynchronous activation was nonetheless possible.
****
Visual Cortex Circuitry and Orientation Tuning
Trevor Mundel, Alexander Dimitrov, Jack Cowan
A simple mathematical model for the large-scale circuitry of pri(cid:173)mary visual cortex is introduced. It is shown that a basic cor(cid:173)tical architecture of recurrent local excitation and lateral inhi(cid:173)bition can account quantitatively for such properties as orien(cid:173)tation tuning. The model can also account for such local ef(cid:173)fects as cross-orientation suppression. It is also shown that non(cid:173)local state-dependent coupling between similar orientation patches, when added to the model, can satisfactorily reproduce such ef(cid:173)fects as non-local iso--orientation suppression, and non-local cross(cid:173)orientation enhancement. Following this an account is given of per(cid:173)ceptual phenomena involving object segmentation, such as "pop(cid:173)out", and the direct and indirect tilt illusions.
****
Orientation Contrast Sensitivity from Long-range Interactions in Visual Cortex

Klaus Pawelzik, Udo Ernst, Fred Wolf, Theo Geisel
Recently Sill ito and coworkers (Nature 378, pp. 492,1995) demon(cid:173) strated that stimulation beyond the classical receptive field (cRF) can not only modulate, but radically change a neuron's response to oriented stimuli. They revealed that patch-suppressed cells when stimulated with contrasting orientations inside and outside their cRF can strongly respond to stimuli oriented orthogonal to their nominal preferred orientation. Here we analyze the emergence of such complex response patterns in a simple model of primary vi(cid:173) sual cortex. We show that the observed sensitivity for orientation contrast can be explained by a delicate interplay between local isotropic interactions and patchy long-range connectivity between distant iso-orientation domains. In particular we demonstrate that the observed properties might arise without specific connections be(cid:173) tween sites with cross-oriented cRFs.
****