Reinforcement Learning with Function Approximation Converges to a Region
Geoffrey J. Gordon
Many algorithms for approximate reinforcement learning are not known to converge. In fact, there are counterexamples showing that the adjustable weights in some algorithms may oscillate within a region rather than convergin g to a point. This paper shows that, for two popular algorithms, such osci llation is the worst that can happen: the weights cannot diverge, but instead must converge to a bounded region. The algorithms are SARSA(O) a nd V(O); the latter algorithm was used in the well-known TD-Gammon program.
****
Redundancy and Dimensionality Reduction in Sparse-Distributed Representations of Natural Objects in Terms of Their Local Features
Penio Penev
Low-dimensional representations are key to solving problems in high(cid:173) lev el vision, such as face compression and recognition. Factorial coding strategie s for reducing the redundancy present in natural images on the basis of their s econd-order statistics have been successful in account(cid:173) ing for both psy chophysical and neurophysiological properties of early vision. Class-specific r epresentations are presumably formed later, at the higher-level stages of corti cal processing. Here we show that when retinotopic factorial codes are derived for ensembles of natural objects, such as human faces, not only redundancy, but also dimensionality is re(cid:173) duced. We also show that objects are built f rom parts in a non-Gaussian fashion which allows these local-feature codes to h ave dimensionalities that are substantially lower than the respective Nyquist s ampling rates.
****
Who Does What? A Novel Algorithm to Determine Function Localization
Ranit Aharonov-Barki, Isaac Meilijson, Eytan Ruppin
We introduce a novel algorithm, termed PPA (Performance Prediction Algo rithm), that quantitatively measures the contributions of elements of a n eural system to the tasks it performs. The algorithm identifies the neurons o r areas which participate in a cognitive or behavioral task, given data about p erformance decrease in a small set of lesions. It also allows the accurate p rediction of performances due to multi-element lesions. The effectiveness of the new algorithm is demonstrated in two models of recurrent neural networks with complex interactions among the ele(cid:173) ments. The algorith m is scalable and applicable to the analysis of large neural networks. Giv en the recent advances in reversible inactivation techniques, it has the potential to significantly contribute to the under(cid:173) standing of the organization of biological nervous systems, and to shed light on the lo ng-lasting debate about local versus distributed computa(cid:173) tion in the b rain.
****
Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopp ing
Rich Caruana, Steve Lawrence, C. Giles
The conventional wisdom is that backprop nets with excess hidden units generali ze poorly. We show that nets with excess capacity generalize well when trained with backprop and early stopping. Experiments sug(cid:173) gest two reasons for this: 1) Overfitting can vary significantly in different regions of the model. Excess capacity allows better fit to regions of high non-lineari ty, and backprop often avoids overfitting the regions of low non-lineari ty. 2) Regardless of size, nets learn task subcomponents in similar seq uence. Big nets pass through stages similar to those learned by smaller net s. Early stopping can stop training the large net when it generalizes com parably to a smaller net. We also show that conjugate gradient can yield worse generalization because it overfits regions of low non-linearity when lea rning to fit regions of high non-linearity.
****
Color Opponency Constitutes a Sparse Representation for the Chromatic Structure

of Natural Scenes
Te-Won Lee, Thomas Wachtler, Terrence J. Sejnowski
The human visual system encodes the chromatic signals conveyed by the three t ypes of retinal cone photoreceptors in an opponent fashion. This color opponency has been shown to constitute an efficient encoding by spectral decorrelation of the receptor signals. We analyze the spatial and chromatic st ructure of natural scenes by decomposing the spectral images into a set of line ar basis functions such that they constitute a representation with minim al redun(cid:173) dancy. Independent component analysis finds the basis functions that transforms the spatiochromatic data such that the output s (activations) are statistically as independent as possible, i.e. least redundant. The resulting basis functions show strong opponency along an achromatic direction (luminance edges), along a blue(cid:173) yellow dir ection, and along a red-blue direction. Furthermore, the resulting activatio ns have very sparse distributions, suggesting that the use of color opponency in the human visual system achieves a highly efficient representation of col ors. Our findings suggest that color opponency is a result of the properties of natural spectra and not solely a consequence of the overlapping cone spec tral sensitiv(cid:173) ities.
****
Learning Segmentation by Random Walks
Marina Meila, Jianbo Shi
We present a new view of image segmentation by pairwise simi(cid:173) larities. We interpret the similarities as edge flows in a Markov random walk and study t he eigenvalues and eigenvectors of the walk's transition matrix. This interpret ation shows that spectral methods for clustering and segmentation have a probab ilistic foun(cid:173) dation. In particular, we prove that the Normalized Cut me thod arises naturally from our framework. Finally, the framework pro(cid:173) v ides a principled method for learning the similarity function as a combination of features.
****
A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work
Ralf Herbrich, Thore Graepel
We present a bound on the generalisation error of linear classifiers in term s of a refined margin quantity on the training set. The result is o btained in a PAC- Bayesian framework and is based on geometrical argume nts in the space of linear classifiers. The new bound constitutes an expon ential improvement of the so far tightest margin bound by Shawe-Taylor et al. [8] and scales logarithmically in the inverse margin. Even in the case of less training examples than input dimensions sufficiently large margins lead to non-trivial bound values and - plexity term. Furthermore, the classical margin is too coarse a measure for the essential quantity that contro ls the generalisation error: the volume ratio between the whole hypothes is space and the subset of consistent hypotheses. The practical relevance of the result lies in the fact that the well-known support vector machine is optimal w.r.t. the new bound only if the feature vectors are all of the same le ngth. As a consequence we recommend to use SVMs on normalised feature vectors only - a recommendation that is well supported by our numerical experiment s on two benchmark data sets.
****
Active Learning for Parameter Estimation in Bayesian Networks
Simon Tong, Daphne Koller
Bayesian networks are graphical representations of probability distributio ns. In virtually all of the work on learning these networks, the assu mption is that we are presented with a data set consisting of randomly generated instances from the underlying distribution. In many situations , however, we also have the option of active learning, where we have th e possibility of guiding the sampling process by querying for certain types of samples. This paper addresses the problem of estimating the parameters of Bayesian networks in an active learning setting. We provide a theoretica

l framework for this problem, and an algorithm that chooses which acti ve learning queries to generate based on the model learned so far. W e present experimental results showing that our active learning algorithm can si gnificantly reduce the need for training data in many situations.
****
A Tighter Bound for Graphical Models
Martijn Leisink, Hilbert Kappen
We present a method to bound the partition function of a Boltz(cid:173) m ann machine neural network with any odd order polynomial. This is a direct e xtension of the mean field bound, which is first order. We show that t he third order bound is strictly better than mean field. Additionally we show the rough outline how this bound is applicable to sigmoid be lief networks. Numerical experiments in(cid:173) dicate that an error redu ction of a factor two is easily reached in the region where expansion based approximations are useful.
****
Occam's Razor
Carl Rasmussen, Zoubin Ghahramani
The Bayesian paradigm apparently only sometimes gives rise to Occam's Razor; a t other times very large models perform well. We give simple examples of both kinds of behaviour. The two views are reconciled when measuring complexity of functions, rather than of the machinery used to implement them. We analyze the complexity of functions for some linear in the parameter models that are equivalent to Gaussian Processes, and always find Occam's Razor at work.
****
Exact Solutions to Time-Dependent MDPs
Justin Boyan, Michael Littman
We describe an extension of the Markov decision process model in which a con tinuous time dimension is included in the state space. This allows for t he representation and exact solution of a wide range of problems in whi ch transitions or rewards vary over time. We examine problems based on rou te planning with public trans(cid:173) portation and telescope observation sche duling.
****
An Information Maximization Approach to Overcomplete and Recurrent Representatio ns
Oren Shriki, Haim Sompolinsky, Daniel Lee
The principle of maximizing mutual information is applied to learning overco mplete and recurrent representations. The underlying model con(cid:173) sists of a network of input units driving a larger number of output units with recurr ent interactions. In the limit of zero noise, the network is de(cid:173) termin istic and the mutual information can be related to the entropy of the out put units. Maximizing this entropy with respect to both the feed(cid:173) forw ard connections as well as the recurrent interactions results in simple learnin g rules for both sets of parameters. The conventional independent components (ICA) learning algorithm can be recovered as a special case where there is an equal number of output units and no recurrent con(cid:173) nections. Th e application of these new learning rules is illustrated on a simple two-dime nsional input example.
****
A Linear Programming Approach to Novelty Detection
Colin Campbell, Kristin Bennett
Novelty detection involves modeling the normal behaviour of a sys(cid:173) tem h ence enabling detection of any divergence from normality. It has potential a pplications in many areas such as detection of ma(cid:173) chine damage o r highlighting abnormal features in medical data. One approach is to bu ild a hypothesis estimating the support of the normal data i.e. construct ing a function which is positive in the region where the data is located a nd negative elsewhere. Recently kernel methods have been proposed for e stimating the support of a distribution and they have performed well in

practice - training involves solution of a quadratic programming problem. I
n this pa(cid:173) per we propose a simpler kernel method for estimating the su
pport based on linear programming. The method is easy to implement an
d can learn large datasets rapidly. We demonstrate the method on medical an
d fault detection datasets.
****
Programmable Reinforcement Learning Agents
David Andre, Stuart Russell
We present an expressive agent design language for reinforcement learn(cid:173)
ing that allows the user to constrain the policies considered by the learn(cid:1
73) ing process.The language includes standard features such as parameter(cid:
173) ized subroutines, temporary interrupts, aborts, and memory variables, but
also allows for unspecified choices in the agent program. For learning t
hat which isn't specified, we present provably convergent learning algo(cid:173)
 rithms. We demonstrate by example that agent programs written in the langua
ge are concise as well as modular. This facilitates state abstraction and the
transferability of learned skills.
****
Learning Joint Statistical Models for Audio-Visual Fusion and Segregation
John W. Fisher III, Trevor Darrell, William Freeman, Paul Viola
People can understand complex auditory and visual information, often using
 one to disambiguate the other. Automated analysis, even at a low(cid:173) leve
l, faces severe challenges, including the lack of accurate statistical mod
els for the signals, and their high-dimensionality and varied sam(cid:17
3) pling rates. Previous approaches [6] assumed simple parametric models for
 the joint distribution which, while tractable, cannot capture the com(cid:173)
plex signal relationships. We learn the joint distribution of the visual and au
ditory signals using a non-parametric approach. First, we project the data i
nto a maximally informative, low-dimensional subspace, suitable for density
 estimation. We then model the complicated stochastic rela(cid:173) tionship
s between the signals using a nonparametric density estimator. These learn
ed densities allow processing across signal modalities. We demonstrate, o
n synthetic and real signals, localization in video of the face that is
 speaking in audio, and, conversely, audio enhancement of a particular speake
r selected from the video.
****
Convergence of Large Margin Separable Linear Classification
Tong Zhang
Large margin linear classification methods have been successfully ap(cid:
173) plied to many applications. For a linearly separable problem, it is known
 that under appropriate assumptions, the expected misclassification error of th
e computed "optimal hyperplane" approaches zero at a rate propor(cid:173) tional
 to the inverse training sample size. This rate is usually charac(cid:
173) terized by the margin and the maximum norm of the input data. In this p
aper, we argue that another quantity, namely the robustness of the in(cid:
173) put data distribution, also plays an important role in characterizing
 the convergence behavior of expected misclassification error. Based on this
 concept of robustness, we show that for a large margin separable linear cla
ssification problem, the expected misclassification error may converge exponent
ially in the number of training sample size.
****
Emergence of Movement Sensitive Neurons' Properties by Learning a Sparse Code fo
r Natural Moving Images
Rafal Bogacz, Malcolm Brown, Christophe Giraud-Carrier
Olshausen & Field demonstrated that a learning algorithm that attempts
 to generate a sparse code for natural scenes develops a complete fam
ily of localised, oriented, bandpass receptive fields, similar to those
 of 'simple cells' in VI. This paper describes an algorithm which fin
ds a sparse code for sequences of images that preserves information abo
ut the input. This algorithm when trained on natural video sequences de

velops bases representing the movement in particular directions with parti cular speeds, similar to the receptive fields of the movement-sensitive cells observed in cortical visual areas. Furthermore, to previous appro aches to learning direction selectivity, the timing of neuronal activity encodes the phase of the movement, so the precise timing of spikes is crucially important to the information encoding.
****
High-temperature Expansions for Learning Models of Nonnegative Data
Oliver Downs
Recent work has exploited boundedness of data in the unsupervised learn ing of new types of generative model. For nonnegative data it was recently sh own that the maximum-entropy generative model is a Non(cid:173) negative B oltzmann Distribution not a Gaussian distribution, when the model is con strained to match the first and second order statistics of the data. Learning for practical sized problems is made difficult by the need to compute expect ations under the model distribution. The computa(cid:173) tional cost of Markov chain Monte Carlo methods and low fidelity of naive mean field techniques has led to increasing interest in advanced mean field theorie s and variational methods. Here I present a second(cid:173) order mean-fie ld approximation for the Nonnegative Boltzmann Machine model, obtained using a "high-temperature" expansion. The theory is tested on learning a bimo dal 2-dimensional model, a high-dimensional translationally invariant distrib ution, and a generative model for hand(cid:173) written digits.
****
A Support Vector Method for Clustering
Asa Ben-Hur, David Horn, Hava Siegelmann, Vladimir Vapnik
We present a novel method for clustering using the support vector ma(cid:173) chine approach. Data points are mapped to a high dimensional feature space , where support vectors are used to define a sphere enclosing them. The boundar y of the sphere forms in data space a set of closed contours containing the dat a. Data points enclosed by each contour are defined as a cluster. As the width parameter of the Gaussian kernel is decreased, these contours fit the data mo re tightly and splitting of contours occurs. The algorithm works by separa ting clusters according to valleys in the un(cid:173) derlying probability distribution, and thus clusters can take on arbitrary geometrical shapes. A s in other SV algorithms, outliers can be dealt with by introducing a soft marg in constant leading to smoother cluster bound(cid:173) aries. The structure of the data is explored by varying the two parame(cid:173) ters. We investigat e the dependence of our method on these parameters and apply it to several dat a sets.
****
Incremental and Decremental Support Vector Machine Learning
Gert Cauwenberghs, Tomaso Poggio
An on-line recursive algorithm for training support vector machines, one vector at a time, is presented. Adiabatic increments retain the Kuhn(cid:173) Tucker conditions on all previously seen training data, in a number of steps each computed analytically. The incremental procedure is re(cid:17 3) versible, and decremental "unlearning" offers an efficient method to ex(cid: 173) actly evaluate leave-one-out generalization performance. Interpretation of decremental unlearning in feature space sheds light on the relationship bet ween generalization and geometry of the data.
****
Active Inference in Concept Learning
Jonathan Nelson, Javier Movellan
People are active experimenters, not just passive observers, constantly seeking new information relevant to their goals. A reasonable approach to active information gathering is to ask questions and conduct expe riments that maximize the expected information gain, given current beli efs (Fedorov 1972, MacKay 1992, Oaksford & Chater 1994). In this pape r we present results on an exploratory experiment designed to study p

eople's active information gathering behavior on a concept task (Tenenbaum 2000). The results of the experiment are analyzed in terms of the expected information gain of the questions asked by subjects.
****

The Interplay of Symbolic and Subsymbolic Processes in Anagram Problem Solving
David Grimes, Michael C. Mozer
Although connectionist models have provided insights into the nature of perception and motor control, connectionist accounts of higher cognition seldom go beyond an implementation of traditional symbol-processing theories. We describe a connectionist constraint satisfaction model of how people solve anagram problems. The model exploits statistics of English orthography, but also addresses the interplay of sub symbolic and symbolic computation by a mechanism that extracts approximate symbolic representations (partial orderings of letters) from sub symbolic structures and injects the extracted representation back into the model to assist in the solution of the anagram. We show the computational benefit of this extraction-injection process and discuss its relationship to conscious mental processes and working memory. We also account for experimental data concerning the difficulty of anagram solution based on the orthographic structure of the anagram string and the target word.
****

Using the Nyström Method to Speed Up Kernel Machines
Christopher Williams, Matthias Seeger
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.
****

Sex with Support Vector Machines
Baback Moghaddam, Ming-Hsuan Yang
Nonlinear Support Vector Machines (SVMs) are investigated for visual sex classification with low resolution "thumbnail" faces (21- by-12 pixels) processed from 1,755 images from the FE RET face database. The performance of SVMs is shown to be superior to traditional pattern classifiers (Linear, Quadratic, Fisher Linear Dis(cid:173) criminant, Nearest-Neighbor) as well as more modern techniques such as Radial Basis Function (RBF) classifiers and large ensemble(cid:173) RBF networks. Furthermore, the SVM performance (3.4% error) is currently the best result reported in the open literature.
****

Recognizing Hand-written Digits Using Hierarchical Products of Experts
Guy Mayraz, Geoffrey E. Hinton
The product of experts learning procedure [1] can discover a set of stochastic binary features that constitute a non-linear generative model of handwritten images of digits. The quality of generative models learned in this way can be assessed by learning a separate model for each class of digit and then comparing the unnormalized probabilities of test images under the 10 different class-specific models. To improve discriminative performance, it is helpful to learn a hierarchy of separate models for each digit class. Each model in the hierarchy has one layer of hidden units and the nth level model is trained on data that consists of the activities of the hidden units in the already trained (n - l)th level model. After train(cid:173) ing, each level produces a separate, unnormalized log probabilty score. With a three-level hierarchy for each of the 10 digit classes, a test image produces 30 scores which can be used as inputs to a supervised, logis(cid:173) tic classification network that is trained on separate data. On the MNIST database, our system is comparable with current state-of-the-art discrimi(cid:173) native methods, demonstrating that the product of experts learning proce(cid:173) dure can produce effective generative models of high-dimensional data.
****

APRICODD: Approximate Policy Construction Using Decision Diagrams

Robert St-Aubin, Jesse Hoey, Craig Boutilier
We propose a method of approximate dynamic programming for Markov decision proc
esses (MDPs) using algebraic decision diagrams (ADDs). We produce near-optimal
 value functions and policies with much lower time and space requirements t
han exact dynamic programming. Our method reduces the sizes of the inte
rmediate value functions generated during value iteration by replacing the val
ues at the terminals of the ADD with ranges of values. Our method is demons
trated on a class of large MDPs (with up to 34 billion states), and we compar
e the results with the optimal value functions.
****
Four-legged Walking Gait Control Using a Neuromorphic Chip Interfaced to a Suppo
rt Vector Learning Algorithm
Susanne Still, Bernhard Schölkopf, Klaus Hepp, Rodney Douglas
To control the walking gaits of a four-legged robot we present a novel neu
romorphic VLSI chip that coordinates the relative phasing of the robot'
s legs similar to how spinal Central Pattern Generators are believed to contro
l vertebrate locomotion [3]. The chip controls the leg move(cid:173) me
nts by driving motors with time varying voltages which are the out(cid:173)
puts of a small network of coupled oscillators. The characteristics of the chi
p's output voltages depend on a set of input parameters. The rela(cid:173
) tionship between input parameters and output voltages can be computed analyti
cally for an idealized system. In practice, however, this ideal re(cid:173)
 lationship is only approximately true due to transistor mismatch and off(cid:17
3) sets. Fine tuning of the chip's input parameters is done automatically by t
he robotic system, using an unsupervised Support Vector (SV) learning algorithm
 introduced recently [7]. The learning requires only that the descriptio
n of the desired output is given. The machine learns from (un(cid:173) labeled
) examples how to set the parameters to the chip in order to obtain a desired m
otor behavior.
****
A Mathematical Programming Approach to the Kernel Fisher Algorithm
Sebastian Mika, Gunnar Rätsch, Klaus-Robert Müller
We investigate a new kernel-based classifier: the Kernel Fisher Discrim(cid:173
) inant (KFD). A mathematical programming formulation based on the ob(cid:173) s
ervation that KFD maximizes the average margin permits an interesting modificat
ion of the original KFD algorithm yielding the sparse KFD. We find that both,
 KFD and the proposed sparse KFD, can be understood in an unifying prob
abilistic context. Furthermore, we show connections to Support Vector Machine
s and Relevance Vector Machines. From this understanding, we are able to
outline an interesting kernel-regression technique based upon the KFD algor
ithm. Simulations support the use(cid:173) fulness of our approach.
****
Using Free Energies to Represent Q-values in a Multiagent Reinforcement Learning
 Task
Brian Sallans, Geoffrey E. Hinton
The problem of reinforcement learning in large factored Markov decision process
es is explored. The Q-value of a state-action pair is approximated by the free
energy of a product of experts network. Network parameters are learned on-line
 using a modified SARSA algorithm which minimizes the inconsistency of the Q-va
lues of consecutive state-action pairs. Ac(cid:173) tions are chosen based on t
he current value estimates by fixing the current state and sampling actions fro
m the network using Gibbs sampling. The algorithm is tested on a co-operativ
e multi-agent task. The product of experts model is found to perform comparabl
y to table-based Q-Iearning for small instances of the task, and continues to
 perform well when the problem becomes too large for a table-based representat
ion.
****
Dopamine Bonuses
Sham Kakade, Peter Dayan
Substantial data support a temporal difference (TO) model of dopamine (OA) neur

on activity in which the cells provide a global error signal for reinforcement learning. However, in certain cir(cid:173) cumstances, OA activity seems anomalous under the TO model, responding to non-rewarding stimuli. We address these anoma(cid:173) lies by suggesting that OA cells multiplex information about re(cid:173) ward bonuses, including Sutton's exploration bonuses and Ng et al's non-distorting shaping bonuses. We interpret this additional role for OA in terms of the unconditional attentional and psy(cid:173) chomotor effects of dopamine, having the computational role of guiding exploration.
****
Sparse Greedy Gaussian Process Regression
Alex Smola, Peter Bartlett
We present a simple sparse greedy technique to approximate the maximum a posteriori estimate of Gaussian Processes with much improved scaling behaviour in the sample size m. In particular, computational requirements are O(n2m), storage is O(nm), the cost for prediction is 0 (n) and the cost to compute confidence bounds is O(nm), where n «: m. We show how to compute a stopping criterion, give bounds on the approximation error, and show applications to large scale problems.
****
Large Scale Bayes Point Machines
Ralf Herbrich, Thore Graepel
also known as the Bayes point -
****
Efficient Learning of Linear Perceptrons
Shai Ben-David, Hans-Ulrich Simon
We consider the existence of efficient algorithms for learning the class of half-spaces in ~n in the agnostic learning model (Le., mak(cid:173) ing no prior assumptions on the example-generating distribution). The resulting combinatorial problem - finding the best agreement half-space over an input sample - is NP hard to approximate to within some constant factor. We suggest a way to circumvent this theoretical bound by introducing a new measure of success for such algorithms. An algorithm is IL-margin successful if the agreement ratio of the half-space it outputs is as good as that of any half-space once training points that are inside the IL-margins of its separating hyper-plane are disregarded. We prove crisp computational com(cid:173) plexity results with respect to this success measure: On one hand, for every positive IL, there exist efficient (poly-time) IL-margin suc(cid:173) cessful learning algorithms. On the other hand, we prove that unless P=NP, there is no algorithm that runs in time polynomial in the sample size and in 1/ IL that is IL-margin successful for all IL> O.
****
Gaussianization
Scott Chen, Ramesh Gopinath
High dimensional data modeling is difficult mainly because the so-called "curse of dimensionality". We propose a technique called "Gaussianiza(cid:173) tion" for high dimensional density estimation, which alleviates the curse of dimensionality by exploiting the independence structures in the data. Gaussianization is motivated from recent developments in the statistics literature: projection pursuit, independent component analysis and Gaus(cid:173) sian mixture models with semi-tied covariances. We propose an iter(cid:173) ative Gaussianization procedure which converges weakly: at each it(cid:173) eration, the data is first transformed to the least dependent coordinates and then each coordinate is marginally Gaussianized by univariate tech(cid:173) niques. Gaussianization offers density estimation sharper than traditional kernel methods and radial basis function methods. Gaussianization can be viewed as efficient solution of nonlinear independent component anal(cid:173) ysis and high dimensional projection pursuit.
****
Error-correcting Codes on a Bethe-like Lattice
Renato Vicente, David Saad, Yoshiyuki Kabashima

We analyze Gallager codes by employing a simple mean-field approxi(cid:173) mat ion that distorts the model geometry and preserves important interac(cid:173) ti ons between sites. The method naturally recovers the probability prop(cid:173) agation decoding algorithm as an extremization of a proper free-energy. We fi nd a thermodynamic phase transition that coincides with informa(cid:173) tion t heoretical upper-bounds and explain the practical code performance in terms of the free-energy landscape.
****
Homeostasis in a Silicon Integrate and Fire Neuron
Shih-Chii Liu, Bradley Minch
In this work, we explore homeostasis in a silicon integrate-and-fire neu(cid:173) ron. The neuron adapts its firing rate over long time periods on the order o f seconds or minutes so that it returns to its spontaneous firing rate after a lasting perturbation. Homeostasis is implemented via two schemes. One scheme looks at the presynaptic activity and adapts the synaptic weigh t depending on the presynaptic spiking rate. The second scheme adapts the synaptic "threshold" depending on the neuron's activity. The threshold is lo wered if the neuron's activity decreases over a long time and is increased for prolonged increase in postsynaptic activity. Both these mechanisms for adaptation use floating-gate technology. The re(cid:173) sults shown h ere are measured from a chip fabricated in a 2-J.lm CMOS process.
****
Incorporating Second-Order Functional Knowledge for Better Option Pricing
Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, René Garcia
Incorporating prior knowledge of a particular task into the architecture of a learning algorithm can greatly improve generalization performance. We study h ere a case where we know that the function to be learned is non-decreasing in two of its arguments and convex in one of them. For this purpose we prop ose a class of functions similar to multi-layer neural networks but (1) that ha s those properties, (2) is a universal approximator of continuous functions wi th these and other properties. We apply this new class of functions to the task of modeling the price of call options. Experiments show improvements on regressing the price of call options using the new types of function classe s that incorporate the a priori con(cid:173) straints.
****
Spike-Timing-Dependent Learning for Oscillatory Networks
Silvia Scarpetta, Zhaoping Li, John Hertz
We apply to oscillatory networks a class of learning rules in which synaptic weights change proportional to pre- and post-synaptic ac(cid:173) tivity, with a kernel A(r) measuring the effect for a postsynaptic spike a time r aft er the presynaptic one. The resulting synaptic ma(cid:173) trices have an outer -product form in which the oscillating patterns are represented as complex vectors. In a simple model, the even part of A(r) enhances the resonant re sponse to learned stimulus by reducing the effective damping, while the odd pa rt determines the frequency of oscillation. We relate our model to the olfacto ry cortex and hippocampus and their presumed roles in forming associative me mories and input representations.
****
On Reversing Jensen's Inequality
Tony Jebara, Alex Pentland
Jensen's inequality is a powerful mathematical tool and one of the workhorses i n statistical learning. Its applications therein include the EM algorithm, Baye sian estimation and Bayesian inference. Jensen com(cid:173) putes simple lower b ounds on otherwise intractable quantities such as products of sums and latent l og-likelihoods. This simplification then per(cid:173) mits operations like integ ration and maximization. Quite often (i.e. in discriminative learning) upper bo unds are needed as well. We derive and prove an efficient analytic inequality t hat provides such variational upper bounds. This inequality holds for latent va riable mixtures of exponential family distributions and thus spans a wide range of contemporary statis(cid:173) tical models. We also discuss applications of t

he upper bounds including maximum conditional likelihood, large margin discriminative models and conditional Bayesian inference. Convergence, efficiency and prediction results are shown. 1
****
From Mixtures of Mixtures to Adaptive Transform Coding
Cynthia Archer, Todd Leen
We establish a principled framework for adaptive transform cod(cid:173) ing. Transform coders are often constructed by concatenating an ad hoc choice of transform with suboptimal bit allocation and quan(cid:173) tizer design. Instead, we start from a probabilistic latent variable model in the form of a mixture of constrained Gaussian mixtures. From this model we derive a transform coding algorithm, which is a constrained version of the generalized Lloyd algorithm for vector quantizer design. A byproduct of our derivation is the introduc(cid:173) tion of a new transform basis, which unlike other transforms (PCA, DCT, etc.) is explicitly optimized for coding. Image compression experiments show adaptive transform coders designed with our al(cid:173) gorithm improve compressed image signal-to-noise ratio up to 3 dB compared to global transform coding and 0.5 to 2 dB compared to other adaptive transform coders.
****
Algorithmic Stability and Generalization Performance
Olivier Bousquet, André Elisseeff
We present a novel way of obtaining PAC-style bounds on the gen(cid:173) eralization error of learning algorithms, explicitly using their stabil(cid:173) ity properties. A stable learner is one for which the learned solution does not change much with small changes in the training set. The bounds we obtain do not depend on any measure of the complexity of the hypothesis space (e.g. VC dimension) but rather depend on how the learning algorithm searches this space, and can thus be applied even when the VC dimension is infinite. We demonstrate that regularization networks possess the required stability property and apply our method to obtain new bounds on their generalization performance.
****
The Kernel Trick for Distances
Bernhard Schölkopf
A method is described which, like the kernel trick in support vector ma(cid:173) chines (SVMs), lets us generalize distance-based algorithms to operate in feature spaces, usually nonlinearly related to the input space. This is done by identifying a class of kernels which can be represented as norm-based distances in Hilbert spaces. It turns out that common kernel algorithms, such as SVMs and kernel PCA, are actually really distance based algorithms and can be run with that class of kernels, too. As well as providing a useful new insight into how these algorithms work, the present work can form the basis for conceiving new algorithms.
****
Higher-Order Statistical Properties Arising from the Non-Stationarity of Natural Signals
Lucas Parra, Clay Spence, Paul Sajda
We present evidence that several higher-order statistical proper(cid:173) ties of natural images and signals can be explained by a stochastic model which simply varies scale of an otherwise stationary Gaus(cid:173) sian process. We discuss two interesting consequences. The first is that a variety of natural signals can be related through a com(cid:173) mon model of spherically invariant random processes, which have the attractive property that the joint densities can be constructed from the one dimensional marginal. The second is that in some cas(cid:173) es the non-stationarity assumption and only second order methods can be explicitly exploited to find a linear basis that is equivalent to independent components obtained with higher-order methods. This is demonstrated on spectro-temporal components of speech.
****
Probabilistic Semantic Video Indexing

Milind Naphade, Igor Kozintsev, Thomas S. Huang
We propose a novel probabilistic framework for semantic video in(cid:173) dex ing. We define probabilistic multimedia objects (multijects) to map low -level media features to high-level semantic labels. A graphical network of such multijects (multinet) captures scene con(cid:173) text by discovering intra-frame as well as inter-frame dependency relations between the con cepts. The main contribution is a novel application of a factor graph framework to model this network. We model relations between semantic c oncepts in terms of their co-occurrence as well as the temporal dependencie s between these concepts within video shots. Using the sum-product algorit hm [1] for approximate or exact inference in these factor graph multinets, w e attempt to correct errors made during isolated concept detec(cid:173) t ion by forcing high-level constraints. This results in a significant impr ovement in the overall detection performance.
****
Accumulator Networks: Suitors of Local Probability Propagation
Brendan J. Frey, Anitha Kannan
One way to approximate inference in richly-connected graphical models i s to apply the sum-product algorithm (a.k.a. probabil(cid:173) ity propag ation algorithm), while ignoring the fact that the graph has cycles. The sum -product algorithm can be directly applied in Gaussian networks and in grap hs for coding, but for many condi(cid:173) tional probability functions - in cluding the sigmoid function - di(cid:173) rect application of the sum-product algorithm is not possible. We introduce "accumulator networks" that have low local complexity (but exponential global complexity) so the sum-product algorithm can be directly applied. In an accumulator network, the probability of a child given its parents is computed by accumulating the inputs from the parents in a Markov chain or more generally a tree. After giving expressions for inference and learning in accumulator net(cid:173) works, we give results on the "bars problem" and on the problem of extracting translated, overlapping faces from an image.
****
Sparse Representation for Gaussian Process Models
Lehel Csató, Manfred Opper
We develop an approach for a sparse representation for Gaussian Process (GP) mo dels in order to overcome the limitations of GPs caused by large data sets. Th e method is based on a combination of a Bayesian online al(cid:173) gorithm toge ther with a sequential construction of a relevant subsample of the data which fully specifies the prediction of the model. Experi(cid:173) mental resu lts on toy examples and large real-world data sets indicate the efficiency of t he approach.
****
Hippocampally-Dependent Consolidation in a Hierarchical Model of Neocortex
Szabolcs Káli, Peter Dayan
In memory consolidation, declarative memories which initially require the hippocampus for their recall, ultimately become independent of it. Consolid ation has been the focus of numerous experimental and qualita(cid:173) tive mod eling studies, but only little quantitative exploration. We present a consolida tion model in which hierarchical connections in the cortex, that initial ly instantiate purely semantic information acquired through probabilistic unsupervised learning, come to instantiate episodic infor(cid:173) mation as well. The hippocampus is responsible for helping complete partial inpu t patterns before consolidation is complete, while also train(cid:173) ing the cortex to perform appropriate completion by itself.
****
Automated State Abstraction for Options using the U-Tree Algorithm
Anders Jonsson, Andrew Barto
Learning a complex task can be significantly facilitated by defining a hierarchy of subtasks. An agent can learn to choose between various temporally abstract actions, each solving an assigned subtask, to accom(cid:173)

plish the overall task. In this paper, we study hierarchical learning using t he framework of options. We argue that to take full advantage of hier(cid:173) archical structure, one should perform option-specific state abstracti on, and that if this is to scale to larger tasks, state abstraction should be au(cid:173) tomated. We adapt McCallum's U-Tree algorithm to automatically build option-specific representations of the state feature space, and we illus(cid:173) trate the resulting algorithm using a simple hierarchica l task. Results suggest that automated option-specific state abstraction is an attractive approach to making hierarchical learning systems more effe ctive.
****
Structure Learning in Human Causal Induction
Joshua Tenenbaum, Thomas Griffiths
We use graphical models to explore the question of how people learn sim(cid:173) ple causal relationships from data. The two leading psychological theo(cid:173) ries can both be seen as estimating the parameters of a fixed graph. We argue that a complete account of causal induction should also consider how p eople learn the underlying causal graph structure, and we propose to model this inductive process as a Bayesian inference. Our argument is supported through the discussion of three data sets.
****
Minimum Bayes Error Feature Selection for Continuous Speech Recognition
George Saon, Mukund Padmanabhan
We consider the problem of designing a linear transformation () E lRPx n, of r ank p ~ n, which projects the features of a classifier x E lRn onto y = ()x E lRP such as to achieve minimum Bayes error (or probabil(cid:173) i ty of misclassification). Two avenues will be explored: the first is t o maximize the ()-average divergence between the class densities and the s econd is to minimize the union Bhattacharyya bound in the range of (). While b oth approaches yield similar performance in practice, they out(cid:173) per form standard LDA features and show a 10% relative improvement in the word error rate over state-of-the-art cepstral features on a large vocabu lary telephony speech recognition task.
****
Bayesian Video Shot Segmentation
Nuno Vasconcelos, Andrew Lippman
Prior knowledge about video structure can be used both as a means to improv e the peiformance of content analysis and to extract features that allow sema ntic classification. We introduce statistical models for two important comp onents of this structure, shot duration and activity, and demonstrate the usefulness of these models by introducing a Bayesian formulation for the shot segmentation problem. The new formulations is shown to extend standa rd thresholding methods in an adaptive and intuitive way, leading to im proved segmentation accuracy.
****
Kernel Expansions with Unlabeled Examples
Martin Szummer, Tommi Jaakkola
Modern classification applications necessitate supplementing the few avai lable labeled examples with unlabeled examples to improve classi(cid:173) f ication performance. We present a new tractable algorithm for exploit(cid:173) ing unlabeled examples in discriminative classification. This is achieved es sentially by expanding the input vectors into longer feature vectors via both l abeled and unlabeled examples. The resulting classification method can be inter preted as a discriminative kernel density estimate and is read(cid:173) ily trai ned via the EM algorithm, which in this case is both discriminative and achieve s the optimal solution. We provide, in addition, a purely dis(cid:173) criminat ive formulation of the estimation problem by appealing to the maximum ent ropy framework. We demonstrate that the proposed ap(cid:173) proach requi res very few labeled examples for high classification accu(cid:173) racy.
****

Universality and Individuality in a Neural Code
Elad Schneidman, Naama Brenner, Naftali Tishby, Robert van Steveninck, William B ialek
The problem of neural coding is to understand how sequences of action potentials (spikes) are related to sensory stimuli, motor out(cid:173) puts, o r (ultimately) thoughts and intentions. One clear question is whether the same coding rules are used by different neurons, or by corresponding neur ons in different individuals. We present a quantitative formulation of th is problem using ideas from informa(cid:173) tion theory, and apply this approa ch to the analysis of experiments in the fly visual system. We find signif icant individual differences in the structure of the code, particularly in the way that tempo(cid:173) ral patterns of spikes are used to convey information beyond that available from variations in spike rate. On the ot her hand, all the flies in our ensemble exhibit a high coding efficiency, so that every spike carries the same amount of information in all the individ uals. Thus the neural code has a quantifiable mixture of individuality and universality.
****
Generalized Belief Propagation
Jonathan S. Yedidia, William Freeman, Yair Weiss
Belief propagation (BP) was only supposed to work for tree-like network s but works surprisingly well in many applications involving networks with lo ops, including turbo codes. However, there has been little understandin g of the algorithm or the nature of the solutions it finds for genera l graphs. We show that BP can only converge to a stationary point of an approximate free energy, known as the Bethe free energy in statis(cid:17 3) tical physics. This result characterizes BP fixed-points and makes connec tions with variational approaches to approximate inference. More importantly, o ur analysis lets us build on the progress made in statistical physics sinc e Bethe's approximation was introduced in 1935. Kikuchi and others have sho wn how to construct more ac(cid:173) curate free energy approximations, of which Bethe's approximation is the simplest. Exploiting the insights from our analysis, we de(cid:173) rive generalized belief propagation (GBP) versions o fthese Kikuchi approximations. These new message passing algorithms can be significantly more accurate than ordinary BP, at an adjustable in(cid:173) c rease in complexity. We illustrate such a new GBP algorithm on a grid Mark ov network and show that it gives much more accurate marginal probabilities tha n those found using ordinary BP.
****
Multiple Timescales of Adaptation in a Neural Code
Adrienne Fairhall, Geoffrey Lewen, William Bialek, Robert van Steveninck
Many neural systems extend their dynamic range by adaptation. We ex(cid:173) a mine the timescales of adaptation in the context of dynamically mod(cid:17 3) ulated rapidly-varying stimuli, and demonstrate in the fly visual system that adaptation to the statistical ensemble of the stimulus dynamically maximizes information transmission about the time-dependent stimulus. Further, while the rate response has long transients, the adaptation takes place on ti mescales consistent with optimal variance estimation.
****
Speech Denoising and Dereverberation Using Probabilistic Models
Hagai Attias, John Platt, Alex Acero, Li Deng
This paper presents a unified probabilistic framework for denoising and derever beration of speech signals. The framework transforms the denois(cid:173) ing and dereverberation problems into Bayes-optimal signal estimation. The key idea is to use a strong speech model that is pre-trained on a large data s et of clean speech. Computational efficiency is achieved by using variational EM, working in the frequency domain, and employing conjugate priors. The fra mework covers both single and multiple micro(cid:173) phones. We apply this a pproach to noisy reverberant speech signals and get results substantially bette r than standard methods.

****
Modelling Spatial Recall, Mental Imagery and Neglect
Suzanna Becker, Neil Burgess
We present a computational model of the neural mechanisms in the pari(cid:173) etal and temporal lobes that support spatial navigation, recall of scenes and imagery of the products of recall. Long term representations are stored in the hippocampus, and are associated with local spatial and object-related features in the parahippocampal region. Viewer-centered representations are dynamically generated from long term memory in the parietal part of the model. The model thereby simulates recall and im(cid:173) agery of locations and objects in complex environments. After parietal damage, the model exhibits hemispatial neglect in mental imagery that rotates with the imagined perspective of the observer, as in the famous Milan Square experiment [1]. Our model makes novel predictions for the neural representations in the parahippocampal and parietal regions and for behavior in healthy volunteers and neuropsychological patients.
****
Adaptive Object Representation with Hierarchically-Distributed Memory Sites
Bosco Tjan
Theories of object recognition often assume that only one representa(cid:173) tion scheme is used within one visual-processing pathway. Versatility of the visual system comes from having multiple visual-processing pathways, each specialized in a different category of objects. We propose a theoretically simpler alternative, capable of explaining the same set of data and more. A single primary visual-processing pathway, loosely modular, is assumed. Memory modules are attached to sites along this pathway. Object-identity decision is made independently at each site. A site's response time is a monotonic-decreasing function of its confidence regarding its decision. An observer's response is the first-arriving response from any site. The effective representation(s) of such a system, determined empirically, can appear to be specialized for different tasks and stimuli, consistent with recent clinical and functional-imaging findings. This, however, merely reflects a decision being made at its appropriate level of abstraction. The system itself is intrinsically flexible and adaptive.
****
Text Classification using String Kernels
Huma Lodhi, John Shawe-Taylor, Nello Cristianini, Christopher Watkins
We introduce a novel kernel for comparing two text documents. The kernel is an inner product in the feature space consisting of all subsequences of length k. A subsequence is any ordered se(cid:173) quence of k characters occurring in the text though not necessarily contiguously. The subsequences are weighted by an exponentially decaying factor of their full length in the text, hence emphasising those occurrences which are close to contiguous. A direct compu(cid:173) tation of this feature vector would involve a prohibitive amount of computation even for modest values of k, since the dimension of the feature space grows exponentially with k. The paper describes how despite this fact the inner product can be efficiently evaluated by a dynamic programming technique. A preliminary experimental comparison of the performance of the kernel compared with a stan(cid:173) dard word feature space kernel results.
****
Model Complexity, Goodness of Fit and Diminishing Returns
Igor Cadez, Padhraic Smyth
We investigate a general characteristic of the trade-off in learning problems between goodness-of-fit and model complexity. Specifi(cid:173) cally we characterize a general class of learning problems where the goodness-of-fit function can be shown to be convex within first(cid:173) order as a function of model complexity. This general property of "diminishing returns" is illustrated on a number of real data sets and learning problems,

including finite mixture modeling and multivariate linear regression.
****
Fast Training of Support Vector Classifiers
Fernando Pérez-Cruz, Pedro Alarcón-Diana, Angel Navia-Vázquez, Antonio Artés-Rodríguez
In this communication we present a new algorithm for solving Support Vector Classifiers (SVC) with large training data sets. The new algorithm is based on an Iterative Re-Weighted Least Squares procedure which is used to optimize the SVc. Moreover, a novel sample selection strategy for the working set is presented, which randomly chooses the working set among the training samples that do not fulfill the stopping criteria. The validity of both proposals, the optimization procedure and sample selection strategy, is shown by means of computer experiments using well-known data sets.
****
Some New Bounds on the Generalization Error of Combined Classifiers
Vladimir Koltchinskii, Dmitriy Panchenko, Fernando Lozano
In this paper we develop the method of bounding the generalization error of a classifier in terms of its margin distribution which was introduced in the recent papers of Bartlett and Schapire, Freund, Bartlett and Lee. The theory of Gaussian and empirical processes allow us to prove the margin type inequalities for the most general functional classes, the complexity of the class being measured via the so called Gaussian complexity func(cid:173) tions. As a simple application of our results, we obtain the bounds of Schapire, Freund, Bartlett and Lee for the generalization error of boost(cid:173) ing. We also substantially improve the results of Bartlett on bounding the generalization error of neural networks in terms of h -norms of the weights of neurons. Furthermore, under additional assumptions on the complexity of the class of hypotheses we provide some tighter bounds, which in the case of boosting improve the results of Schapire, Freund, Bartlett and Lee.
****
Beyond Maximum Likelihood and Density Estimation: A Sample-Based Criterion for Unsupervised Learning of Complex Models
Sepp Hochreiter, Michael C. Mozer
The goal of many unsupervised learning procedures is to bring two probability distributions into alignment. Generative models such as Gaussian mixtures and Boltzmann machines can be cast in this light, as can recoding models such as ICA and projection pursuit. We propose a novel sample-based error measure for these classes of models, which applies even in situations where maximum likelihood (ML) and probability density estimation-based formulations can (cid:173) not be applied, e.g., models that are nonlinear or have intractable posteriors. Furthermore, our sample-based error measure avoids the difficulties of approximating a density function. We prove that with an unconstrained model, (1) our approach converges on the correct solution as the number of samples goes to infinity, and (2) the expected solution of our approach in the generative framework is the ML solution. Finally, we evaluate our approach via simula(cid:173) tions of linear and nonlinear models on mixture of Gaussians and ICA problems. The experiments show the broad applicability and generality of our approach.
****
A Neural Probabilistic Language Model
Yoshua Bengio, Réjean Ducharme, Pascal Vincent
A goal of statistical language modeling is to learn the joint probability function of sequences of words. This is intrinsically difficult because of the curse of dimensionality: we propose to fight it with its own weapons. In the proposed approach one learns simultaneously (1) a distributed rep(cid:173) resentation for each word (i.e. a similarity between words) along with (2) the probability function for word sequences, expressed with these repre(cid:173) sentations. Generalization is obtained because a sequence of words that has never been seen before gets high probability if it is made of words that are similar to words forming an already seen sentence. We report on experiments using neur

al networks for the probability function, showing on two text corpora that the proposed approach very significantly im(cid:173) proves on a state-of-th e-art trigram model.
****
Support Vector Novelty Detection Applied to Jet Engine Vibration Spectra
Paul Hayton, Bernhard Schölkopf, Lionel Tarassenko, Paul Anuzis
A system has been developed to extract diagnostic information from jet engine carcass vibration data. Support Vector Machines applied to nov(cid:173) elty detection provide a measure of how unusual the shape of a vibra(cid:173) t ion signature is, by learning a representation of normality. We describe a nov el method for Support Vector Machines of including information from a second class for novelty detection and give results from the appli(cid:173) cation to J et Engine vibration analysis.
****
`N-Body' Problems in Statistical Learning
Alexander Gray, Andrew Moore
We present efficient algorithms for all-point-pairs problems , or 'N(cid: 173) body '-like problems, which are ubiquitous in statistical learning. We focus on six examples, including nearest-neighbor classification, kernel densit y estimation, outlier detection , and the two-point correlation. These i nclude any problem which abstractly requires a comparison of each of the N points in a dataset with each other point and would naively be solve d using N 2 distance computations. In practice N is often large enoug h to make this infeasible. We present a suite of new geometric t echn iques which are applicable in principle to any 'N-body' computation in cluding large-scale mixtures of Gaussians, RBF neural networks, and HMM 's . Our algorithms exhibit favorable asymptotic scaling and are empirically several orders of magnitude faster than the naive computation, even f or small datasets. We are aware of no exact algorithms for these problems which are more effi(cid:173) cient either empirically or theoretically. In addition, our framework yields simple and elegant algorithms. It also permits two important generalizations beyond the standard all-point-pairs problems, which are more difficult. These are represented by our final examples, the multiple two-point correlation and the notorious n-point cor relation.
****
A Silicon Primitive for Competitive Learning
David Hsu, Miguel Figueroa, Chris Diorio
Competitive learning is a technique for training classification and clu stering networks. We have designed and fabricated an 11- transistor prim itive, that we term an automaximizing bump circuit, that implements co mpetitive learning dynamics. The circuit per(cid:173) forms a similarity computation, affords nonvolatile storage, and implements simultaneous lo cal adaptation and computation. We show that our primitive is suitable for implementing competitive learning in VLSI, and demonstrate its ef fectiveness in a standard clustering task.
****
Automatic Choice of Dimensionality for PCA
Thomas Minka
A central issue in principal component analysis (PCA) is choosing the numb er of principal components to be retained. By interpreting PCA as density es timation, we show how to use Bayesian model selection to es(cid:173) timate the true dimensionality of the data. The resulting estimate is sim(cid:173) ple to compute yet guaranteed to pick the correct dimensionality, given enough data. The estimate involves an integral over the Steifel manifold of k-frames, whic h is difficult to compute exactly. But after choosing an appropriate parameter ization and applying Laplace's method, an accu(cid:173) rate and practical estimator is obtained. In simulations, it is convincingly better than cross-va lidation and other proposed algorithms, plus it runs much faster.
****

Propagation Algorithms for Variational Bayesian Learning
Zoubin Ghahramani, Matthew Beal
Variational approximations are becoming a widespread tool for Bayesian l earning of graphical models. We provide some theoret(cid:173) ical results f or the variational updates in a very general family of conjugate-exponentia l graphical models. We show how the belief propagation and the junction tree algorithms can be used in the inference step of variational Bayesi an learning. Applying these re(cid:173) sults to the Bayesian analysis of linea r-Gaussian state-space models we obtain a learning procedure that exploits th e Kalman smooth(cid:173) ing propagation, while integrating over all model pa rameters. We demonstrate how this can be used to infer the hidden state dimen (cid:173) sionality of the state-space model in a variety of synthetic problems and one real high-dimensional data set.
****
An Adaptive Metric Machine for Pattern Classification
Carlotta Domeniconi, Jing Peng, Dimitrios Gunopulos
Nearest neighbor classification assumes locally constant class con(cid:173) dit ional probabilities. This assumption becomes invalid in high dimensions with finite samples due to the curse of dimensionality. Severe bias can be introduced under these conditions when using the nearest neighbor rule. We propose a locally adaptive nearest neighbor classification method to try to minimize bias. We use a Chi-squared distance analysis to compute a flexible metric for pro(cid:173) ducing neighborhoods that are elongated along less relevant feature dimensions and constricted along most influential ones. As a result, the class conditional probabilities tend to be smoother in the mo d(cid:173) ified neighborhoods, whereby better classification performance can b e achieved. The efficacy of our method is validated and compared against other techniques using a variety of real world data.
****
On Iterative Krylov-Dogleg Trust-Region Steps for Solving Neural Networks Nonlin ear Least Squares Problems
Eiji Mizutani, James Demmel
This paper describes a method of dogleg trust-region steps, or re(cid:173) stricted Levenberg-Marquardt steps, based on a projection pro(cid:173) cess onto the Krylov subspaces for neural networks nonlinear least square s problems. In particular, the linear conjugate gradient (CG) method works a s the inner iterative algorithm for solving the lin(cid:173) earized Gaus s-Newton normal equation, whereas the outer nonlin(cid:173) ear algorithm repe atedly takes so-called "Krylov-dogleg" steps, re(cid:173) lying only on matri x-vector multiplication without explicitly form(cid:173) ing the Jacobian matrix or the Gauss-Newton model Hessian. That is, our iterative dogleg algorith m can reduce both operational counts and memory space by a factor of O(n) (the number of pa(cid:173) rameters) in comparison with a direct linear -equation solver. This memory-less property is useful for large-scale prob lems.
****
Stability and Noise in Biochemical Switches
William Bialek
Many processes in biology, from the regulation of gene expression in bacteria t o memory in the brain, involve switches constructed from networks of biochemica l reactions. Crucial molecules are present in small numbers, raising questions about noise and stability. Analysis of noise in simple reaction schemes indic ates that switches stable for years and switchable in milliseconds can be built from fewer than one hundred molecules. Prospects for direct tests of this prediction, as well as implications, are discussed.
****
Computing with Finite and Infinite Networks
Ole Winther
Using statistical mechanics results, I calculate learning curves (average ge neralization error) for Gaussian processes (GPs) and Bayesian neural network

s (NNs) used for regression. Applying the results to learning a teacher defin ed by a two-layer network, I can directly compare GP and Bayesian NN learning. I find that a GP in general requires CJ (d S)-training examples to learn i nput features of order s (d is the input dimension), whereas a NN can learn the task with order the number of adjustable weights training exam ples. Since a GP can be considered as an infinite NN, the results show that even in the Bayesian approach, it is important to limit the complexity of the learning machine. The theoretical findings are confirmed in simulations with a nalytical GP learning and a NN mean field algorithm.
****
Hierarchical Memory-Based Reinforcement Learning
Natalia Hernandez-Gardiol, Sridhar Mahadevan
Sridhar Mahadevan
****
Second Order Approximations for Probability Models
Hilbert Kappen, Wim Wiegerinck
In this paper, we derive a second order mean field theory for directe d graphical probability models. By using an information theoretic argu(cid:173) ment it is shown how this can be done in the absense of a partition function. This method is a direct generalisation of the well-known TAP approximation for Boltzmann Machines. In a numerical example, it is s hown that the method greatly improves the first order mean field ap(cid :173) proximation. For a restricted class of graphical models, so-called single overlap graphs, the second order method has comparable complexity to the firs t order method. For sigmoid belief networks, the method is shown to be particu larly fast and effective.
****
Feature Selection for SVMs
Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Pog gio, Vladimir Vapnik
We introduce a method of feature selection for Support Vector Machines. The met hod is based upon finding those features which minimize bounds on the leave-on e-out error. This search can be efficiently performed via gradient descent. The resulting algorithms are shown to be superior to some standard feature selection algorithms on both toy data and real-life problems of face recognit ion, pedestrian detection and analyzing DNA micro array data.
****
Direct Classification with Indirect Data
Timothy Brown
We classify an input space according to the outputs of a real-valued function. The function is not given, but rather examples of the function. We c ontribute a consistent classifier that avoids the un(cid:173) necessary comple xity of estimating the function.
****
Constrained Independent Component Analysis
Wei Lu, Jagath Rajapakse
The paper presents a novel technique of constrained independent componen t analysis (CICA) to introduce constraints into the clas(cid:173) sical ICA an d solve the constrained optimization problem by using Lagrange multiplier met hods. This paper shows that CICA can be used to order the resulted indepe ndent components in a specific manner and normalize the demixing matrix in the signal separation procedure. It can systematically eliminate the ICA's indeter minacy on permutation and dilation. The experiments demonstrate the use of CI CA in ordering of independent components while providing normalized demix ing processes. Keywords: Independent component analysis, constrained indepen (cid:173) dent component analysis, constrained optimization, Lagrange mul(cid:17 3) tiplier methods
****
A Gradient-Based Boosting Algorithm for Regression Problems
Richard Zemel, Toniann Pitassi

In adaptive boosting, several weak learners trained sequentially are combined to boost the overall algorithm performance. Re(cid:173) cently adaptive boo sting methods for classification problems have been derived as gradient descent algorithms. This formulation jus(cid:173) tifies key elements and parameters i n the methods, all chosen to optimize a single common objective function. We p ropose an anal(cid:173) ogous formulation for adaptive boosting of regression pr oblems, utilizing a novel objective function that leads to a simple boosting a lgorithm. We prove that this method reduces training error, and compare its pe rformance to other regression methods.
****
The Early Word Catches the Weights
Mark Smith, Garrison Cottrell, Karen Anderson
The strong correlation between the frequency of words and their naming latency has been well documented. However, as early as 1973, the Age of Acquisition (AoA) of a word was alleged to be the actual variable of interest, but these studies seem to have been ignored in most of the lit(cid:173) erature. Rece ntly, there has been a resurgence of interest in AoA. While some studies have s hown that frequency has no effect when AoA is con(cid:173) trolled for, more re cent studies have found independent contributions of frequency and AoA. Connect ionist models have repeatedly shown strong effects of frequency, but little attention has been paid to whether they can also show AoA effects. Ind eed, several researchers have explicitly claimed that they cannot show AoA effects. In this work, we explore these claims using a simple feed forwa rd neural network. We find a sig(cid:173) nificant contribution of AoA to naming latency, as well as conditions under which frequency provides an in dependent contribution.
****
The Manhattan World Assumption: Regularities in Scene Statistics which Enable Ba yesian Inference
James Coughlan, Alan L. Yuille
Preliminary work by the authors made use of the so-called "Man(cid:173) hattan world" assumption about the scene statistics of city and indoor scenes . This assumption stated that such scenes were built on a cartesian grid whi ch led to regularities in the image edge gra(cid:173) dient statistics. In thi s paper we explore the general applicability of this assumption and show th at, surprisingly, it holds in a large variety of less structured environmen ts including rural scenes. This enables us, from a single image, to determine the orientation of the viewer relative to the scene structure and also to dete ct target ob(cid:173) jects which are not aligned with the grid. These inferences are performed using a Bayesian model with probability distri butions (e.g. on the image gradient statistics) learnt from real data.
****
Processing of Time Series by Neural Circuits with Biologically Realistic Synapti c Dynamics
Thomas Natschläger, Wolfgang Maass, Eduardo Sontag, Anthony Zador
Experimental data show that biological synapses behave quite differently from t he symbolic synapses in common artificial neural network models. Biological syn apses are dynamic, i.e., their "weight" changes on a short time scale by se veral hundred percent in dependence of the past input to the synapse. In this article we explore the consequences that these synaptic dynamics entail for the computational power of feedforward neural networks. We show that gradient descent suffices to approximate a given (quadratic) filter by a rather small neural system with dynamic synapses. We also compar e our network model to artificial neural net(cid:173) works designed for ti me series processing. Our numerical results are complemented by theoreti cal analysis which show that even with just a single hidden layer such networ ks can approximate a surprisingly large large class of nonlinear filters: al l filters that can be characterized by Volterra series. This result is ro bust with regard to various changes in the model for synaptic dynamics.
****

Machine Learning for Video-Based Rendering
Arno Schödl, Irfan Essa
We present techniques for rendering and animation of realistic scenes by analyzing and training on short video sequences. This work extends the new paradigm for computer animation, video tex(cid:173) tures, which uses recorded video to generate novel animations by replaying the video samples in a new order. Here we concentrate on video sprites, which are a special type of video texture. In video sprites, instead of storing whole images, the object of inter(cid:173) est is separated from the background and the video samples are stored as a sequence of alpha-matted sprites with associated veloc(cid:173) ity information. They can be rendered anywhere on the screen to create a novel animation of the object. We present methods to cre(cid:173) ate such animations by finding a sequence of sprite samples that is both visually smooth and follows a desired path. To estimate visual smoothness, we train a linear classifier to estimate visual similarity between video samples. If the motion path is known in advance, we use beam search to find a good sample sequence. We can specify the motion interactively by precomputing the sequence cost function using Q-Iearning.
****
Discovering Hidden Variables: A Structure-Based Approach
Gal Elidan, Noam Lotner, Nir Friedman, Daphne Koller
A serious problem in learning probabilistic models is the presence of hid(cid:173) den variables. These variables are not observed, yet interact with several of the observed variables. As such, they induce seemingly complex de(cid:173) pendencies among the latter. In recent years, much attention has been devoted to the development of algorithms for learning parameters, and in some cases structure, in the presence of hidden variables. In this pa(cid:173) per, we address the related problem of detecting hidden variables that interact with the observed variables. This problem is of interest both for improving our understanding of the domain and as a preliminary step that guides the learning procedure towards promising models. A very natural approach is to search for "structural signatures" of hidden variables - substructures in the learned network that tend to suggest the presence of a hidden variable. We make this basic idea concrete, and show how to integrate it with structure-search algorithms. We evaluate this method on several synthetic and real-life datasets, and show that it performs surpris(cid:173) ingly well.
****
Natural Sound Statistics and Divisive Normalization in the Auditory System
Odelia Schwartz, Eero Simoncelli
We explore the statistical properties of natural sound stimuli pre(cid:173) processed with a bank of linear filters. The responses of such filters exhibit a striking form of statistical dependency, in which the response variance of each filter grows with the response amplitude of filters tuned for nearby frequencies. These dependencies may be substantially re(cid:173) duced using an operation known as divisive normalization, in which the response of each filter is divided by a weighted sum of the recti(cid:173) fied responses of other filters. The weights may be chosen to maximize the independence of the normalized responses for an ensemble of natu(cid:173) ral sounds. We demonstrate that the resulting model accounts for non(cid:173) linearities in the response characteristics of the auditory nerve, by com(cid:173) paring model simulations to electrophysiological recordings. In previous work (NIPS, 1998) we demonstrated that an analogous model derived from the statistics of natural images accounts for non-linear properties of neurons in primary visual cortex. Thus, divisive normalization appears to be a generic mechanism for eliminating a type of statistical dependency that is prevalent in natural signals of different modalities.
****
Regularized Winnow Methods
Tong Zhang

In theory, the Winnow multiplicative update has certain advantages over the Per ceptron additive update when there are many irrelevant attributes. Recently, t here has been much effort on enhancing the Perceptron algo(cid:173) rithm by u sing regularization, leading to a class of linear classification methods cal led support vector machines. Similarly, it is also possible to apply the regul arization idea to the Winnow algorithm, which gives meth(cid:173) ods we call regularized Winnows. We show that the resulting methods compare with the bas ic Winnows in a similar way that a support vector machine compares with th e Perceptron. We investigate algorithmic is(cid:173) sues and learning prope rties of the derived methods. Some experimental results will also be provided to illustrate different methods.
****
FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks
Malcolm Slaney, Michele Covell
FaceSync is an optimal linear algorithm that finds the degree of syn(c id:173) chronization between the audio and image recordings of a human speaker. Using canonical correlation, it finds the best direction to com(ci d:173) bine all the audio and image data, projecting them onto a singl e axis. FaceSync uses Pearson's correlation to measure the degree of synchro (cid:173) nization between the audio and image data. We derive the optimal linea r transform to combine the audio and visual information and describe an implem entation that avoids the numerical problems caused by comput(cid:173) ing the correlation matrices.
****
Mixtures of Gaussian Processes
Volker Tresp
We introduce the mixture of Gaussian processes (MGP) model which is useful for applications in which the optimal bandwidth of a map is input dependent. The M GP is derived from the mixture of experts model and can also be used for model ing general conditional probability densities. We discuss how Gaussian process es -in particular in form of Gaussian process classification, the support vecto r machine and the MGP model(cid:173) can be used for quantifying the dependencie s in graphical models.
****
What Can a Single Neuron Compute?
Blaise Agüera y Arcas, Adrienne Fairhall, William Bialek
In this paper we formulate a description of the computation per(cid:173) formed by a neuron as a combination of dimensional reduction and nonl inearity. We implement this description for the Hodgkin(cid:173) Huxley model , identify the most relevant dimensions and find the nonlinearity. A two dimensional description already captures a significant fraction of the information that spikes carry about dy(cid:173) namic inputs. This descripti on also shows that computation in the Hodgkin-Huxley model is more complex than a simple integrate(cid:173) and-fire or perceptron model.
****
Tree-Based Modeling and Estimation of Gaussian Processes on Graphs with Cycles
Martin J. Wainwright, Erik Sudderth, Alan Willsky
We present the embedded trees algorithm, an iterative technique for esti mation of Gaussian processes defined on arbitrary graphs. By exactly solv ing a series of modified problems on embedded span(cid:173) ning trees, it co mputes the conditional means with an efficiency comparable to or better t han other techniques. Unlike other meth(cid:173) ods, the embedded trees alg orithm also computes exact error co(cid:173) variances. The error covaria nce computation is most efficient for graphs in which removing a small number of edges reveals an em(cid:173) bedded tree. In this context, we d emonstrate that sparse loopy graphs can provide a significant increase in modeling power rela(cid:173) tive to trees, with only a minor increase in estimation complexity.
****

Learning Winner-take-all Competition Between Groups of Neurons in Lateral Inhibitory Networks
Xiaohui Xie, Richard Hahnloser, H. Sebastian Seung
It has long been known that lateral inhibition in neural networks can lead to a winner-take-all competition, so that only a single neuron is active at a steady state. Here we show how to organize lateral inhibition so that groups of neurons compete to be active. Given a collection of poten(cid:173) tially overlapping groups, the inhibitory connectivity is set by a formula that can be interpreted as arising from a simple learning rule. Our analy(cid:173) sis demonstrates that such inhibition generally results in winner-take-all competition between the given groups, with the exception of some de(cid:173) generate cases. In a broader context, the network serves as a particular illustration of the general distinction between permitted and forbidden sets, which was introduced recently. From this viewpoint, the computa(cid:173) tional function of our network is to store and retrieve memories as per(cid:173) mitted sets of coactive neurons.
****
Foundations for a Circuit Complexity Theory of Sensory Processing
Robert Legenstein, Wolfgang Maass
We introduce total wire length as salient complexity measure for an anal(cid:173) ysis of the circuit complexity of sensory processing in biological neural systems and neuromorphic engineering. This new complexity measure is applied to a set of basic computational problems that apparently need to be solved by circuits for translation- and scale-invariant sensory process(cid:173) ing. We exhibit new circuit design strategies for these new benchmark functions that can be implemented within realistic complexity bounds, in particular with linear or almost linear total wire length.
****
Bayes Networks on Ice: Robotic Search for Antarctic Meteorites
Liam Pedersen, Dimitrios Apostolopoulos, William Whittaker
A Bayes network based classifier for distinguishing terrestrial rocks from meteorites is implemented onboard the Nomad robot. Equipped with a camera, spectrometer and eddy current sensor, this robot searched the ice sheets of Antarctica and autonomously made the first robotic identification of a meteorite, in January 2000 at the Elephant Moraine. This paper discusses rock classification from a robotic platform, and describes the system onboard Nomad.
****
Learning and Tracking Cyclic Human Motion
Dirk Ormoneit, Hedvig Sidenbladh, Michael Black, Trevor Hastie
We present methods for learning and tracking human motion in video. We estimate a statistical model of typical activities from a large set of 3D periodic human motion data by segmenting these data automatically into "cycles". Then the mean and the princi(cid:173) pal components of the cycles are computed using a new algorithm that accounts for missing information and enforces smooth tran(cid:173) sitions between cycles. The learned temporal model provides a prior probability distribution over human motions that can be used in a Bayesian framework for tracking human subjects in complex monocular video sequences and recovering their 3D motion.
****
The Use of MDL to Select among Computational Models of Cognition
In Myung, Mark Pitt, Shaobo Zhang, Vijay Balasubramanian
How should we decide among competing explanations of a cognitive process given limited observations? The problem of model selection is at the heart of progress in cognitive science. In this paper, Minimum Description Length (MDL) is introduced as a method for selecting among computational models of cognition. We also show that differential geometry provides an intuitive understanding of what drives model selection in MDL. Finally, adequacy of MDL is demonstrated in two areas of cognitive modeling.

****
Active Support Vector Machine Classification
Olvi Mangasarian, David Musicant
An active set strategy is applied to the dual of a simple reformula(cid:173) tion of the standard quadratic program of a linear support vector machine. This application generates a fast new dual algorithm that consists of solving a finite number of linear equations, with a typically large dimensionality equal to the number of points to be classified. However, by making novel use of the Sherman-Morrison(cid:173) Woodbury formula, a much smaller matrix of the order of the orig(cid:173) inal input space is inverted at each step. Thus, a problem with a 32-dimensional input space and 7 million points required inverting positive definite symmetric matrices of size 33 x 33 with a total run(cid:173) ning time of 96 minutes on a 400 MHz Pentium II. The algorithm requires no specialized quadratic or linear programming code, but merely a linear equation solver which is publicly available.
****
Decomposition of Reinforcement Learning for Admission Control of Self-Similar Call Arrival Processes
Jakob Carlström
This paper presents predictive gain scheduling, a technique for simplify(cid:173) ing reinforcement learning problems by decomposition. Link admission control of self-similar call traffic is used to demonstrate the technique. The control problem is decomposed into on-line prediction of near-fu(cid:173) ture call arrival rates, and precomputation of policies for Poisson call ar(cid:173) rival processes. At decision time, the predictions are used to select among the policies. Simulations show that this technique results in sig(cid:173) nificantly faster learning without any performance loss, compared to a reinforcement learning controller that does not decompose the problem.
****
Generalizable Singular Value Decomposition for Ill-posed Datasets
Ulrik Kjems, Lars Hansen, Stephen Strother
We demonstrate that statistical analysis of ill-posed data sets is subject to a bias, which can be observed when projecting indepen(cid:173) dent test set examples onto a basis defined by the training exam(cid:173) ples. Because the training examples in an ill-posed data set do not fully span the signal space the observed training set variances in each basis vector will be too high compared to the average vari(cid:173) ance of the test set projections onto the same basis vectors. On basis of this understanding we introduce the Generalizable Singu(cid:173) lar Value Decomposition (GenSVD) as a means to reduce this bias by re-estimation of the singular values obtained in a conventional Singular Value Decomposition, allowing for a generalization perfor(cid:173) mance increase of a subsequent statistical model. We demonstrate that the algorithm succesfully corrects bias in a data set from a functional PET activation study of the human brain.
****
Vicinal Risk Minimization
Olivier Chapelle, Jason Weston, Léon Bottou, Vladimir Vapnik
The Vicinal Risk Minimization principle establishes a bridge between generative models and methods derived from the Structural Risk Mini(cid:173) mization Principle such as Support Vector Machines or Statistical Reg(cid:173) ularization. We explain how VRM provides a framework which inte(cid:173) grates a number of existing algorithms, such as Parzen windows, Support Vector Machines, Ridge Regression, Constrained Logistic Classifiers and Tangent-Prop. We then show how the approach implies new algorithm(cid:173) s for solving problems usually associated with generative models. New algorithms are described for dealing with pattern recognition problems with very different pattern distributions and dealing with unlabeled data. Preliminary empirical results are presented.
****
Sparse Kernel Principal Component Analysis
Michael Tipping

'Kernel' principal component analysis (PCA) is an elegant non(cid:173) linear generalisation of the popular linear data analysis method, where a kernel function implicitly defines a nonlinear transforma(cid:173) tion into a feature space wherein standard PCA is performed. Un(cid:173) fortunately, the technique is not 'sparse', since the components thus obtained are expressed in terms of kernels associated with ev(cid:173) ery training vector. This paper shows that by approximating the covariance matrix in feature space by a reduced number of exam(cid:173) ple vectors, using a maximum-likelihood approach, we may obtain a highly sparse form of kernel PCA without loss of effectiveness.
****
Sparsity of Data Representation of Optimal Kernel Machine and Leave-one-out Estimator
Adam Kowalczyk
Vapnik's result that the expectation of the generalisation error ofthe opti(cid:173) mal hyperplane is bounded by the expectation of the ratio of the number of support vectors to the number of training examples is extended to a broad class of kernel machines. The class includes Support Vector Ma(cid:173) chines for soft margin classification and regression, and Regularization Networks with a variety of kernels and cost functions. We show that key inequalities in Vapnik's result become equalities once "the classification error" is replaced by "the margin error", with the latter defined as an in(cid:173) stance with positive cost. In particular we show that expectations of the true margin error and the empirical margin error are equal, and that the sparse solutions for kernel machines are possible only if the cost function is "partially" insensitive.
****
Rate-coded Restricted Boltzmann Machines for Face Recognition
Yee Whye Teh, Geoffrey E. Hinton
We describe a neurally-inspired, unsupervised learning algorithm that builds a non-linear generative model for pairs of face images from the same individual. Individuals are then recognized by finding the highest relative probability pair among all pairs that consist of a test image and an image whose identity is known. Our method compares favorably with other methods in the literature. The generative model consists of a single layer of rate-coded, non-linear feature detectors and it has the property that, given a data vector, the true posterior probability distribution over the feature detector activities can be inferred rapidly without iteration or approximation. The weights of the feature detectors are learned by com(cid:173) paring the correlations of pixel intensities and feature activations in two phases: When the network is observing real data and when it is observing reconstructions of real data generated from the feature activations.
****
Shape Context: A New Descriptor for Shape Matching and Object Recognition
Serge Belongie, Jitendra Malik, Jan Puzicha
We develop an approach to object recognition based on match(cid:173) ing shapes and using a resulting measure of similarity in a nearest neighbor classifier. The key algorithmic problem here is that of finding pointwise correspondences between an image shape and a stored prototype shape. We introduce a new shape descriptor, the shape context, which makes this possible, using a simple and robust algorithm. The shape context at a point captures the distri(cid:173) bution over relative positions of other shape points and thus sum(cid:173) marizes global shape in a rich, local descriptor. We demonstrate that shape contexts greatly simplify recovery of correspondences between points of two given shapes. Once shapes are aligned, shape contexts are used to define a robust score for measuring shape sim(cid:173) ilarity. We have used this score in a nearest-neighbor classifier for recognition of hand written digits as well as 3D objects, using exactly the same distance function. On the benchmark MNIST dataset of handwritten digits, this yields an error rate of 0.63%, outperforming other

published techniques.
****
Position Variance, Recurrence and Perceptual Learning
Zhaoping Li, Peter Dayan
Stimulus arrays are inevitably presented at different positions on the retina in visual tasks, even those that nominally require fixation. In par(cid:173) ticular, this applies to many perceptual learning tasks. We show that per(cid:173) ceptual inference or discrimination in the face of positional variance has a structurally different quality from inference about fixed position stimuli, involving a particular, quadratic, non-linearity rather than a purely lin(cid:173) ear discrimination. We show the advantage taking this non-linearity into account has for discrimination, and suggest it as a role for recurrent con(cid:173) nections in area VI, by demonstrating the superior discrimination perfor(cid:173) mance of a recurrent network. We propose that learning the feedforward and recurrent neural connections for these tasks corresponds to the fast and slow components of learning observed in perceptual learning tasks.
****
Permitted and Forbidden Sets in Symmetric Threshold-Linear Networks
Richard Hahnloser, H. Sebastian Seung
Ascribing computational principles to neural feedback circuits is an important problem in theoretical neuroscience. We study symmet(cid:173) ric threshold-linear networks and derive stability results that go beyond the insights that can be gained from Lyapunov theory or energy functions. By applying linear analysis to subnetworks com(cid:173) posed of coactive neurons, we determine the stability of potential steady states. We find that stability depends on two types of eigen(cid:173) modes. One type determines global stability and the other type determines whether or not multistability is possible. We can prove the equivalence of our stability criteria with criteria taken from quadratic programming. Also, we show that there are permitted sets of neurons that can be coactive at a steady state and forbid(cid:173) den sets that cannot. Permitted sets are clustered in the sense that subsets of permitted sets are permitted and supersets of forbidden sets are forbidden. By viewing permitted sets as memories stored in the synaptic connections, we can provide a formulation of long(cid:173) term memory that is more general than the traditional perspective of fixed point attractor networks.
****
Competition and Arbors in Ocular Dominance
Peter Dayan
Hebbian and competitive Hebbian algorithms are almost ubiquitous in modeling pattern formation in cortical development. We analyse in the(cid:173) oretical detail a particular model (adapted from Piepenbrock & Ober(cid:173) mayer, 1999) for the development of Id stripe-like patterns, which places competitive and interactive cortical influences, and free and restricted ini(cid:173) tial arborisation onto a common footing.
****
Learning Switching Linear Models of Human Motion
Vladimir Pavlovic, James M. Rehg, John MacCormick
The human figure exhibits complex and rich dynamic behavior that is both nonlinear and time-varying. Effective models of human dynamics can be learned from motion capture data using switching linear dynamic system (SLDS) models. We present results for human motion synthe(cid:173) sis, classification, and visual tracking using learned SLDS models. Since exact inference in SLDS is intractable, we present three approximate in(cid:173) ference algorithms and compare their performance. In particular, a new variational inference algorithm is obtained by casting the SLDS model as a Dynamic Bayesian Network. Classification experiments show the superiority of SLDS over conventional HMM's for our problem domain.
****
New Approaches Towards Robust and Adaptive Speech Recognition
Hervé Bourlard, Samy Bengio, Katrin Weber

In this paper, we discuss some new research directions in automatic speech reco gnition (ASR), and which somewhat deviate from the usual approaches. Mor e specifically, we will motivate and briefly describe new approaches b ased on multi-stream and multi/band ASR. These approaches extend the standa rd hidden Markov model (HMM) based approach by assuming that the different (fr equency) channels representing the speech signal are processed by different (i ndependent) "experts", each expert focusing on a different char(cid:173) act eristic of the signal, and that the different stream likelihoods (or posteriors) are combined at some (temporal) stage to yield a global recognition output. As a further extension to multi-stream ASR, we will finally introduce a new approach, referred to as HMM2, where the HMM emission probabilities a re estimated via state spe(cid:173) cific feature based HMMs responsible for mer ging the stream infor(cid:173) mation and modeling their possible correlation.
****

Place Cells and Spatial Navigation Based on 2D Visual Feature Extraction, Path I ntegration, and Reinforcement Learning
Angelo Arleo, Fabrizio Smeraldi, Stéphane Hug, Wulfram Gerstner
We model hippocampal place cells and head-direction cells by combin(cid:173) ing allothetic (visual) and idiothetic (proprioceptive) stimuli. Visual in(cid :173) put, provided by a video camera on a miniature robot, is preprocessed by a set of Gabor filters on 31 nodes of a log-polar retinotopic graph. Unsu(cid:17 3) pervised Hebbian learning is employed to incrementally build a popula(cid:17 3) tion of localized overlapping place fields. Place cells serve as basis func(cid:173) tions for reinforcement learning. Experimental results for goal-orient ed navigation of a mobile robot are presented.
****

Kernel-Based Reinforcement Learning in Average-Cost Problems: An Application to Optimal Portfolio Choice
Dirk Ormoneit, Peter W. Glynn
Many approaches to reinforcement learning combine neural net(cid:173) wor ks or other parametric function approximators with a form of temporal-dif ference learning to estimate the value function of a Markov Decision Process. A significant disadvantage of those pro(cid:173) cedures is that th e resulting learning algorithms are frequently un(cid:173) stable. In this wo rk, we present a new, kernel-based approach to reinforcement learning wh ich overcomes this difficulty and provably converges to a unique solution. B y contrast to existing algorithms, our method can also be shown to be c onsistent in the sense that its costs converge to the optimal costs a symptotically. Our focus is on learning in an average-cost framework and on a practical ap(cid:173) plication to the optimal portfolio choice problem.
****

A New Approximate Maximal Margin Classification Algorithm
Claudio Gentile
A new incremental learning algorithm is described which approximates the maxi mal margin hyperplane w.r.t. norm p ~ 2 for a set of linearly separable data . Our algorithm, called ALMAp (Approximate Large Mar- gin algorithm w.r.t. norm p), takes 0 ((P~21;;2) corrections to sepa(cid:173) rate the data wit h p-norm margin larger than (1 - 0:) ,,(, where,,(is the p-norm margin of the data and X is a bound on the p-norm of the in(cid:173) stances. ALM Ap avoids quadratic (or higher-order) programming meth(cid:173) ods. It is ve ry easy to implement and is as fast as on-line algorithms, such as Rosenblatt' s perceptron. We report on some experiments comparing ALMAp to two increme ntal algorithms: Perceptron and Li and Long's ROMMA. Our algorithm seems to perform quite better than both. The accuracy levels achieved by ALMAp a re slightly inferior to those obtained by Support vector Machines (SVMs). On t he other hand, ALMAp is quite faster and easier to implement than standard SVMs training algorithms.
****

Ensemble Learning and Linear Response Theory for ICA
Pedro Højen-Sørensen, Ole Winther, Lars Hansen

We propose a general Bayesian framework for performing independent component a nalysis (leA) which relies on ensemble learning and lin(cid:173) ear resp onse theory known from statistical physics. We apply it to both discrete an d continuous sources. For the continuous source the underde(cid:173) termined (o vercomplete) case is studied. The naive mean-field approach fails in this cas e whereas linear response theory-which gives an improved estimate of covarianc es-is very efficient. The examples given are for sources without tempora l correlations. However, this derivation can eas(cid:173) ily be extended to treat temporal correlations. Finally, the framework offers a simple way of generating new leA algorithms without needing to define the prior distribut ion of the sources explicitly.
****
Regularization with Dot-Product Kernels
Alex Smola, Zoltán Óvári, Robert C. Williamson
In this paper we give necessary and sufficient conditions under which kernels of dot product type k(x, y) = k(x . y) satisfy Mer(cid:173) cer' s condition and thus may be used in Support Vector Ma(cid:173) chines (SVM), Regularization Networks (RN) or Gaussian Pro(cid:173) cesses (GP) . In particular, we show that if the kernel is analytic (i.e. can be expanded in a Taylor series), all expansion coefficients have to be nonnegati ve. We give an explicit functional form for the feature map by calculating i ts eigenfunctions and eigenvalues.
****
From Margin to Sparsity
Thore Graepel, Ralf Herbrich, Robert C. Williamson
We present an improvement of Novikoff's perceptron convergence theorem. R einterpreting this mistake bound as a margin dependent sparsity guarantee allow s us to give a PAC-style generalisation er(cid:173) ror bound for the classifier learned by the perceptron learning algo(cid:173) rithm. The bound value crucia lly depends on the margin a support vector machine would achieve on the same da ta set using the same kernel. Ironically, the bound yields better guarantees t han are cur(cid:173) rently available for the support vector solution itself.
****
On a Connection between Kernel PCA and Metric Multidimensional Scaling
Christopher Williams
In this paper we show that the kernel peA algorithm of Sch6lkopf et al (1998) can be interpreted as a form of metric multidimensional scaling (MDS) when the kernel function k(x, y) is isotropic, i.e. it depends only on Ilx - yll. This leads to a metric MDS algorithm where the desired configuration of points is found via the solution of an eigenproblem rather than through t he iterative optimization of the stress objective function. The question o f kernel choice is also discussed.
****
One Microphone Source Separation
Sam Roweis
Source separation, or computational auditory scene analysis , attempts to extr act individual acoustic objects from input which contains a mixture of sound s from different sources, altered by the acoustic environment. Unmixing algorithms such as lCA and its extensions recover sources by reweight ing multiple obser(cid:173) vation sequences, and thus cannot operate when onl y a single observation signal is available. I present a technique called refiltering which recovers sources by a nonstationary reweighting ("maski ng") of frequency sub-bands from a single recording, and argue for the application of statistical algorithms to learning this masking function . I present results of a simple factorial HMM system which learns on recor dings of single speakers and can then separate mixtures using only one observat ion signal by computing the masking function and then refiltering.
****
Interactive Parts Model: An Application to Recognition of On-line Cursive Script
Predrag Neskovic, Philip Davis, Leon Cooper

In this work, we introduce an Interactive Parts (IP) model as an alte
rnative to Hidden Markov Models (HMMs). We tested both models on a d
atabase of on-line cursive script. We show that im(cid:173) plementations
of HMMs and the IP model, in which all letters are assumed to have the sam
e average width, give comparable results. However , in contrast to HMMs, the
IP model can handle duration modeling without an increase in computational
complexity.
****
A New Model of Spatial Representation in Multimodal Brain Areas
Sophie Denève, Jean-René Duhamel, Alexandre Pouget
Most models of spatial representations in the cortex assume cells with lim
ited receptive fields that are defined in a particular egocen(cid:173) tric fram
e of reference. However, cells outside of primary sensory cortex are eit
her gain modulated by postural input or partially shifting. We show t
hat solving classical spatial tasks, like sen(cid:173) sory prediction, mu
lti-sensory integration, sensory-motor transfor(cid:173) mation and motor contro
l requires more complicated intermediate representations that are not invar
iant in one frame of reference. We present an iterative basis functio
n map that performs these spatial tasks optimally with gain modulated and p
artially shifting units, and tests it against neurophysiological and neuropsy
cholog(cid:173) ical data.
****
Feature Correspondence: A Markov Chain Monte Carlo Approach
Frank Dellaert, Steven Seitz, Sebastian Thrun, Charles Thorpe
When trying to recover 3D structure from a set of images, the most d
ifficult problem is establishing the correspondence between the measurement
s. Most existing approaches assume that features can be tracked across frame
s, whereas methods that exploit rigidity constraints to facilitate matching do
so only under restricted cam(cid:173) era motion. In this paper we propo
se a Bayesian approach that avoids the brittleness associated with sin
gling out one "best" cor(cid:173) respondence, and instead consider the dist
ribution over all possible correspondences. We treat both a fully Bayesia
n approach that yields a posterior distribution, and a MAP approach t
hat makes use of EM to maximize this posterior. We show how Markov chain M
onte Carlo methods can be used to implement these techniques in practice,
and present experimental results on real data.
****
Stagewise Processing in Error-correcting Codes and Image Restoration
K. Y. Michael Wong, Hidetoshi Nishimori
We introduce stagewise processing in error-correcting codes and image re
storation, by extracting information from the former stage and using it sele
ctively to improve the performance of the latter one. Both mean-field
analysis using the cavity method and sim(cid:173) ulations show that it
has the advantage of being robust against uncertainties in hyperparamet
er estimation.
****
The Kernel Gibbs Sampler
Thore Graepel, Ralf Herbrich
We present an algorithm that samples the hypothesis space of ker(cid:173) nel c
lassifiers. Given a uniform prior over normalised weight vectors and a likel
ihood based on a model of label noise leads to a piece(cid:173) wise cons
tant posterior that can be sampled by the kernel Gibbs sampler (KGS).
The KGS is a Markov Chain Monte Carlo method that chooses a random directio
n in parameter space and samples from the resulting piecewise constant dens
ity along the line chosen. The KGS can be used as an analytical tool for th
e exploration of Bayesian transduction, Bayes point machines, active learning,
and evidence-based model selection on small data sets that are contam(cid:173)
inated with label noise. For a simple toy example we demonstrate experime
ntally how a Bayes point machine based on the KGS out(cid:173) performs an S
VM that is incapable of taking into account label noise.

****
Learning Sparse Image Codes using a Wavelet Pyramid Architecture
Bruno Olshausen, Phil Sallee, Michael Lewicki
We show how a wavelet basis may be adapted to best represent natural images in terms of sparse coefficients. The wavelet basis, which may be either complete or overcomplete, is specified by a small number of spatial functions which are repeated across space and combined in a recursive fashion so as to be self-similar across scale. These functions are adapted to minimize the estimated code length under a model that assumes images are composed of a linear superposition of sparse, independent components. When adapted to natural images, the wavelet bases take on different orientations and they evenly tile the orientation domain, in stark contrast to the standard, non-oriented wavelet bases used in image compression. When the basis set is allowed to be overcomplete, it also yields higher coding efficiency than standard wavelet bases.
****
Weak Learners and Improved Rates of Convergence in Boosting
Shie Mannor, Ron Meir
The problem of constructing weak classifiers for boosting algo(cid:173) rithms is studied. We present an algorithm that produces a linear classifier that is guaranteed to achieve an error better than random guessing for any distribution on the data. While this weak learner is not useful for learning in general, we show that under reasonable conditions on the distribution it yields an effective weak learner for one-dimensional problems. Preliminary simulations suggest that similar behavior can be expected in higher dimensions, a result which is corroborated by some recent theoretical bounds. Addi(cid:173) tionally, we provide improved convergence rate bounds for the gen(cid:173) eralization error in situations where the empirical error can be made small, which is exactly the situation that occurs if weak learners with guaranteed performance that is better than random guessing can be established.
****
Keeping Flexible Active Contours on Track using Metropolis Updates
Trausti Kristjansson, Brendan J. Frey
Condensation, a form of likelihood-weighted particle filtering, has been successfully used to infer the shapes of highly constrained "active" con(cid:173) tours in video sequences. However, when the contours are highly flexible (e.g. for tracking fingers of a hand), a computationally burdensome num(cid:173) ber of particles is needed to successfully approximate the contour distri(cid:173) bution. We show how the Metropolis algorithm can be used to update a particle set representing a distribution over contours at each frame in a video sequence. We compare this method to condensation using a video sequence that requires highly flexible contours, and show that the new algorithm performs dramatically better that the condensation algorithm. We discuss the incorporation of this method into the "active contour" framework where a shape-subspace is used constrain shape variation.
****
Data Clustering by Markovian Relaxation and the Information Bottleneck Method
Naftali Tishby, Noam Slonim
We introduce a new, non-parametric and principled, distance based clustering method. This method combines a pairwise based ap(cid:173) proach with a vector-quantization method which provide a mean(cid:173) ingful interpretation to the resulting clusters. The idea is based on turning the distance matrix into a Markov process and then examine the decay of mutual-information during the relaxation of this process. The clusters emerge as quasi-stable structures dur(cid:173) ing this relaxation, and then are extracted using the information bottleneck method. These clusters capture the information about the initial point of the relaxation in the most effective way. The method can cluster data with no geometric or other bias and makes no assumption about the underlying distribution.

****
Balancing Multiple Sources of Reward in Reinforcement Learning
Christian Shelton
For many problems which would be natural for reinforcement learning, the rewar d signal is not a single scalar value but has multiple scalar com(cid:173) ponen ts. Examples of such problems include agents with multiple goals and agents with multiple users. Creating a single reward value by com(cid:173) bining th e multiple components can throwaway vital information and can lead to inc orrect solutions. We describe the multiple reward source problem and discuss the problems with applying traditional reinforce(cid:173) ment learning. We then present an new algorithm for finding a solution and results on simu lated environments.
****
Temporally Dependent Plasticity: An Information Theoretic Account
Gal Chechik, Naftali Tishby
The paradigm of Hebbian learning has recently received a novel in(cid:173) terpr etation with the discovery of synaptic plasticity that depends on the relative timing of pre and post synaptic spikes. This paper derives a temporally depen dent learning rule from the basic princi(cid:173) ple of mutual information maxi mization and studies its relation to the experimentally observed plasticity. We find that a supervised spike-dependent learning rule sharing similar str ucture with the ex(cid:173) perimentally observed plasticity increases mutual in formation to a stable near optimal level. Moreover, the analysis reveals how the temporal structure of time-dependent learning rules is determined b y the temporal filter applied by neurons over their inputs. These results sug gest experimental prediction as to the dependency of the learning rule on neuro nal biophysical parameters
****
Analysis of Bit Error Probability of Direct-Sequence CDMA Multiuser Demodulators
Toshiyuki Tanaka
We analyze the bit error probability of multiuser demodulators for direct(cid:17 3) sequence binary phase-shift-keying (DSIBPSK) CDMA channel with ad(cid:173) di tive gaussian noise. The problem of multiuser demodulation is cast into the fin ite-temperature decoding problem, and replica analysis is ap(cid:173) plied to e valuate the performance of the resulting MPM (Marginal Pos(cid:173) terior Mode) demodulators, which include the optimal demodulator and the MAP demodulator as special cases. An approximate implementa(cid:173) tion of demodulators is propo sed using analog-valued Hopfield model as a naive mean-field approximation to t he MPM demodulators, and its performance is also evaluated by the replica analy sis. Results of the per(cid:173) formance evaluation shows effectiveness of the optimal demodulator and the mean-field demodulator compared with the convention al one, espe(cid:173) cially in the cases of small information bit rate and low noise level.
****
A Variational Mean-Field Theory for Sigmoidal Belief Networks
Chiranjib Bhattacharyya, S. Keerthi
A variational derivation of Plefka's mean-field theory is presented. This the ory is then applied to sigmoidal belief networks with the aid of furthe r approximations. Empirical evaluation on small scale networks show that th e proposed approximations are quite com(cid:173) petitive.
****
Robust Reinforcement Learning
Jun Morimoto, Kenji Doya
This paper proposes a new reinforcement learning (RL) paradigm that explic itly takes into account input disturbance as well as mod(cid:173) eling errors. The use of environmental models in RL is quite pop(cid:173) ular for both off-line learning by simulations and for on-line ac(cid:173) tion plann ing. However, the difference between the model and the real environment can le ad to unpredictable, often unwanted results. Based on the theory of H oocontrol , we consider a differential game in which a 'disturbing' agent (disturbe

r) tries to make the worst possible disturbance while a 'control' age
nt (actor) tries to make the best control input. The problem is formulated
 as finding a min(cid:173) max solution of a value function that takes into acc
ount the norm of the output deviation and the norm of the disturbance. We deri
ve on-line learning algorithms for estimating the value function and f
or calculating the worst disturbance and the best control in refer(cid:173) en
ce to the value function. We tested the paradigm, which we call "Robust Rein
forcement Learning (RRL)," in the task of inverted pendulum. In the lin
ear domain, the policy and the value func(cid:173) tion learned by the on
-line algorithms coincided with those derived analytically by the linear H oo
theory. For a fully nonlinear swing(cid:173) up task, the control by RRL ac
hieved robust performance against changes in the pendulum weight and frictio
n while a standard RL control could not deal with such environmental changes
.
****
Explaining Away in Weight Space
Peter Dayan, Sham Kakade
Explaining away has mostly been considered in terms of inference of st
ates in belief networks. We show how it can also arise in a Bayesian co
ntext in inference about the weights governing relationships such as thos
e between stimuli and reinforcers in conditioning experiments such as bacA,
'Ward blocking. We show how explaining away in weight space can be accoun
ted for using an extension of a Kalman filter model; pro(cid:173) vide a new a
pproximate way of looking at the Kalman gain matrix as a whitener for the c
orrelation matrix of the observation process; suggest a network implementat
ion of this whitener using an architecture due to Goodall; and show that the
resulting model exhibits backward blocking.
****
Improved Output Coding for Classification Using Continuous Relaxation
Koby Crammer, Yoram Singer
Output coding is a general method for solving multiclass problems by redu
cing them to multiple binary classification problems. Previous re(cid:173) sea
rch on output coding has employed, almost solely, predefined discrete codes. We
 describe an algorithm that improves the performance of output codes by relaxin
g them to continuous codes. The relaxation procedure is cast as an optim
ization problem and is reminiscent of the quadratic program for support vec
tor machines. We describe experiments with the proposed algorithm, comparing i
t to standard discrete output codes. The experimental results indicate that c
ontinuous relaxations of output codes often improve the generalization performa
nce, especially for short codes.
****
Learning Curves for Gaussian Processes Regression: A Framework for Good Approxim
ations
Dörthe Malzahn, Manfred Opper
Based on a statistical mechanics approach, we develop a method for appr
oximately computing average case learning curves for Gaus(cid:173) sian process
  regression models. The approximation works well in the large sample si
ze limit and for arbitrary dimensionality of the input space. We explain h
ow the approximation can be systemati(cid:173) cally improved and argue that sim
ilar techniques can be applied to general likelihood models.
****
Sequentially Fitting ``Inclusive'' Trees for Inference in Noisy-OR Networks
Brendan J. Frey, Relu Patrascu, Tommi Jaakkola, Jodi Moran
An important class of problems can be cast as inference in noisy(cid:173
) OR Bayesian networks, where the binary state of each variable is a logical
  OR of noisy versions of the states of the variable's par(cid:173) ents.
For example, in medical diagnosis, the presence of a symptom can be expressed a
s a noisy-OR of the diseases that may cause the symptom - on some occasions,
  a disease may fail to activate the symptom. Inference in richly-conne
cted noisy-OR networks is in(cid:173) tractable, but approximate methods (e

.g., variational techniques) are showing increasing promise as practical solutions. One prob(cid:173) lem with most approximations is that they tend to concentrate on a relatively small number of modes in the tr ue posterior, ig(cid:173) noring other plausible configurations of the hidd en variables. We introduce a new sequential variational method for biparti te noisy(cid:173) OR networks, that favors including all modes of the true pos terior and models the posterior distribution as a tree. We compare this method with other approximations using an ensemble of networks with network sta tistics that are comparable to the QMR-DT med(cid:173) ical diagnostic network.
****
Whence Sparseness?
Carl van Vreeswijk
It has been shown that the receptive fields of simple cells in VI can be ex(cid:173) plained by assuming optimal encoding, provided that an extra constraint of sparseness is added. This finding suggests that there is a reason, in(cid:173) dependent of optimal representation, for sparseness. However this work u sed an ad hoc model for the noise. Here I show that, if a biologically m ore plausible noise model, describing neurons as Poisson processes, is used s parseness does not have to be added as a constraint. Thus I con(cid:173) cl ude that sparseness is not a feature that evolution has striven for, but is si mply the result of the evolutionary pressure towards an optimal repre(cid:173) sentation.
****
Periodic Component Analysis: An Eigenvalue Method for Representing Periodic Stru cture in Speech
Lawrence Saul, Jont Allen
An eigenvalue method is developed for analyzing periodic structure in s peech. Signals are analyzed by a matrix diagonalization reminiscent of method s for principal component analysis (PCA) and independent com(cid:173) ponent analysis (ICA). Our method-called periodic component analysis (1l"CA)-uses c onstructive interference to enhance periodic components of the frequency spe ctrum and destructive interference to cancel noise. The front end emulates important aspects of auditory processing, such as cochlear filtering, nonlinear compression, and insensitivity to phase, with the aim of approaching the rob ustness of human listeners. The method avoids the inefficiencies of autocorr elation at the pitch period: it does not require long delay lines, and it correlates signals at a clock rate on the order of the actual pitch, a s opposed to the original sampling rate. We derive its cost function and present some experimental results.
****
Partially Observable SDE Models for Image Sequence Recognition Tasks
Javier Movellan, Paul Mineiro, Ruth Williams
This paper explores a framework for recognition of image sequences using partia lly observable stochastic differential equation (SDE) models. Monte-Carlo impor tance sampling techniques are used for efficient estimation of sequence likelih oods and sequence likelihood gradients. Once the network dynamics are learned, we apply the SDE models to sequence recognition tasks in a manner similar to t he way Hidden Markov models (HMMs) are commonly applied. The potential advantag e of SDEs over HMMS is the use of contin(cid:173) uous state dynamics. We presen t encouraging results for a video sequence recognition task in which SDE models provided excellent performance when compared to hidden Markov models.
****
Combining ICA and Top-Down Attention for Robust Speech Recognition
Un-Min Bae, Soo-Young Lee
We present an algorithm which compensates for the mismatches between ch aracteristics of real-world problems and assumptions of independent component analysis algorithm. To provide additional information to the ICA network, we incorporate top-down selec(cid:173) tive attention. An MLP classifier is added to the separated signal channel and the error of the classifie r is backpropagated to the ICA network. This backpropagation process resu

lts in estimation of expected ICA output signal for the top-down attentio
n. Then, the unmixing matrix is retrained according to a new cost function
 representing the backpropagated error as well as independence. It modifies t
he density of recovered signals to the density appropriate for classification.
 For noisy speech signal recorded in real environ(cid:173) ments, the algori
thm improved the recognition performance and showed robustness against par
ametric changes.
****
A Comparison of Image Processing Techniques for Visual Speech Recognition Applic
ations
Michael Gray, Terrence J. Sejnowski, Javier Movellan
We examine eight different techniques for developing visual rep(cid:173)
 resentations in machine vision tasks. In particular we compare differe
nt versions of principal component and independent com(cid:173) ponent a
nalysis in combination with stepwise regression methods for variable sel
ection. We found that local methods, based on the statistics of image patch
es, consistently outperformed global meth(cid:173) ods based on the statistics o
f entire images. This result is consistent with previous work on emotion
and facial expression recognition. In addition, the use of a stepwise regres
sion technique for selecting variables and regions of interest substantially b
oosted performance.
****
Learning Continuous Distributions: Simulations With Field Theoretic Priors
Ilya Nemenman, William Bialek
Learning of a smooth but nonparametric probability density can be reg(cid:173)
ularized using methods of Quantum Field Theory. We implement a field theoretic
 prior numerically, test its efficacy, and show that the free pa(cid:173
) rameter of the theory (,smoothness scale') can be determined self con(cid:1
73) sistently by the data; this forms an infinite dimensional generalization of
 the MDL principle. Finally, we study the implications of one's choice of
the prior and the parameterization and conclude that the smoothness scale det
ermination makes density estimation very weakly sensitive to the choice
of the prior, and that even wrong choices can be advantageous for small data se
ts.
****
Algebraic Information Geometry for Learning Machines with Singularities
Sumio Watanabe
Algebraic geometry is essential to learning theory. In hierarchical learni
ng machines such as layered neural networks and gaussian mixtures, the
 asymptotic normality does not hold, since Fisher in(cid:173) formation matrice
s are singular. In this paper , the rigorous asymp(cid:173) totic form of the s
tochastic complexity is clarified based on resolu(cid:173) tion of singulariti
es and two different problems are studied. (1) If the prior is positive
, then the stochastic complexity is far smaller than BIO, resulting in the
 smaller generalization error than regular statistical models, even when the tr
ue distribution is not contained in the parametric model. nate free and
 equal to zero at singularities, is employed then the stochastic comple
xity has the same form as BIO. It is useful for model selection, but
 not for generalization.
****
The Missing Link - A Probabilistic Model of Document Content and Hypertext Conne
ctivity
David Cohn, Thomas Hofmann
We describe a joint probabilistic model for modeling the contents and i
nter-connectivity of document collections such as sets of web pages or rese
arch paper archives. The model is based on a probabilistic factor deco
mposition and allows identifying principal topics of the collection as w
ell as authoritative documents within those topics. Furthermore, the relation
ships between topics is mapped out in order to build a predictive model of link
 content. Among the many applications of this approach are information retriev

al and search, topic identification, query disambigua(cid:173) tion, focused web crawling, web authoring, and bibliometric analysis.
****
The Use of Classifiers in Sequential Inference
Vasin Punyakanok, Dan Roth
We study the problem of combining the outcomes of several different classifiers in a way that provides a coherent inference that satisfies some constraints. In particular, we develop two general approaches for an im(cid:173) portant subproblem - identifying phrase structure. The first is a Marko(cid:173) vian approach that extends standard HMMs to allow the use of a rich ob(cid:173) servation structure and of general classifiers to model state-observation dependencies. The second is an extension of constraint satisfaction for(cid:173) malisms. We develop efficient combination algorithms under both mod(cid:173) els and study them experimentally in the context of shallow parsing.
****
Finding the Key to a Synapse
Thomas Natschläger, Wolfgang Maass
Experimental data have shown that synapses are heterogeneous: different synapses respond with different sequences of amplitudes of postsynaptic responses to the same spike train. Neither the role of synaptic dynamics itself nor the role of the heterogeneity of synaptic dynamics for com(cid:173) putations in neural circuits is well understood. We present in this article methods that make it feasible to compute for a given synapse with known synaptic parameters the spike train that is optimally fitted to the synapse, for example in the sense that it produces the largest sum of postsynap(cid:173) tic responses. To our surprise we find that most of these optimally fitted spike trains match common firing patterns of specific types of neurons that are discussed in the literature.
****
The Unscented Particle Filter
Rudolph van der Merwe, Arnaud Doucet, Nando de Freitas, Eric Wan
In this paper, we propose a new particle filter based on sequential importance sampling. The algorithm uses a bank of unscented fil(cid:173) ters to obtain the importance proposal distribution. This proposal has two very "nice" properties. Firstly, it makes efficient use of the latest available information and, secondly, it can have heavy tails. As a result, we find that the algorithm outperforms stan(cid:173) dard particle filtering and other nonlinear filtering methods very substantially. This experimental finding is in agreement with the theoretical convergence proof for the algorithm. The algorithm also includes resampling and (possibly) Markov chain Monte Carlo (MCMC) steps.
****
Algorithms for Non-negative Matrix Factorization
Daniel Lee, H. Sebastian Seung
Non-negative matrix factorization (NMF) has previously been shown to be a useful decomposition for multivariate data. Two different multi- plicative algorithms for NMF are analyzed. They differ only slightly in the multiplicative factor used in the update rules. One algorithm can be shown to minimize the conventional least squares error while the other minimizes the generalized Kullback-Leibler divergence. The monotonic convergence of both algorithms can be proven using an auxiliary func- tion analogous to that used for proving convergence of the Expectation- Maximization algorithm. The algorithms can also be interpreted as diag- onally rescaled gradient descent, where the rescaling factor is optimally chosen to ensure convergence.
****
Divisive and Subtractive Mask Effects: Linking Psychophysics and Biophysics
Barbara Zenger, Christof Koch
We describe an analogy between psychophysically measured effects in contrast masking, and the behavior of a simple integrate-and(cid:173) fire neuron

that receives time-modulated inhibition. In the psy(cid:173) chophysical experiments, we tested observers ability to discriminate contrasts of periphera l Gabor patches in the presence of collinear Gabor flankers. The data rev eal a complex interaction pattern that we account for by assuming that fl ankers provide divisive inhibi(cid:173) tion to the target unit for low target contrasts, but provide sub(cid:173) tractive inhibition to the target unit for higher target contrasts. A similar switch from divisive to subtractive inhibition is observed in an integrate-and-fire unit that receiv es inhibition modulated in time such that the cell spends part of the time in a high-inhibition state and part of the time in a low-inhibition sta te. The simi(cid:173) larity between the effects suggests that one may cause the other. The biophysical model makes testable predictions for physio logical single-cell recordings.
****
Factored Semi-Tied Covariance Matrices
Mark Gales
A new form of covariance modelling for Gaussian mixture models and hidden Mar kov models is presented. This is an extension to an efficient form of cov ariance modelling used in speech recognition, semi-tied co(cid:173) variance ma trices. In the standard form of semi-tied covariance matrices the covariance m atrix is decomposed into a highly shared decorrelating transform and a compone nt-specific diagonal covariance matrix. The use of a factored decorrelating tra nsform is presented in this paper. This fac(cid:173) toring effectively increase s the number of possible transforms without in(cid:173) creasing the number of f ree parameters. Maximum likelihood estimation schemes for all the model paramet ers are presented including the compo(cid:173) nent/transform assignment, trans form and component parameters. This new model form is evaluated on a larg e vocabulary speech recognition task. It is shown that using this factored form of covariance modelling reduces the word error rate.
****
Noise Suppression Based on Neurophysiologically-motivated SNR Estimation for Rob ust Speech Recognition
Jürgen Tchorz, Michael Kleinschmidt, Birger Kollmeier
A novel noise suppression scheme for speech signals is proposed which is based on a neurophysiologically-motivated estimation of the local si gnal-to-noise ratio (SNR) in different frequency chan(cid:173) nels. For SNR-estimation, the input signal is transformed into so-called Amplitud e Modulation Spectrograms (AMS), which rep(cid:173) resent both spectral and t emporal characteristics of the respective analysis frame, and which imitate the representation of modula(cid:173) tion frequencies in higher stages of the mammalian auditory sys(cid:173) tem. A neural network is used to ana lyse AMS patterns generated from noisy speech and estimates the local S NR. Noise suppres(cid:173) sion is achieved by attenuating frequency cha nnels according to their SNR. The noise suppression algorithm is evaluated in speaker(cid:173) independent digit recognition experiments and compared to noise suppression by Spectral Subtraction.
****
Dendritic Compartmentalization Could Underlie Competition and Attentional Biasin g of Simultaneous Visual Stimuli
Kevin Archie, Bartlett Mel
Neurons in area V4 have relatively large receptive fields (RFs), so multi(cid:1 73) ple visual features are simultaneously "seen" by these cells. Recordings from single V 4 neurons suggest that simultaneously presented stimuli c ompete to set the output firing rate, and that attention acts to iso late individual features by biasing the competition in favor of the att ended object. We propose that both stimulus competition and attentional bias(c id:173) ing arise from the spatial segregation of afferent synapses onto differe nt regions of the excitable dendritic tree of V 4 neurons. The pattern of feed(cid:173) forward, stimulus-driven inputs follows from a Hebbian rule: excitato ry afferents with similar RFs tend to group together on the dendritic

tree, avoiding randomly located inhibitory inputs with similar RFs. The same principle guides the formation of inputs that mediate attentional mod(cid:173) ulation. Using both biophysically detailed compartmental models and simplified models of computation in single neurons, we demonstrate that such an architecture could account for the response properties and atten(cid:173) tional modulation of V 4 neurons. Our results suggest an important role for nonlinear dendritic conductances in extrastriate cortical processing.
****

Smart Vision Chip Fabricated Using Three Dimensional Integration Technology
Hiroyuki Kurino, M. Nakagawa, Kang Lee, Tomonori Nakamura, Yuusuke Yamada, Ki Park, Mitsumasa Koyanagi
The smart VISIOn chip has a large potential for application in general purpose high speed image processing systems. In order to fabricate smart vision chips including photo detector compactly, we have proposed the application of three dimensional LSI technology for smart vision chips. Three dimensional technology has great potential to realize new neuromorphic systems inspired by not only the biological function but also the biological structure. In this paper, we describe our three dimensional LSI technology for neuromorphic circuits and the design of smart vision chips .
****

A Productive, Systematic Framework for the Representation of Visual Structure
Shimon Edelman, Nathan Intrator
We describe a unified framework for the understanding of struc(cid:173) ture representation in primate vision. A model derived from this framework is shown to be effectively systematic in that it has the ability to interpret and associate together objects that are related through a rearrangement of common "middle-scale" parts, repre(cid:173) sented as image fragments. The model addresses the same concerns as previous work on compositional representation through the use of what+where receptive fields and attentional gain modulation. It does not require prior exposure to the individual parts, and avoids the need for abstract symbolic binding.
****