Regular and Irregular Gallager-zype Error-Correcting Codes
Yoshiyuki Kabashima, Tatsuto Murayama, David Saad, Renato Vicente
The performance of regular and irregular Gallager-type error(cid:173) cor recting code is investigated via methods of statistical physics. The tran smitted codeword comprises products of the original mes(cid:173) sage bits sel ected by two randomly-constructed sparse matrices; the number of non-ze ro row/column elements in these matrices constitutes a family of codes. We show that Shannon's channel capacity may be saturated in equilibrium for many of the regular codes while slightly lower performance is obtained fo r others which may be of higher practical relevance. Decoding aspects ar e con(cid:173) sidered by employing the TAP approach which is identical t o the commonly used belief-propagation-based decoding. We show that irregul ar codes may saturate Shannon's capacity but with improved dynamical properties .
****
An MEG Study of Response Latency and Variability in the Human Visual System Duri ng a Visual-Motor Integration Task
Akaysha Tang, Barak Pearlmutter, Tim Hely, Michael Zibulevsky, Michael Weisend
Human reaction times during sensory-motor tasks vary consider(cid:173) ab ly. To begin to understand how this variability arises, we exam(cid:173) ine d neuronal populational response time variability at early versus late visual processing stages. The conventional view is that pre(cid:173) cise tempo ral information is gradually lost as information is passed through a layered n etwork of mean-rate "units." We tested in hu(cid:173) mans whether neuronal populations at different processing stages behave like mean-rate "units". A blind source separation algorithm was applied to MEG signals from sensory-motor integration tasks. Response time latency and variability for multip le visual sources were estimated by detecting single-trial stimulus-locked e vents for each source. In two subjects tested on four visual reaction time tasks, we reliably identified sources belonging to early and late vi(c id:173) sual processing stages. The standard deviation of response latency was smaller for early rather than late processing stages. This sup(cid:173) ports the hypothesis that human populational response time vari(cid:173) ability in creases from early to late visual processing stages.
****
The Entropy Regularization Information Criterion
Alex Smola, John Shawe-Taylor, Bernhard Schölkopf, Robert C. Williamson
Effective methods of capacity control via uniform convergence bounds for functi on expansions have been largely limited to Support Vector ma(cid:173) chines, w here good bounds are obtainable by the entropy number ap(cid:173) proach . We extend these methods to systems with expansions in terms of arbitrary (par ametrized) basis functions and a wide range of regulariza(cid:173) tion methods covering the whole range of general linear additive models. This is achieved b y a data dependent analysis of the eigenvalues of the corresponding design mat rix.
****
Invariant Feature Extraction and Classification in Kernel Spaces
Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, Alex Smola, Kla us-Robert Müller
In hyperspectral imagery one pixel typically consists of a mixture of the r eflectance spectra of several materials, where the mixture coefficients c orrespond to the abundances of the constituting ma(cid:173) terials. We a ssume linear combinations of reflectance spectra with some additive normal sens or noise and derive a probabilistic MAP framework for analyzing hyperspectr al data. As the material re(cid:173) flectance characteristics are not know a priori, we face the problem of unsupervised linear unmixing. The incor poration of different prior information (e.g. positivity and normalizati on of the abun(cid:173) dances) naturally leads to a family of interest ing algorithms, for example in the noise-free case yielding an algorit hm that can be understood as constrained independent component analysis (ICA

). Simulations underline the usefulness of our theory.
****
Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology
Yair Weiss, William Freeman
Local "belief propagation" rules of the sort proposed by Pearl [15] are guaranteed to converge to the correct posterior probabilities in singly connected graphical models. Recently, a number of researchers have em(cid:173) pirically demonstrated good performance of "loopy belief propagation"(cid:173) using these same rules on graphs with loops. Perhaps the most dramatic instance is the near Shannon-limit performance of "Turbo codes", whose decoding algorithm is equivalent to loopy belief propagation. Except for the case of graphs with a single loop, there has been little theo(cid:173) retical understanding of the performance of loopy propagation. Here we analyze belief propagation in networks with arbitrary topologies when the nodes in the graph describe jointly Gaussian random variables. We give an analytical formula relating the true posterior probabilities with those calculated using loopy propagation. We give sufficient conditions for convergence and show that when belief propagation converges it gives the correct posterior means for all graph topologies, not just networks with a single loop. The related "max-product" belief propagation algorithm finds the max(cid:173) imum posterior probability estimate for singly connected networks. We show that, even for non-Gaussian probability distributions, the conver(cid:173) gence points of the max-product algorithm in loopy networks are max(cid:173) ima over a particular large local neighborhood of the posterior proba(cid:173) bility. These results help clarify the empirical performance results and motivate using the powerful belief propagation algorithm in a broader class of networks.
****
Population Decoding Based on an Unfaithful Model
Si Wu, Hiroyuki Nakahara, Noboru Murata, Shun-ichi Amari
We study a population decoding paradigm in which the maximum likeli(cid:173) hood inference is based on an unfaithful decoding model (UMLI). This is usually the case for neural population decoding because the encoding process of the brain is not exactly known, or because a simplified de(cid:173) coding model is preferred for saving computational cost. We consider an unfaithful decoding model which neglects the pair-wise correlation between neuronal activities, and prove that UMLI is asymptotically effi(cid:173) cient when the neuronal correlation is uniform or of limited-range. The performance of UMLI is compared with that of the maximum likelihood inference based on a faithful model and that of the center of mass de(cid:173) coding method. It turns out that UMLI has advantages of decreasing the computational complexity remarkablely and maintaining a high-level decoding accuracy at the same time. The effect of correlation on the decoding accuracy is also discussed.
****
Broadband Direction-Of-Arrival Estimation Based on Second Order Statistics
Justinian Rosca, Joseph Ruanaidh, Alexander Jourjine, Scott Rickard
N wideband sources recorded using N closely spaced receivers can feasibly be separated based only on second order statistics when using a physical model of the mixing process. In this case we show that the parameter estimation problem can be essentially reduced to considering directions of arrival and attenuations of each signal. The paper presents two demixing methods operating in the time and frequency domain and experimentally shows that it is always possible to demix signals arriving at different angles. Moreover, one can use spatial cues to solve the channel selection problem and a post-processing Wiener filter to ameliorate the artifacts caused by demixing.
****
Emergence of Topography and Complex Cell Properties from Natural Images using Extensions of ICA
Aapo Hyvärinen, Patrik Hoyer
Independent component analysis of natural images leads to emer(cid:173) gence

of simple cell properties, Le. linear filters that resemble wavelets or Gabor functions. In this paper, we extend ICA to explain further properties of VI cells. First, we decompose natural images into independent subspaces instead of scalar components. This model leads to emergence of phase and shift invariant fea(cid:173) tures, similar to those in VI complex cells. Second, we define a topography between the linear components obtained by ICA. The topographic distance between two components is defined by their higher-order correlations, so that two components are close to each other in the topography if they are strongly dependent on each other. This leads to simultaneous emergence of both topography and invariances similar to complex cell properties.
****
Reinforcement Learning Using Approximate Belief States
Andres Rodriguez, Ronald Parr, Daphne Koller
The problem of developing good policies for partially observable Markov decision problems (POMDPs) remains one of the most challenging ar(cid:173) eas of research in stochastic planning. One line of research in this area involves the use of reinforcement learning with belief states, probabil(cid:173) ity distributions over the underlying model states. This is a promis(cid:173) ing method for small problems, but its application is limited by the in(cid:173) tractability of computing or representing a full belief state for large prob(cid:173) lems. Recent work shows that, in many settings, we can maintain an approximate belief state, which is fairly close to the true belief state. In particular, great success has been shown with approximate belief states that marginalize out correlations between state variables. In this paper, we investigate two methods of full belief state reinforcement learning and one novel method for reinforcement learning using factored approximate belief states. We compare the performance of these algorithms on several well-known problem from the literature. Our results demonstrate the im(cid:173) portance of approximate belief state representations for large problems.
****
Building Predictive Models from Fractal Representations of Symbolic Sequences
Peter Tiño, Georg Dorffner
We propose a novel approach for building finite memory predictive mod(cid:173) els similar in spirit to variable memory length Markov models (VLMMs). The models are constructed by first transforming the n-block structure of the training sequence into a spatial structure of points in a unit hypercube, such that the longer is the common suffix shared by any two n-blocks, the closer lie their point representations. Such a transformation embodies a Markov assumption - n-blocks with long common suffixes are likely to produce similar continuations. Finding a set of prediction contexts is formulated as a resource allocation problem solved by vector quantizing the spatial n-block representation. We compare our model with both the classical and variable memory length Markov models on three data sets with different memory and stochastic components. Our models have a superior performance, yet, their construction is fully automatic, which is shown to be problematic in the case of VLMMs.
****
Neural Computation with Winner-Take-All as the Only Nonlinear Operation
Wolfgang Maass
Everybody "knows" that neural networks need more than a single layer of nonlinear units to compute interesting functions. We show that this is false if one employs winner-take-all as nonlinear unit:
****
Support Vector Method for Multivariate Density Estimation
Vladimir Vapnik, Sayan Mukherjee
A new method for multivariate density estimation is developed based on the Support Vector Method (SVM) solution of inverse ill-posed problems. The solution has the form of a mixture of den(cid:173) sities. This method with Gaussian kernels compared favorably to both Parzen's method and the Gaussian Mixture Model method. For synthetic data we achieve more accu

rate estimates for densities of 2, 6, 12, and 40 dimensions.
****

Leveraged Vector Machines
Yoram Singer
We describe an iterative algorithm for building vector machines used in classif ication tasks. The algorithm builds on ideas from support vector machines, boosting, and generalized additive models. The algorithm can be used with va rious continuously differential functions that bound the discrete (0-1) classi fication loss and is very simple to implement. We test the proposed algorithm w ith two different loss functions on synthetic and natural data. We also describ e a norm-penalized version of the algorithm for the exponential loss function u sed in AdaBoost. The performance of the algorithm on natural data is compara ble to support vector machines while typically its running time is shorter tha n of SVM.
****

Learning Factored Representations for Partially Observable Markov Decision Proce sses
Brian Sallans
The problem of reinforcement learning in a non-Markov environment is explored u sing a dynamic Bayesian network, where conditional indepen(cid:173) dence assump tions between random variables are compactly represented by network parameters. The parameters are learned on-line, and approx(cid:173) imations are used to p erform inference and to compute the optimal value function. The relative eff ects of inference and value function approxi(cid:173) mations on the quality of the final policy are investigated, by learning to solve a moderately diffi cult driving task. The two value function approx(cid:173) imations, linear and quadratic, were found to perform similarly, but the quadratic model was more s ensitive to initialization. Both performed be(cid:173) low the level of human p erformance on the task. The dynamic Bayesian network performed comparably to a model using a localist hidden state representation, while requiring exponent ially fewer parameters.
****

Variational Inference for Bayesian Mixtures of Factor Analysers
Zoubin Ghahramani, Matthew Beal
We present an algorithm that infers the model structure of a mix(cid:173) ture of factor analysers using an efficient and deterministic varia(cid:173) ti onal approximation to full Bayesian integration over model pa(cid:173) ra meters. This procedure can automatically determine the opti(cid:173) mal number of components and the local dimensionality of each component (Le . the number of factors in each factor analyser) . Alternatively it c an be used to infer posterior distributions over number of components an d dimensionalities. Since all parameters are integrated out the method is n ot prone to overfitting. Using a stochastic procedure for adding components it is possible to per(cid:173) form the variational optimisation increment ally and to avoid local maxima. Results show that the method works very well in practice and correctly infers the number and dimensionality of nontr ivial synthetic examples. By importance sampling from the variational ap proximation we show how to obtain unbiased estimates of the true evide nce, the exact predictive density, and the KL divergence between the varia(ci d:173) tional posterior and the true posterior, not only in this model but fo r variational approximations in general.
****

Topographic Transformation as a Discrete Latent Variable
Nebojsa Jojic, Brendan J. Frey
Invariance to topographic transformations such as translation and shearing in an image has been successfully incorporated into feed(cid:173) forward mechani sms, e.g., "convolutional neural networks", "tan(cid:173) gent propagation". We describe a way to add transformation invari(cid:173) ance to a generative d ensity model by approximating the nonlinear transformation manifold by a dis crete set of transformations. An EM algorithm for the original model can

be extended to the new model by computing expectations over the set of trans formations. We show how to add a discrete transformation variable to Gaussian mixture modeling, factor analysis and mixtures of factor analysis. We give results on filtering microscopy images, face and facial pose clustering, and handwritten digit modeling and recognition.
****

Channel Noise in Excitable Neural Membranes
Amit Manwani, Peter Steinmetz, Christof Koch
Stochastic fluctuations of voltage-gated ion channels generate current an d voltage noise in neuronal membranes. This noise may be a criti(cid: 173) cal determinant of the efficacy of information processing within neural systems. Using Monte-Carlo simulations, we carry out a systematic in(cid:173) v estigation of the relationship between channel kinetics and the result(cid:1 73) ing membrane voltage noise using a stochastic Markov version of the Mainen-Sejnowski model of dendritic excitability in cortical neurons. Ou r simulations show that kinetic parameters which lead to an increase in memb rane excitability (increasing channel densities, decreasing tem(cid:173) peratur e) also lead to an increase in the magnitude of the sub-threshold voltage noise. Noise also increases as the membrane is depolarized from rest towards th reshold. This suggests that channel fluctuations may in(cid:173) terfere wit h a neuron's ability to function as an integrator of its synaptic inputs and ma y limit the reliability and precision of neural information processing.
****

Efficient Approaches to Gaussian Process Classification
Lehel Csató, Ernest Fokoué, Manfred Opper, Bernhard Schottky, Ole Winther
We present three simple approximations for the calculation of the post erior mean in Gaussian Process classification. The first two methods a re related to mean field ideas known in Statistical Physics. The third a pproach is based on Bayesian online approach which was motivated by recent res ults in the Statistical Mechanics of Neural Networks. We present simulation re sults showing: 1. that the mean field Bayesian evidence may be used for hype rparameter tuning and 2. that the online approach may achieve a low training error fast.
****

Optimal Sizes of Dendritic and Axonal Arbors
Dmitri Chklovskii
I consider a topographic projection between two neuronal layers with dif(cid:173) ferent densities of neurons. Given the number of output neurons con(cid:17 3) nected to each input neuron (divergence or fan-out) and the number of input neurons synapsing on each output neuron (convergence or fan-in) I determine the widths of axonal and dendritic arbors which minimize the total volume ofaxons and dendrites. My analytical results can be sum(cid:173) marized qualitatively in the following rule: neurons of the sparser layer should have arbors wider than those of the denser layer. This agrees with the anatomical data from reti nal and cerebellar neurons whose morphol(cid:173) ogy and connectivity are known . The rule may be used to infer connec(cid:173) tivity of neurons from their mo rphology.
****

v-Arc: Ensemble Learning in the Presence of Outliers
Gunnar Rätsch, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Takashi Onod a, Sebastian Mika
AdaBoost and other ensemble methods have successfully been ap(cid:173) plied t o a number of classification tasks, seemingly defying prob(cid:173) lems o f overfitting. AdaBoost performs gradient descent in an error function with respect to the margin, asymptotically concentrating on the patterns which are hardest to learn. For very noisy prob(cid:173) lems, however, this c an be disadvantageous. Indeed, theoretical analysis has shown that the marg in distribution, as opposed to just the minimal margin, plays a crucial role i n understanding this phe(cid:173) nomenon. Loosely speaking, some outliers should be tolerated if this has the benefit of substantially increasing

the margin on the remaining points. We propose a new boosting algorith m which al(cid:173) lows for the possibility of a pre-specified fraction of p oints to lie in the margin area Or even on the wrong side of the decision boun dary.
****
Monte Carlo POMDPs
Sebastian Thrun
We present a Monte Carlo algorithm for learning to act in partially observ able Markov decision processes (POMDPs) with real-valued state and action space s. Our approach uses importance sampling for representing beliefs, and Monte Ca rlo approximation for belief propagation. A reinforcement learning algorithm, value iteration, is employed to learn value functions over belief states. Final ly, a sample(cid:173) based version of nearest neighbor is used to genera lize across states. Initial empirical results suggest that our approach work s well in practical applications.
****
A Recurrent Model of the Interaction Between Prefrontal and Inferotemporal Corte x in Delay Tasks
Alfonso Renart, Néstor Parga, Edmund Rolls
A very simple model of two reciprocally connected attractor neural net(cid:173) works is studied analytically in situations similar to those encountered in del ay match-to-sample tasks with intervening stimuli and in tasks of memory guided attention. The model qualitatively reproduces many of the experimental data on these types of tasks and provides a framework for the understanding of the exp erimental observations in the context of the attractor neural network scenario.
****
Information Factorization in Connectionist Models of Perception
Javier Movellan, James McClelland
We examine a psychophysical law that describes the influence of stimul us and context on perception. According to this law choice probability rat ios factorize into components independently con(cid:173) trolled by stimulu s and context. It has been argued that this pat(cid:173) tern of results is i ncompatible with feedback models of perception. In this paper we examine this claim using neural network models defined via stochastic differential equat ions. We show that the law is related to a condition named channel separabi lity and has little to do with the existence of feedback connections. In essen ce, chan(cid:173) nels are separable if they converge into the response units wi thout direct lateral connections to other channels and if their sensors are no t directly contaminated by external inputs to the other chan(cid:173) n els. Implications of the analysis for cognitive and computational neurosicenc e are discussed.
****
Hierarchical Image Probability (H1P) Models
Clay Spence, Lucas Parra
We formulate a model for probability distributions on image spaces. We show th at any distribution of images can be factored exactly into condi(cid:173) tional distributions of feature vectors at one resolution (pyramid level) conditi oned on the image information at lower resolutions. We would like to factor this over positions in the pyramid levels to make it tractable, but such facto ring may miss long-range dependencies. To fix this, we in(cid:173) troduce hidde n class labels at each pixel in the pyramid. The result is a hierarchi cal mixture of conditional probabilities, similar to a hidden Markov model on a tree. The model parameters can be found with max(cid:173) imum likelihood estimation using the EM algorithm. We have obtained encouraging preliminary r esults on the problems of detecting various ob(cid:173) jects in SAR images and target recognition in optical aerial images.
****
Reinforcement Learning for Spoken Dialogue Systems
Satinder Singh, Michael Kearns, Diane Litman, Marilyn Walker
Recently, a number of authors have proposed treating dialogue systems as Markov

decision processes (MDPs). However, the practical application ofMDP algorithms to dialogue systems faces a number of severe technical challenges. We have buil t a general software tool (RLDS, for Reinforcement Learning for Dialogue System s) based on the MDP framework, and have applied it to dialogue corpora gathered from two dialogue systems built at AT&T Labs. Our experiments demonstrate that RLDS holds promise as a tool for "browsing" and understanding correlations in complex, temporally dependent dialogue corpora.
****
Distributed Synchrony of Spiking Neurons in a Hebbian Cell Assembly
David Horn, Nir Levy, Isaac Meilijson, Eytan Ruppin
We investigate the behavior of a Hebbian cell assembly of spiking neurons f ormed via a temporal synaptic learning curve. This learn(cid:173) ing function is based on recent experimental findings . It includes potentiation for short time delays between pre- and post-synaptic neuronal spiking, and de pression for spiking events occuring in the reverse order. The coupling betw een the dynamics of the synaptic learning and of the neuronal activation leads to interesting results. We find that the cell assembly can fire asynchr onously, but may also function in complete synchrony, or in distribute d synchrony. The latter implies spontaneous division of the Hebbian cell asse m(cid:173) bly into groups of cells that fire in a cyclic manner. We inveti gate the behavior of distributed synchrony both by simulations and by analyti c calculations of the resulting synaptic distributions.
****
Image Representations for Facial Expression Coding
Marian Bartlett, Gianluca Donato, Javier Movellan, Joseph Hager, Paul Ekman, Ter rence J. Sejnowski
The Facial Action Coding System (FACS) (9) is an objective method for quantifying facial movement in terms of component actions. This syste m is widely used in behavioral investigations of emotion, cognitive pro cesses, and social interaction. The cod(cid:173) ing is presently perform ed by highly trained human experts. This paper explores and compares tech niques for automatically recog(cid:173) nizing facial actions in sequences of images. These methods include unsupervised learning techniques for finding basis images such as principal component analysis, independent component a nalysis and local feature analysis, and supervised learning techniques suc h as Fisher's linear discriminants. These data-driven bases are com(cid :173) pared to Gabor wavelets, in which the basis images are predefined. Best performances were obtained using the Gabor wavelet repre(cid:173) sentati on and the independent component representation, both of which achieved 96% accuracy for classifying 12 facial actions. The ICA representation employ s 2 orders of magnitude fewer basis im(cid:173) ages than the Gabor represent ation and takes 90% less CPU time to compute for new images. The results pr ovide converging support for using local basis images, high spatial frequencies , and statistical independence for classifying facial actions.
****
Algorithms for Independent Components Analysis and Higher Order Statistics
Daniel Lee, Uri Rokni, Haim Sompolinsky
A latent variable generative model with finite noise is used to de(ci d:173) scribe several different algorithms for Independent Components Anal (cid:173) ysis (lCA). In particular, the Fixed Point ICA algorithm is shown to be equivalent to the Expectation-Maximization algorithm for maximum likelihood under certain constraints, allowing the conditions for global convergence to be elucidated. The algorithms can also be explained by thei r generic behavior near a singular point where the size of the opti(cid:1 73) mal generative bases vanishes. An expansion of the likelihood about this s ingular point indicates the role of higher order correlations in determin(cid:1 73) ing the features discovered by ICA. The application and convergence of the se algorithms are demonstrated on a simple illustrative example.
****
A SNoW-Based Face Detector

Ming-Hsuan Yang, Dan Roth, Narendra Ahuja
A novel learning approach for human face detection using a network of linear units is presented. The SNoW learning architecture is a sparse network of linear functions over a pre-defined or incremen(cid:173) tally learned feature space and is specifically tailored for learning in the presence of a very large number of features. A wide range of face images in different poses, with different expressions and under different lighting conditions are used as a training set to capture the variations of human faces. Experimental results on commonly used benchmark data sets of a wide range of face images show that the SNoW-based approach outperforms methods that use neural networks, Bayesian methods, support vector machines and oth(cid:173) ers. Furthermore, learning and evaluation using the SNoW-based method are significantly more efficient than with other methods.
****

A Winner-Take-All Circuit with Controllable Soft Max Property
Shih-Chii Liu
I describe a silicon network consisting of a group of excitatory neu(cid:173) rons and a global inhibitory neuron. The output of the inhibitory neuron is normalized with respect to the input strengths. This out(cid:173) put models the normalization property of the wide-field direction(cid:173) selective cells in the fly visual system. This normalizing property is also useful in any system where we wish the output signal to code only the strength of the inputs, and not be dependent on the num(cid:173) ber of inputs. The circuitry in each neuron is equivalent to that in Lazzaro's winner-take-all (WTA) circuit with one additional tran(cid:173) sistor and a voltage reference. Just as in Lazzaro's circuit, the outputs of the excitatory neurons code the neuron with the largest input. The difference here is that multiple winners can be chosen. By varying the voltage reference of the neuron, the network can transition between a soft-max behavior and a hard WTA behav(cid:173) ior. I show results from a fabricated chip of 20 neurons in a 1.2J.Lm CMOS technology.
****

Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers
Matthias Seeger
We present a variational Bayesian method for model selection over families of kernels classifiers like Support Vector machines or Gaus(cid:173) sian processes. The algorithm needs no user interaction and is able to adapt a large number of kernel parameters to given data without having to sacrifice training cases for validation. This opens the pos(cid:173) sibility to use sophisticated families of kernels in situations where the small "standard kernel" classes are clearly inappropriate. We relate the method to other work done on Gaussian processes and clarify the relation between Support Vector machines and certain Gaussian process models.
****

Kirchoff Law Markov Fields for Analog Circuit Design
Richard Golden
Three contributions to developing an algorithm for assisting engi(cid:173) neers in designing analog circuits are provided in this paper. First, a method for representing highly nonlinear and non-continuous analog circuits using Kirchoff current law potential functions within the context of a Markov field is described. Second, a relatively effi(cid:173) cient algorithm for optimizing the Markov field objective function is briefly described and the convergence proof is briefly sketched. And third, empirical results illustrating the strengths and limita(cid:173) tions of the approach are provided within the context of a JFET transistor design problem. The proposed algorithm generated a set of circuit components for the JFET circuit model that accurately generated the desired characteristic curves.
****

Can VI Mechanisms Account for Figure-Ground and Medial Axis Effects?

Zhaoping Li
When a visual image consists of a figure against a background, V1 cells are ph
ysiologically observed to give higher responses to image regions corresponding
  to the figure relative to their responses to the background. The med
ial axis of the figure also induces rela(cid:173) tively higher responses
  compared to responses to other locations in the figure (except for the
boundary between the figure and the background). Since the receptive fields
 of V1 cells are very smal(cid:173) l compared with the global scale of th
e figure-ground and medial axis effects, it has been suggested that these eff
ects may be caused by feedback from higher visual areas. I show how these ef
fects can be accounted for by V1 mechanisms when the size of the figu
re is small or is of a certain scale. They are a manifestation of the
 processes of pre-attentive segmentation which detect and highlight the boundar
ies between homogeneous image regions.
****
Policy Gradient Methods for Reinforcement Learning with Function Approximation
Richard S. Sutton, David McAllester, Satinder Singh, Yishay Mansour
Function approximation is essential to reinforcement learning, but the s
tandard approach of approximating a value function and deter(cid:173) mining a
  policy from it has so far proven theoretically intractable. In this pap
er we explore an alternative approach in which the policy is explicitly represe
nted by its own function approximator, indepen(cid:173) dent of the value funct
ion, and is updated according to the gradient of expected reward with respect
 to the policy parameters. Williams's REINFORCE method and actor-critic method
s are examples of this approach. Our main new result is to show that t
he gradient can be written in a form suitable for estimation from ex
perience aided by an approximate action-value or advantage function. Us
ing this result, we prove for the first time that a version of policy i
teration with arbitrary differentiable function approximation is convergent to
 a locally optimal policy.
****
Lower Bounds on the Complexity of Approximating Continuous Functions by Sigmoida
l Neural Networks
Michael Schmitt
We calculate lower bounds on the size of sigmoidal neural networks that approx
imate continuous functions. In particular, we show that for the approx
imation of polynomials the network size has to grow as O((logk)1/4) whe
re k is the degree of the polynomials. This bound is valid for any input d
imension, i.e. independently of the number of variables. The result is obt
ained by introducing a new method employing upper bounds on the Vapnik-Chervo
nenkis dimension for proving lower bounds on the size of networks that
 approximate continuous functions.
****
Evolving Learnable Languages
Bradley Tonkes, Alan Blair, Janet Wiles
Recent theories suggest that language acquisition is assisted by the evolution o
f languages towards forms that are easily learnable. In this paper, we evolve co
mbinatorial languages which can be learned by a recurrent neural network quickly
 and from relatively few ex(cid:173) amples. Additionally, we evolve languages f
or generalization in different "worlds", and for generalization from specific ex
amples. We find that languages can be evolved to facilitate different forms of i
mpressive generalization for a minimally biased, general pur(cid:173) pose learn
er. The results provide empirical support for the theory that the language itsel
f, as well as the language environment of a learner, plays a substantial role in
 learning: that there is far more to language acquisition than the language acqu
isition device.
****
Large Margin DAGs for Multiclass Classification
John Platt, Nello Cristianini, John Shawe-Taylor
We present a new learning architecture: the Decision Directed Acyclic G

raph (DDAG), which is used to combine many two-class classifiers into a multiclass classifier. For an N -class problem, the DDAG con(cid:173) tains N(N - 1)/2 classifiers, one for each pair of classes. We present a VC analysis of the case when the node classifiers are hyperplanes; the re(cid:173) sulting bound on the test error depends on N and on the margin achieved at th e nodes, but not on the dimension of the space. This motivates an algor ithm, DAGSVM, which operates in a kernel-induced feature space and uses two -class maximal margin hyperplanes at each decision-node of the DDAG. The D AGSVM is substantially faster to train and evalu(cid:173) ate than eithe r the standard algorithm or Max Wins, while maintaining comparable accuracy to both of these algorithms.
****
Approximate Planning in Large POMDPs via Reusable Trajectories
Michael Kearns, Yishay Mansour, Andrew Ng
We consider the problem of reliably choosing a near-best strategy from a restri cted class of strategies TI in a partially observable Markov deci(cid:173) si on process (POMDP). We assume we are given the ability to simulate the POMDP, and study what might be called the sample complexity - that is, the amount o f data one must generate in the POMDP in order to choose a good strategy. We prove upper bounds on the sample com(cid:173) plexity showing that, eve n for infinitely large and arbitrarily complex POMDPs, the amount of da ta needed can be finite, and depends only linearly on the complexity of the restricted strategy class TI, and expo(cid:173) nentially on the horiz on time. This latter dependence can be eased in a variety of ways, includin g the application of gradient and local search algorithms. Our measure of complexity generalizes the classical super(cid:173) vised learning notion of V C dimension to the settings of reinforcement learning and planning.
****
Maximum Entropy Discrimination
Tommi Jaakkola, Marina Meila, Tony Jebara
We present a general framework for discriminative estimation based on the maxim um entropy principle and its extensions. All calcula(cid:173) tions involve di stributions over structures and/or parameters rather than specific settings and reduce to relative entropy projections. This holds even when the da ta is not separable within the chosen parametric class, in the context of anomaly detection rather than classification, or when the labels in the train ing set are uncertain or incomplete. Support vector machines are naturally subsumed un(cid:173) der this class and we provide several extensions. We are also able to estimate exactly and efficiently discriminative distributio ns over tree structures of class-conditional models within this framework. Preliminary experimental results are indicative of the potential in these tec hniques.
****
The Relaxed Online Maximum Margin Algorithm
Yi Li, Philip Long
We describe a new incremental algorithm for training linear thresh(cid: 173) old functions: the Relaxed Online Maximum Margin Algorithm, or ROM MA. ROMMA can be viewed as an approximation to the algorithm that repeatedly c hooses the hyperplane that classifies previously seen ex(cid:173) amples correc tly with the maximum margin. It is known that such a maximum-margin h ypothesis can be computed by minimizing the length of the weight vector subject to a number of linear constraints. ROMMA works by maintaining a relatively simple relaxation of these constraints that can be efficiently updated. We pro ve a mistake bound for ROMMA that is the same as that proved for the perceptron algorithm. Our analysis implies that the more computationally intensive maxim um-margin algo(cid:173) rithm also satisfies this mistake bound; this is the fir st worst-case perfor(cid:173) mance guarantee for this algorithm. We describe s ome experiments us(cid:173) ing ROMMA and a variant that updates its hypothesis more aggressively as batch algorithms to recognize handwritten digits. The computational complexity and simplicity of these algorithms is similar to t

hat of per(cid:173) ceptron algorithm , but their generalization is much better. We describe a sense in which the performance of ROMMA converges to that of SVM in the limit if bias isn't considered.
****

Bayesian Modelling of fMRI lime Series
Pedro Højen-Sørensen, Lars Hansen, Carl Rasmussen
We present a Hidden Markov Model (HMM) for inferring the hidden psychological state (or neural activity) during single trial tMRI activa(cid:173) tion experiments with blocked task paradigms. Inference is based on Bayesian methodology, using a combination of analytical and a variety of Markov Chain Monte Carlo (MCMC) sampling techniques. The ad(cid:173) vantage of this method is that detection of short time learning effects be(cid:173) tween repeated trials is possible since inference is based only on single trial experiments.
****

Bayesian Averaging is Well-Temperated
Lars Hansen
Bayesian predictions are stochastic just like predictions of any other inference scheme that generalize from a finite sample. While a sim(cid:173) ple variational argument shows that Bayes averaging is generaliza(cid:173) tion optimal given that the prior matches the teacher parameter distribution the situation is less clear if the teacher distribution is unknown. I define a class of averaging procedures, the temperated likelihoods, including both Bayes averaging with a uniform prior and maximum likelihood estimation as special cases. I show that Bayes is generalization optimal in this family for any teacher dis(cid:173) tribution for two learning problems that are analytically tractable: learning the mean of a Gaussian and asymptotics of smooth learn(cid:173) ers.
****

Policy Search via Density Estimation
Andrew Ng, Ronald Parr, Daphne Koller
We propose a new approach to the problem of searching a space of stochastic controllers for a Markov decision process (MDP) or a partially observable Markov decision process (POMDP). Following several other authors, our approach is based on searching in parameterized families of policies (for example, via gradient descent) to optimize solution qual(cid:173) ity. However, rather than trying to estimate the values and derivatives of a policy directly, we do so indirectly using estimates for the proba(cid:173) bility densities that the policy induces on states at the different points in time. This enables our algorithms to exploit the many techniques for efficient and robust approximate density propagation in stochastic sys(cid:173) tems. We show how our techniques can be applied both to deterministic propagation schemes (where the MDP's dynamics are given explicitly in compact form,) and to stochastic propagation schemes (where we have access only to a generative model, or simulator, of the MDP). We present empirical results for both of these variants on complex problems.
****

Low Power Wireless Communication via Reinforcement Learning
Timothy Brown
This paper examines the application of reinforcement learning to a wire(cid:173) less communication problem. The problem requires that channel util(cid:173) ity be maximized while simultaneously minimizing battery usage. We present a solution to this multi-criteria problem that is able to signifi(cid:173) cantly reduce power consumption. The solution uses a variable discount factor to capture the effects of battery usage.
****

Learning to Parse Images
Geoffrey E. Hinton, Zoubin Ghahramani, Yee Whye Teh
We describe a class of probabilistic models that we call credibility networks. Using parse trees as internal representations of images, credibility networks are able to perform segmentation and recog(cid:173) nition simultane

ously, removing the need for ad hoc segmentation heuristics. Promising resu lts in the problem of segmenting hand(cid:173) written digits were obtained .
****
Robust Recognition of Noisy and Superimposed Patterns via Selective Attention
Soo-Young Lee, Michael C. Mozer
In many classification tasks, recognition accuracy is low because input pattern s are corrupted by noise or are spatially or temporally overlapping. W e propose an approach to overcoming these limitations based on a model of human selective attention. The model, an early selection filter guided by top-down attentional control, entertains each candidate output class in se quence and adjusts attentional gain coefficients in order to produce a st rong response for that class. The chosen class is then the one that obtains the strongest response with the least modulation of attention. We present simulation results on classification of corrupted and superimposed handwritt en digit patterns, showing a significant improvement in recognition rates. The algorithm has also been applied in the domain of speech recognition, with comparable results.
****
Bayesian Network Induction via Local Neighborhoods
Dimitris Margaritis, Sebastian Thrun
In recent years, Bayesian networks have become highly successful tool for di(cid:173) agnosis, analysis, and decision making in real-world doma ins. We present an efficient algorithm for learning Bayes networks from d ata. Our approach con(cid:173) structs Bayesian networks by first identifying each node's Markov blankets, then connecting nodes in a maximally consistent way. In contrast to the majority of work, which typically uses hill-climbing approaches that may produce dense and causally incorrect nets, our approach yie lds much more compact causal networks by heeding independencies in the data. C ompact causal networks facilitate fast in(cid:173) ference and are also easier t o understand. We prove that under mild assumptions, our approach requires time polynomial in the size of the data and the number of nodes. A randomized varia nt, also presented here, yields comparable results at much higher speeds.
****
Spiking Boltzmann Machines
Geoffrey E. Hinton, Andrew Brown
We first show how to represent sharp posterior probability distribu(cid:173) tio ns using real valued coefficients on broadly-tuned basis functions. Then we sh ow how the precise times of spikes can be used to con(cid:173) vey the real-val ued coefficients on the basis functions quickly and accurately. Finally we describe a simple simulation in which spik(cid:173) ing neurons learn to mode l an image sequence by fitting a dynamic generative model.
****
Actor-Critic Algorithms
Vijay Konda, John Tsitsiklis
We propose and analyze a class of actor-critic algorithms for simulati on-based optimization of a Markov decision process over a parameterized family of randomized stationary policies. These are two-time-scale algor ithms in which the critic uses TD learning with a linear approximation arch itecture and the actor is updated in an approximate gradient direction ba sed on information pro(cid:173) vided by the critic. We show that the featu res for the critic should span a subspace prescribed by the choice of paramete rization of the actor. We conclude by discussing convergence properties and s ome open problems.
****
Training Data Selection for Optimal Generalization in Trigonometric Polynomial N etworks
Masashi Sugiyama, Hidemitsu Ogawa
In this paper, we consider the problem of active learning in trigonomet(cid: 173) ric polynomial networks and give a necessary and sufficient conditi

on of sample points to provide the optimal generalization capability. By an a(cid:173) lyzing the condition from the functional analytic point of view, we clarify the mechanism of achieving the optimal generalization capabili ty. We also show that a set of training examples satisfying the cond ition does not only provide the optimal generalization but also reduces the compu(cid:173) tational complexity and memory required for the calculation o f learning results. Finally, examples of sample points satisfying the con dition are given and computer simulations are performed to demonstrate the eff ec(cid:173) tiveness of the proposed active learning method.
****
Image Recognition in Context: Application to Microscopic Urinalysis
Xubo Song, Joseph Sill, Yaser Abu-Mostafa, Harvey Kasdan
We propose a new and efficient technique for incorporating contextual infor mation into object classification. Most of the current techniques face the prob lem of exponential computation cost. In this paper, we propose a new general f ramework that incorporates partial context at a linear cost. This technique is applied to microscopic urinalysis image recognition, resulting in a sign ificant improvement of recognition rate over the context free approach. This ga in would have been impossible using conventional context incorporation techniqu es.
****
Bayesian Map Learning in Dynamic Environments
Kevin P. Murphy
We consider the problem of learning a grid-based map using a robot with noisy s ensors and actuators. We compare two approaches: online EM, where the map is tr eated as a fixed parameter, and Bayesian inference, where the map is a (matrix-valued) random variable. We show that even on a very simple example, online EM can get stuck in local minima, which causes the robot to get "lost" and the re sulting map to be useless. By contrast, the Bayesian approach, by maintaining m ultiple hypotheses, is much more ro(cid:173) bust. We then introduce a method fo r approximating the Bayesian solution, called Rao-Blackwellised particle filter ing. We show that this approximation, when coupled with an active learning stra tegy, is fast but accurate.
****
Better Generative Models for Sequential Data Problems: Bidirectional Recurrent M ixture Density Networks
Mike Schuster
This paper describes bidirectional recurrent mixture density net(cid:173) works, which can model multi-modal distributions of the type P(Xt Iyf) and P(Xt lXI, X2 , ... ,Xt-l, yf) without any explicit as(cid:173) sumptio ns about the use of context . These expressions occur fre(cid:173) quent ly in pattern recognition problems with sequential data, for example i n speech recognition. Experiments show that the pro(cid:173) posed genera tive models give a higher likelihood on test data com(cid:173) pared to a trad itional modeling approach, indicating that they can summarize the statistical properties of the data better.
****
Statistical Dynamics of Batch Learning
Song Li, K. Y. Michael Wong
An important issue in neural computing concerns the description of learning dyn amics with macroscopic dynamical variables. Recen(cid:173) t progress on on-line learning only addresses the often unrealistic case of an infinite tra ining set. We introduce a new framework to model batch learning of restric ted sets of examples, widely applica(cid:173) ble to any learning cost function , and fully taking into account the temporal correlations introduced by the r ecycling of the examples. For illustration we analyze the effects of weig ht decay and early stopping during the learning of teacher-generated example s.
****
Scale Mixtures of Gaussians and the Statistics of Natural Images

Martin J. Wainwright, Eero Simoncelli
The statistics of photographic images, when represented using multiscale (wavelet) bases, exhibit two striking types of non(cid:173) Gaussian behavior. First, the marginal densities of the coefficients have extended heavy tails. Second, the joint densities exhibit vari(cid:173) ance dependencies not captured by second-order models. We ex(cid:173) amine properties of the class of Gaussian scale mixtures, and show that these densities can accurately characterize both the marginal and joint distributions of natural image wavelet coefficients. This class of model suggests a Markov structure, in which wavelet coeffi(cid:173) cients are linked by hidden scaling variables corresponding to local image structure. We derive an estimator for these hidden variables, and show that a nonlinear "normalization" procedure can be used to Gaussianize the coefficients.
****
Independent Factor Analysis with Temporally Structured Sources
Hagai Attias
We present a new technique for time series analysis based on dy(cid:173) namic probabilistic networks. In this approach, the observed data are modeled in terms of unobserved, mutually independent factors, as in the recently introduced technique of Independent Factor Anal(cid:173) ysis (IFA). However, unlike in IFA, the factors are not Li.d.; each factor has its own temporal statistical characteristics. We derive a family of EM algorithms that learn the structure of the underlying factors and their relation to the data. These algorithms perform source separation and noise reduction in an integrated manner, and demonstrate superior performance compared to IFA.
****
Managing Uncertainty in Cue Combination
Zhiyong Yang, Richard Zemel
We develop a hierarchical generative model to study cue combi(cid:173) nation. The model maps a global shape parameter to local cue(cid:173) specific parameters, which in tum generate an intensity image. Inferring shape from images is achieved by inverting this model. Inference produces a probability distribution at each level; using distributions rather than a single value of underlying variables at each stage preserves information about the validity of each local cue for the given image. This allows the model, unlike standard combination models, to adaptively weight each cue based on gen(cid:173) eral cue reliability and specific image context. We describe the results of a cue combination psychophysics experiment we con(cid:173) ducted that allows a direct comparison with the model. The model provides a good fit to our data and a natural account for some in(cid:173) teresting aspects of cue combination.
****
Potential Boosters?
Nigel Duffy, David Helmbold
Recent interpretations of the Adaboost algorithm view it as per(cid:173) forming a gradient descent on a potential function. Simply chang(cid:173) ing the potential function allows one to create new algorithms re(cid:173) lated to AdaBoost. However, these new algorithms are generally not known to have the formal boosting property. This paper ex(cid:173) amines the question of which potential functions lead to new al(cid:173) gorithms that are boosters. The two main results are general sets of conditions on the potential; one set implies that the resulting algorithm is a booster, while the other implies that the algorithm is not. These conditions are applied to previously studied potential functions , such as those used by LogitBoost and Doom II.
****
Resonance in a Stochastic Neuron Model with Delayed Interaction
Toru Ohira, Yuzuru Sato, Jack Cowan
We study here a simple stochastic single neuron model with delayed self-feedback capable of generating spike trains. Simulations show that its spike trains exhibit resonant behavior between "noise" and "delay". In order to gain insight into this resonance, we simplify the model and study a stochastic bin

ary element whose transition probability depends on its state at a fixed interval in the past. With this simplified model we can analytically compute interspike interval histograms, and show how the resonance between noise and delay arises. The resonance is also observed when such elements are coupled through delayed interaction.
****
Wiring Optimization in the Brain
Dmitri Chklovskii, Charles Stevens
The complexity of cortical circuits may be characterized by the number of synapses per neuron. We study the dependence of complexity on the fraction of the cortical volume that is made up of "wire" (that is, ofaxons and dendrites), and find that complexity is maximized when wire takes up about 60% of the cortical volume. This prediction is in good agree(cid:173) ment with experimental observations. A consequence of our arguments is that any rearrangement of neurons that takes more wire would sacrifice computational power.
****
Learning from User Feedback in Image Retrieval Systems
Nuno Vasconcelos, Andrew Lippman
We formulate the problem of retrieving images from visual databases as a problem of Bayesian inference. This leads to natural and effective solutions for two of the most challenging issues in the design of a retrieval system: providing support for region-based queries without requiring prior image segmentation, and accounting for user-feedback during a retrieval session. We present a new learning algorithm that relies on belief propagation to account for both positive and negative examples of the user's interests.
****
Online Independent Component Analysis with Local Learning Rate Adaptation
Nicol Schraudolph, Xavier Giannakopoulos
Stochastic meta-descent (SMD) is a new technique for online adap(cid:173) tation of local learning rates in arbitrary twice-differentiable sys(cid:173) tems. Like matrix momentum it uses full second-order information while retaining O(n) computational complexity by exploiting the efficient computation of Hessian-vector products. Here we apply SMD to independent component analysis, and employ the result(cid:173) ing algorithm for the blind separation of time-varying mixtures. By matching individual learning rates to the rate of change in each source signal's mixture coefficients, our technique is capable of si(cid:173) multaneously tracking sources that move at very different, a priori unknown speeds.
****
A Variational Baysian Framework for Graphical Models
Hagai Attias
This paper presents a novel practical framework for Bayesian model averaging and model selection in probabilistic graphical models. Our approach approximates full posterior distributions over model parameters and structures, as well as latent variables, in an analyt(cid:173) ical manner. These posteriors fall out of a free-form optimization procedure, which naturally incorporates conjugate priors. Unlike in large sample approximations, the posteriors are generally non(cid:173) Gaussian and no Hessian needs to be computed. Predictive quanti(cid:173) ties are obtained analytically. The resulting algorithm generalizes the standard Expectation Maximization algorithm, and its conver(cid:173) gence is guaranteed. We demonstrate that this approach can be applied to a large class of models in several domains, including mixture models and source separation.
****
Algebraic Analysis for Non-regular Learning Machines
Sumio Watanabe
Hierarchical learning machines are non-regular and non-identifiable statistical models, whose true parameter sets are analytic sets with singularities. Using algebraic analysis, we rigorously prove that the stochastic complexity

of a non-identifiable learning machine (ml - 1) log log n + const., is asymptotically equal to >'1 log n - where n is the number of training sam ples. Moreover we show that the rational number >'1 and the integer ml can be algorithmically calculated using resolution of singularities in algebra ic geometry. Also we obtain inequalities 0 < >'1 ~ d/2 and 1 ~ ml ~ d, whe re d is the number of parameters.
****
Model Selection in Clustering by Uniform Convergence Bounds
Joachim Buhmann, Marcus Held
Unsupervised learning algorithms are designed to extract struc(cid:173) t ure from data samples. Reliable and robust inference requires a guarantee that extracted structures are typical for the data source, Le., similar stru ctures have to be inferred from a second sample set of the same data sou rce. The overfitting phenomenon in max(cid:173) imum entropy based annealing algorithms is exemplarily studied for a class of histogram clustering m odels. Bernstein's inequality for large deviations is used to determine the maximally achievable approximation quality parameterized by a minimal tem perature. Monte Carlo simulations support the proposed model selection cri(cid: 173) terion by finite temperature annealing.
****
Unmixing Hyperspectral Data
Lucas Parra, Clay Spence, Paul Sajda, Andreas Ziehe, Klaus-Robert Müller
In hyperspectral imagery one pixel typically consists of a mixture of the r eflectance spectra of several materials, where the mixture coefficients c orrespond to the abundances of the constituting ma(cid:173) terials. We a ssume linear combinations of reflectance spectra with some additive normal sens or noise and derive a probabilistic MAP framework for analyzing hyperspectr al data. As the material re(cid:173) flectance characteristics are not know a priori, we face the problem of unsupervised linear unmixing. The incor poration of different prior information (e.g. positivity and normalizati on of the abun(cid:173) dances) naturally leads to a family of interest ing algorithms, for example in the noise-free case yielding an algorit hm that can be understood as constrained independent component analysis (ICA). Simulations underline the usefulness of our theory.
****
Some Theoretical Results Concerning the Convergence of Compositions of Regulariz ed Linear Functions
Tong Zhang
Recently, sample complexity bounds have been derived for problems in(cid:173) v olving linear functions such as neural networks and support vector ma(cid:173) c hines. In this paper, we extend some theoretical results in this area by deri ving dimensional independent covering number bounds for regular(cid:173) ized l inear functions under certain regularization conditions. We show that such bo unds lead to a class of new methods for training linear clas(cid:173) sifiers wi th similar theoretical advantages of the support vector machine. Furthermore, we also present a theoretical analysis for these new meth(cid:173) ods from the asymptotic statistical point of view. This technique provides better descripti on for large sample behaviors of these algorithms.
****
Inference for the Generalization Error
Claude Nadeau, Yoshua Bengio
In order to to compare learning algorithms, experimental results reported in t he machine learning litterature often use statistical tests of signifi(ci d:173) cance. Unfortunately, most of these tests do not take into account the variability due to the choice of training set. We perform a theor etical investigation of the variance of the cross-validation estimate of the ge n(cid:173) eralization error that takes into account the variability due to th e choice of training sets. This allows us to propose two new ways to estimate this variance. We show, via simulations, that these new statistics pe rform well relative to the statistics considered by Dietterich (Dietterich, 199

8).
****
LTD Facilitates Learning in a Noisy Environment
Paul Munro, Gerardina Hernández
Long-term potentiation (LTP) has long been held as a biological substrate for a ssociative learning. Recently, evidence has emerged that long-term depression (LTD) results when the presynaptic cell fires after the postsynaptic cell. The c omputational utility of LTD is explored here. Synaptic modification kernels for both LTP and LTD have been proposed by other laboratories based studies of one postsynaptic unit. Here, the interaction between time-dependent LTP and LTD i s studied in small networks.
****
An Oculo-Motor System with Multi-Chip Neuromorphic Analog VLSI Control
Oliver Landolt, Steve Gyger
A system emulating the functionality of a moving eye-hence the name oculo-motor system-has been built and successfully tested. It is made of an optical devi ce for shifting the field of view of an image sensor by up to 45 ° in any dire ction, four neuromorphic analog VLSI circuits imple(cid:173) menting an oculo-mo tor control loop, and some off-the-shelf electronics. The custom integrated cir cuits communicate with each other primarily by non-arbitrated address-event bus es. The system implements the behav(cid:173) iors of saliency-based saccadic e xploration, and smooth pursuit of light spots. The duration of saccades ranges from 45 ms to 100 ms, which is comparable to human eye performance. Smoot h pursuit operates on light sources moving at up to 50 0 /s in the visual field .
****
Understanding Stepwise Generalization of Support Vector Machines: a Toy Model
Sebastian Risau-Gusman, Mirta Gordon
In this article we study the effects of introducing structure in the input distribution of the data to be learnt by a simple perceptron. We determine th e learning curves within the framework of Statis(cid:173) tical Mechanics. Ste pwise generalization occurs as a function of the number of examples when the distribution of patterns is highly anisotropic. Although extremely simple , the model seems to cap(cid:173) ture the relevant features of a class of Support Vector Machines which was recently shown to present this behav ior.
****
A Geometric Interpretation of v-SVM Classifiers
David Crisp, Christopher J. C. Burges
We show that the recently proposed variant of the Support Vector machine (SVM) algorithm, known as v-SVM, can be interpreted as a maximal separation b etween subsets of the convex hulls of the data, which we call soft convex hulls. The soft convex hulls are controlled by choice of the parameter v . If the intersection of the convex hulls is empty, the hyperplane is position ed halfway between them such that the distance between convex hulls, measured a long the normal, is maximized; and if it is not, the hyperplane's normal is similarly determined by the soft convex hulls, but its position (perpen dicular distance from the origin) is adjusted to minimize the error s um. The proposed geometric interpretation of v-SVM also leads to necessary and sufficient conditions for the existence of a choice of v for which the v -SVM solution is nontrivial.
****
Robust Learning of Chaotic Attractors
Rembrandt Bakker, Jaap Schouten, Marc-Olivier Coppens, Floris Takens, C. Giles, Cor van den Bleek
A fundamental problem with the modeling of chaotic time series data is that min imizing short-term prediction errors does not guarantee a match between the rec onstructed attractors of model and experiments. We introduce a modeling paradig m that simultaneously learns to short-tenn predict and to locate the outlines o f the attractor by a new way of nonlinear principal component analysis. Closed-

loop predictions are constrained to stay within these outlines, to prevent dive rgence from the attractor. Learning is exceptionally fast: parameter estimation for the 1000 sample laser data from the 1991 Santa Fe time series competition took less than a minute on a 166 MHz Pentium PC.
****
Greedy Importance Sampling
Dale Schuurmans
I present a simple variation of importance sampling that explicitly search(cid:1 73) es for important regions in the target distribution. I prove that the tech (cid:173) nique yields unbiased estimates, and show empirically it can reduce t he variance of standard Monte Carlo estimators. This is achieved by con(cid: 173) centrating samples in more significant regions of the sample space.
****
Recurrent Cortical Competition: Strengthen or Weaken?
Péter Adorján, Lars Schwabe, Christian Piepenbrock, Klaus Obermayer
We investigate the short term .dynamics of the recurrent competition and neural activity in the primary visual cortex in terms of information pro(cid:173) ces sing and in the context of orientation selectivity. We propose that af(cid:173) ter stimulus onset, the strength of the recurrent excitation decreases due to fast synaptic depression. As a consequence, the network shifts from an in itially highly nonlinear to a more linear operating regime. Sharp orientat ion tuning is established in the first highly competitive phase. In the second and less competitive phase, precise signaling of multiple ori(cid:173) entation s and long range modulation, e.g., by intra- and inter-areal con(cid:173) necti ons becomes possible (surround effects). Thus the network first ex(cid:173) tra cts the salient features from the stimulus, and then starts to proces s the details. We show that this signal processing strategy is optimal if the neurons have limited bandwidth and their objective is to transmit the maximum amount of information in any time interval beginning with the stimulus onset.
****
Constrained Hidden Markov Models
Sam Roweis
By thinking of each state in a hidden Markov model as corresponding to some spatial region of a fictitious topology space it is possible to naturally define neigh(cid:173) bouring states as those which are connected in that space . The transition matrix can then be constrained to allow transitions only betw een neighbours; this means that all valid state sequences correspond to connect ed paths in the topology space. I show how such constrained HMMs can learn t o discover underlying structure in complex sequences of high dimensional dat a, and apply them to the problem of recovering mouth movements from acousti cs in continuous speech.
****
Approximate Inference A lgorithms for Two-Layer Bayesian Networks
Andrew Ng, Michael Jordan
We present a class of approximate inference algorithms for graphical mo dels of the QMR-DT type. We give convergence rates for these al(cid:173) gorithms and for the Jaakkola and Jordan (1999) algorithm, and verify t hese theoretical predictions empirically. We also present empirical re(cid:17 3) sults on the difficult QMR-DT network problem, obtaining performance of the new algorithms roughly comparable to the Jaakkola and Jordan algorithm.
****
Perceptual Organization Based on Temporal Dynamics
Xiuwen Liu, DeLiang Wang
A figure-ground segregation network is proposed based on a novel boundary pair representation. Nodes in the network are bound(cid:173) ary segmen ts obtained through local grouping. Each node is ex(cid:173) citatorily coupled with the neighboring nodes that belong to the same region, and inhibitorily coupled with the corresponding paired node. Gestalt grouping rul es are incorporated by modulating con(cid:173) nections. The status of a n

ode represents its probability being figural and is updated according to a differential equation. The system solves the figure-ground segregation problem through tem(cid:173) poral evolution. Different perceptual phenomen a, such as modal and amodal completion, virtual contours, grouping and s hape de(cid:173) composition are then explained through local diffusion. The sy stem eliminates combinatorial optimization and accounts for many psy(cid:173) chophysical results with a fixed set of parameters.
****

Learning Informative Statistics: A Nonparametnic Approach
John W. Fisher III, Alexander Ihler, Paul Viola
We discuss an information theoretic approach for categorizing and mod(cid:173) e ling dynamic processes. The approach can learn a compact and informa(cid:173) ti ve statistic which summarizes past states to predict future observations. Furth ermore, the uncertainty of the prediction is characterized nonpara(cid:173) met rically by a joint density over the learned statistic and present obser(cid:173) vation. We discuss the application of the technique to both noise driven dy namical systems and random processes sampled from a density which is conditione d on the past. In the first case we show results in which both the dynamics of random walk and the statistics of the driving noise are captured. In the s econd case we present results in which a summarizing statistic is learned on noisy random telegraph waves with differing de(cid:173) pendencies on pa st states. In both cases the algorithm yields a principled approach for discri minating processes with differing dynamics and/or de(cid:173) pendencies. The m ethod is grounded in ideas from information theory and nonparametric statis tics.
****

Rules and Similarity in Concept Learning
Joshua Tenenbaum
This paper argues that two apparently distinct modes of generalizing con(cid:173) cepts - abstracting rules and computing similarity to exemplars - should both be seen as special cases of a more general Bayesian learning frame(cid:173) wor k. Bayes explains the specific workings of these two modes - which rules are a bstracted, how similarity is measured - as well as why gener(cid:173) alizat ion should appear rule- or similarity-based in different situations. This anal ysis also suggests why the rules/similarity distinction, even if not computat ionally fundamental, may still be useful at the algorithmic level as part of a principled approximation to fully Bayesian learning.
****

Support Vector Method for Novelty Detection
Bernhard Schölkopf, Robert C. Williamson, Alex Smola, John Shawe-Taylor, John Pl att
Suppose you are given some dataset drawn from an underlying probabil(cid:173) it y distribution P and you want to estimate a "simple" subset S of input space such that the probability that a test point drawn from P lies outside of S equa ls some a priori specified l/ between 0 and 1. We propose a method to approac h this problem by trying to estimate a function f which is positive on S and negative on the complement. The functional form of f is given by a kernel ex pansion in terms of a poten(cid:173) tially small subset of the training data; i t is regularized by controlling the length of the weight vector in an associat ed feature space. We provide a theoretical analysis of the statistical perform ance of our algorithm. The algorithm is a natural extension of the support ve ctor algorithm to the case of unlabelled data.
****

Generalized Model Selection for Unsupervised Learning in High Dimensions
Shivakumar Vaithyanathan, Byron Dom
We describe a Bayesian approach to model selection in unsupervised learning that determines both the feature set and the number of clusters. We th en evaluate this scheme (based on marginal likelihood) and one based on cro ss-validated likelihood. For the Bayesian scheme we derive a closed-for m solution of the marginal likelihood by assuming appropriate forms of

the likelihood function and prior. Extensive experiments compare these a pproaches and all results are verified by comparison against ground truth. In these experiments the Bayesian scheme using our objective function gave b etter results than cross-validation.
****
An Improved Decomposition Algorithm for Regression Support Vector Machines
Pavel Laskov
A new decomposition algorithm for training regression Support Vector Ma chines (SVM) is presented. The algorithm builds on the basic principle s of decomposition proposed by Osuna et. al., and addresses the issue of o ptimal working set selection. The new criteria for testing optimality of a working set are derived. Based on these criteria, the principle of "maximal inconsistency" is pro(cid:173) posed to form (approximately) optimal working sets. Experimental results show superior performance of the new algorithm in c ompar(cid:173) ison with traditional training of regression SVM without decompo(cid:173) sition. Similar results have been previously reported on decomposi(cid :173) tion algorithms for pattern recognition SVM. The new algorithm is also a pplicable to advanced SVM formulations based on regression, such as density est imation and integral equation SVM.
****
An Analog VLSI Model of Periodicity Extraction
André van Schaik
that extracts
****
From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regu lation among Gene Classes from Large-Scale Expression Data
Eric Mjolsness, Tobias Mann, Rebecca Castaño, Barbara Wold
small-scale gene
****
Data Visualization and Feature Selection: New Algorithms for Nongaussian Data
Howard Yang, John Moody
Data visualization and feature selection methods are proposed based on t he)oint mutual information and ICA. The visualization methods can find many good 2-D projections for high dimensional data interpretation, which cannot be easily found by the other ex(cid:173) isting methods. The new variable se lection method is found to be better in eliminating redundancy in the inputs than other methods based on simple mutual information. The efficacy of the methods is illustrated on a radar signal analysis problem to find 2-D viewing coordinates for data visualization and to select inputs for a neural netw ork classifier. Keywords: feature selection, joint mutual information, ICA , vi(cid:173) sualization, classification.
****
An Information-Theoretic Framework for Understanding Saccadic Eye Movements
Tai Sing Lee, Stella Yu
In this paper, we propose that information maximization can pro(cid:173) vid e a unified framework for understanding saccadic eye move(cid:173) ments . In this framework, the mutual information among the cor(cid:173) tical rep resentations of the retinal image, the priors constructed from our long term visual experience, and a dynamic short-term internal representati on constructed from recent saccades provides a map for guiding eye navi gation . By directing the eyes to loca(cid:173) tions of maximum complexit y in neuronal ensemble responses at each step, the automatic saccadic e ye movement system greedily collects information about the external world, wh ile modifying the neural representations in the process. This framework attempts to connect several psychological phenomena, such as pop-out an d inhibition of return, to long term visual experience and short term workin g memory. It also provides an interesting perspective on contextual com putation and formation of neural representation in the visual system.
****
Noisy Neural Networks and Generalizations

Hava Siegelmann, Alexander Roitershtein, Asa Ben-Hur
In this paper we define a probabilistic computational model which generalize s many noisy neural network models, including the recent work of Maass and Son tag [5]. We identify weak ergodicjty as the mechanism responsible for rest riction of the computational power of probabilistic models to definite la nguages, independent of the characteristics of the noise: whether it is d iscrete or analog, or if it depends on the input or not, and independ ent of whether the variables are discrete or continuous. We give example s of weakly ergodic models including noisy computational systems with noise depending on the current state and inputs, aggregate models, and computation al systems which update in continuous time.
****
The Nonnegative Boltzmann Machine
Oliver Downs, David MacKay, Daniel Lee
The nonnegative Boltzmann machine (NNBM) is a recurrent neural net(cid:173) work model that can describe multimodal nonnegative data. Application of maximum l ikelihood estimation to this model gives a learning rule that is analogous to t he binary Boltzmann machine. We examine the utility of the mean field approxim ation for the NNBM, and describe how Monte Carlo sampling techniques can be used to learn its parameters. Reflec(cid:173) tive slice sampling is part icularly well-suited for this distribution, and can efficiently be implemen ted to sample the distribution. We illustrate learning of the NNBM on a tran siationally invariant distribution, as well as on a generative model for images of human faces.
****
Boosting Algorithms as Gradient Descent
Llew Mason, Jonathan Baxter, Peter Bartlett, Marcus Frean
We provide an abstract characterization of boosting algorithms as gradient de csent on cost-functionals in an inner-product function space. We prove convergence of these functional-gradient-descent algorithms under quite weak conditions. Following previous theo(cid:173) retical results bounding the gen eralization performance of convex combinations of classifiers in terms of gene ral cost functions of the margin, we present a new algorithm (DOOM II) for performing a gradient descent optimization of such cost functions. Exper iments on several data sets from the UC Irvine repository demonstrate that DOOM II generally outperforms AdaBoost, especially in high noise situation s, and that the overfitting behaviour of AdaBoost is predicted by our cost func tions.
****
Local Probability Propagation for Factor Analysis
Brendan J. Frey
Ever since Pearl's probability propagation algorithm in graphs with cycles was shown to produce excellent results for error-correcting decoding a few years ago, we have been curious about whether local probability prop agation could be used successfully for ma(cid:173) chine learning. One of the simplest adaptive models is the factor analyzer, which is a two-l ayer network that models bottom layer sensory inputs as a linear combinat ion of top layer factors plus in(cid:173) dependent Gaussian sensor noise. W e show that local probability propagation in the factor analyzer network usua lly takes just a few iterations to perform accurate inference, even in networks with 320 sensors and 80 factors. We derive an expression for the algorithm's fixed point and show that this fixed point matches the exact solu(cid:173) t ion in a variety of networks, even when the fixed point is unstable. We also s how that this method can be used successfully to perform inference for approxim ate EM and we give results on an online face recognition task. 1 Factor analy sis A simple way to encode input patterns is to suppose that each input c an be well(cid:173) approximated by a linear combination of component vectors, where the amplitudes of the vectors are modulated to match the input. For a given training set, the most appropriate set of component vectors will depe nd on how we expect the modula(cid:173) tion levels to behave and how

we measure the distance between the input and its approximation. These ef fects can be captured by a generative probabilit~ model that specifies a distribution p(z) over modulation levels z = (Zl, ... ,ZK) and a d istribution p(xlz) over sensors x = (Xl, ... ,XN)T given the modulation levels. Principal component analysis, independent component analysis and fac tor analysis can be viewed as maximum likelihood learning in a model of this ty pe, where we as(cid:173) sume that over the training set, the appropriate modula tion levels are independent and the overall distortion is given by the sum of the individual sensor distortions.
****
A MCMC Approach to Hierarchical Mixture Modelling
Christopher Williams
There are many hierarchical clustering algorithms available, but these l ack a firm statistical basis. Here we set up a hierarchical probabilistic mi xture model, where data is generated in a hierarchical tree-structured manner . Markov chain Monte Carlo (MCMC) methods are demonstrated which can be used t o sample from the posterior distribution over trees containing variable num bers of hidden units.
****
The Infinite Gaussian Mixture Model
Carl Rasmussen
In a Bayesian mixture model it is not necessary a priori to limit the num(cid:17 3) ber of components to be finite. In this paper an infinite Gaussian mixture model is presented which neatly sidesteps the difficult problem of find(cid:173) ing the "right" number of mixture components. Inference in the model is done using an efficient parameter-free Markov Chain that relies entirely on Gibbs sa mpling.
****
Reconstruction of Sequential Data with Probabilistic Models and Continuity Const raints
Miguel Carreira-Perpiñán
We consider the problem of reconstructing a temporal discrete sequence of multi dimensional real vectors when part of the data is missing, under the assumpti on that the sequence was generated by a continuous pro(cid:173) cess. A particular case of this problem is multivariate regression, which is very d ifficult when the underlying mapping is one-to-many. We pro(cid:173) pose a n algorithm based on a joint probability model of the variables of inter est, implemented using a nonlinear latent variable model. Each point in the sequence is potentially reconstructed as any of the modes of the con ditional distribution of the missing variables given the present variables (co mputed using an exhaustive mode search in a Gaussian mix(cid:173) ture). Mode s election is determined by a dynamic programming search that minimises a geome tric measure of the reconstructed sequence, de(cid:173) rived from continuity c onstraints. We illustrate the algorithm with a toy example and apply it to a real-world inverse problem, the acoustic-to(cid:173) articulatory mapping . The results show that the algorithm outperforms conditional mean imputati on and multilayer perceptrons.
****
Learning the Similarity of Documents: An Information-Geometric Approach to Docum ent Retrieval and Categorization
Thomas Hofmann
The project pursued in this paper is to develop from first informatio n-geometric principles a general method for learning the similarity bet ween text documents. Each individual docu(cid:173) ment is modeled as a memoryless information source. Based on a latent class decomposition of the term-document matrix, a low(cid:173) dimensional (curved) multinomial subfamily is learned. From this model a canonical similarity function - know n as the Fisher kernel - is derived. Our approach can be applied for u nsupervised and supervised learning problems alike. This in particular cove rs inter(cid:173) esting cases where both, labeled and unlabeled data are

available. Experiments in automated indexing and text categorization verify the advantages of the proposed method.
****
Bayesian Reconstruction of 3D Human Motion from Single-Camera Video
Nicholas Howe, Michael Leventon, William Freeman
The three-dimensional motion of humans is underdetermined when the observatio n is limited to a single camera, due to the inherent 3D ambi(cid:173) guity of 2D video. We present a system that reconstructs the 3D motion of human subject s from single-camera video, relying on prior knowledge about human motion, le arned from training data, to resolve those am(cid:173) biguities. After i nitialization in 2D, the tracking and 3D reconstruction is automatic; we sho w results for several video sequences. The results show the power of treatin g 3D body tracking as an inference problem.
****
Mixture Density Estimation
Jonathan Li, Andrew Barron
Gaussian mixtures (or so-called radial basis function networks) for density es timation provide a natural counterpart to sigmoidal neu(cid:173) ral networks fo r function fitting and approximation. In both cases, it is possible to give simple expressions for the iterative improve(cid:173) ment of performance as co mponents of the network are introduced one at a time. In particular, for mixtu re density estimation we show that a k-component mixture estimated by maximum l ikelihood (or by an iterative likelihood improvement that we introduce) achiev es log-likelihood within order 1/k of the log-likelihood achievable by any convex combination. Consequences for approximation and es(cid:173) timation u sing Kullback-Leibler risk are also given. A Minimum Description Length principle selects the optimal number of compo(cid:173) nents k that minimizes th e risk bound.
****
Information Capacity and Robustness of Stochastic Neuron Models
Elad Schneidman, Idan Segev, Naftali Tishby
The reliability and accuracy of spike trains have been shown to depen d on the nature of the stimulus that the neuron encodes. Adding ion channel stochasticity to neuronal models results in a macroscopic behav ior that replicates the input-dependent reliabili(cid:173) ty and precision of r eal neurons. We calculate the amount of infor(cid:173) mation that an ion chan nel based stochastic Hodgkin-Huxley (HH) neuron model can encode about a wide s et of stimuli. We show that both the information rate and the information pe r spike of the s(cid:173) tochastic model are similar to the values reported experimentally. Moreover, the amount of information that the neuron enco des is correlated with the amplitude of fluctuations in the input, and less s o with the average firing rate of the neuron. We also show that for the HH io n channel density, the information capacity is robust to changes in the d ensity of ion channels in the membrane, whereas changing the ratio bet ween the Na+ and K+ ion channels has a considerable effect on the infor mation that the neuron can encode. Finally, we suggest that neurons may max imize their information capacity by appropriately balancing the density of the different ion channels that underlie neuronal excitability.
****
Bayesian Transduction
Thore Graepel, Ralf Herbrich, Klaus Obermayer
Transduction is an inference principle that takes a training sam(cid:173) ple and aims at estimating the values of a function at given points contain ed in the so-called working sample as opposed to the whole of input space fo r induction. Transduction provides a confidence measure on single predi ctions rather than classifiers - a feature particularly important for ris k-sensitive applications. The possibly infinite number of functions is redu ced to a finite number of equiv(cid:173) alence classes on the working sample. A rigorous Bayesian analysis reveals that for standard classification loss we cannot benefit from considering more than one test point at a tim

e. The probability of the label of a given test point is determined as the posterior measure of the corresponding subset of hypothesis space. We con(cid:173) sider the PAC setting of binary classification by linear discri minant functions (perceptrons) in kernel space such that the probability of labels is determined by the volume ratio in version space. We suggest to sample this region by an ergodic billiard. Experimen(cid:173) tal re sults on real world data indicate that Bayesian Transduction compares favou rably to the well-known Support Vector Machine, in particular if the p osterior probability of labellings is used as a confidence measure to ex clude test points of low confidence.
****

Constructing Heterogeneous Committees Using Input Feature Grouping: Application to Economic Forecasting
Yuansong Liao, John Moody
The committee approach has been proposed for reducing model uncertainty and improving generalization performance. The ad(cid:173) vantage of comm ittees depends on (1) the performance of individ(cid:173) ual members and (2) the correlational structure of errors between members. This paper presents an input grouping technique for de(cid:173) signing a heterogeneous committe e. With this technique, all input variables are first grouped based on thei r mutual information. Sta(cid:173) tistically similar variables are assigne d to the same group. Each member's input set is then formed by inpu t variables extracted from different groups. Our designed committees have l ess error cor(cid:173) relation between its members, since each member observes different input variable combinations. The individual member's feature sets c ontain less redundant information, because highly correlated vari(cid:173) ables will not be combined together. The member feature sets con(cid:173) tain almos t complete information, since each set contains a feature from each informati on group. An empirical study for a noisy and nonstationary economic fore casting problem shows that commit(cid:173) tees constructed by our proposed technique outperform committees formed using several existing techniques.
****

An Analysis of Turbo Decoding with Gaussian Densities
Paat Rusmevichientong, Benjamin Van Roy
We provide an analysis of the turbo decoding algorithm (TDA) in a se tting involving Gaussian densities. In this context, we are able to sho w that the algorithm converges and that - somewhat surprisingly - though the density generated by the TDA may differ significantly from the desired po sterior density, the means of these two densities coincide.
****

Probabilistic Methods for Support Vector Machines
Peter Sollich
I describe a framework for interpreting Support Vector Machines (SVMs) as maximum a posteriori (MAP) solutions to inference problems with Gauss ian Process priors. This can provide intuitive guidelines for choosing a ' good' SVM kernel. It can also assign (by evidence maximization) optima l values to parameters such as the noise level C which cannot be determined u nambiguously from properties of the MAP solution alone (such as cross-valida tion er(cid:173) ror) . I illustrate this using a simple approximate expression for the SVM evidence. Once C has been determined, error bars on SVM pred ictions can also be obtained.
****

Neural System Model of Human Sound Localization
Craig Jin, Simon Carlile
This paper examines the role of biological constraints in the human audi(cid:173) tory localization process. A psychophysical and neural system modeling appro ach was undertaken in which performance comparisons between competing mod els and a human subject explore the relevant biologi(cid:173) cally plau sible "realism constraints". The directional acoustical cues, upon which sound localization is based, were derived from the human subject's head-

related transfer functions (HRTFs). Sound stimuli were generated by convolving bandpass noise with the HRTFs and were pre(cid:173) sented to both the subject and the model. The input stimuli to the model was processed using the Auditory Image Model of cochlear processing. The cochlear data was then analyzed by a time-delay neural network which integrated temporal and spectral information to determine the spa(cid:173) tial location of the sound source. The combined cochlear model and neural network provided a system model of the sound localization pro(cid:173) cess. Human-like localization performance was qualitatively achieved for broadband and bandpass stimuli when the model architecture incor(cid:173) porated frequency division (or tonotopicity), and was trained using vari(cid:173) able bandwidth and center-frequency sounds.
****
A Neuromorphic VLSI System for Modeling the Neural Control of Axial Locomotion
Girish Patel, Edgar Brown, Stephen DeWeerth
We have developed and tested an analog/digital VLSI system that mod(cid:173) els the coordination of biological segmental oscillators underlying axial locomotion in animals such as leeches and lampreys. In its current form the system consists of a chain of twelve pattern generating circuits that are capable of arbitrary contralateral inhibitory synaptic coupling. Each pattern generating circuit is implemented with two independent silicon Morris-Lecar neurons with a total of 32 programmable (floating-gate based) inhibitory synapses, and an asynchronous address-event inter(cid:173) connection element that provides synaptic connectivity and implements axonal delay. We describe and analyze the data from a set of experi(cid:173) ments exploring the system behavior in terms of synaptic coupling.
****
Graded Grammaticality in Prediction Fractal Machines
Shan Parfitt, Peter Tiño, Georg Dorffner
We introduce a novel method of constructing language models, which avoids some of the problems associated with recurrent neu(cid:173) ral networks. The method of creating a Prediction Fractal Machine (PFM) [1] is briefly described and some experiments are presented which demonstrate the suitability of PFMs for language modeling. PFMs distinguish reliably between minimal pairs, and their be(cid:173) havior is consistent with the hypothesis [4] that wellformedness is 'graded' not absolute. A discussion of their potential to offer fresh insights into language acquisition and processing follows.
****
Learning Sparse Codes with a Mixture-of-Gaussians Prior
Bruno Olshausen, K. Millman
We describe a method for learning an overcomplete set of basis functions for the purpose of modeling sparse structure in images. The sparsity of the basis function coefficients is modeled with a mixture-of-Gaussians distribution. One Gaussian captures non(cid:173) active coefficients with a small-variance distribution centered at zero, while one or more other Gaussians capture active coefficients with a large-variance distribution. We show that when the prior is in such a form, there exist efficient methods for learning the basis functions as well as the parameters of the prior. The performance of the algorithm is demonstrated on a number of test cases and also on natural images. The basis functions learned on natural images are similar to those obtained with other methods, but the sparse form of the coefficient distribution is much better described. Also, since the parameters of the prior are adapted to the data, no assumption about sparse structure in the images need be made a priori, rather it is learned from the data.
****
Semiparametric Approach to Multichannel Blind Deconvolution of Nonminimum Phase Systems
Liqing Zhang, Shun-ichi Amari, Andrzej Cichocki
In this paper we discuss the semi parametric statistical model for blind deconvolution. First we introduce a Lie Group to the manifold of non(cid:17

3) causal FIR filters. Then blind deconvolution problem is formulated in the framework of a semiparametric model, and a family of estimating func tions is derived for blind deconvolution. A natural gradient learn(cid:17 3) ing algorithm is developed for training noncausal filters. Stability of the natural gradient algorithm is also analyzed in this framework.
****
Application of Blind Separation of Sources to Optical Recording of Brain Activit y
Holger Schoner, Martin Stetter, Ingo Schießl, John Mayhew, Jennifer Lund, Niall McLoughlin, Klaus Obermayer
In the analysis of data recorded by optical imaging from intrinsic signals (mea surement of changes of light reflectance from cortical tissue) the re(cid:173) m oval of noise and artifacts such as blood vessel patterns is a serious problem. Often bandpass filtering is used, but the underlying assumption that a spatial frequency exists, which separates the mapping component from oth er components (especially the global signal), is questionable. Here we pr opose alternative ways of processing optical imaging data, us(cid:173) ing blind source separation techniques based on the spatial decorre1ation of the data. We first perform benchmarks on artificial data in order to select the way of processing, which is most robust with respect to sen(cid:173) sor no ise. We then apply it to recordings of optical imaging experiments from macaq ue primary visual cortex. We show that our BSS technique is able to extract o cular dominance and orientation preference maps from single condition stacks , for data, where standard post-processing pro(cid:173) cedures fail. Art ifacts, especially blood vessel patterns, can often be completely remove d from the maps. In summary, our method for blind source separation using extended spatial decorrelation is a superior tech(cid:173) nique for the analysi s of optical recording data.
****
Spectral Cues in Human Sound Localization
Craig Jin, Anna Corderoy, Simon Carlile, André van Schaik
The differential contribution of the monaural and interaural spectral cu es to human sound localization was examined using a combined psy(cid:173) chophy sical and analytical approach. The cues to a sound's location were co rrelated on an individual basis with the human localization re(cid:173) spo nses to a variety of spectrally manipulated sounds. The spectral cues derive f rom the acoustical filtering of an individual's auditory periphery which is c haracterized by the measured head-related transfer functions (HRTFs). Auditory localization performance was determined in virtual auditory space (VAS). Psy choacoustical experiments were conducted in which the amplitude spectra of the sound stimulus was varied indepen(cid:173) dentlyat each ear while preserving t he normal timing cues, an impossibil(cid:173) ity in the free-field environment. Virtual auditory noise stimuli were gen(cid:173) erated over earphones for a sp ecified target direction such that there was a "false" flat spectrum at the le ft eardrum. Using the subject's HRTFs, the sound spectrum at the right eardrum was then adjusted so that either the true right monaural spectral cue or the true interaural spectral cue was preserved. All subjects showed syste matic mislocalizations in both the true right and true interaural spectral cond itions which was absent in their control localization performance. The analysis of the different cues along with the subjects' localization responses suggests there are signif(cid:173) icant differences in the use of the monaural and inte raural spectral cues and that the auditory system's reliance on the spectral c ues varies with the sound condition.
****
Robust Full Bayesian Methods for Neural Networks
Christophe Andrieu, João de Freitas, Arnaud Doucet
In this paper, we propose a full Bayesian model for neural networks. This model treats the model dimension (number of neurons), model parameters, regularisati on parameters and noise parameters as ran(cid:173) dom variables that need t o be estimated. We then propose a re(cid:173) versible jump Markov chain Mo

nte Carlo (MCMC) method to per(cid:173) form the necessary computations. We find that the results are not only better than the previously reported ones, but also appear to be robust with respect to the prior specification . Moreover, we present a geometric convergence theorem for the algorithm.
****

A Neurodynamical Approach to Visual Attention
Gustavo Deco, Josef Zihl
The psychophysical evidence for "selective attention" originates mainly from v isual search experiments. In this work, we formulate a hierarchi(cid:173) cal s ystem of interconnected modules consisting in populations of neu(cid:173) rons for modeling the underlying mechanisms involved in selective visual at tention. We demonstrate that our neural system for visual search works across the visual field in parallel but due to the different intrinsic dynamics can show the two experimentally observed modes of visual attention, namely: the serial and the parallel search mode. In other words, neith er explicit model of a focus of attention nor saliencies maps are used. The focus of attention appears as an emergent property of the dynamic behavior of the system. The neural population dynamics are handled in the framework of the mean-field approximation. Conse(cid:173) quently, the whole process can be expressed as a system of coupled dif(cid:173) ferential equations.
****

Dynamics of Supervised Learning with Restricted Training Sets and Noisy Teachers
Anthony Coolen, C. Mace
We generalize a recent formalism to describe the dynamics of supervised learnin g in layered neural networks, in the regime where data recycling is inevitable , to the case of noisy teachers. Our theory generates reliable predictions for the evolution in time of training- and generalization er(cid:173) rors, and e xtends the class of mathematically solvable learning processes in large neural networks to those situations where overfitting can occur.
****

Audio Vision: Using Audio-Visual Synchrony to Locate Sounds
John Hershey, Javier Movellan
Psychophysical and physiological evidence shows that sound local(cid:173) izatio n of acoustic signals is strongly influenced by their synchrony with visual si gnals. This effect, known as ventriloquism, is at work when sound coming f rom the side of a TV set feels as if it were coming from the mouth of the actors. The ventriloquism effect suggests that there is important information about sound location encoded in the synchrony between the audio an d video signals. In spite of this evidence, audiovisual synchrony is ra rely used as a source of information in computer vision tasks. In this paper we explore the use of audio visual synchrony to locate sound source s. We developed a system that searches for regions of the visual land(cid:173) scape that correlate highly with the acoustic signals and tags them as likely to contain an acoustic source. We discuss our experience implementing the system, present results on a speaker localization task and discuss potential applications of the approach.
****

Predictive Sequence Learning in Recurrent Neocortical Circuits
Rajesh Rao, Terrence J. Sejnowski
Neocortical circuits are dominated by massive excitatory feedback: more than e ighty percent of the synapses made by excitatory cortical neurons are onto othe r excitatory cortical neurons. Why is there such massive re(cid:173) current ex citation in the neocortex and what is its role in cortical compu(cid:173) tation ? Recent neurophysiological experiments have shown that the plas(cid:173) ticity of recurrent neocortical synapses is governed by a temporally asym(cid:173) met ric Hebbian learning rule. We describe how such a rule may allow the cortex to modify recurrent synapses for prediction of input sequences. The goal is to predict the next cortical input from the recent past based on previous experien ce of similar input sequences. We show that a temporal difference learning rule for prediction used in conjunction with dendritic back-propagating action pote

ntials reproduces the temporally asymmet(cid:173) ric Hebbian plasticity observe d physiologically. Biophysical simulations demonstrate that a network of cortic al neurons can learn to predict mov(cid:173) ing stimuli and develop direction s elective responses as a consequence of learning. The space-time response proper ties of model neurons are shown to be similar to those of direction selective c ells in alert monkey VI.
****
Differentiating Functions of the Jacobian with Respect to the Weights
Gary Flake, Barak Pearlmutter
For many problems, the correct behavior of a model depends not only on its inpu t-output mapping but also on properties of its Jacobian matrix, the matrix of p artial derivatives of the model's outputs with respect to its in(cid:173) puts. We introduce the J-prop algorithm, an efficient general method for computing t he exact partial derivatives of a variety of simple functions of the Jacobian o f a model with respect to its free parameters. The algorithm applies to any par ametrized feedforward model, including nonlinear re(cid:173) gression, multilay er perceptrons, and radial basis function networks.
****
Effects of Spatial and Temporal Contiguity on the Acquisition of Spatial Informa tion
Thea Ghiselli-Crippa, Paul Munro
Spatial information comes in two forms: direct spatial information (for examp le, retinal position) and indirect temporal contiguity information, since objec ts encountered sequentially are in general spatially close. The acquisition of spatial information by a neural network is investigated here. Given a s patial layout of several objects, networks are trained on a prediction task. N etworks using temporal sequences with no direct spa(cid:173) tial information ar e found to develop internal representations that show distances correlated wit h distances in the external layout. The influence of spatial information is an alyzed by providing direct spatial information to the system during training th at is either consistent with the layout or inconsistent with it. This appro ach allows examination of the relative contributions of spatial and temporal contiguity.
****
Agglomerative Information Bottleneck
Noam Slonim, Naftali Tishby
We introduce a novel distributional clustering algorithm that max(cid:173) imiz es the mutual information per cluster between data and giv(cid:173) en categories. This algorithm can be considered as a bottom up hard versi on of the recently introduced "Information Bottleneck Method". The algor ithm is compared with the top-down soft ver(cid:173) sion of the information bottleneck method and a relationship be(cid:173) tween the hard and soft results is established. We demonstrate the algorithm on the 20 Newsgroups dat a set. For a subset of two news(cid:173) groups we achieve compression by 3 orders of magnitudes loosing only 10% of the original mutual information.
****
Spike-based Learning Rules and Stabilization of Persistent Neural Activity
Xiaohui Xie, H. Sebastian Seung
We analyze the conditions under which synaptic learning rules based on action potential timing can be approximated by learning rules based on f iring rates. In particular, we consider a form of plasticity in which syn apses depress when a presynaptic spike is followed by a postsynaptic spike, and potentiate with the opposite temporal ordering. Such differen(cid:173) tial an ti-Hebbian plasticity can be approximated under certain conditions by a learnin g rule that depends on the time derivative of the postsynaptic firing rate. Su ch a learning rule acts to stabilize persistent neural activity patterns in rec urrent neural networks.
****
Nonlinear Discriminant Analysis Using Kernel Functions
Volker Roth, Volker Steinhage

Fishers linear discriminant analysis (LDA) is a classical multivari(cid:173) ate technique both for dimension reduction and classification. The data vect ors are transformed into a low dimensional subspace such that the class cent roids are spread out as much as possible. In this subspace LDA works as a simple prototype classifier with lin(cid:173) ear decision boundaries. However, in many applications the linear boundaries do not adequately sep arate the classes. We present a nonlinear generalization of discriminant an alysis that uses the ker(cid:173) nel trick of representing dot products by kern el functions. The pre(cid:173) sented algorithm allows a simple formulation of the EM-algorithm in terms of kernel functions which leads to a unique concept f or un(cid:173) supervised mixture analysis, supervised discriminant analysis and semi-supervised discriminant analysis with partially unlabelled ob(cid:17 3) servations in feature spaces.
****

On Input Selection with Reversible Jump Markov Chain Monte Carlo Sampling
Peter Sykacek
In this paper we will treat input selection for a radial basis function (RBF) l ike classifier within a Bayesian framework. We approximate the a-posteriori di stribution over both model coefficients and input subsets by samples drawn wi th Gibbs updates and reversible jump moves. Using some public datasets, we compare the classification accuracy of the method with a conventional AR D scheme. These datasets are also used to infer the a-posteriori probab ilities of dif(cid:173) ferent input subsets.
****

Uniqueness of the SVM Solution
Christopher J. C. Burges, David Crisp
We give necessary and sufficient conditions for uniqueness of the suppo rt vector solution for the problems of pattern recognition and regression estim ation, for a general class of cost functions. We show that if the solution is not unique, all support vectors are necessarily at bound, and we give some simple examples of non-unique solu(cid:173) tions. We note that uniqueness o f the primal (dual) solution does not necessarily imply uniqueness of the dual (primal) solution. We show how to compute the threshold b when the solution is unique, but when all support vectors are at bound, in which case the usual method for determining b does not work.
****

The Parallel Problems Server: an Interactive Tool for Large Scale Machine Learni ng
Charles Isbell, Parry Husbands
Imagine that you wish to classify data consisting of tens of thousands of ex(cid :173) amples residing in a twenty thousand dimensional space. How can one ap(ci d:173) ply standard machine learning algorithms? We describe the Parallel Prob(c id:173) lems Server (PPServer) and MATLAB*P. In tandem they allow users of n etworked computers to work transparently on large data sets from within Matlab. This work is motivated by the desire to bring the many benefits of scienti fic computing algorithms and computational power to machine learning researche rs. We demonstrate the usefulness of the system on a number of tasks. For example, we perform independent components analysis on very large text corpora consisting of tens of thousands of documents, making minimal changes to the original Bell and Sejnowski Matlab source (Bell and Se(cid:173) jnowski , 1995). Applying ML techniques to data previously beyond their reach leads t o interesting analyses of both data and algorithms.
****

Search for Information Bearing Components in Speech
Howard Yang, Hynek Hermansky
In this paper, we use mutual information to characterize the dis(cid:173) tributions of phonetic and speaker/channel information in a time(cid:173) f requency space. The mutual information (MI) between the pho(cid:173) neti c label and one feature, and the joint mutual information (JMI) between the ph onetic label and two or three features are estimated . The Miller's bias formul

as for entropy and mutual information es(cid:173) timates are extended to include higher order terms. The MI and the JMI for speaker/channel recognition are also estimated. The results are complementary to those for phonetic classification. Our results show how the phonetic information is locally spread and how the speaker/channel information is globally spread in time and frequency.
****

An Oscillatory Correlation Frame work for Computational Auditory Scene Analysis
Guy Brown, DeLiang Wang
A neural model is described which uses oscillatory correlation to segregate speech from interfering sound sources. The core of the model is a two-layer neural oscillator network. A sound stream is represented by a synchronized population of oscillators, and different streams are represented by desynchronized oscillator populations. The model has been evaluated using a corpus of speech mixed with interfering sounds, and produces an improvement in signal-to-noise ratio for every mixture.
****

Manifold Stochastic Dynamics for Bayesian Learning
Mark Zlochin, Yoram Baram
We propose a new Markov Chain Monte Carlo algorithm which is a gen(cid:173) eralization of the stochastic dynamics method. The algorithm performs exploration of the state space using its intrinsic geometric structure, facil(cid:173) itating efficient sampling of complex distributions. Applied to Bayesian learning in neural networks, our algorithm was found to perform at least as well as the best state-of-the-art method while consuming considerably less time.
****

Bifurcation Analysis of a Silicon Neuron
Girish Patel, Gennady Cymbalyuk, Ronald Calabrese, Stephen DeWeerth
We have developed a VLSI silicon neuron and a corresponding mathe(cid:173) matical model that is a two state-variable system. We describe the cir(cid:173) cuit implementation and compare the behaviors observed in the silicon neuron and the mathematical model. We also perform bifurcation analy(cid:173) sis of the mathematical model by varying the externally applied current and show that the behaviors exhibited by the silicon neuron under corre(cid:173) sponding conditions are in good agreement to those predicted by the bifurcation analysis.
****

Speech Modelling Using Subspace and EM Techniques
Gavin Smith, João de Freitas, Tony Robinson, Mahesan Niranjan
The speech waveform can be modelled as a piecewise-stationary linear stochastic state space system, and its parameters can be estimated using an expectation-maximisation (EM) algorithm. One problem is the ini(cid:173) tialisation of the EM algorithm. Standard initialisation schemes can lead to poor formant trajectories. But these trajectories however are impor(cid:173) tant for vowel intelligibility. The aim of this paper is to investigate the suitability of subspace identification methods to initialise EM. The paper compares the subspace state space system identification (4SID) method with the EM algorithm. The 4SID and EM methods are similar in that they both estimate a state sequence (but using Kalman fil(cid:173) ters and Kalman smoothers respectively), and then estimate parameters (but using least-squares and maximum likelihood respectively). The sim(cid:173) ilarity of 4SID and EM motivates the use of 4SID to initialise EM. Also, 4SID is non-iterative and requires no initialisation, whereas EM is itera(cid:173) tive and requires initialisation. However 4SID is sub-optimal compared to EM in a probabilistic sense. During experiments on real speech, 4SID methods compare favourably with conventional initialisation techniques. They produce smoother formant trajectories, have greater frequency res(cid:173) olution, and produce higher likelihoods.
****

Optimal Kernel Shapes for Local Linear Regression
Dirk Ormoneit, Trevor Hastie

Local linear regression performs very well in many low-dimensional forecasting problems. In high-dimensional spaces, its performance typically decays due to the well-known "curse-of-dimensionality". A possible way to approach this problem is by varying the "shape" of the weighting kernel. In this work we s uggest a new, data-driven method to estimating the optimal kernel shape. Exp eriments us(cid:173) ing an artificially generated data set and data from the UC Irvine repository show the benefits of kernel shaping.
****
Robust Neural Network Regression for Offline and Online Learning
Thomas Briegel, Volker Tresp
We replace the commonly used Gaussian noise model in nonlinear regress ion by a more flexible noise model based on the Student-t(cid:173) dis tribution. The degrees of freedom of the t-distribution can be chosen such th at as special cases either the Gaussian distribution or the Cauchy distributio n are realized. The latter is commonly used in robust regres(cid:173) sion. Since the t-distribution can be interpreted as being an infinite mix(cid:173) ture of Gaussians, parameters and hyperparameters such as the degrees of freedom of the t-distribution can be learned from the data based on an EM-learn ing algorithm. We show that modeling using the t-distribution leads to imp roved predictors on real world data sets. In particular, if outliers a re present, the t-distribution is superior to the Gaussian noise model. In effect, by adapting the degrees of freedom, the system can "learn" t o distinguish between outliers and non-outliers. Especially for online lear ning tasks, one is interested in avoiding inappropriate weight changes due to measurement outliers to maintain stable online learn(cid:173) ing cap ability. We show experimentally that using the t-distribution as a noise mo del leads to stable online learning algorithms and outperforms state-of-the art online learning methods like the extended Kalman filter algorithm.
****
Neural Network Based Model Predictive Control
Stephen Piche, James Keeler, Greg Martin, Gene Boe, Doug Johnson, Mark Gerules
Model Predictive Control (MPC), a control algorithm which uses an optim izer to solve for the optimal control moves over a future time horizon based upon a model of the process, has become a stan(cid:173) dard control techn ique in the process industries over the past two decades. In most indus trial applications, a linear dynamic model developed using empirical data is used even though the process it(cid:173) self is often nonlinear. Linear mo dels have been used because of the difficulty in developing a generic nonl inear model from empirical data and the computational expense often invol ved in using non(cid:173) linear models. In this paper, we present a g eneric neural network based technique for developing nonlinear dynamic models from em(cid:173) pirical data and show that these models can be efficiently used in a model predictive control framework. This nonlinear MPC based appr oach has been successfully implemented in a number of indus(cid:173) trial app lications in the refining, petrochemical, paper and food industries. Pe rformance of the controller on a nonlinear industrial process, a polyethylen e reactor, is presented.
****
Coastal Navigation with Mobile Robots
Nicholas Roy, Sebastian Thrun
The problem that we address in this paper is how a mobile robot can pl an in order to arrive at its goal with minimum uncertainty. Tradition al motion planning algo(cid:173) rithms often assume that a mobile robot can t rack its position reliably, however, in real world situations, reliable localiz ation may not always be feasible. Partially Observable Markov Decision Process es (POMDPs) provide one way to maximize the certainty of reaching the goal st ate, but at the cost of computational intractability for large state spaces. The method we propose explicitly models the uncertainty of the robot's p osition as a state variable, and generates trajectories through the aug mented pose-uncertainty space. By minimizing the positional uncertainty

at the goal, the robot reduces the likelihood it becomes lost. We dem onstrate experimentally that coastal navigation reduces the uncertainty at the goal, especially with degraded localization.
****
Boosting with Multi-Way Branching in Decision Trees
Yishay Mansour, David McAllester
It is known that decision tree learning can be viewed as a form of boosting. However, existing boosting theorems for decision tree learning all ow only binary-branching trees and the generalization to multi-branching trees is not immediate. Practical decision tree al(cid:173) gorithms, such as CART and C4.5, implement a trade-off between the number of branches and the i mprovement in tree quality as measured by an index function. Here we gi ve a boosting justifica(cid:173) tion for a particular quantitative trade-off c urve. Our main theorem states, in essence, that if we require an improve ment proportional to the log of the number of branches then top-down gree dy con(cid:173) struction of decision trees remains an effective boosting algo rithm.
****
Neural Representation of Multi-Dimensional Stimuli
Christian Eurich, Stefan Wilke, Helmut Schwegler
Spatial information comes in two forms: direct spatial information (for examp le, retinal position) and indirect temporal contiguity information, since objec ts encountered sequentially are in general spatially close. The acquisition of spatial information by a neural network is investigated here. Given a s patial layout of several objects, networks are trained on a prediction task. N etworks using temporal sequences with no direct spa(cid:173) tial information ar e found to develop internal representations that show distances correlated wit h distances in the external layout. The influence of spatial information is an alyzed by providing direct spatial information to the system during training th at is either consistent with the layout or inconsistent with it. This appro ach allows examination of the relative contributions of spatial and temporal contiguity.
****
Model Selection for Support Vector Machines
Olivier Chapelle, Vladimir Vapnik
New functionals for parameter (model) selection of Support Vector Ma(cid:173) ch ines are introduced based on the concepts of the span of support vec(cid:173) to rs and rescaling of the feature space. It is shown that using these func(cid:17 3) tionals, one can both predict the best choice of parameters of the model and the relative quality of performance for any value of parameter.
****
Memory Capacity of Linear vs. Nonlinear Models of Dendritic Integration
Panayiota Poirazi, Bartlett Mel
Previous biophysical modeling work showed that nonlinear interac(cid:173) tions among nearby synapses located on active dendritic trees can provide a large boo st in the memory capacity of a cell (Mel, 1992a, 1992b). The aim of our pres ent work is to quantify this boost by estimating the capacity of (1) a neuron model with passive den(cid:173) dritic integration where inputs are combined linearly across the entire cell followed by a single globa l threshold, and (2) an active dendrite model in which a threshold is applied separately to the output of each branch, and the branch subtotals are combined lin(cid:173) early. We focus here on the limiting case of binary-valued synaptic weights, and derive expressions which measure model capacity by estimating the number of distinct input-output functions available to bot h neuron types. We show that (1) the application of a fixed nonlinearity to each dendritic compartment substantially increases the model's flexibility, (2) for a neuron of realistic size, the capacity of the nonlinear cell can excee d that of the same-sized linear cell by more than an order of magnitude, and (3) the largest capacity boost occurs for cells with a relatively large number o f dendritic subunits of relatively small size. We validated the analysis

by empirically measuring memory capacity with randomized two-class classifi ca(cid:173) tion problems, where a stochastic delta rule was used to train both linear and nonlinear models. We found that large capacity boosts predicted for the nonlinear dendritic model were readily achieved in practice.
****

State Abstraction in MAXQ Hierarchical Reinforcement Learning
Thomas Dietterich
Many researchers have explored methods for hierarchical reinforce(cid:173) ment learning (RL) with temporal abstractions, in which abstract actions are defined that can perform many primitive actions before terminating. However, little is known about learning with state ab(cid:173) stractions, in which aspec ts of the state space are ignored. In previ(cid:173) ous work, we developed th e MAXQ method for hierarchical RL. In this paper, we define five conditions under which state abstraction can be combined with the MAXQ value funct ion decomposition. We prove that the MAXQ-Q learning algorithm converge s under these conditions and show experimentally that state abstraction is im portant for the successful application of MAXQ-Q learning.
****

Modeling High-Dimensional Discrete Data with Multi-Layer Neural Networks
Yoshua Bengio, Samy Bengio
The curse of dimensionality is severe when modeling high-dimensional discrete data: the number of possible combinations of the variables ex(cid:173) plodes e xponentially. In this paper we propose a new architecture for modeling hig h-dimensional data that requires resources (parameters and computations) that g row only at most as the square of the number of vari(cid:173) ables, using a mu lti-layer neural network to represent the joint distribu(cid:173) tion of the v ariables as the product of conditional distributions. The neu(cid:173) ral netw ork can be interpreted as a graphical model without hidden ran(cid:173) dom va riables, but in which the conditional distributions are tied through the hidden units. The connectivity of the neural network can be pruned by using dependenc y tests between the variables. Experiments on modeling the distribution of seve ral discrete data sets show statistically significant improvements over other m ethods such as naive Bayes and comparable Bayesian networks, and show that s ignificant improvements can be ob(cid:173) tained by pruning the network.
****

A Generative Model for Attractor Dynamics
Richard Zemel, Michael C. Mozer
Attractor networks, which map an input space to a discrete out(cid:173) put spac e, are useful for pattern completion. However, designing a net to have a given set of attractors is notoriously tricky; training procedures are CPU intensive and often produce spurious afuac(cid:173) tors and ill-conditioned attractor bas ins. These difficulties occur because each connection in the network particip ates in the encod(cid:173) ing of multiple attractors. We describe an alternati ve formulation of attractor networks in which the encoding of knowledge is loca l, not distributed. Although localist attractor networks have similar dynamics to their distributed counterparts, they are much easier to work with and inter pret. We propose a statistical formulation of localist attract or net dynamics, which yields a convergence proof and a mathematical interpretation of model pa rameters.
****

An Environment Model for Nonstationary Reinforcement Learning
Samuel Choi, Dit-Yan Yeung, Nevin Zhang
Reinforcement learning in nonstationary environments is generally regarded as an important and yet difficult problem. This paper partially addresses the problem by formalizing a subclass of nonsta(cid:173) tionary environments. The environment model, called hidden-mode Markov decision process (HM-MDP), assumes that environmental changes are always confined to a small numb er of hidden modes. A mode basically indexes a Markov decision process (MDP) and evolves with time according to a Markov chain. While HM-MDP is a special case of partially observable Markov decision processes (P

OMDP), modeling an HM-MDP environment via the more gen(cid:173) eral POMDP model unnecessarily increases the problem complex(cid:173) ity. A variant of the Baum-Welch algorithm is developed for model learning requiring less data and time.
****
Predictive App roaches for Choosing Hyperparameters in Gaussian Processes
S. Sundararajan, S. Keerthi
Gaussian Processes are powerful regression models specified by parametrized mean and covariance functions. Standard approaches to estimate these parameters (known by the name Hyperparam(cid:173) eters) are Maximum Likelihood (ML) and Maximum APosterior (MAP) approaches. In this paper, we propose and investigate pre(cid:173) dictive approaches, namely, maximization of Geisser's Surrogate Predictive Probability (GPP) and minimization of mean square er(cid:173) ror with respect to GPP (referred to as Geisser's Predictive mean square Error (GPE)) to estimate the hyperparameters. We also derive results for the standard Cross-Validation (CV) error and make a comparison. These approaches are tested on a number of problems and experimental results show that these approaches are strongly competitive to existing approaches.
****
Acquisition in Autoshaping
Sham Kakade, Peter Dayan
Quantitative data on the speed with which animals acquire behav(cid:173) ioral responses during classical conditioning experiments should provide strong constraints on models of learning. However, most models have simply ignored these data; the few that have attempt(cid:173) ed to address them have failed by at least an order of magnitude. We discuss key data on the speed of acquisition, and show how to account for them using a statistically sound model of learning, in which differential reliabilities of stimuli playa crucial role.
****
Transductive Inference for Estimating Values of Functions
Olivier Chapelle, Vladimir Vapnik, Jason Weston
We introduce an algorithm for estimating the values of a function at a set of test points Xe+!, ... , xl+m given a set of training points (XI,YI), ... ,(xe,Ye) without estimating (as an intermediate step) the regression function . We demonstrate that this direct (transduc(cid:173) ti ve) way for estimating values of the regression (or classification in pattern recognition) can be more accurate than the tradition(cid:173) alone based on two steps, first estimating the function and then calculating the values of this function at the points of interest.
****
Effective Learning Requires Neuronal Remodeling of Hebbian Synapses
Gal Chechik, Isaac Meilijson, Eytan Ruppin
This paper revisits the classical neuroscience paradigm of Hebbian learning. We find that a necessary requirement for effective as(cid:173) sociative memory learning is that the efficacies of the incoming synapses should be uncorrelated. This requirement is difficult to achieve in a robust manner by Hebbian synaptic learning, since it depends on network level information. Effective learning can yet be obtained by a neuronal process that maintains a zero sum of the in(cid:173) coming synaptic efficacies. This normalization drastically improves the memory capacity of associative networks, from an essentially bounded capacity to one that linearly scales with the network's size. It also enables the effective storage of patterns with heterogeneous coding levels in a single network. Such neuronal normalization can be successfully carried out by activity-dependent homeostasis of the neuron's synaptic efficacies, which was recently observed in cortical tissue. Thus, our findings strongly suggest that effective associa(cid:173) tive learning with Hebbian synapses alone is biologically implausi(cid:173) ble and that Hebbian synapses must be continuously remodeled by neuronally-driven regulatory processes in the brain.

****
The Relevance Vector Machine
Michael Tipping
The support vector machine (SVM) is a state-of-the-art technique for regress ion and classification, combining excellent generalisation properties with a sparse kernel representation. However, it does suffer from a number of di sadvantages, notably the absence of prob(cid:173) abilistic outputs, the require ment to estimate a trade-off parameter and the need to utilise 'Mercer' kernel functions. In this paper we introduce the Relevance Vector Machine (RVM), a Bayesian treat(cid:173) ment of a generalised linear model of identical functional form to the SVM. The RVM suffers from none of the above disadv antages, and examples demonstrate that for comparable generalisation per(cid:1 73) formance, the RVM requires dramatically fewer kernel functions.
****
Dual Estimation and the Unscented Transformation
Eric Wan, Rudolph van der Merwe, Alex Nelson
Dual estimation refers to the problem of simultaneously estimating the stat e of a dynamic system and the model which gives rise to the dynam(cid:173) ics. Algorithms include expectation-maximization (EM), dual Kalman filtering, and joint Kalman methods. These methods have recently been explored in the c ontext of nonlinear modeling, where a neural network is used as the functi onal form of the unknown model. Typically, an ex(cid:173) tended Kalman fil ter (EKF) or smoother is used for the part of the al(cid:173) gorithm th at estimates the clean state given the current estimated model. An EKF may al so be used to estimate the weights of the network. This paper points out the flaws in using the EKF, and proposes an improve(cid:173) ment based on a new approach called the unscented transformation (UT) [3]. A substantial per formance gain is achieved with the same order of computational complexity as that of the standard EKF. The approach is illustrated on several dual estima tion methods.
****
Learning Statistically Neutral Tasks without Expert Guidance
Ton Weijters, Antal van den Bosch, Eric Postma
Eric Postma
****
Gaussian Fields for Approximate Inference in Layered Sigmoid Belief Networks
David Barber, Peter Sollich
Local "belief propagation" rules of the sort proposed by Pearl [15] are guar anteed to converge to the correct posterior probabilities in singly conn ected graphical models. Recently, a number of researchers have em(cid:173) piric ally demonstrated good performance of "loopy belief propagation"(cid:173) using these same rules on graphs with loops. Perhaps the most dramatic instance is t he near Shannon-limit performance of "Turbo codes", whose decoding algorithm is equivalent to loopy belief propagation. Except for the case of graphs with a s ingle loop, there has been little theo(cid:173) retical understanding of the per formance of loopy propagation. Here we analyze belief propagation in networks with arbitrary topologies when the nodes in the graph describe jointly Gau ssian random variables. We give an analytical formula relating the true poste rior probabilities with those calculated using loopy propagation. We give suff icient conditions for convergence and show that when belief propagation converg es it gives the correct posterior means for all graph topologies, not just ne tworks with a single loop. The related "max-product" belief propagation algori thm finds the max(cid:173) imum posterior probability estimate for singly conne cted networks. We show that, even for non-Gaussian probability distributions , the conver(cid:173) gence points of the max-product algorithm in loopy netw orks are max(cid:173) ima over a particular large local neighborhood of the posterior proba(cid:173) bility. These results help clarify the empirical performance results and motivate using the powerful belief propagation alg orithm in a broader class of networks.
****

A Multi-class Linear Learning Algorithm Related to Winnow
Chris Mesterharm
In this paper, we present Committee, a new multi-class learning algo(cid:173) rithm related to the Winnow family of algorithms. Committee is an al(cid:173) gorithm for combining the predictions of a set of sub-experts in the on(cid:173) line mistake-bounded model oflearning. A sub-expert is a special type of attribute that predicts with a distribution over a finite number of classes. Committee learns a linear function of sub-experts and uses this function to make class predictions. We provide bounds for Committee that show it performs well when the target can be represented by a few relevant sub-experts. We also show how Committee can be used to solve more traditional problems composed of attributes. This leads to a natural ex(cid:173) tension that learns on multi-class problems that contain both traditional attributes and sub-experts.
****
Churn Reduction in the Wireless Industry
Michael C. Mozer, Richard Wolniewicz, David Grimes, Eric Johnson, Howard Kaushansky
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Recognizing Evoked Potentials in a Virtual Environment
Jessica Bayliss, Dana Ballard
Virtual reality (VR) provides immersive and controllable experimen(cid:173) tal environments. It expands the bounds of possible evoked potential (EP) experiments by providing complex, dynamic environments in or(cid:173) der to study cognition without sacrificing environmental control. VR also serves as a safe dynamic testbed for brain-computer .interface (BCl) research. However, there has been some concern about detecting EP sig(cid:173) nals in a complex VR environment. This paper shows that EPs exist at red, green, and yellow stop lights in a virtual driving environment. Ex(cid:173) perimental results show the existence of the P3 EP at "go" and "stop" lights and the contingent negative variation (CNY) EP at "slow down" lights. In order to test the feasibility of on-line recognition in VR, we looked at recognizing the P3 EP at red stop tights and the absence of this signal at yellow slow down lights. Recognition results show that the P3 may successfully be used to control the brakes of a VR car at stop lights.
****