Real Time Image Saliency for Black Box Classifiers
Piotr Dabkowski, Yarin Gal

In this work we develop a fast saliency detection method that can be applied to any differentiable image classifier. We train a masking model to manipulate the scores of the classifier by masking salient parts of the input image. Our model generalises well to unseen images and requires a single forward pass to perform saliency detection, therefore suitable for use in real-time systems. We test our approach on CIFAR-10 and ImageNet datasets and show that the produced saliency maps are easily interpretable, sharp, and free of artifacts. We suggest a new metric for saliency and test our method on the ImageNet object localisation task. We achieve results outperforming other weakly supervised methods.

************************************

Joint distribution optimal transportation for domain adaptation
Nicolas Courty, Rémi Flamary, Amaury Habrard, Alain Rakotomamonjy

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.

************************************

Learning A Structured Optimal Bipartite Graph for Co-Clustering
Feiping Nie, Xiaoqian Wang, Cheng Deng, Heng Huang

Co-clustering methods have been widely applied to document clustering and gene expression analysis. These methods make use of the duality between features and samples such that the co-occurring structure of sample and feature clusters can be extracted. In graph based co-clustering methods, a bipartite graph is constructed to depict the relation between features and samples. Most existing co-clustering methods conduct clustering on the graph achieved from the original data matrix, which doesn’t have explicit cluster structure, thus they require a post-processing step to obtain the clustering results. In this paper, we propose a novel co-clustering method to learn a bipartite graph with exactly k connected components, where k is the number of clusters. The new bipartite graph learned in our model approximates the original graph but maintains an explicit cluster structure, from which we can immediately get the clustering results without post-processing. Extensive empirical results are presented to verify the effectiveness and robustness of our model.

************************************

Learning to Inpaint for Image Compression
Mohammad Haris Baig, Vladlen Koltun, Lorenzo Torresani

We study the design of deep architectures for lossy image compression. We present two architectural recipes in the context of multi-stage progressive encoders and empirically demonstrate their importance on compression performance. Specifically, we show that: 1) predicting the original image data from residuals in a multi-stage progressive architecture facilitates learning and leads to improved performance at approximating the original content and 2) learning to inpaint (from neighboring image pixels) before performing compression reduces the amount of information that must be stored to achieve a high-quality approximation. Incorporating these design choices in a baseline progressive encoder yields an average reduction of over 60% in file size with similar quality compared to the original residual encoder.

************************************

Inverse Filtering for Hidden Markov Models
Robert Mattila, Cristian Rojas, Vikram Krishnamurthy, Bo Wahlberg

This paper considers a number of related inverse filtering problems for hidden Markov models (HMMs). In particular, given a sequence of state posteriors and the system dynamics; i) estimate the corresponding sequence of observations, ii) estimate the observation likelihoods, and iii) jointly estimate the observation likelihoods and the observation sequence. We show how to avoid a computationally expensive mixed integer linear program (MILP) by exploiting the algebraic structure of the HMM filter using simple linear algebra operations, and provide conditions for when the quantities can be uniquely reconstructed. We also propose a sol ution to the more general case where the posteriors are noisily observed. Finally, the proposed inverse filtering algorithms are evaluated on real-world polysomnographic data used for automatic sleep segmentation.

************************************

On clustering network-valued data

Soumendu Sundar Mukherjee, Purnamrita Sarkar, Lizhen Lin

Community detection, which focuses on clustering nodes or detecting communities in (mostly) a single network, is a problem of considerable practical interest and has received a great deal of attention in the  research community. While being able to cluster within a network is important, there are emerging needs to be able to \emph{cluster multiple networks}. This is largely motivated by the routine collection of network data that are generated from potentially different populations. These networks may or may not have node correspondence. When node correspondence is present, we cluster networks by summarizing a network by its graphon estimate, whereas when node correspondence is not present, we propose a novel solution for clustering such networks by associating a computationally feasible feature vector to each network based on trace of powers of the adjacency matrix. We illustrate our methods using both simulated and real data sets, and theoretical justifications are provided in terms of consistency.

************************************

Langevin Dynamics with Continuous Tempering for Training Deep Neural Networks

Nanyang Ye, Zhanxing Zhu, Rafal Mantiuk

Minimizing non-convex and high-dimensional objective functions is challenging, especially when training modern deep neural networks.  In this paper, a novel approach is proposed which divides the training process into two consecutive phases to obtain better generalization performance: Bayesian sampling and stochastic optimization. The first phase is to explore the energy landscape and to capture the `fat'' modes; and the second one is to fine-tune the parameter learned from the first phase. In the Bayesian learning phase, we apply continuous tempering and stochastic approximation into the Langevin dynamics to create an efficient and effective sampler, in which the temperature is adjusted automatically according to the designed ``temperature dynamics''.  These strategies can overcome the challenge of early trapping into bad local minima and have achieved remarkable improvements in various types of neural networks as shown in our theoretical analysis and empirical experiments.

************************************

Beyond Worst-case: A Probabilistic Analysis of Affine Policies in Dynamic Optimization

Omar El Housni, Vineet Goyal

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.

************************************

Few-Shot Learning Through an Information Retrieval  Lens

Eleni Triantafillou, Richard Zemel, Raquel Urtasun

Few-shot learning refers to understanding new concepts from only a few examples. We propose an information retrieval-inspired approach for this problem that is motivated by the increased importance of maximally leveraging all the available information in this low-data regime. We define a training objective that aims to extract as much information as possible from each training batch by effectively optimizing over all relative orderings of the batch points simultaneously. In particular, we view each batch point as a `query' that ranks the remaining ones based on its predicted relevance to them and we define a model within the framework of structured prediction to optimize mean Average Precision over these rankings. Our method achieves impressive results on the standard few-shot classification benchmarks while is also capable of few-shot retrieval.

************************************

Accelerated consensus via Min-Sum Splitting

Patrick Rebeschini, Sekhar C. Tatikonda

We apply the Min-Sum message-passing protocol to solve the consensus problem in distributed optimization. We show that while the ordinary Min-Sum algorithm does not converge, a modified version of it known as Splitting yields convergence to the problem solution. We prove that a proper choice of the tuning parameters allows Min-Sum Splitting to yield subdiffusive accelerated convergence rates, matching the rates obtained by shift-register methods. The acceleration scheme embodied by Min-Sum Splitting for the consensus problem bears similarities with lifted Markov chains techniques and with multi-step first order methods in convex optimization.

************************************

Saliency-based Sequential Image Attention with Multiset Prediction

Sean Welleck, Jialin Mao, Kyunghyun Cho, Zheng Zhang

Humans process visual scenes selectively and sequentially using attention. Central to models of human visual attention is the saliency map. We propose a hierarchical visual architecture that operates on a saliency map and uses a novel attention mechanism to sequentially focus on salient regions and take additional glimpses within those regions. The architecture is motivated by human visual attention, and is used for multi-label image classification on a novel multiset task, demonstrating that it achieves high precision and recall while localizing objects with its attention. Unlike conventional multi-label image classification models, the model supports multiset prediction due to a reinforcement-learning based training process that allows for arbitrary label permutation and multiple instances per label.

************************************

Adaptive Bayesian Sampling with Monte Carlo EM

Anirban Roychowdhury, Srinivasan Parthasarathy

We present a novel technique for learning the mass matrices in samplers obtained from discretized dynamics that preserve some energy function. Existing adaptive samplers use Riemannian preconditioning techniques, where the mass matrices are functions of the parameters being sampled. This leads to significant complexities in the energy reformulations and resultant dynamics, often leading to implicit systems of equations and requiring inversion of high-dimensional matrices in the leapfrog steps. Our approach provides a simpler alternative, by using existing dynamics in the sampling step of a Monte Carlo EM framework, and learning the mass matrices in the M step with a novel online technique. We also propose a way to adaptively set the number of samples gathered in the E step, using sampling error estimates from the leapfrog dynamics. Along with a novel stochastic sampler based on Nos\'{e}-Poincar\'{e} dynamics, we use this framework with standard Hamiltonian Monte Carlo (HMC) as well as newer stochastic algorithms such as SGHMC and SGNHT, and show strong performance on synthetic and real high-dimensional sampling scenarios; we achieve sampling accuracies comparable to Riemannian samplers while being significantly faster.

************************************

Scalable Levy Process Priors for Spectral Kernel Learning

Phillip A. Jang, Andrew Loeb, Matthew Davidow, Andrew G. Wilson

Gaussian processes are rich distributions over functions, with generalization properties determined by a kernel function. When used for long-range extrapolation, predictions are particularly sensitive to the choice of kernel parameters. It is therefore critical to account for kernel uncertainty in our predictive distributions. We propose a distribution over kernels formed by modelling a spectral mixture density with a Levy process. The resulting distribution has support for all stationary covariances---including the popular RBF, periodic, and Matern kernels---combined with inductive biases which enable automatic and data efficient learning, long-range extrapolation, and state of the art predictive performance. The proposed model also presents an approach to spectral regularization, as the Levy process introduces a sparsity-inducing prior over mixture components, allowing automatic selection over model order and pruning of extraneous components. We exploit the algebraic structure of the proposed process for O(n) training and O(1) predictions. We perform extrapolations having reasonable uncertainty estimates on several benchmarks, show that the proposed model can recover flexible gro und truth covariances and that it is robust to errors in initialization.

************************************

Model-Powered Conditional Independence Test

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G. Dimakis, Sanjay Shakkottai

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.

************************************

Learning Multiple Tasks with Multilinear Relationship Networks

Mingsheng Long, ZHANGJIE CAO, Jianmin Wang, Philip S. Yu

Deep networks trained on large-scale data can learn transferable features to promote learning multiple tasks. Since deep features eventually transition from general to specific along deep networks, a fundamental problem of multi-task learning is how to exploit the task relatedness underlying  parameter tensors and improve feature transferability in the multiple task-specific layers. This paper presents Multilinear Relationship Networks (MRN) that discover the task relationships based on novel tensor normal priors over parameter tensors of multiple task-specific layers in deep convolutional networks. By jointly learning transferable features and multilinear relationships of tasks and features, MRN is able to alleviate the dilemma of negative-transfer in the feature layers and under-transfer in the classifier layer. Experiments show that MRN yields state-of-the-art results on three multi-task learning datasets.

************************************

Query Complexity of Clustering with Side Information

Arya Mazumdar, Barna Saha

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.

************************************

Non-parametric Structured Output Networks

Andreas Lehrmann, Leonid Sigal

Deep neural networks (DNNs) and probabilistic graphical models (PGMs) are the two main tools for statistical modeling. While DNNs provide the ability to model rich and complex relationships between input and output variables, PGMs provide the ability to encode dependencies among the output variables themselves. End-to-end training methods for models with structured graphical dependencies on top of neural predictions have recently emerged as a principled way of combining these two paradigms. While these models have proven to be powerful in discriminative settings with discrete outputs, extensions to structured continuous spaces, as well as performing efficient inference in these spaces, are lacking. We propose non-parametric structured output networks (NSON), a modular approach that cleanly separates a non-parametric, structured posterior representation from a discriminative inference scheme but allows joint end-to-end training of both components. Our experiments evaluate the ability of NSONs to capture structured posterior densities (modeling) and to compute complex statistics of those densities (inference). We compare our model to output spaces of varying expressiveness and popular variational and sampling-based inference algorithms.

************************************

Robust Imitation of Diverse Behaviors

Ziyu Wang, Josh S. Merel, Scott E. Reed, Nando de Freitas, Gregory Wayne, Nicolas Heess

Deep generative models have recently shown great promise in imitation learning for motor control. Given enough data, even supervised approaches can do one-shot imitation learning; however, they are vulnerable to cascading failures when the agent trajectory diverges from the demonstrations. Compared to purely supervised methods, Generative Adversarial Imitation Learning (GAIL) can learn more robust controllers from fewer demonstrations, but is inherently mode-seeking and more difficult to train. In this paper, we show how to combine the favourable aspects of these two approaches. The base of our model is a new type of variational autoencoder on demonstration trajectories that learns semantic policy embeddings. We show that these embeddings can be learned on a 9 DoF Jaco robot arm in reaching tasks, and then smoothly interpolated with a resulting smooth interpolation of reaching behavior. Leveraging these policy representations, we develop a new version of GAIL that (1) is much more robust than the purely-supervised controller, especially with few demonstrations, and (2) avoids mode collapse, capturing many diverse behaviors when GAIL on its own does not. We demonstrate our approach on learning diverse gaits from demonstration on a 2D biped and a 62 DoF 3D humanoid in the MuJoCo physics environment.

************************************

High-Order Attention Models for Visual Question Answering

Idan Schwartz, Alexander Schwing, Tamir Hazan

The quest for algorithms that enable cognitive abilities is an important part of machine learning. A common trait in many recently investigated cognitive-like tasks is that they take into account different data modalities, such as visual and textual input. In this paper we propose a novel and generally applicable form of attention mechanism that learns high-order correlations between various data modalities. We show that high-order correlations effectively direct the appropriate attention to the relevant elements in the different data modalities that are required to solve the joint task. We demonstrate the effectiveness of our high-order attention mechanism on the task of visual question answering (VQA), where we achieve state-of-the-art performance on the standard VQA dataset.

************************************

FALKON: An Optimal Large Scale Kernel Method

Alessandro Rudi, Luigi Carratino, Lorenzo Rosasco

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.

************************************

Generalized Linear Model Regression under Distance-to-set Penalties

Jason Xu, Eric Chi, Kenneth Lange

Estimation in generalized linear models (GLM) is complicated by the presence of constraints. One can handle constraints by maximizing a penalized log-likelihood. Penalties such as the lasso are effective in high dimensions but often lead to severe shrinkage. This paper explores instead penalizing the squared distance to constraint sets. Distance penalties are more flexible than algebraic and regularization penalties, and avoid the drawback of shrinkage. To optimize distance penalized objectives, we make use of the majorization-minimization principle. Resulting algorithms constructed within this framework are amenable to acceleration and come with global convergence guarantees. Applications to shape constraints, sparse regression, and rank-restricted matrix regression on synthetic and real data showcase the strong empirical performance of distance penalization, even under non-convex constraints.

************************************

Fisher GAN

Youssef Mroueh, Tom Sercu

Generative Adversarial Networks (GANs) are powerful models for learning complex distributions. Stable training of GANs has been addressed in many recent works which explore different metrics between distributions. In this paper we introduce Fisher GAN that fits within the Integral Probability Metrics (IPM) framework for training GANs. Fisher GAN defines a data dependent constraint on the second order moments of the critic. We show in this paper that Fisher GAN allows for stable and time efficient training that does not compromise the capacity of the critic, and does not need data independent constraints such as weight clipping. We analyze our Fisher IPM theoretically and provide an algorithm based on Augmented Lagrangian for Fisher GAN. We validate our claims on both image sample generation and semi-supervised classification using Fisher GAN.

************************************

Minimax Estimation of Bandable Precision Matrices

Addison Hu, Sahand Negahban

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.

************************************

Kernel functions based on triplet comparisons

Matthäus Kleindessner, Ulrike von Luxburg

Given only information in the form of similarity triplets "Object A is more similar to object B than to object C" about a data set, we propose two ways of defining a kernel function on the data set. While previous approaches construct a low-dimensional Euclidean embedding of the data set that reflects the given similarity triplets, we aim at defining kernel functions that correspond to high-dimensional embeddings. These kernel functions can subsequently be used to apply any kernel method to the data set.

************************************

Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization

Fabian Pedregosa, Rémi Leblond, Simon Lacoste-Julien

Due to their simplicity and excellent performance, parallel asynchronous variants of stochastic gradient descent have become popular methods to solve a wide range of large-scale optimization problems on multi-core architectures. Yet, despite their practical success, support for nonsmooth objectives is still lacking, making them unsuitable for many problems of interest in machine learning, such as the Lasso, group Lasso or empirical risk minimization with convex constraints.  In this work, we propose and analyze ProxASAGA, a fully asynchronous sparse method inspired by SAGA, a variance reduced incremental gradient algorithm. The proposed method is easy to implement and significantly outperforms the state of the art on several nonsmooth, large-scale problems. We prove that our method achieves a theoretical linear speedup with respect to the sequential version under assumptions on the sparsity of gradients and block-separability of the proximal term. Empirical benchmarks on a multi-core architecture illustrate practical speedups of up to 12x on a 20-core machine.

************************************

A New Theory for Matrix Completion

Guangcan Liu, Qingshan Liu, Xiaotong Yuan

Prevalent matrix completion theories reply on an assumption that the locations of the missing data are distributed uniformly and randomly (i.e., uniform sampling). Nevertheless, the reason for observations being missing often depends on the unseen observations themselves, and thus the missing data in practice usually occurs in a nonuniform and deterministic fashion rather than randomly. To break through the limits of random sampling, this paper introduces a new hypothesis called \emph{isomeric condition}, which is provably weaker than the assumption of uniform sampling and arguably holds even when the missing data is placed irregularly. Equipped with this new tool, we prove a series of theorems for missing data recovery and matrix completion. In particular, we prove that the exact solutions that identify the target matrix are included as critical points by the commonly used nonconvex programs. Unlike the existing theories for nonconvex matrix completion, which are built upon the same condition as convex programs, our theory shows that nonconvex programs have the potential to work with a much weaker condition. Comparing to the existing studies on nonuniform sampling, our setup is more general.

************************************

A Bayesian Data Augmentation Approach for Learning Deep Models

Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, Ian Reid

Data augmentation is an essential part of the training process applied to deep learning models.  The motivation is that a robust training process for deep learning models depends on large annotated datasets, which are expensive to be acquir ed, stored and processed.  Therefore a reasonable alternative is to be able to automatically generate new annotated training samples using a process known as data augmentation. The dominant data augmentation approach in the field assumes that new training samples can be obtained via random geometric or appearance transformations applied to annotated training samples, but this is a strong assumption because it is unclear if this is a reliable generative model for producing new training samples. In this paper, we provide a novel Bayesian formulation to data augmentation, where new annotated training points are treated as missing variables and generated based on the distribution learned from the training set. For learning, we introduce a theoretically sound algorithm --- generalised Monte Carlo expectation maximisation, and demonstrate one possible implementation via an extension of the Generative Adversarial Network (GAN). Classification results on MNIST, CIFAR-10 and CIFAR-100 show the better performance of our proposed method compared to the current dominant data augmentation approach mentioned above --- the results also show that our approach produces better classification results than similar GAN models.

************************************

Deep Hyperalignment

Muhammad Yousefnezhad, Daoqiang Zhang

This paper proposes Deep Hyperalignment (DHA) as a regularized, deep extension, scalable Hyperalignment (HA) method, which is well-suited for applying functional alignment to fMRI datasets with nonlinearity, high-dimensionality (broad ROI), and a large number of subjects. Unlink previous methods, DHA is not limited by a restricted fixed kernel function. Further, it uses a parametric approach, rank-m Singular Value Decomposition (SVD), and stochastic gradient descent for optimization. Therefore, DHA has a suitable time complexity for large datasets, and DHA does not require the training data when it computes the functional alignment for a new subject. Experimental studies on multi-subject fMRI analysis confirm that the DHA method achieves superior performance to other state-of-the-art HA algorithms.

************************************

Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, Dhruv Batra

We present a novel training framework for neural sequence models, particularly for grounded dialog generation. The standard training paradigm for these models is maximum likelihood estimation (MLE), or minimizing the cross-entropy of the human responses. Across a variety of domains, a recurring problem with MLE trained generative neural dialog models (G) is that they tend to produce 'safe' and generic responses like "I don't know", "I can't tell"). In contrast, discriminative dialog models (D) that are trained to rank a list of candidate human responses outperform their generative counterparts; in terms of automatic metrics, diversity, and informativeness of the responses. However, D is not useful in practice since it can not be deployed to have real conversations with users.   Our work aims to achieve the best of both worlds -- the practical usefulness of G and the strong performance of D -- via knowledge transfer from D to G. Our primary contribution is an end-to-end trainable generative visual dialog model, where G receives gradients from D as a perceptual (not adversarial) loss of the sequence sampled from G. We leverage the recently proposed Gumbel-Softmax (GS) approximation to the discrete distribution -- specifically, a RNN is augmented with a sequence of GS samplers, which coupled with the straight-through gradient estimator enables end-to-end differentiability. We also introduce a stronger encoder for visual dialog, and employ a self-attention mechanism for answer encoding along with a metric learning loss to aid D in better capturing semantic similarities in answer responses. Overall, our proposed model outperforms state-of-the-art on the VisDial dataset by a significant margin (2.67% on recall@10). The source code can be downloaded from https://github.com/jiasenlu/visDial.pytorch

************************************

PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference

Jonathan Huggins, Ryan P. Adams, Tamara Broderick

Generalized linear models (GLMs)---such as logistic regression, Poisson regression, and robust regression---provide interpretable models for diverse data types. Probabilistic approaches, particularly Bayesian ones, allow coherent estimates of uncertainty, incorporation of prior information, and sharing of power across experiments via hierarchical models. In practice, however, the approximate Bayesian methods necessary for inference have either failed to scale to large data sets or failed to provide theoretical guarantees on the quality of inference. We propose a new approach based on constructing polynomial approximate sufficient statistics for GLMs (PASS-GLM). We demonstrate that our method admits a simple algorithm as well as trivial streaming and distributed extensions that do not compound error across computations. We provide theoretical guarantees on the quality of point (MAP) estimates, the approximate posterior, and posterior mean and uncertainty estimates. We validate our approach empirically in the case of logistic regression using a quadratic approximation and show competitive performance with stochastic gradient descent, MCMC, and the Laplace approximation in terms of speed and multiple measures of accuracy---including on an advertising data set with 40 million data points and 20,000 covariates.

************************************

Online multiclass boosting

Young Hun Jung, Jack Goetz, Ambuj Tewari

Recent work has extended the theoretical analysis of boosting algorithms to multiclass problems and to online settings. However, the multiclass extension is in the batch setting and the online extensions only consider binary classification. We fill this gap in the literature by defining, and justifying, a weak learning condition for online multiclass boosting. This condition leads to an optimal boosting algorithm that requires the minimal number of weak learners to achieve a certain accuracy. Additionally, we propose an adaptive algorithm which is near optimal and enjoys an excellent performance on real data due to its adaptive property.

************************************

State Aware Imitation Learning

Yannick Schroecker, Charles L. Isbell

Imitation learning is the study of learning how to act given a set of demonstrations provided by a human expert. It is intuitively apparent that learning to take optimal actions is a simpler undertaking in situations that are similar to the ones shown by the teacher. However, imitation learning approaches do not tend to use this insight directly. In this paper, we introduce State Aware Imitation Learning (SAIL), an imitation learning algorithm that allows an agent to learn how to remain in states where it can confidently take the correct action and how to recover if it is lead astray. Key to this algorithm is a gradient learned using a temporal difference update rule which leads the agent to prefer states similar to the demonstrated states. We show that estimating a linear approximation of this gradient yields similar theoretical guarantees to online temporal difference learning approaches and empirically show that SAIL can effectively be used for imitation learning in continuous domains with non-linear function approximators used for both the policy representation and the gradient estimate.

************************************

Adaptive SVRG Methods under Error Bound Conditions with Unknown Growth Parameter

Yi Xu, Qihang Lin, Tianbao Yang

Error bound, an inherent property of an optimization problem, has recently revived in the development of algorithms with improved global convergence without strong convexity. The most studied error bound is the quadratic error bound, which generalizes strong convexity and is satisfied by a large family of machine learning problems. Quadratic error bound have been leveraged to achieve linear convergence in many first-order methods including the stochastic variance reduced gradient (SVRG) method, which is one of the most important stochastic optimization methods in machine learning. However, the studies along this direction face the critical issue that the algorithms must depend on an unknown growth parameter (a generalization of strong convexity modulus) in the error bound. This parameter is difficult to estimate exactly and the algorithms choosing this parameter heuristically do not have theoretical convergence guarantee. To address this issue, we propose novel SVRG methods that automatically search for this unknown parameter on the fly of optimization while still obtain almost the same convergence rate as when this parameter is known. We also analyze the convergence property of SVRG methods under H\"{o}lderian error bound, which generalizes the quadratic error bound.

************************************

Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data

Wei-Ning Hsu, Yu Zhang, James Glass

We present a factorized hierarchical variational autoencoder, which learns disentangled and interpretable representations from sequential data without supervision. Specifically, we exploit the multi-scale nature of information in sequential data by formulating it explicitly within a factorized hierarchical graphical model that imposes sequence-dependent priors and sequence-independent priors to different sets of latent variables. The model is evaluated on two speech corpora to demonstrate, qualitatively, its ability to transform speakers or linguistic content by manipulating different sets of latent variables; and quantitatively, its ability to outperform an i-vector baseline for speaker verification and reduce the word error rate by as much as 35% in mismatched train/test scenarios for automatic speech recognition tasks.

************************************

Recurrent Ladder Networks

Isabeau Prémont-Schwarz, Alexander Ilin, Tele Hao, Antti Rasmus, Rinu Boney, Harri Valpola

We propose a recurrent extension of the Ladder networks whose structure is motivated by the inference required in hierarchical latent variable models. We demonstrate that the recurrent Ladder is able to handle a wide variety of complex learning tasks that benefit from iterative inference and temporal modeling. The architecture shows close-to-optimal results on temporal modeling of video data, competitive results on music modeling, and improved perceptual grouping based on higher order abstractions, such as stochastic textures and motion cues. We present results for fully supervised, semi-supervised, and unsupervised tasks. The results suggest that the proposed architecture and principles are powerful tools for learning a hierarchy of abstractions, learning iterative inference and handling temporal information.

************************************

Distral: Robust multitask reinforcement learning

Yee Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, Razvan Pascanu

Most deep reinforcement learning algorithms are data inefficient in complex and rich environments, limiting their applicability to many scenarios. One direction for improving data efficiency is multitask learning with shared neural network parameters, where efficiency may be improved through transfer across related tasks. In practice, however,  this is not usually observed, because gradients from different tasks can interfere negatively, making learning unstable and sometimes even less data efficient. Another issue is the different reward schemes between tasks, which can easily lead to one task dominating the learning of a shared model. We propose a new  approach for joint training of multiple tasks, which we refer to as Distral (DIStill & TRAnsfer Learning). Instead of sharing parameters between the different workers, we propose to share a distilled policy that captures common behaviour across tasks. Each worker is trained to solve its own task while constrained to stay close to the shared policy, while the shared policy is trained by distillation to be the centroid of all task policies. Both aspects of the learning process are derived by optimizing a joint objective function. We show that our approach supports efficient transfer on complex 3D environments, outperforming several related methods. Moreover, the proposed learning process is more robust and more stable---attributes that are critical in deep reinforcement learning.

************************************

Real-Time Bidding with Side Information

arthur flajolet, Patrick Jaillet

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.

************************************

Learning Spherical Convolution for Fast Features from 360° Imagery

Yu-Chuan Su, Kristen Grauman

While 360° cameras offer tremendous new possibilities in vision, graphics, and augmented reality, the spherical images they produce make core feature extraction non-trivial. Convolutional neural networks (CNNs) trained on images from perspective cameras yield “flat" filters, yet 360° images cannot be projected to a single plane without significant distortion. A naive solution that repeatedly projects the viewing sphere to all tangent planes is accurate, but much too computationally intensive for real problems. We propose to learn a spherical convolutional network that translates a planar CNN to process 360° imagery directly in its equirectangular projection. Our approach learns to reproduce the flat filter outputs on 360° data, sensitive to the varying distortion effects across the viewing sphere. The key benefits are 1) efficient feature extraction for 360° images and video, and 2) the ability to leverage powerful pre-trained networks researchers have carefully honed (together with massive labeled image training sets) for perspective images. We validate our approach compared to several alternative methods in terms of both raw CNN output accuracy as well as applying a state-of-the-art “flat" object detector to 360° data. Our method yields the most accurate results while saving orders of magnitude in computation versus the existing exact reprojection solution.

************************************

Approximate Supermodularity Bounds for Experimental Design

Luiz Chamon, Alejandro Ribeiro

This work provides performance guarantees for the greedy solution of experimental design problems. In particular, it focuses on A- and E-optimal designs, for which typical guarantees do not apply since the mean-square error and the maximum eigenvalue of the estimation error covariance matrix are not supermodular. To do so, it leverages the concept of approximate supermodularity to derive non-asymptotic worst-case suboptimality bounds for these greedy solutions. These bounds reveal that as the SNR of the experiments decreases, these cost functions behave increasingly as supermodular functions. As such, greedy A- and E-optimal designs approach (1-1/e)-optimality. These results reconcile the empirical success of greedy experimental design with the non-supermodularity of the A- and E-optimality criteria.

************************************

Differentiable Learning of Logical Rules for Knowledge Base Reasoning

Fan Yang, Zhilin Yang, William W. Cohen

We study the problem of learning probabilistic first-order logical rules for knowledge base reasoning. This learning problem is difficult because it requires learning the parameters in a continuous space as well as the structure in a discrete space. We propose a framework, Neural Logic Programming, that combines the parameter and structure learning of first-order logical rules in an end-to-end differentiable model. This approach is inspired by a recently-developed differentiable logic called TensorLog [5], where inference tasks can be compiled into sequences of differentiable operations. We design a neural controller system that learns to compose these operations. Empirically, our method outperforms prior work on multiple knowledge base benchmark datasets, including Freebase and WikiMovies.

************************************

When Cyclic Coordinate Descent Outperforms Randomized Coordinate Descent

Mert Gurbuzbalaban, Asuman Ozdaglar, Pablo A. Parrilo, Nuri Vanli

The coordinate descent (CD) method is a classical optimization algorithm that ha s seen a revival of interest because of its competitive performance in machine learning applications. A number of recent papers provided convergence rate estimates for their deterministic (cyclic) and randomized variants that differ in the selection of update coordinates. These estimates suggest randomized coordinate descent (RCD) performs better than cyclic coordinate descent (CCD), although numerical experiments do not provide clear justification for this comparison. In this paper, we provide examples and more generally problem classes for which CCD (or CD with any deterministic order) is faster than RCD in terms of asymptotic worst-case convergence. Furthermore, we provide lower and upper bounds on the amount of improvement on the rate of CCD relative to RCD, which depends on the deterministic order used. We also provide a characterization of the best deterministic order (that leads to the maximum improvement in convergence rate) in terms of the combinatorial properties of the Hessian matrix of the objective function.

************************************

Principles of Riemannian Geometry  in Neural Networks

Michael Hauser, Asok Ray

This study deals with neural networks in the sense of geometric transformations acting on the coordinate representation of the underlying data manifold which the data is sampled from. It forms part of an attempt to construct a formalized general theory of neural networks in the setting of Riemannian geometry. From this perspective, the following theoretical results are developed and proven for feedforward networks. First it is shown that residual neural networks are finite difference approximations to dynamical systems of first order differential equations, as opposed to ordinary networks that are static. This implies that the network is learning systems of differential equations governing the coordinate transformations that represent the data. Second it is shown that a closed form solution of the metric tensor on the underlying data manifold can be found by backpropagating the coordinate representations learned by the neural network itself. This is formulated in a formal abstract sense as a sequence of Lie group actions on the metric fibre space in the principal and associated bundles on the data manifold. Toy experiments were run to confirm parts of the proposed theory, as well as to provide intuitions as to how neural networks operate on data.

************************************

Continual Learning with Deep Generative Replay

Hanul Shin, Jung Kwon Lee, Jaehong Kim, Jiwon Kim

Attempts to train a comprehensive artificial intelligence capable of solving multiple tasks have been impeded by a chronic problem called catastrophic forgetting. Although simply replaying all previous data alleviates the problem, it requires large memory and even worse, often infeasible in real world applications where the access to past data is limited. Inspired by the generative nature of the hippocampus as a short-term memory system in primate brain, we propose the Deep Generative Replay, a novel framework with a cooperative dual model architecture consisting of a deep generative model (“generator”) and a task solving model (“solver”). With only these two models, training data for previous tasks can easily be sampled and interleaved with those for a new task. We test our methods in several sequential learning settings involving image classification tasks.

************************************

Nonlinear random matrix theory for deep learning

Jeffrey Pennington, Pratik Worah

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.

************************************

Identification of Gaussian Process State Space Models

Stefanos Eleftheriadis, Tom Nicholson, Marc Deisenroth, James Hensman

The Gaussian process state space model (GPSSM) is a non-linear dynamical system, where unknown transition and/or measurement mappings are described by GPs. Most research in GPSSMs has focussed on the state estimation problem, i.e., computing a posterior of the latent state given the model. However, the key challenge in

GPSSMs has not been satisfactorily addressed yet: system identification, i.e., learning the model. To address this challenge, we impose a structured Gaussian variational posterior distribution over the latent states, which is parameterised by a recognition model in the form of a bi-directional recurrent neural network. Inference with this structure allows us to recover a posterior smoothed over sequences of data. We provide a practical algorithm for efficiently computing a lower bound on the marginal likelihood using the reparameterisation trick. This further allows for the use of arbitrary kernels within the GPSSM. We demonstrate that the learnt GPSSM can efficiently generate plausible future trajectories of the identified system after only observing a small number of episodes from the true system.

************************************

Estimation of the covariance structure of heavy-tailed distributions
Xiaohan Wei, Stanislav Minsker
We propose and analyze a new estimator of the covariance matrix that admits strong theoretical guarantees under weak assumptions on the underlying distribution, such as existence of moments of only low order. While estimation of covariance matrices corresponding to sub-Gaussian distributions is well-understood, much less in known in the case of heavy-tailed data. As K. Balasubramanian and M. Yuan write, data from real-world experiments oftentimes tend to be corrupted with outliers and/or exhibit heavy tails. In such cases, it is not clear that those covariance matrix estimators .. remain optimal'' and..what are the other possible strategies to deal with heavy tailed distributions warrant further studies.'' We make a step towards answering this question and prove tight deviation inequalities for the proposed estimator that depend only on the parameters controlling the ``intrinsic dimension'' associated to the covariance matrix (as opposed to the dimension of the ambient space); in particular, our results are applicable in the case of high-dimensional observations.

************************************

Robust Optimization for Non-Convex Objectives
Robert S. Chen, Brendan Lucier, Yaron Singer, Vasilis Syrgkanis
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.

************************************

Exploring Generalization in Deep Learning
Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, Nati Srebro
With a goal of understanding what drives generalization in deep networks, we consider several recently suggested explanations, including norm-based control, sharpness and robustness. We study how these measures can ensure generalization, highlighting the importance of scale normalization, and making a connection between sharpness and PAC-Bayes theory. We then investigate how well the measures explain different observed phenomena.

************************************

Spherical convolutions and their application in molecular modelling
Wouter Boomsma, Jes Frellsen
Convolutional neural networks are increasingly used outside the domain of image analysis, in particular in various areas of the natural sciences concerned with spatial data. Such networks often work out-of-the box, and in some cases entire model architectures from image analysis can be carried over to other problem domains almost unaltered. Unfortunately, this convenience does not trivially extend to data in non-euclidean spaces, such as spherical data. In this paper, we introduce two strategies for conducting convolutions on the sphere, using either a spherical-polar grid or a grid based on the cubed-sphere representation. We investigate the challenges that arise in this setting, and extend our discussion to include scenarios of spherical volumes, with several strategies for parameterizing the radial dimension. As a proof of concept, we conclude with an assessment of the performance of spherical convolutions in the context of molecular modelling, by considering structural environments within proteins. We show that the model

s are capable of learning non-trivial functions in these molecular environments, and that our spherical convolutions generally outperform standard 3D convolutions in this setting. In particular, despite the lack of any domain specific feature-engineering, we demonstrate performance comparable to state-of-the-art methods in the field, which build on decades of domain-specific knowledge.

************************************

Safe Adaptive Importance Sampling

Sebastian U. Stich, Anant Raj, Martin Jaggi

Importance sampling has become an indispensable strategy to speed up optimization algorithms for large-scale applications. Improved adaptive variants -- using importance values defined by the complete gradient information which changes during optimization -- enjoy favorable theoretical properties, but are typically computationally infeasible. In this paper we propose an efficient approximation of gradient-based sampling, which is based on safe bounds on the gradient. The proposed sampling distribution is  (i) provably the \emph{best sampling} with respect to the given bounds,  (ii) always better than uniform sampling and fixed importance sampling and  (iii) can efficiently be computed -- in many applications  at negligible extra cost. The proposed sampling scheme is generic and can easily be integrated into existing algorithms. In particular, we show that coordinate-descent (CD) and stochastic gradient descent (SGD) can enjoy significant a speed-up under the novel scheme. The proven efficiency of the proposed sampling is verified by extensive numerical testing.

************************************

Introspective Classification with Convolutional Nets

Long Jin, Justin Lazarow, Zhuowen Tu

We propose introspective convolutional networks (ICN) that emphasize the importance of having convolutional neural networks empowered with generative capabilities. We employ a reclassification-by-synthesis algorithm to perform training using a formulation stemmed from the Bayes theory. Our ICN tries to iteratively: (1) synthesize pseudo-negative samples; and (2) enhance itself by improving the classification. The single CNN classifier learned is at the same time generative --- being able to directly synthesize new samples within its own discriminative model. We conduct experiments on benchmark datasets including MNIST, CIFAR-10, and SVHN using state-of-the-art CNN architectures, and observe improved classification results.

************************************

Hybrid Reward Architecture for Reinforcement Learning

Harm Van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, Jeffrey Tsang

One of the main challenges in reinforcement learning (RL) is generalisation. In typical deep RL methods this is achieved by approximating the optimal value function with a low-dimensional representation using a deep network.  While this approach works well in many domains, in domains where the optimal value function cannot easily be reduced to a low-dimensional representation, learning can be very slow and unstable. This paper contributes towards tackling such challenging domains, by proposing a new method, called Hybrid Reward Architecture (HRA). HRA takes as input a decomposed reward function and learns a separate value function for each component reward function. Because each component typically only depends on a subset of all features, the corresponding value function can be approximated more easily by a low-dimensional representation, enabling more effective learning. We demonstrate HRA on a toy-problem and the Atari game Ms. Pac-Man, where HRA achieves above-human performance.

************************************

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness

Chris Russell, Matt J. Kusner, Joshua Loftus, Ricardo Silva

Machine learning is now being used to make crucial decisions about people's lives. For nearly all of these decisions there is a risk that individuals of a certain race, gender, sexual orientation, or any other subpopulation are unfairly discriminated against. Our recent method has demonstrated how to use techniques fro m counterfactual inference to make predictions fair across different subpopulations. This method requires that one provides the causal model that generated the data at hand. In general, validating all causal implications of the model is not possible without further assumptions. Hence, it is desirable to integrate competing causal models to provide counterfactually fair decisions, regardless of which causal "world" is the correct one. In this paper, we show how it is possible to make predictions that are approximately fair with respect to multiple possible causal models at once, thus mitigating the problem of exact causal specification. We frame the goal of learning a fair classifier as an optimization problem with fairness constraints entailed by competing causal explanations. We show how this optimization problem can be efficiently solved using gradient-based methods. We demonstrate the flexibility of our model on two real-world fair classification problems. We show that our model can seamlessly balance fairness in multiple worlds with prediction accuracy.

************************************

Dualing GANs

Yujia Li, Alexander Schwing, Kuan-Chieh Wang, Richard Zemel

Generative adversarial nets (GANs) are a promising technique for modeling a distribution from samples. It is however well known that GAN training suffers from instability due to the nature of its saddle point formulation. In this paper, we explore ways to tackle the instability problem by dualizing the discriminator. We start from linear discriminators in which case conjugate duality provides a mechanism to reformulate the saddle point objective into a maximization problem, such that both the generator and the discriminator of this ‘dualing GAN’ act in concert. We then demonstrate how to extend this intuition to non-linear formulations. For GANs with linear discriminators our approach is able to remove the instability in training, while for GANs with nonlinear discriminators our approach provides an alternative to the commonly used GAN training algorithm.

************************************

A Universal Analysis of Large-Scale Regularized Least Squares Solutions

Ashkan Panahi, Babak Hassibi

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.

************************************

Diffusion Approximations for Online Principal Component Estimation and Global Convergence

Chris Junchi Li, Mengdi Wang, Han Liu, Tong Zhang

In this paper, we propose to adopt the diffusion approximation tools to study the dynamics of Oja's iteration which is an online stochastic gradient method for the principal component analysis. Oja's iteration maintains a running estimate of the true principal component from streaming data and enjoys less temporal and spatial complexities. We show that the Oja's iteration for the top eigenvector generates a continuous-state discrete-time Markov chain over the unit sphere. We characterize the Oja's iteration in three phases using diffusion approximation and weak convergence tools. Our three-phase analysis further provides a finite-sample error bound for the running estimate, which matches the minimax information lower bound for PCA under the additional assumption of bounded samples.

************************************

k-Support and Ordered Weighted Sparsity for Overlapping Groups: Hardness and Algorithms

Cong Han Lim, Stephen Wright

The k-support and OWL norms generalize the l1 norm, providing better prediction accuracy and better handling of correlated variables. We study the norms obtained from extending the k-support norm and OWL norms to the setting in which there are overlapping groups. The resulting norms are in general NP-hard to compute, but they are tractable for certain collections of groups. To demonstrate this fact, we develop a dynamic program for the problem of projecting onto the set of vectors supported by a fixed number of groups. Our dynamic program utilizes tree decompositions and its complexity scales with the treewidth. This program can be converted to an extended formulation which, for the associated group structure, models the k-group support norms and an overlapping group variant of the ordered weighted l1 norm. Numerical results demonstrate the efficacy of the new penalties.

************************************

Learning to Model the Tail

Yu-Xiong Wang, Deva Ramanan, Martial Hebert

We describe an approach to learning from long-tailed, imbalanced datasets that are prevalent in real-world settings.  Here, the challenge is to learn accurate "few-shot'' models for classes in the tail of the class distribution, for which little data is available. We cast this problem as transfer learning, where knowledge from the data-rich classes in the head of the distribution is transferred to the data-poor classes in the tail. Our key insights are as follows. First, we propose to transfer meta-knowledge about learning-to-learn from the head classes. This knowledge is encoded with a meta-network that operates on the space of model parameters, that is trained to predict many-shot model parameters from few-shot model parameters.  Second, we transfer this meta-knowledge in a progressive manner, from classes in the head to the "body'', and from the "body'' to the tail. That is, we transfer knowledge in a gradual fashion, regularizing meta-networks for few-shot regression with those trained with more training data. This allows our final network to capture a notion of model dynamics, that predicts how model parameters are likely to change as more training data is gradually added. We demonstrate results on image classification datasets (SUN, Places, and ImageNet) tuned for the long-tailed setting, that significantly outperform common heuristics, such as data resampling or reweighting.

************************************

Neural Variational Inference and Learning in Undirected Graphical Models

Volodymyr Kuleshov, Stefano Ermon

Many problems in machine learning are naturally expressed in the language of undirected graphical models. Here, we propose black-box learning and inference algorithms for undirected models that optimize a variational approximation to the log-likelihood of the model. Central to our approach is an upper bound on the log-partition function parametrized by a function q that we express as a flexible neural network. Our bound makes it possible to track the partition function during learning, to speed-up sampling, and to train a broad class of hybrid directed/undirected models via a unified variational inference framework. We empirically demonstrate the effectiveness of our method on several popular generative modeling datasets.

************************************

Aggressive Sampling for Multi-class to Binary Reduction with Applications to Text Classification

Bikash Joshi, Massih R. Amini, Ioannis Partalas, Franck Iutzeler, Yury Maximov

We address the problem of multi-class classification in the case where the number of classes is very large. We propose a double sampling strategy on top of a multi-class to binary reduction strategy, which transforms the original multi-class problem into a binary classification problem over pairs of examples. The aim of the sampling strategy is to overcome the curse of long-tailed class distributions exhibited in majority  of  large-scale  multi-class classification problems and to reduce the number of pairs of examples in the expanded data.  We show that this strategy does not alter the consistency of the empirical risk minimization principle defined over the double sample reduction. Experiments are carried out on DMOZ and Wikipedia collections with 10,000 to 100,000 classes where we show the efficiency of the proposed approach in terms of training and prediction time, memory consumption, and predictive performance with respect to state-of-the-art approaches.

************************************

Learning Linear Dynamical Systems via Spectral Filtering

Elad Hazan, Karan Singh, Cyril Zhang

We present an efficient and practical algorithm for the online prediction of dis crete-time linear dynamical systems with a symmetric transition matrix. We circumvent the non-convex optimization problem using improper learning: carefully overparameterize the class of LDSs by a polylogarithmic factor, in exchange for convexity of the loss functions. From this arises a polynomial-time algorithm with a near-optimal regret guarantee, with an analogous sample complexity bound for agnostic learning. Our algorithm is based on a novel filtering technique, which may be of independent interest: we convolve the time series with the eigenvectors of a certain Hankel matrix.

************************************

Efficient Modeling of Latent Information in Supervised Learning using Gaussian Processes

Zhenwen Dai, Mauricio Álvarez, Neil Lawrence

Often in machine learning, data are collected as a combination of multiple conditions, e.g., the voice recordings of multiple persons, each labeled with an ID. How could we build a model that captures the latent information related to  these conditions and generalize to a new one with few data? We present a new model called Latent Variable Multiple Output Gaussian Processes (LVMOGP) that allows to jointly model multiple conditions for regression and generalize to a new condition with a few data points at test time. LVMOGP infers the posteriors of Gaussian processes together with a latent space representing the information about different conditions. We derive an efficient variational inference method for LVMOGP for which the computational complexity is as low as sparse Gaussian processes. We show that LVMOGP significantly outperforms related Gaussian process methods on various tasks with both synthetic and real data.

************************************

Semi-Supervised Learning for Optical Flow with Generative Adversarial Networks

Wei-Sheng Lai, Jia-Bin Huang, Ming-Hsuan Yang

Convolutional neural networks (CNNs) have recently been applied to the optical flow estimation problem. As training the CNNs requires sufficiently large ground truth training data, existing approaches resort to synthetic, unrealistic datasets. On the other hand, unsupervised methods are capable of leveraging real-world videos for training where the ground truth flow fields are not available. These methods, however, rely on the fundamental assumptions of brightness constancy and spatial smoothness priors which do not hold near motion boundaries. In this paper, we propose to exploit unlabeled videos for semi-supervised learning of optical flow with a Generative Adversarial Network. Our key insight is that the adversarial loss can capture the structural patterns of flow warp errors without making explicit assumptions. Extensive experiments on benchmark datasets demonstrate that the proposed semi-supervised algorithm performs favorably against purely supervised and semi-supervised learning schemes.

************************************

Phase Transitions in the Pooled Data Problem

Jonathan Scarlett, Volkan Cevher

In this paper, we study the {\em pooled data} problem of identifying the labels associated with a large collection of items, based on a sequence of pooled tests revealing the counts of each label within the pool.  In the noiseless setting, we identify an exact asymptotic threshold on the required number of tests with optimal decoding, and prove a {\em phase transition} between complete success and complete failure.  In addition, we present a novel {\em noisy} variation of the problem, and provide an information-theoretic framework for characterizing the required number of tests for general random noise models.  Our results reveal that noise can make the problem considerably more difficult, with strict increases in the scaling laws even at low noise levels.  Finally, we demonstrate similar behavior in an {\em approximate recovery} setting, where a given number of errors is allowed in the decoded labels.

************************************

Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning

Christoph Dann, Tor Lattimore, Emma Brunskill

Statistical performance bounds for reinforcement learning (RL) algorithms can be critical for high-stakes applications like healthcare. This paper introduces a new framework for theoretically measuring the performance of such algorithms called Uniform-PAC, which is a strengthening of the classical Probably Approximately Correct (PAC) framework. In contrast to the PAC framework, the uniform version may be used to derive high probability regret guarantees and so forms a bridge between the two setups that has been missing in the literature. We demonstrate the benefits of the new framework for finite-state episodic MDPs with a new algorithm that is Uniform-PAC and simultaneously achieves optimal regret and PAC guarantees except for a factor of the horizon.

************************************

Stein Variational Gradient Descent as Gradient Flow

Qiang Liu

Stein variational gradient descent (SVGD) is a deterministic sampling algorithm that iteratively transports a set of particles to approximate given distributions, based on a  gradient-based update constructed to optimally decrease the KL divergence within a function space. This paper develops the first theoretical analysis on SVGD. We establish that the empirical measures of the SVGD samples weakly converge to the target distribution, and show that the asymptotic behavior of SVGD is characterized by a nonlinear Fokker-Planck equation known as Vlasov equation in physics. We develop a geometric perspective that views SVGD as a gradient flow of the KL divergence functional under a new metric structure on the space of distributions induced by Stein operator.

************************************

Expectation Propagation for t-Exponential Family Using q-Algebra

Futoshi Futami, Issei Sato, Masashi Sugiyama

Exponential family distributions are highly useful in machine learning since their calculation can be performed efficiently through natural parameters. The exponential family has recently been extended to the t-exponential family, which contains Student-t distributions as family members and thus allows us to handle noisy data well. However, since the t-exponential family is defined by the deformed exponential, an efficient learning algorithm for the t-exponential family such as expectation propagation (EP) cannot be derived in the same way as the ordinary exponential family. In this paper, we borrow the mathematical tools of q-algebra from statistical physics and show that the pseudo additivity of distributions allows us to perform calculation of t-exponential family distributions through natural parameters. We then develop an expectation propagation (EP) algorithm for the t-exponential family, which provides a deterministic approximation to the posterior or predictive distribution with simple moment matching. We finally apply the proposed EP algorithm to the Bayes point machine and Student-t process classification, and demonstrate their performance numerically.

************************************

Collaborative PAC Learning

Avrim Blum, Nika Haghtalab, Ariel D. Procaccia, Mingda Qiao

We introduce a collaborative PAC learning model, in which k players attempt to learn the same underlying concept. We ask how much more information is required to learn an accurate classifier for all players simultaneously. We refer to the ratio between the sample complexity of collaborative PAC learning and its non-collaborative (single-player) counterpart as the overhead. We design learning algorithms with O(ln(k)) and O(ln^2(k)) overhead in the personalized and centralized variants our model. This gives an exponential improvement upon the naive algorithm that does not share information among players. We complement our upper bounds with an Omega(ln(k)) overhead lower bound, showing that our results are tight up to a logarithmic factor.

************************************

Polynomial time algorithms for dual volume sampling

Chengtao Li, Stefanie Jegelka, Suvrit Sra

We study dual volume sampling, a method for selecting k columns from an n*m short and wide matrix (n <= k <= m) such that the probability of selection is proportional to the volume spanned by the rows of the induced submatrix. This method was proposed by Avron and Boutsidis (2013), who showed it to be a promising method for column subset selection and its multiple applications. However, its wider adoption has been hampered by the lack of polynomial time sampling algorithms. We remove this hindrance by developing an exact (randomized) polynomial time sampling algorithm as well as its derandomization. Thereafter, we study dual volume sampling via the theory of real stable polynomials and prove that its distribution satisfies the “Strong Rayleigh” property. This result has numerous consequences, including a provably fast-mixing Markov chain sampler that makes dual volume sampling much more attractive to practitioners. This sampler is closely related to classical algorithms for popular experimental design methods that are to date lacking theoretical analysis but are known to empirically work well.

************************************

Premise Selection for Theorem Proving by Deep Graph Embedding

Mingzhe Wang, Yihe Tang, Jian Wang, Jia Deng

We propose a deep learning-based approach to the problem of premise selection: selecting mathematical statements relevant for proving a given conjecture. We represent a higher-order logic formula as a graph that is invariant to variable renaming but still fully preserves syntactic and semantic information. We then embed the graph into a vector via a novel embedding method that preserves the information of edge ordering. Our approach achieves state-of-the-art results on the HolStep dataset, improving the classification accuracy from 83% to 90.3%.

************************************

Differentiable Learning of Submodular Models

Josip Djolonga, Andreas Krause

Can we incorporate discrete optimization algorithms within modern machine learning models? For example, is it possible to use in deep architectures a layer whose output is the minimal cut of a parametrized graph? Given that these models are trained end-to-end by leveraging gradient information, the introduction of such layers seems very challenging due to their non-continuous output. In this paper we focus on the problem of submodular minimization, for which we show that such layers are indeed possible. The key idea is that we can continuously relax the output without sacrificing guarantees. We provide an easily computable approximation to the Jacobian complemented with a complete theoretical analysis. Finally, these contributions let us experimentally learn probabilistic log-supermodular models via a bi-level variational inference formulation.

************************************

YASS: Yet Another Spike Sorter

Jin Hyung Lee, David E. Carlson, Hooshmand Shokri Razaghi, Weichi Yao, Georges A. Goetz, Espen Hagen, Eleanor Batty, E.J. Chichilnisky, Gaute T. Einevoll, Liam Paninski

Spike sorting is a critical first step in extracting neural signals from large-scale electrophysiological data.  This manuscript describes an efficient, reliable pipeline for spike sorting on dense multi-electrode arrays (MEAs), where neural signals appear across many electrodes and spike sorting currently represents a major computational bottleneck. We present several new techniques that make dense MEA spike sorting more robust and scalable. Our pipeline is based on an efficient multi-stage ''triage-then-cluster-then-pursuit'' approach that initially extracts only clean, high-quality waveforms from the electrophysiological time series by temporarily skipping noisy or ''collided'' events (representing two neurons firing synchronously). This is accomplished by developing a neural network detection method followed by efficient outlier triaging. The clean waveforms are then used to infer the set of neural spike waveform templates through nonparametric Bayesian clustering. Our clustering approach adapts a ''coreset'' approach for data reduction and uses efficient inference methods in a Dirichlet process mixture model framework to dramatically improve the scalability and reliability of the entire pipeline. The ''triaged'' waveforms are then finally recovered with matching-pursuit deconvolution techniques. The proposed methods improve on the state-of-the-art in terms of accuracy and stability on both real and biophysically-realistic simulated MEA data. Furthermore, the proposed pipeline is efficient, learning templates and clustering faster than real-time for a 500-electrode dataset, largely on a single CPU core.

************************************

## Variational Laws of Visual Attention for Dynamic Scenes

Dario Zanca, Marco Gori

Computational models of visual attention are at the crossroad of disciplines like cognitive science, computational neuroscience, and computer vision. This paper proposes a model of attentional scanpath that is based on the principle that there are foundational laws that drive the emergence of visual attention. We devise variational laws of the eye-movement that rely on a generalized view of the Least Action Principle in physics. The potential energy  captures details  as well as peripheral visual features, while the kinetic energy corresponds with the classic interpretation in analytic mechanics. In addition, the Lagrangian contains a brightness invariance term, which characterizes significantly the scanpath trajectories. We obtain differential equations of visual attention as the stationary point of the generalized action, and we propose an algorithm to estimate the model parameters.  Finally, we report experimental results to validate the model in tasks of saliency detection.

************************************

## How regularization affects the critical points in linear networks

Amirhossein Taghvaei, Jin W. Kim, Prashant Mehta

This paper is concerned with the problem of representing and learning a linear transformation using a linear neural network.  In recent years, there is a growing interest in the study of such networks, in part due to the successes of deep learning.  The main question of this body of research (and also of our paper) is related to the existence and optimality properties of the critical points of the mean-squared loss function.  An additional primary concern of our paper pertains to the robustness of these critical points in the face of (a small amount of) regularization.  An optimal control model is introduced for this purpose and a learning algorithm (backprop with weight decay) derived for the same using the Hamilton's formulation of optimal control.  The formulation is used to provide a complete characterization of the critical points in terms of the solutions of a nonlinear matrix-valued equation, referred to as the characteristic equation.  Analytical and numerical tools from bifurcation theory are used to compute the critical points via the solutions of the characteristic equation.

************************************

## On Tensor Train Rank Minimization : Statistical Efficiency and Scalable Algorithm

Masaaki Imaizumi, Takanori Maehara, Kohei Hayashi

Tensor train (TT) decomposition provides a space-efficient representation for higher-order tensors. Despite its advantage, we face two crucial limitations when we apply the TT decomposition to machine learning problems: the lack of statistical theory and of scalable algorithms. In this paper, we address the limitations. First, we introduce a convex relaxation of the TT decomposition problem and derive its error bound for the tensor completion task. Next, we develop a randomized optimization method, in which the time complexity is as efficient as the space complexity is. In experiments, we numerically confirm the derived bounds and empirically demonstrate the performance of our method with a real higher-order tensor.

************************************

## EX2: Exploration with Exemplar Models for Deep Reinforcement Learning

Justin Fu, John Co-Reyes, Sergey Levine

Deep reinforcement learning algorithms have been shown to learn complex tasks using highly general policy classes. However, sparse reward problems remain a significant challenge. Exploration methods based on novelty detection have been particularly successful in such settings but typically require generative or predictive models of the observations, which can be difficult to train when the observations are very high-dimensional and complex, as in the case of raw images. We propose a novelty detection algorithm for exploration that is based entirely on discriminatively trained exemplar models, where classifiers are trained to discriminate each visited state against all others. Intuitively, novel states are easier to distinguish against other states seen during training. We show that this kind of discriminative modeling corresponds to implicit density estimation, and th

at it can be combined with count-based exploration to produce competitive results on a range of popular benchmark tasks, including state-of-the-art results on challenging egocentric observations in the vizDoom benchmark.
************************************

Training Quantized Nets: A Deeper Understanding
Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, Tom Goldstein
Currently, deep neural networks are deployed on low-power portable devices by first training a full-precision model using powerful hardware, and then deriving a corresponding low-precision model for efficient inference on such systems. However, training models directly with coarsely quantized weights is a key step towards learning on embedded platforms that have limited computing resources, memory capacity, and power consumption.  Numerous recent publications have studied methods for training quantized networks, but these studies have mostly been empirical. In this work, we investigate training methods for quantized neural networks from a theoretical viewpoint.  We first explore accuracy guarantees for training methods under convexity assumptions.  We then look at the behavior of these algorithms for non-convex problems, and show that training algorithms that exploit high-precision representations have an important greedy search phase that purely quantized training methods lack, which explains the difficulty of training using low-precision arithmetic.
************************************

Convolutional Gaussian Processes
Mark van der Wilk, Carl Edward Rasmussen, James Hensman
We present a practical way of introducing convolutional structure into Gaussian processes, making them more suited to high-dimensional inputs like images. The main contribution of our work is the construction of an inter-domain inducing point approximation that is well-tailored to the convolutional kernel. This allows us to gain the generalisation benefit of a convolutional kernel, together with fast but accurate posterior inference. We investigate several variations of the convolutional kernel, and apply it to MNIST and CIFAR-10, where we obtain significant improvements over existing Gaussian process models. We also show how the marginal likelihood can be used to find an optimal weighting between convolutional and RBF kernels to further improve performance. This illustration of the usefulness of the marginal likelihood may help automate discovering architectures in larger models.
************************************

Best Response Regression
Omer Ben-Porat, Moshe Tennenholtz
In a regression task, a predictor is given a set of instances, along with a real value for each point. Subsequently, she has to identify the value of a new instance as accurately as possible. In this work, we initiate the study of strategic predictions in machine learning.  We consider a regression task tackled by two players, where the payoff of each player is the proportion of the points she predicts more accurately than the other player. We first revise the probably approximately correct learning framework to deal with the case of a duel between two predictors.  We then devise an algorithm which finds a linear regression predictor that is a best response to any (not necessarily linear) regression algorithm. We show that it has linearithmic sample complexity, and polynomial time complexity when the dimension of the instances domain is fixed. We also test our approach in a high-dimensional setting, and show it significantly defeats classical regression algorithms in the prediction duel. Together, our work introduces a novel machine learning task that lends itself well to current competitive online settings, provides its theoretical foundations, and illustrates its applicability.
************************************

Elementary Symmetric Polynomials for Optimal Experimental Design
Zelda E. Mariet, Suvrit Sra
We revisit the classical problem of optimal experimental design (OED) under a new mathematical model grounded in a geometric motivation. Specifically, we introduce models based on elementary symmetric polynomials; these polynomials capture "partial volumes" and offer a graded interpolation between the widely used A-opt

imal and D-optimal design models, obtaining each of them as special cases. We analyze properties of our models, and derive both greedy and convex-relaxation algorithms for computing the associated designs. Our analysis establishes approximation guarantees on these algorithms, while our empirical results substantiate our claims and demonstrate a curious phenomenon concerning our greedy algorithm. Finally, as a byproduct, we obtain new results on the theory of elementary symmetric polynomials that may be of independent interest.

************************************

Learning from Complementary Labels

Takashi Ishida, Gang Niu, Weihua Hu, Masashi Sugiyama

Collecting labeled data is costly and thus a critical bottleneck in real-world classification tasks. To mitigate this problem, we propose a novel setting, namely learning from complementary labels for multi-class classification. A complementary label specifies a class that a pattern does not belong to. Collecting complementary labels would be less laborious than collecting ordinary labels, since users do not have to carefully choose the correct class from a long list of candidate classes. However, complementary labels are less informative than ordinary labels and thus a suitable approach is needed to better learn from them. In this paper, we show that an unbiased estimator to the classification risk can be obtained only from complementarily labeled data, if a loss function satisfies a particular symmetric condition. We derive estimation error bounds for the proposed method and prove that the optimal parametric convergence rate is achieved. We further show that learning from complementary labels can be easily combined with learning from ordinary labels (i.e., ordinary supervised learning), providing a highly practical implementation of the proposed method. Finally, we experimentally demonstrate the usefulness of the proposed methods.

************************************

Dynamic Importance Sampling for Anytime Bounds of the Partition Function

Qi Lou, Rina Dechter, Alexander T. Ihler

Computing the partition function is a key inference task in many graphical models. In this paper, we propose a dynamic importance sampling scheme that provides anytime finite-sample bounds for the partition function. Our algorithm balances the advantages of the three major inference strategies, heuristic search, variational bounds, and Monte Carlo methods, blending sampling with search to refine a variationally defined proposal. Our algorithm combines and generalizes recent work on anytime search and probabilistic bounds of the partition function. By using an intelligently chosen weighted average over the samples, we construct an unbiased estimator of the partition function with strong finite-sample confidence intervals that inherit both the rapid early improvement rate of sampling and the long-term benefits of an improved proposal from search. This gives significantly improved anytime behavior, and more flexible trade-offs between memory, time, and solution quality. We demonstrate the effectiveness of our approach empirically  on real-world problem instances taken from recent UAI competitions.

************************************

Process-constrained batch Bayesian optimisation

Pratibha Vellanki, Santu Rana, Sunil Gupta, David Rubin, Alessandra Sutti, Thomas Dorin, Murray Height, Paul Sanders, Svetha Venkatesh

Abstract Prevailing batch Bayesian optimisation methods allow all control variables to be freely altered at each iteration. Real-world experiments, however, often have physical limitations making it time-consuming to alter all settings for each recommendation in a batch. This gives rise to a unique problem in BO: in a recommended batch, a set of variables that are expensive to experimentally change need to be fixed, while the remaining control variables can be varied. We formulate this as a process-constrained batch Bayesian optimisation problem. We propose two algorithms, pc-BO(basic) and pc-BO(nested). pc-BO(basic) is simpler but lacks convergence guarantee. In contrast pc-BO(nested) is slightly more complex, but admits convergence analysis. We show that the regret of pc-BO(nested) is sublinear. We demonstrate the performance of both pc-BO(basic) and pc-BO(nested) by optimising benchmark test functions, tuning hyper-parameters of the SVM classifier, optimising the heat-treatment process for an Al-Sc alloy to achieve target

hardness, and optimising the short polymer fibre production process.
*************************************

Uprooting and Rerooting Higher-Order Graphical Models
Mark Rowland, Adrian Weller
The idea of uprooting and rerooting graphical models was introduced specifically for binary pairwise models by Weller (2016) as a way to transform a model to any of a whole equivalence class of related models, such that inference on any one model yields inference results for all others. This is very helpful since inference, or relevant bounds, may be much easier to obtain or more accurate for some model in the class. Here we introduce methods to extend the approach to models with higher-order potentials and develop theoretical insights. In particular, we show that the triplet-consistent polytope TRI is unique in being `universally rooted'. We demonstrate  empirically that rerooting can significantly improve accuracy of methods of inference for higher-order models at negligible computational cost.
*************************************

Learned in Translation: Contextualized Word Vectors
Bryan McCann, James Bradbury, Caiming Xiong, Richard Socher
Computer vision has benefited from initializing multiple deep layers with weights pretrained on large supervised training sets like ImageNet. Natural language processing (NLP) typically sees initialization of only the lowest layer of deep models with pretrained word vectors. In this paper, we use a deep LSTM encoder from an attentional sequence-to-sequence model trained for machine translation (MT) to contextualize word vectors. We show that adding these context vectors (CoVe) improves performance over using only unsupervised word and character vectors on a wide variety of common NLP tasks: sentiment analysis (SST, IMDb), question classification (TREC), entailment (SNLI), and question answering (SQuAD). For fine-grained sentiment analysis and entailment, CoVe improves performance of our baseline models to the state of the art.
*************************************

Semisupervised Clustering, AND-Queries and Locally Encodable Source Coding
Arya Mazumdar, Soumyabrata Pal
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.
*************************************

Regularizing Deep Neural Networks by Noise: Its Interpretation and Optimization
Hyeonwoo Noh, Tackgeun You, Jonghwan Mun, Bohyung Han
Overfitting is one of the most critical challenges in deep neural networks, and there are various types of regularization methods to improve generalization performance. Injecting noises to hidden units during training, e.g., dropout, is known as a successful regularizer, but it is still not clear enough why such training techniques work well in practice and how we can maximize their benefit in the presence of two conflicting objectives---optimizing to true data distribution and preventing overfitting by regularization. This paper addresses the above issues by 1) interpreting that the conventional training methods with regularization by noise injection optimize the lower bound of the true objective and 2) proposing a technique to achieve a tighter lower bound using multiple noise samples per training example  in a stochastic gradient descent iteration. We demonstrate the effectiveness of our idea in several computer vision applications.
*************************************

Few-Shot Adversarial Domain Adaptation
Saeid Motiian, Quinn Jones, Seyed Iranmanesh, Gianfranco Doretto
This work provides a framework for addressing the problem of supervised domain adaptation with deep models. The main idea is to exploit adversarial learning to learn an embedded subspace that simultaneously maximizes the confusion between two domains while semantically aligning their embedding. The supervised setting becomes attractive especially when there are only a few target data samples that need to be labeled. In this few-shot learning scenario, alignment and separation

of semantic probability distributions is difficult because of the lack of data. We found that by carefully designing a training scheme whereby the typical binary adversarial discriminator is augmented to distinguish between four different classes, it is possible to effectively address the supervised adaptation problem. In addition, the approach has a high “speed” of adaptation, i.e. it requires an extremely low number of labeled target training samples, even one per category can be effective. We then extensively compare this approach to the state of the art in domain adaptation in two experiments: one using datasets for handwritten digit recognition, and one using datasets for visual object recognition.
************************************

Robust and Efficient Transfer Learning with Hidden Parameter Markov Decision Processes
Taylor W. Killian, Samuel Daulton, George Konidaris, Finale Doshi-Velez
We introduce a new formulation of the Hidden Parameter Markov Decision Process (HiP-MDP), a framework for modeling families of related tasks using low-dimensional latent embeddings.  Our new framework correctly models the joint uncertainty in the latent parameters and the state space.  We also replace the original Gaussian Process-based model with a Bayesian Neural Network, enabling more scalable inference.  Thus, we expand the scope of the HiP-MDP to applications with higher dimensions and more complex dynamics.
************************************

Multi-View Decision Processes: The Helper-AI Problem
Christos Dimitrakakis, David C. Parkes, Goran Radanovic, Paul Tylkin
We consider a  two-player sequential game in which agents have the same reward function but may disagree on the transition probabilities of an underlying Markovian model of the world. By committing to play a specific policy, the agent with the correct model can steer the behavior of the other agent, and seek to improve utility. We model this setting as a multi-view decision process, which we use to formally analyze the positive effect of steering policies. Furthermore, we develop an algorithm for computing the agents' achievable joint policy, and we experimentally show that it can lead to a large utility increase when the agents' models diverge.
************************************

Maximum Margin Interval Trees
Alexandre Drouin, Toby Hocking, Francois Laviolette
Learning a regression function using censored or interval-valued output data is an important problem in fields such as genomics and medicine. The goal is to learn a real-valued prediction function, and the training output labels indicate an interval of possible values. Whereas most existing algorithms for this task are linear models, in this paper we investigate learning nonlinear tree models. We propose to learn a tree by minimizing a margin-based discriminative objective function, and we provide a dynamic programming algorithm for computing the optimal solution in log-linear time. We show empirically that this algorithm achieves state-of-the-art speed and prediction accuracy in a benchmark of several data sets.
************************************

Online Learning with a Hint
Ofer Dekel, arthur flajolet, Nika Haghtalab, Patrick Jaillet
We study a variant of online linear optimization where the player receives a hint about the loss function at the beginning of each round. The hint is given in the form of a vector that is weakly correlated with the loss vector on that round. We show that the player can benefit from such a hint if the set of feasible actions is sufficiently round. Specifically, if the set is strongly convex, the hint can be used to guarantee a regret of O(log(T)), and if the set is q-uniformly convex for q\in(2,3), the hint can be used to guarantee a regret of o(sqrt{T}). In contrast, we establish Omega(sqrt{T}) lower bounds on regret when the set of feasible actions is a polyhedron.
************************************

DPSCREEN: Dynamic Personalized Screening
Kartik Ahuja, William Zame, Mihaela van der Schaar

Screening is important for the diagnosis and treatment of a wide variety of diseases. A good screening policy should be personalized to the disease, to the features of the patient and to the dynamic history of the patient (including the history of screening). The growth of electronic health records data has led to the development of many models to predict the onset and progression of different diseases. However, there has been limited work to address the personalized screening for these different diseases. In this work, we develop the first framework to construct screening policies for a large class of disease models. The disease is modeled as a finite state stochastic process with an absorbing disease state. The patient observes an external information process (for instance, self-examinations, discovering comorbidities, etc.) which can trigger the patient to arrive at the clinician earlier than scheduled screenings. The clinician carries out the tests; based on the test results and the external information it schedules the next arrival. Computing the exactly optimal screening policy that balances the delay in the detection against the frequency of screenings is computationally intractable; this paper provides a computationally tractable construction of an approximately optimal policy. As an illustration, we make use of a large breast cancer data set. The constructed policy screens patients more or less often according to their initial risk -- it is personalized to the features of the patient -- and according to the results of previous screens – it is personalized to the history of the patient. In comparison with existing clinical policies, the constructed policy leads to large reductions (28-68 %) in the number of screens performed while achieving the same expected delays in disease detection.

************************************

Online Learning of Optimal Bidding Strategy in Repeated Multi-Commodity Auctions

M. Sevi Baltaoglu, Lang Tong, Qing Zhao

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.

************************************

A-NICE-MC: Adversarial Training for MCMC

Jiaming Song, Shengjia Zhao, Stefano Ermon

Existing Markov Chain Monte Carlo (MCMC) methods are either based on general-purpose and domain-agnostic schemes, which can lead to slow convergence, or require hand-crafting of problem-specific proposals by an expert. We propose A-NICE-MC, a novel method to train flexible parametric Markov chain kernels to produce samples with desired properties. First, we propose an efficient likelihood-free adversarial training method to train a Markov chain and mimic a given data distribution. Then, we leverage flexible volume preserving flows to obtain parametric kernels for MCMC. Using a bootstrap approach, we show how to train efficient Markov Chains to sample from a prescribed posterior distribution by iteratively improving the quality of both the model and the samples. A-NICE-MC provides the first framework to automatically design efficient domain-specific MCMC proposals. Empirical results demonstrate that A-NICE-MC combines the strong guarantees of MCMC with the expressiveness of deep neural networks, and is able to significantly outperform competing methods such as Hamiltonian Monte Carlo.

************************************

Question Asking as Program Generation

Anselm Rothe, Brenden M. Lake, Todd Gureckis

A hallmark of human intelligence is the ability to ask rich, creative, and revealing questions. Here we introduce a cognitive model capable of constructing human-like questions. Our approach treats questions as formal programs that, when executed on the state of the world, output an answer. The model specifies a probability distribution over a complex, compositional space of programs, favoring concise programs that help the agent learn in the current context. We evaluate our approach by modeling the types of open-ended questions generated by humans who were attempting to learn about an ambiguous situation in a game. We find that our model predicts what questions people will ask, and can creatively produce novel questions that were not present in the training set. In addition, we compare a number of model variants, finding that both question informativeness and complexity are important for producing human-like questions.

************************************

Gradient Methods for Submodular Maximization

Hamed Hassani, Mahdi Soltanolkotabi, Amin Karbasi

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.

************************************

Recycling Privileged Learning and Distribution Matching for Fairness

Novi Quadrianto, Viktoriia Sharmanska

Equipping machine learning models with ethical and legal constraints is a serious issue; without this, the future of machine learning is at risk. This paper takes a step forward in this direction and focuses on ensuring machine learning models deliver fair decisions. In legal scholarships, the notion of fairness itself is evolving and multi-faceted. We set an overarching goal to develop a unified machine learning framework that is able to handle any definitions of fairness, their combinations, and also new definitions that might be stipulated in the future. To achieve our goal, we recycle two well-established machine learning techniques, privileged learning and distribution matching, and harmonize them for satisfying multi-faceted fairness definitions. We consider protected characteristics such as race and gender as privileged information that is available at training but not at test time; this accelerates model training and delivers fairness through unawareness. Further, we cast demographic parity, equalized odds, and equality of opportunity as a classical two-sample problem of conditional distributions, which can be solved in a general form by using distance measures in Hilbert Space. We show several existing models are special cases of ours. Finally, we advocate returning the Pareto frontier of multi-objective minimization of error and unfairness in predictions. This will facilitate decision makers to select an operating point and to be accountable for it.

************************************

Collecting Telemetry Data Privately

Bolin Ding, Janardhan Kulkarni, Sergey Yekhanin

The collection and analysis of telemetry data from user's devices is routinely performed by many software companies. Telemetry collection leads to improved user experience but poses significant risks to users' privacy. Locally differentially private (LDP) algorithms have recently emerged as the main tool that allows data collectors to estimate various population statistics, while preserving privacy. The guarantees provided by such algorithms are typically very strong for a single round of telemetry collection, but degrade rapidly when telemetry is collected regularly. In particular, existing LDP algorithms are not suitable for repeated collection of counter data such as daily app usage statistics. In this paper, we develop new LDP mechanisms geared towards repeated collection of counter data, with formal privacy guarantees even after being executed for an arbitrarily long period of time. For two basic analytical tasks, mean estimation and histogram estimation, our LDP mechanisms for repeated data collection provide estimates with comparable or even the same accuracy as existing single-round LDP collection mechanisms. We conduct empirical evaluation on real-world counter datasets to verify our theoretical results. Our mechanisms have been deployed by Microsoft to collect telemetry across millions of devices.

************************************

Parallel Streaming Wasserstein Barycenters

Matthew Staib, Sebastian Claici, Justin M. Solomon, Stefanie Jegelka

Efficiently aggregating data from different sources is a challenging problem, particularly when samples from each source are distributed differently. These differences can be inherent to the inference task or present for other reasons: sensors in a sensor network may be placed far apart, affecting their individual measurements. Conversely, it is computationally advantageous to split Bayesian inference tasks across subsets of data, but data need not be identically distributed across subsets. One principled way to fuse probability distributions is via the lens of optimal transport: the Wasserstein barycenter is a single distribution that summarizes a collection of input measures while respecting their geometry. However, computing the barycenter scales poorly and requires discretization of all input distributions and the barycenter itself. Improving on this situation, we present a scalable, communication-efficient, parallel algorithm for computing the Wasserstein barycenter of arbitrary distributions. Our algorithm can operate directly on continuous input distributions and is optimized for streaming data. Our method is even robust to nonstationary input distributions and produces a barycenter estimate that tracks the input measures over time. The algorithm is semi-discrete, needing to discretize only the barycenter estimate. To the best of our knowledge, we also provide the first bounds on the quality of the approximate barycenter as the discretization becomes finer. Finally, we demonstrate the practical effectiveness of our method, both in tracking moving distributions on a sphere, as well as in a large-scale Bayesian inference task.

************************************

Adaptive Accelerated Gradient Converging Method under H\"{o}lderian Error Bound Condition

Mingrui Liu, Tianbao Yang

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.

************************************

What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?

Alex Kendall, Yarin Gal

There are two major types of uncertainty one can model. Aleatoric uncertainty captures noise inherent in the observations. On the other hand, epistemic uncertainty accounts for uncertainty in the model - uncertainty which can be explained away given enough data. Traditionally it has been difficult to model epistemic uncertainty in computer vision, but with new Bayesian deep learning tools this is now possible. We study the benefits of modeling epistemic vs. aleatoric uncertainty in Bayesian deep learning models for vision tasks. For this we present a Bayesian deep learning framework combining input-dependent aleatoric uncertainty together with epistemic uncertainty. We study models under the framework with per-pixel semantic segmentation and depth regression tasks. Further, our explicit uncertainty formulation leads to new loss functions for these tasks, which can be interpreted as learned attenuation. This makes the loss more robust to noisy data, also giving new state-of-the-art results on segmentation and depth regression benchmarks.

************************************

Reconstruct & Crush Network

Erinc Merdivan, Mohammad Reza Loghmani, Matthieu Geist

This article introduces an energy-based model that is adversarial regarding data: it minimizes the energy for a given data distribution (the positive samples) while maximizing the energy for another given data distribution (the negative or unlabeled samples). The model is especially instantiated with autoencoders where the energy, represented by the reconstruction error, provides a general distance measure for unknown data. The resulting neural network thus learns to reconstruct data from the first distribution while crushing data from the second distribution. This solution can handle different problems such as Positive and Unlabeled (PU) learning or covariate shift, especially with imbalanced data. Using autoencoders allows handling a large variety of data, such as images, text or even dialogues. Our experiments show the flexibility of the proposed approach in dealing with different types of data in different settings: images with CIFAR-10 and CIFAR-100 (not-in-training setting), text with Amazon reviews (PU learning) and dialogues with Facebook bAbI (next response classification and dialogue completion).

************************************

Permutation-based Causal Inference Algorithms with Interventions

Yuhao Wang, Liam Solus, Karren Yang, Caroline Uhler

Learning directed acyclic graphs using both observational and interventional data is now a fundamentally important problem due to recent technological developments in genomics that generate such single-cell gene expression data at a very large scale. In order to utilize this data for learning gene regulatory networks, efficient and reliable causal inference algorithms are needed that can make use of both observational and interventional data. In this paper, we present two algorithms of this type and prove that both are consistent under the faithfulness assumption. These algorithms are interventional adaptations of the Greedy SP algorithm and are the first algorithms using both observational and interventional data with consistency guarantees. Moreover, these algorithms have the advantage that they are nonparametric, which makes them useful also for analyzing non-Gaussian data. In this paper, we present these two algorithms and their consistency guarantees, and we analyze their performance on simulated data, protein signaling data, and single-cell gene expression data.

************************************

Deep Dynamic Poisson Factorization Model

Chengyue Gong, win-bin huang

A new model, named as deep dynamic poisson factorization model, is proposed in this paper for analyzing sequential count vectors. The model based on the Poisson Factor Analysis method captures dependence among time steps by neural networks, representing the implicit distributions. Local complicated relationship is obtained from local implicit distribution, and deep latent structure is exploited to get the long-time dependence. Variational inference on latent variables and gradient descent based on the loss functions derived from variational distribution is performed in our inference. Synthetic datasets and real-world datasets are applied to the proposed model and our results show good predicting and fitting performance with interpretable latent structure.

************************************

Scalable Generalized Linear Bandits: Online Computation and Hashing

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, Rebecca Willett

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.

************************************

Experimental Design for Learning Causal Graphs with Latent Variables

Murat Kocaoglu, Karthikeyan Shanmugam, Elias Bareinboim

We consider the problem of learning causal structures with latent variables using interventions. Our objective is not only to learn the causal graph between the observed variables, but to locate unobserved variables that could confound the relationship between observables. Our approach is stage-wise: We first learn the observable graph, i.e., the induced graph between observable variables. Next we learn the existence and location of the latent variables given the observable graph. We propose an efficient randomized algorithm that can learn the observable graph using O(d\log^2 n) interventions where d is the degree of the graph. We further propose an efficient deterministic variant which uses O(log n + l) interventions, where l is the longest directed path in the graph. Next, we propose an algorithm that uses only O(d^2 log n) interventions that can learn the latents between both non-adjacent and adjacent variables. While a naive baseline approach would require O(n^2) interventions, our combined algorithm can learn the causal graph with latents using O(d log^2 n + d^2 log (n)) interventions.

************************************

Lower bounds on the robustness to adversarial perturbations

Jonathan Peck, Joris Roels, Bart Goossens, Yvan Saeys

The input-output mappings learned by state-of-the-art neural networks are significantly discontinuous. It is possible to cause a neural network used for image recognition to misclassify its input by applying very specific, hardly perceptible perturbations to the input, called adversarial perturbations. Many hypotheses have been proposed to explain the existence of these peculiar samples as well as several methods to mitigate them. A proven explanation remains elusive, however. In this work, we take steps towards a formal characterization of adversarial perturbations by deriving lower bounds on the magnitudes of perturbations necessary to change the classification of neural networks. The bounds are experimentally verified on the MNIST and CIFAR-10 data sets.

************************************

Reliable Decision Support using Counterfactual Models
Peter Schulam, Suchi Saria

Decision-makers are faced with the challenge of estimating what is likely to happen when they take an action. For instance, if I choose not to treat this patient, are they likely to die? Practitioners commonly use supervised learning algorithms to fit predictive models that help decision-makers reason about likely future outcomes, but we show that this approach is unreliable, and sometimes even dangerous. The key issue is that supervised learning algorithms are highly sensitive to the policy used to choose actions in the training data, which causes the model to capture relationships that do not generalize. We propose using a different learning objective that predicts counterfactuals instead of predicting outcomes under an existing action policy as in supervised learning. To support decision-making in temporal settings, we introduce the Counterfactual Gaussian Process (CGP) to predict the counterfactual future progression of continuous-time trajectories under sequences of future actions. We demonstrate the benefits of the CGP on two important decision-support tasks: risk prediction and “what if?” reasoning for individualized treatment planning.

************************************

Group Additive Structure Identification for Kernel Nonparametric Regression
Chao Pan, Michael Zhu

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.

************************************

A multi-agent reinforcement learning model of common-pool resource appropriation
Julien Pérolat, Joel Z. Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, Thore Graepel

Humanity faces numerous problems of common-pool resource appropriation. This class of multi-agent social dilemma includes the problems of ensuring sustainable use of fresh water, common fisheries, grazing pastures, and irrigation systems. Abstract models of common-pool resource appropriation based on non-cooperative game theory predict that self-interested agents will generally fail to find socially positive equilibria---a phenomenon called the tragedy of the commons. However, in reality, human societies are sometimes able to discover and implement stable cooperative solutions. Decades of behavioral game theory research have sought to uncover aspects of human behavior that make this possible. Most of that work was based on laboratory experiments where participants only make a single choice: how much to appropriate. Recognizing the importance of spatial and temporal resource dynamics, a recent trend has been toward experiments in more complex real-time video game-like environments. However, standard methods of non-cooperative game theory can no longer be used to generate predictions for this case. Here we show that deep reinforcement learning can be used instead. To that end, we study the emergent behavior of groups of independently learning agents in a partially observed Markov game modeling common-pool resource  appropriation. Our experiments highlight the importance of trial-and-error learning in common-pool resource appropriation and shed light on the relationship between exclusion, sustainability, and inequality.

************************************

Decoding with Value Networks for Neural Machine Translation
Di He, Hanqing Lu, Yingce Xia, Tao Qin, Liwei Wang, Tie-Yan Liu

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-auth ors prior to requesting a name change in the electronic proceedings.

************************************

Population Matching Discrepancy and Applications in Deep Learning
Jianfei Chen, Chongxuan LI, Yizhong Ru, Jun Zhu

A differentiable estimation of the distance between two distributions based on samples is important for many deep learning tasks. One such estimation  is maximum mean discrepancy (MMD). However, MMD suffers from its sensitive kernel bandwidth hyper-parameter, weak gradients, and large mini-batch size when used as a training objective. In this paper, we propose population matching discrepancy (PMD) for estimating the distribution distance based on samples, as well as an algorithm to learn the parameters of the distributions using PMD as an objective. PMD is defined as the minimum weight matching of sample populations from each distribution, and we prove that PMD is a strongly consistent estimator of the first Wasserstein metric. We apply PMD to two deep learning tasks, domain adaptation and generative modeling. Empirical results demonstrate that PMD overcomes the aforementioned drawbacks of MMD, and outperforms MMD on both tasks in terms of the performance as well as the convergence speed.

************************************

Predictive State Recurrent Neural Networks
Carlton Downey, Ahmed Hefny, Byron Boots, Geoffrey J. Gordon, Boyue Li

We present a new model, Predictive State Recurrent Neural Networks (PSRNNs), for filtering and prediction in dynamical systems. PSRNNs draw on insights from both Recurrent Neural Networks (RNNs) and Predictive State Representations (PSRs), and inherit advantages from both types of models. Like many successful RNN architectures, PSRNNs use (potentially deeply composed) bilinear transfer functions to combine information from multiple sources. We show that such bilinear functions arise naturally from state updates in Bayes filters like PSRs, in which observations can be viewed as gating belief states. We also show that PSRNNs can be learned effectively by combining Backpropogation Through Time (BPTT) with an initialization  derived from a statistically consistent learning algorithm for PSRs called two-stage regression (2SR). Finally, we show that PSRNNs can be  factorized using tensor decomposition, reducing model size and suggesting interesting connections to existing multiplicative architectures such as LSTMs and GRUs. We apply PSRNNs to 4 datasets, and show that we outperform several popular alternative approaches to modeling dynamical systems in all cases.

************************************

Robust Hypothesis Test for Nonlinear Effect with Gaussian Processes
Jeremiah Liu, Brent Coull

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.

************************************

Sharpness, Restart and Acceleration
Vincent Roulet, Alexandre d'Aspremont

The {\L}ojasiewicz inequality shows that H\"olderian error bounds on the minimum of convex optimization problems hold almost generically. Here, we clarify results of \citet{Nemi85} who show that H\"olderian error bounds directly controls the performance of restart schemes. The constants quantifying error bounds are of course unobservable, but we show that optimal restart strategies are robust, and searching for the best scheme only increases the complexity by a logarithmic factor compared to the optimal bound. Overall then, restart schemes generically accelerate accelerated methods.

************************************

Dynamic Routing Between Capsules
Sara Sabour, Nicholas Frosst, Geoffrey E. Hinton

A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or object part. We use the length of the activity vector to represent the probability that the entity exists and its orientation to represent the instantiation parameters. Active capsules at one level make predictions, via transformation matrices, for the instantiation parameters of higher-level capsules.  When multiple predictions agree, a higher level capsule becomes active. We show that a discrimininatively trained, multi-layer capsule system achieves state-of-the-art performance on MNIST and is considerably better than a convolutional net at recognizing highly overlapping digits. To achieve these results we use an iterative routing-by-agreement mechanism: A lower-level capsule prefers to send its output to higher level capsules whose activity vectors have a big scalar product with the prediction coming from the lower-level capsule.

************************************

InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations
Yunzhu Li, Jiaming Song, Stefano Ermon

The goal of imitation learning is to mimic expert behavior without access to an explicit reward signal. Expert demonstrations provided by humans, however, often show significant variability due to latent factors that are typically not explicitly modeled. In this paper, we propose a new algorithm that can infer the latent structure of expert demonstrations in an unsupervised way. Our method, built on top of Generative Adversarial Imitation Learning, can not only imitate complex behaviors, but also learn interpretable and meaningful representations of complex behavioral data, including visual demonstrations. In the driving domain, we show that a model learned from human demonstrations is able to both accurately reproduce a variety of behaviors and accurately anticipate human actions using raw visual inputs. Compared with various baselines, our method can better capture the latent structure underlying expert demonstrations, often recovering semantically meaningful factors of variation in the data.

************************************

A Regularized Framework for Sparse and Structured Neural Attention
Vlad Niculae, Mathieu Blondel

Modern neural networks are often augmented with an attention mechanism, which tells the network where to focus within the input.  We propose in this paper a new framework for sparse and structured attention, building upon a smoothed max operator. We show that the gradient of this operator defines a mapping from real values to probabilities, suitable as an attention mechanism. Our framework includes softmax and a slight generalization of the recently-proposed sparsemax as special cases. However, we also show how our framework can incorporate modern structured penalties, resulting in more interpretable attention mechanisms, that focus on entire segments or groups of an input.  We derive efficient algorithms to compute the forward and backward passes of our attention mechanisms, enabling their use in a neural network trained with backpropagation.  To showcase their potential as a drop-in replacement for existing ones, we evaluate our attention mechanisms on three large-scale tasks: textual entailment, machine translation, and sentence summarization.  Our attention mechanisms improve interpretability without sacrificing performance; notably, on textual entailment and summarization, we outperform the standard attention mechanisms based on softmax and sparsemax.

************************************

Style Transfer from Non-Parallel Text by Cross-Alignment
Tianxiao Shen, Tao Lei, Regina Barzilay, Tommi Jaakkola

This paper focuses on style transfer on the basis of non-parallel text. This is an instance of a broad family of problems including machine translation, decipherment, and sentiment modification. The key challenge is to separate the content from other aspects such as style. We assume a shared latent content distribution across different text corpora, and propose a method that leverages refined alignment of latent representations to perform style transfer. The transferred sentences from one style should match example sentences from the other style as a population. We demonstrate the effectiveness of this cross-alignment method on three tasks: sentiment modification, decipherment of word substitution ciphers, and recovery of word order.

************************************

Unsupervised Learning of Disentangled Representations from Video
Emily L. Denton, vighnesh Birodkar

We present a new model DRNET that learns disentangled image representations from video. Our approach leverages the temporal coherence of video and a novel adversarial loss to learn a representation that factorizes each frame into a stationary part and a temporally varying  component. The disentangled representation can be used for a range of tasks. For example, applying a standard LSTM to the time-vary components enables prediction of future frames. We evaluating our approach on a range of synthetic and real videos. For the latter, we demonstrate the ability to coherently generate up to several hundred steps into the future.
************************************

Countering Feedback Delays in Multi-Agent Learning
Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter W. Glynn, Claire Tomlin

We consider a model of game-theoretic learning based on online mirror descent (OMD) with asynchronous and delayed feedback information. Instead of focusing on specific games, we consider a broad class of continuous games defined by the general equilibrium stability notion, which we call λ-variational stability. Our first contribution is that, in this class of games, the actual sequence of play induced by OMD-based learning converges to Nash equilibria provided that the feedback delays faced by the players are synchronous and bounded. Subsequently, to tackle fully decentralized, asynchronous environments with (possibly) unbounded delays between actions and feedback, we propose a variant of OMD which we call delayed mirror descent (DMD), and which relies on the repeated leveraging of past information. With this modification, the algorithm converges to Nash equilibria with no feedback synchronicity assumptions and even when the delays grow superlinearly relative to the horizon of play.
************************************

Affinity Clustering: Hierarchical Clustering at Scale
Mohammadhossein Bateni, Soheil Behnezhad, Mahsa Derakhshan, MohammadTaghi Hajiaghayi, Raimondas Kiveris, Silvio Lattanzi, Vahab Mirrokni

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.
************************************

Geometric Matrix Completion with Recurrent Multi-Graph Neural Networks
Federico Monti, Michael Bronstein, Xavier Bresson

Matrix completion models are among the most common formulations of recommender systems. Recent works have showed a boost of performance of these techniques when introducing the pairwise relationships between users/items in the form of graphs, and imposing smoothness priors on these graphs. However, such techniques do not fully exploit the local stationary structures on user/item graphs, and the number of parameters to learn is linear w.r.t. the number of users and items. We propose a novel approach to overcome these limitations by using geometric deep learning on graphs. Our matrix completion architecture combines a novel multi-graph convolutional neural network that can learn meaningful statistical graph-structured patterns from users and items, and a recurrent neural network that applies a learnable diffusion on the score matrix. Our neural network system is computationally attractive as it requires a constant number of parameters independent of the matrix size. We apply our method on several standard datasets, showing that it outperforms state-of-the-art matrix completion techniques.
************************************

Maximizing Subset Accuracy with Recurrent Neural Networks in Multi-label Classification
Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J. Kim, Johannes Fürnkranz

Multi-label classification is the task of predicting a set of labels for a given input instance. Classifier chains are a state-of-the-art method for tackling such problems, which essentially converts this problem into a sequential prediction problem, where the labels are first ordered in an arbitrary fashion, and the task is to predict a sequence of binary values for these labels. In this paper, we replace classifier chains with recurrent neural networks, a sequence-to-sequen ce prediction algorithm which has recently been successfully applied to sequential prediction tasks in many domains. The key advantage of this approach is that it allows to focus on the prediction of the positive labels only, a much smaller set than the full set of possible labels. Moreover, parameter sharing across all classifiers allows to better exploit information of previous decisions. As both, classifier chains and recurrent neural networks depend on a fixed ordering of the labels, which is typically not part of a multi-label problem specification, we also compare different ways of ordering the label set, and give some recommendations on suitable ordering strategies.

************************************

f-GANs in an Information Geometric Nutshell

Richard Nock, Zac Cranko, Aditya K. Menon, Lizhen Qu, Robert C. Williamson

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.

************************************

Active Bias: Training More Accurate Neural Networks by Emphasizing High Variance Samples

Haw-Shiuan Chang, Erik Learned-Miller, Andrew McCallum

Self-paced learning and hard example mining re-weight training instances to improve learning accuracy. This paper presents two improved alternatives based on lightweight estimates of sample uncertainty in stochastic gradient descent (SGD): the variance in predicted probability of the correct class across iterations of mini-batch SGD, and the proximity of the correct class probability to the decision threshold. Extensive experimental results on six datasets show that our methods reliably improve accuracy in various network architectures, including additional gains on top of other popular training techniques, such as residual learning, momentum, ADAM, batch normalization, dropout, and distillation.

************************************

SchNet: A continuous-filter convolutional neural network for modeling quantum interactions

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, Klaus-Robert Müller

Deep learning has the potential to revolutionize quantum chemistry as it is ideally suited to learn representations for structured data and speed up the exploration of chemical space. While convolutional neural networks have proven to be the first choice for images, audio and video data, the atoms in molecules are not restricted to a grid. Instead, their precise locations contain essential physical information, that would get lost if discretized. Thus, we propose to use continuous-filter convolutional layers to be able to model local correlations without requiring the data to lie on a grid. We apply those layers in SchNet: a novel deep learning architecture modeling quantum interactions in molecules. We obtain a joint model for the total energy and interatomic forces that follows fundamental quantum-chemical principles. Our architecture achieves state-of-the-art performance for benchmarks of equilibrium molecules and molecular dynamics trajectories. Finally, we introduce a more challenging benchmark with chemical and structural variations that suggests the path for further work.

************************************

GibbsNet: Iterative Adversarial Inference for Deep Graphical Models

Alex M. Lamb, Devon Hjelm, Yaroslav Ganin, Joseph Paul Cohen, Aaron C. Courville, Yoshua Bengio

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.

************************************

Bayesian GAN

Yunus Saatci, Andrew G. Wilson

Generative adversarial networks (GANs) can implicitly learn rich distributions o ver images, audio, and data which are hard to model with an explicit likelihood. We present a practical Bayesian formulation for unsupervised and semi-supervised learning with GANs. Within this framework, we use stochastic gradient Hamiltonian Monte Carlo to marginalize the weights of the generator and discriminator networks. The resulting approach is straightforward and obtains good performance without any standard interventions such as feature matching or mini-batch discrimination. By exploring an expressive posterior over the parameters of the generator, the Bayesian GAN avoids mode-collapse, produces interpretable and diverse candidate samples, and provides state-of-the-art quantitative results for semi-supervised learning on benchmarks including SVHN, CelebA, and CIFAR-10, outperforming DCGAN, Wasserstein GANs, and DCGAN ensembles.

************************************

Alternating minimization for dictionary learning with random initialization
Niladri Chatterji, Peter L. Bartlett

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.

************************************

Sparse Embedded $k$-Means Clustering
Weiwei Liu, Xiaobo Shen, Ivor Tsang

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.

************************************

Reducing Reparameterization Gradient Variance
Andrew Miller, Nick Foti, Alexander D'Amour, Ryan P. Adams

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.

************************************

Min-Max Propagation
Christopher Srinivasa, Inmar Givoni, Siamak Ravanbakhsh, Brendan J. Frey

We study the application of min-max propagation, a variation of belief propagation, for approximate min-max inference in factor graphs. We show that for “any” high-order function that can be minimized in O(ω), the min-max message update can be obtained using an efficient O(K(ω + log(K)) procedure, where K is the number of variables. We demonstrate how this generic procedure, in combination with efficient updates for a family of high-order constraints, enables the application of min-max propagation to efficiently approximate the NP-hard problem of makespan minimization, which seeks to distribute a set of tasks on machines, such that the worst case load is minimized.

************************************

Statistical Cost Sharing
Eric Balkanski, Umar Syed, Sergei Vassilvitskii

We study the cost sharing problem for cooperative games in situations where the cost function C is not available via oracle queries, but must instead be learned from samples drawn from a distribution, represented as tuples (S, C(S)), for different subsets S of players. We formalize this approach, which we call statistical cost sharing, and consider the computation of the core and the Shapley value. Expanding on the work by Balcan et al, we give precise sample complexity bounds for computing cost shares that satisfy the core property with high probability for any function with a non-empty core. For the Shapley value, which has never been studied in this setting, we show that for submodular cost functions with curvature bounded curvature kappa it can be approximated from samples from the uniform distribution to a sqrt{1 - kappa} factor, and that the bound is tight. We then define statistical analogues of the Shapley axioms, and derive a notion of statistical Shapley value and that these can be approximated arbitrarily well f rom samples from any distribution and for any function.

************************************

### Dilated Recurrent Neural Networks

Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark A. Hasegawa-Johnson, Thomas S. Huang

Learning with recurrent neural networks (RNNs) on long sequences is a notoriously difficult task.  There are three major challenges: 1) complex dependencies, 2) vanishing and exploding gradients, and 3) efficient parallelization. In this paper, we introduce a simple yet effective RNN connection structure, the DilatedRNN, which simultaneously tackles all of these challenges.  The proposed architecture is characterized by multi-resolution dilated recurrent skip connections and can be combined flexibly with diverse RNN cells.  Moreover, the DilatedRNN reduces the number of parameters needed and enhances training efficiency significantly, while matching state-of-the-art performance (even with standard RNN cells) in tasks involving very long-term dependencies.  To provide a theory-based quantification of the architecture's advantages, we introduce a memory capacity measure, the mean recurrent length, which is more suitable for RNNs with long skip connections than existing measures.  We rigorously prove the advantages of the DilatedRNN over other recurrent neural architectures.  The code for our method is publicly available at https://github.com/code-terminator/DilatedRNN.

************************************

### The Expressive Power of Neural Networks: A View from the Width

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, Liwei Wang

The expressive power of neural networks is important for understanding deep learning. Most existing works consider this problem from the view of the depth of a network. In this paper, we study how width affects the expressiveness of neural networks. Classical results state that depth-bounded (e.g. depth-2) networks with suitable activation functions are universal approximators. We show a universal approximation theorem for width-bounded ReLU networks: width-(n + 4) ReLU networks, where n is the input dimension, are universal approximators. Moreover, except for a measure zero set, all functions cannot be approximated by width-n ReLU networks, which exhibits a phase transition. Several recent works demonstrate the benefits of depth by proving the depth-efficiency of neural networks. That is, there are classes of deep networks which cannot be realized by any shallow network whose size is no more than an exponential bound. Here we pose the dual question on the width-efficiency of ReLU networks: Are there wide networks that cannot be realized by narrow networks whose size is not substantially larger? We show that there exist classes of wide networks which cannot be realized by any narrow network whose depth is no more than a polynomial bound. On the other hand, we demonstrate by extensive experiments that narrow networks whose size exceed the polynomial bound by a constant factor can approximate wide and shallow network with high accuracy. Our results provide more comprehensive evidence that depth may be more effective than width for the expressiveness of ReLU networks.

************************************

### Inverse Reward Design

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J. Russell, Anca Dragan

Autonomous agents optimize the reward function we give them. What they don't know is how hard it is for us to design a reward function that actually captures what we  want. When designing the reward, we might think of some specific training scenarios, and make sure that the reward will lead to the right behavior in those scenarios. Inevitably, agents encounter new scenarios (e.g., new types of terrain) where optimizing that same reward may lead to undesired behavior. Our insight is that reward functions are merely observations about what the designer actually wants, and that they should be interpreted in the context in which they were designed. We introduce inverse reward design (IRD) as the problem of inferring the true objective based on the designed reward and the training MDP. We introduce approximate methods for solving IRD problems, and use their solution to plan risk-averse behavior in test MDPs. Empirical results suggest that this approach can help alleviate negative side effects of misspecified reward functions and mitigate reward hacking.

************************************

The power of absolute discounting: all-dimensional distribution estimation

Moein Falahatgar, Mesrob I. Ohannessian, Alon Orlitsky, Venkatadheeraj Pichapati

Categorical models are a natural fit for many problems. When learning the distribution of categories from samples, high-dimensionality may dilute the data. Minimax optimality is too pessimistic to remedy this issue. A serendipitously discovered estimator, absolute discounting, corrects empirical frequencies by subtracting a constant from observed categories, which it then redistributes among the unobserved. It outperforms classical estimators empirically, and has been used extensively in natural language modeling. In this paper, we rigorously explain the prowess of this estimator using less pessimistic notions. We show  that (1) absolute discounting recovers classical minimax KL-risk rates, (2) it is \emph{adaptive} to an effective dimension rather than the true dimension, (3) it is strongly related to the Good-Turing estimator and inherits its \emph{competitive} properties. We use power-law distributions as the cornerstone of these results. We validate the theory via synthetic data and an application to the Global Terrorism Database.

************************************

A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, Thore Graepel

There has been a resurgence of interest in multiagent reinforcement learning (MARL), due partly to the recent success of deep neural networks. The simplest form of MARL is independent reinforcement learning (InRL), where each agent treats all of its experience as part of its (non stationary) environment. In this paper, we first observe that policies learned using InRL can overfit to the other agents' policies during training, failing to sufficiently generalize during execution. We introduce a new metric, joint-policy correlation, to quantify this effect. We describe a meta-algorithm for general MARL, based on approximate best responses to mixtures of policies generated using deep reinforcement learning, and empirical game theoretic analysis to compute meta-strategies for policy selection. The meta-algorithm generalizes previous algorithms such as InRL, iterated best response, double oracle, and fictitious play. Then, we propose a scalable implementation which reduces the memory requirement using decoupled meta-solvers. Finally, we demonstrate the generality of the resulting policies in three partially observable settings: gridworld coordination problems, emergent language games, and poker.

************************************

Spectral Mixture Kernels for Multi-Output Gaussian Processes

Gabriel Parra, Felipe Tobar

Early approaches to multiple-output Gaussian processes (MOGPs) relied on linear combinations of independent, latent, single-output Gaussian processes (GPs). This resulted in cross-covariance functions with limited parametric interpretation, thus conflicting with the ability  of single-output GPs to understand lengthscales, frequencies and magnitudes to name a few. On the contrary, current approaches to MOGP are able to better interpret the relationship between different channels by directly modelling the cross-covariances as a spectral mixture kernel with a phase shift. We extend this rationale and propose a parametric family of complex-valued cross-spectral densities and then build on Cramér's Theorem (the multivariate version of Bochner's Theorem) to provide a principled approach to design multivariate covariance functions. The so-constructed kernels are able to model delays among channels in addition to phase differences and are thus more expressive than previous methods, while also providing full parametric interpretation of the relationship across channels. The proposed method is first validated on synthetic data and then compared to existing MOGP methods on two real-world examples.

************************************

Affine-Invariant Online Optimization and the Low-rank Experts Problem

Tomer Koren, Roi Livni

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.

************************************

Pose Guided Person Image Generation

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, Luc Van Gool

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.

************************************

Successor Features for Transfer in Reinforcement Learning

Andre Barreto, Will Dabney, Remi Munos, Jonathan J. Hunt, Tom Schaul, Hado P. van Hasselt, David Silver

Transfer in reinforcement learning refers to the notion that generalization should occur not only within a task but also across tasks. We propose a transfer framework for the scenario where the reward function changes between tasks but the environment's dynamics remain the same. Our approach rests on two key ideas: "successor features", a value function representation that decouples the dynamics of the environment from the rewards, and "generalized policy improvement", a generalization of dynamic programming's policy improvement operation that considers a set of policies rather than a single one. Put together, the two ideas lead to an approach that integrates seamlessly within the reinforcement learning framework and allows the free exchange of information across tasks. The proposed method also provides performance guarantees for the transferred policy even before any learning has taken place. We derive two theorems that set our approach in firm theoretical ground and present experiments that show that it successfully promotes transfer in practice, significantly outperforming alternative methods in a sequence of navigation tasks and in the control of a simulated robotic arm.

************************************

On Quadratic Convergence of DC Proximal Newton Algorithm in Nonconvex Sparse Learning

Xingguo Li, Lin Yang, Jason Ge, Jarvis Haupt, Tong Zhang, Tuo Zhao

We propose a DC proximal Newton algorithm for solving nonconvex regularized sparse learning problems in high dimensions. Our proposed algorithm integrates the proximal newton algorithm with multi-stage convex relaxation based on the difference of convex (DC) programming,  and enjoys both strong computational and statistical guarantees. Specifically, by leveraging a sophisticated characterization of sparse modeling structures (i.e., local restricted strong convexity and Hessian smoothness), we prove that within each stage of convex relaxation, our proposed algorithm achieves (local) quadratic convergence, and eventually obtains a sparse approximate local optimum with optimal statistical properties after only a few convex relaxations. Numerical experiments are provided to support our theory.

************************************

Hypothesis Transfer Learning via Transformation Functions

Simon S. Du, Jayanth Koushik, Aarti Singh, Barnabas Poczos

We consider the Hypothesis Transfer Learning (HTL) problem where one incorporates a hypothesis trained on the source domain into the learning procedure of the target domain. Existing theoretical analysis either only studies specific algorithms or only presents upper bounds on the generalization error but not on the excess risk. In this paper, we propose a unified algorithm-dependent framework for HTL through a novel notion of transformation functions, which characterizes the relation between the source and the target domains. We conduct a general risk analysis of this framework and in particular, we show for the first time, if two domains are related, HTL enjoys faster convergence rates of excess risks for Kernel Smoothing and Kernel Ridge Regression than those of the classical non-transfer learning settings. We accompany this framework with an analysis of cross-validation for HTL to search for the best transfer technique and gracefully reduce to non-transfer learning when HTL is not helpful. Experiments on robotics and neur al imaging data demonstrate the effectiveness of our framework.

************************************

Finite Sample Analysis of the GTD Policy Evaluation Algorithms in Markov Setting

Yue Wang, Wei Chen, Yuting Liu, Zhi-Ming Ma, Tie-Yan Liu

In reinforcement learning (RL), one of the key components is policy evaluation, which aims to estimate the value function (i.e., expected long-term accumulated reward) of a policy. With a good policy evaluation method, the RL algorithms will estimate the value function more accurately and find a better policy. When the state space is large or continuous \emph{Gradient-based Temporal Difference(GTD)} policy evaluation algorithms with linear function approximation are widely used. Considering that the collection of the evaluation data is both time and reward consuming, a clear understanding of the finite sample performance of the policy evaluation algorithms is very important to reinforcement learning. Under the assumption that data are i.i.d. generated, previous work provided the finite sample analysis of the GTD algorithms with constant step size by converting  them into  convex-concave saddle point problems. However, it is well-known that, the data are generated from Markov processes rather than i.i.d in RL problems.. In this paper, in the realistic Markov setting, we derive the finite sample bounds for the general convex-concave saddle point problems, and hence for the GTD algorithms. We have the following discussions based on our bounds. (1) With variants of step size, GTD algorithms converge. (2) The convergence rate is determined by the step size, with the mixing time of the Markov process as the coefficient. The faster the Markov processes mix, the faster the convergence. (3) We explain that the experience replay trick is effective by improving the mixing property of the Markov process.  To the best of our knowledge, our analysis is the first to provide finite sample bounds for the GTD algorithms in Markov setting.

************************************

Variational Inference via $\chi$ Upper Bound Minimization

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, David Blei

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.

************************************

A Probabilistic Framework for Nonlinearities in Stochastic Neural Networks

Qinliang Su, xuejun Liao, Lawrence Carin

We present a probabilistic framework for nonlinearities, based on doubly truncated Gaussian distributions. By setting the truncation points appropriately, we are able to generate various types of nonlinearities within a unified framework, including sigmoid, tanh and ReLU, the most commonly used nonlinearities in neural networks. The framework readily integrates into existing stochastic neural networks (with hidden units characterized as random variables), allowing one for the first time to learn the nonlinearities alongside model weights in these networks. Extensive experiments demonstrate the performance improvements brought about by the proposed framework when integrated with the restricted Boltzmann machine (RBM), temporal RBM and the truncated Gaussian graphical model (TGGM).

************************************

Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation

Yuhuai Wu, Elman Mansimov, Roger B. Grosse, Shun Liao, Jimmy Ba

In this work, we propose to apply trust region optimization to deep reinforcement learning using a recently proposed Kronecker-factored approximation to the curvature. We extend the framework of natural policy gradient and propose to optimize both the actor and the critic using Kronecker-factored approximate curvature (K-FAC) with trust region; hence we call our method Actor Critic using Kronecker-Factored Trust Region (ACKTR). To the best of our knowledge, this is the first scalable trust region natural gradient method for actor-critic methods. It is also the method that learns non-trivial tasks in continuous control as well as discrete control policies directly from raw pixel inputs. We tested our approach across discrete domains in Atari games as well as continuous domains in the MuJoCo environment. With the proposed methods, we are able to achieve higher rewards and a 2- to 3-fold improvement in sample efficiency on average, compared to previous state-of-the-art on-policy actor-critic methods. Code is available at https://github.com/openai/baselines

************************************

Optimistic posterior sampling for reinforcement learning: worst-case regret bounds

Shipra Agrawal, Randy Jia

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.

************************************

Efficient Second-Order Online Kernel Learning with Adaptive Embedding

Daniele Calandriello, Alessandro Lazaric, Michal Valko

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.

************************************

Solving Most Systems of Random Quadratic Equations

Gang Wang, Georgios Giannakis, Yousef Saad, Jie Chen

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.

************************************

Online Reinforcement Learning in Stochastic Games

Chen-Yu Wei, Yi-Te Hong, Chi-Jen Lu

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.

************************************

Independence clustering (without a matrix)

Daniil Ryabko

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.

************************************

Effective Parallelisation for Machine Learning

Michael Kamp, Mario Boley, Olana Missura, Thomas Gärtner

We present a novel parallelisation scheme that simplifies the adaptation of learning algorithms to growing amounts of data as well as growing needs for accurate and confident predictions in critical applications. In contrast to other parallelisation techniques, it can be applied to a broad class of learning algorithms without further mathematical derivations and without writing dedicated code, while at the same time maintaining theoretical performance guarantees. Moreover, our parallelisation scheme is able to reduce the runtime of many learning algorithms to polylogarithmic time on quasi-polynomially many processing units. This is a significant step towards a general answer to an open question on efficient parallelisation of machine learning algorithms in the sense of Nick's Class (NC). The cost of this parallelisation is in the form of a larger sample complexity. Our empirical study confirms the potential of our parallelisation scheme with fixed numbers of processors and instances in realistic application scenarios.

************************************

Deep Mean-Shift Priors for Image Restoration

Siavash Arjomand Bigdeli, Matthias Zwicker, Paolo Favaro, Meiguang Jin

In this paper we introduce a natural image prior that directly represents a Gaus sian-smoothed version of the natural image distribution. We include our prior in a formulation of image restoration as a Bayes estimator that also allows us to solve noise-blind image restoration problems. We show that the gradient of our prior corresponds to the mean-shift vector on the natural image distribution. In addition, we learn the mean-shift vector field using denoising autoencoders, and use it in a gradient descent approach to perform Bayes risk minimization. We demonstrate competitive results for noise-blind deblurring, super-resolution, and demosaicing.

************************************

On Structured Prediction Theory with Calibrated Convex Surrogate Losses
Anton Osokin, Francis Bach, Simon Lacoste-Julien
We provide novel theoretical insights on structured prediction in the context of efficient convex surrogate loss minimization with consistency guarantees. For any task loss, we construct a convex surrogate that can be optimized via stochastic gradient descent and we prove tight bounds on the so-called "calibration function" relating the excess surrogate risk to the actual risk. In contrast to prior related work, we carefully monitor the effect of the exponential number of classes in the learning guarantees as well as on the optimization complexity. As an interesting consequence, we formalize the intuition that some task losses make learning harder than others, and that the classical 0-1 loss is ill-suited for structured prediction.

************************************

Invariance and Stability of Deep Convolutional Representations
Alberto Bietti, Julien Mairal
In this paper, we study deep signal representations that are near-invariant to groups of transformations and stable to the action of diffeomorphisms without losing signal information. This is achieved by generalizing the multilayer kernel introduced in the context of convolutional kernel networks and by studying the geometry of the corresponding reproducing kernel Hilbert space. We show that the signal representation is stable, and that models from this functional space, such as a large class of convolutional neural networks, may enjoy the same stability.

************************************

Variational Memory Addressing in Generative Models
Jörg Bornschein, Andriy Mnih, Daniel Zoran, Danilo Jimenez Rezende
Aiming to augment generative models with external memory, we interpret the output of a memory module with stochastic addressing as a conditional mixture distribution, where a read operation corresponds to sampling a discrete memory address and retrieving the corresponding content from memory. This perspective allows us to apply variational inference to memory addressing, which enables effective training of the memory module by using the target information to guide memory lookups. Stochastic addressing is particularly well-suited for generative models as it naturally encourages multimodality which is a prominent aspect of most high-dimensional datasets. Treating the chosen address as a latent variable also allows us to quantify the amount of information gained with a memory lookup and measure the contribution of the memory module to the generative process. To illustrate the advantages of this approach we incorporate it into a variational autoencoder and apply the resulting model to the task of generative few-shot learning. The intuition behind this architecture is that the memory module can pick a relevant template from memory and the continuous part of the model can concentrate on modeling remaining variations. We demonstrate empirically that our model is able to identify and access the relevant memory contents even with hundreds of unseen Omniglot characters in memory.

************************************

Shallow Updates for Deep Reinforcement Learning
Nir Levine, Tom Zahavy, Daniel J. Mankowitz, Aviv Tamar, Shie Mannor
Deep reinforcement learning (DRL) methods such as the Deep Q-Network (DQN) have achieved state-of-the-art results in a variety of challenging, high-dimensional domains. This success is mainly attributed to the power of deep neural networks to learn rich domain representations for approximating the value function or pol

icy. Batch reinforcement learning methods with linear representations, on the other hand, are more stable and require less hyper parameter tuning. Yet, substantial feature engineering is necessary to achieve good results. In this work we propose a hybrid approach -- the Least Squares Deep Q-Network (LS-DQN), which combines rich feature representations learned by a DRL algorithm with the stability of a linear least squares method. We do this by periodically re-training the last hidden layer of a DRL network with a batch least squares update. Key to our approach is a Bayesian regularization term for the least squares update, which prevents over-fitting to the more recent data. We tested LS-DQN on five Atari games and demonstrate significant improvement over vanilla DQN and Double-DQN. We also investigated the reasons for the superior performance of our method. Interestingly, we found that the performance improvement can be attributed to the large batch size used by the LS method when optimizing the last layer.

************************************

Learning with Bandit Feedback in Potential Games

Amélie Heliou, Johanne Cohen, Panayotis Mertikopoulos

This paper examines the equilibrium convergence properties of no-regret learning with exponential weights in potential games. To establish convergence with minimal information requirements on the players' side, we focus on two frameworks: the semi-bandit case (where players have access to a noisy estimate of their payoff vectors, including strategies they did not play), and the bandit case (where players are only able to observe their in-game, realized payoffs). In the semi-bandit case, we show that the induced sequence of play converges almost surely to a Nash equilibrium at a quasi-exponential rate. In the bandit case, the same result holds for approximate Nash equilibria if we introduce a constant exploration factor that guarantees that action choice probabilities never become arbitrarily small. In particular, if the algorithm is run with a suitably decreasing exploration factor, the sequence of play converges to a bona fide Nash equilibrium with probability 1.

************************************

A Greedy Approach for Budgeted Maximum Inner Product Search

Hsiang-Fu Yu, Cho-Jui Hsieh, Qi Lei, Inderjit S. Dhillon

Maximum Inner Product Search (MIPS) is an important task in many machine learning applications such as the prediction phase of low-rank matrix factorization models and deep learning models. Recently, there has been substantial research on how to perform MIPS in sub-linear time, but most of the existing work does not have the flexibility to control the trade-off between search efficiency and search quality. In this paper, we study the important problem of MIPS with a computational budget. By carefully studying the problem structure of MIPS, we develop a novel Greedy-MIPS algorithm, which can handle budgeted MIPS by design. While simple and intuitive, Greedy-MIPS yields surprisingly superior performance compared to state-of-the-art approaches. As a specific example, on a candidate set containing half a million vectors of dimension 200, Greedy-MIPS runs 200x faster than the naive approach while yielding search results with the top-5 precision greater than 75%.

************************************

Riemannian approach to batch normalization

Minhyung Cho, Jaehyung Lee

Batch normalization (BN) has proven to be an effective algorithm for deep neural network training by normalizing the input to each neuron and reducing the internal covariate shift. The space of weight vectors in the BN layer can be naturally interpreted as a Riemannian manifold, which is invariant to linear scaling of weights. Following the intrinsic geometry of this manifold provides a new learning rule that is more efficient and easier to analyze. We also propose intuitive and effective gradient clipping and regularization methods for the proposed algorithm by utilizing the geometry of the manifold. The resulting algorithm consistently outperforms the original BN on various types of network architectures and datasets.

************************************

Adaptive Clustering through Semidefinite Programming

Martin Royer

We analyze the clustering problem through a flexible probabilistic model that aims to identify an optimal partition on the sample X1,...,Xn. We perform exact clustering with high probability using a convex semidefinite estimator that interprets as a corrected, relaxed version of K-means. The estimator is analyzed through a non-asymptotic framework and showed to be optimal or near-optimal in recovering the partition. Furthermore, its performances are shown to be adaptive to the problem’s effective dimension, as well as to K the unknown number of groups in this partition. We illustrate the method’s performances in comparison to other classical clustering algorithms with numerical experiments on simulated high-dimensional data.

************************************

#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, Pieter Abbeel

Count-based exploration algorithms are known to perform near-optimally when used in conjunction with tabular reinforcement learning (RL) methods for solving small discrete Markov decision processes (MDPs). It is generally thought that count-based methods cannot be applied in high-dimensional state spaces, since most states will only occur once. Recent deep RL exploration strategies are able to deal with high-dimensional continuous state spaces through complex heuristics, often relying on optimism in the face of uncertainty or intrinsic motivation. In this work, we describe a surprising finding: a simple generalization of the classic count-based approach can reach near state-of-the-art performance on various high-dimensional and/or continuous deep RL benchmarks. States are mapped to hash codes, which allows to count their occurrences with a hash table. These counts are then used to compute a reward bonus according to the classic count-based exploration theory. We find that simple hash functions can achieve surprisingly good results on many challenging tasks. Furthermore, we show that a domain-dependent learned hash code may further improve these results. Detailed analysis reveals important aspects of a good hash function: 1) having appropriate granularity and 2) encoding information relevant to solving the MDP. This exploration strategy achieves near state-of-the-art performance on both continuous control tasks and Atari 2600 games, hence providing a simple yet powerful baseline for solving MDPs that require considerable exploration.

************************************

Learning Koopman Invariant Subspaces for Dynamic Mode Decomposition

Naoya Takeishi, Yoshinobu Kawahara, Takehisa Yairi

Spectral decomposition of the Koopman operator is attracting attention as a tool for the analysis of nonlinear dynamical systems. Dynamic mode decomposition is a popular numerical algorithm for Koopman spectral analysis; however, we often need to prepare nonlinear observables manually according to the underlying dynamics, which is not always possible since we may not have any a priori knowledge about them. In this paper, we propose a fully data-driven method for Koopman spectral analysis based on the principle of learning Koopman invariant subspaces from observed data. To this end, we propose minimization of the residual sum of squares of linear least-squares regression to estimate a set of functions that transforms data into a form in which the linear regression fits well. We introduce an implementation with neural networks and evaluate performance empirically using nonlinear dynamical systems and applications.

************************************

Online Prediction with Selfish Experts

Tim Roughgarden, Okke Schrijvers

We consider the problem of binary prediction with expert advice in settings where experts have agency and seek to maximize their credibility. This paper makes three main contributions.  First, it defines a model to reason formally about settings with selfish experts, and demonstrates that  ``incentive compatible'' (IC) algorithms are closely related to the design of proper scoring rules.   Second, we design IC algorithms with good performance guarantees for the absolute loss function. Third, we give a formal separation between the power of online predict ion with selfish experts and online prediction with honest experts by proving lower bounds for both IC and non-IC algorithms.  In particular, with selfish experts and the absolute loss function, there is no (randomized) algorithm for online prediction---IC or otherwise---with asymptotically vanishing regret.

************************************

Streaming Robust Submodular Maximization: A Partitioned Thresholding Approach
Slobodan Mitrovic, Ilija Bogunovic, Ashkan Norouzi-Fard, Jakub M. Tarnawski, Volkan Cevher

We study the classical problem of maximizing a monotone submodular function subject to a cardinality constraint k, with two additional twists: (i) elements arrive in a streaming fashion, and (ii) m items from the algorithm’s memory are removed after the stream is finished. We develop a robust submodular algorithm STAR-T. It is based on a novel partitioning structure and an exponentially decreasing thresholding rule. STAR-T makes one pass over the data and retains a short but robust summary. We show that after the removal of any m elements from the obtained summary, a simple greedy algorithm STAR-T-GREEDY that runs on the remaining elements achieves a constant-factor approximation guarantee. In two different data summarization tasks, we demonstrate that it matches or outperforms existing greedy and streaming methods, even if they are allowed the benefit of knowing the removed subset in advance.

************************************

Neural Program Meta-Induction
Jacob Devlin, Rudy R. Bunel, Rishabh Singh, Matthew Hausknecht, Pushmeet Kohli

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.

************************************

The Scaling Limit of High-Dimensional Online Independent Component Analysis
Chuang Wang, Yue Lu

We analyze the dynamics of an online algorithm for independent component analysis in the high-dimensional scaling limit. As the ambient dimension tends to infinity, and with proper time scaling, we show that the time-varying joint empirical measure of the target feature vector and the estimates provided by the algorithm will converge weakly to a deterministic measured-valued process that can be characterized as the unique solution of a nonlinear PDE. Numerical solutions of this PDE, which involves two spatial variables and one time variable, can be efficiently obtained. These solutions provide detailed information about the performance of the ICA algorithm, as many practical performance metrics are functionals of the joint empirical measures. Numerical simulations show that our asymptotic analysis is accurate even for moderate dimensions. In addition to providing a tool for understanding the performance of the algorithm, our PDE analysis also provides useful insight. In particular, in the high-dimensional limit, the original coupled dynamics associated with the algorithm will be asymptotically “decoupled”, with each coordinate independently solving a 1-D effective minimization problem via stochastic gradient descent. Exploiting this insight to design new algorithms for achieving optimal trade-offs between computational and statistical efficiency may prove an interesting line of future research.

************************************

Practical Locally Private Heavy Hitters
Raef Bassily, Kobbi Nissim, Uri Stemmer, Abhradeep Guha Thakurta

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.

************************************

Mixture-Rank Matrix Approximation for Collaborative Filtering
Dongsheng Li, Chao Chen, Wei Liu, Tun Lu, Ning Gu, Stephen Chu

Low-rank matrix approximation (LRMA) methods have achieved excellent accuracy among today's collaborative filtering (CF) methods. In existing LRMA methods, the rank of user/item feature matrices is typically fixed, i.e., the same rank is adopted to describe all users/items. However, our studies show that submatrices with different ranks could coexist in the same user-item rating matrix, so that approximations with fixed ranks cannot perfectly describe the internal structures of the rating matrix, therefore leading to inferior recommendation accuracy. In this paper, a mixture-rank matrix approximation (MRMA) method is proposed, in which user-item ratings can be characterized by a mixture of LRMA models with different ranks. Meanwhile, a learning algorithm capitalizing on iterated condition modes is proposed to tackle the non-convex optimization problem pertaining to MRMA. Experimental studies on MovieLens and Netflix datasets demonstrate that MRMA can outperform six state-of-the-art LRMA-based CF methods in terms of recommendation accuracy.

************************************

Higher-Order Total Variation Classes on Grids: Minimax Theory and Trend Filtering Methods

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James L. Sharpnack, Ryan J. Tibshirani

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.

************************************

Robust Conditional Probabilities

Yoav Wald, Amir Globerson

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.

************************************

Attention is All you Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ■ukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent orconvolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attentionm echanisms.  We propose a novel, simple network architecture based solely onan attention mechanism, dispensing with recurrence and convolutions entirely.Experiments on two machine translation tasks show these models to be superiorin quality while being more parallelizable and requiring significantly less timeto train. Our single model with 165 million parameters, achieves 27.5 BLEU onEnglish-to-German translation, improving over the existing best ensemble result by over 1 BLEU. On English-to-French translation, we outperform the previoussingle state-of-the-art with model by 0.7 BLEU, achieving a BLEU score of 41.1.

************************************

A General Framework for Robust Interactive Learning

Ehsan Emamjomeh-Zadeh, David Kempe

We propose a general framework for interactively learning models, such as (binary or non-binary) classifiers, orderings/rankings of items, or clusterings of data points. Our framework is based on a generalization of Angluin's equivalence query model and Littlestone's online learning model: in each iteration, the algorithm proposes a model, and the user either accepts it or reveals a specific mistake in the proposal. The feedback is correct only with probability p > 1/2 (and adversarially incorrect with probability 1 - p), i.e., the algorithm must be able to learn in the presence of arbitrary noise. The algorithm's goal is to learn the ground truth model using few iterations.  Our general framework is based on a graph representation of the models and user feedback. To be able to learn efficiently, it is sufficient that there be a graph G whose nodes are the models, and (weighted) edges capture the user feedback, with the property that if s, s* are the proposed and target models, respectively, then any (correct) user feedback s' must lie on a shortest s-s* path in G. Under this one assumption, there is a natural algorithm, reminiscent of the Multiplicative Weights Update algorithm, w hich will efficiently learn s* even in the presence of noise in the user's feedback.  From this general result, we rederive with barely any extra effort classic results on learning of classifiers and a recent result on interactive clustering; in addition, we easily obtain new interactive learning algorithms for ordering/ranking.

************************************

Sample and Computationally Efficient Learning Algorithms under S-Concave Distributions

Maria-Florina F. Balcan, Hongyang Zhang

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.

************************************

Net-Trim: Convex Pruning of Deep Neural Networks with Performance Guarantee

Alireza Aghasi, Afshin Abdi, Nam Nguyen, Justin Romberg

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.

************************************

ELF: An Extensive, Lightweight and Flexible Research Platform for Real-time Strategy Games

Yuandong Tian, Qucheng Gong, Wenling Shang, Yuxin Wu, C. Lawrence Zitnick

In this paper, we propose ELF, an Extensive, Lightweight and Flexible platform for fundamental reinforcement learning research. Using ELF, we implement a highly customizable real-time strategy (RTS) engine with three game environments (Mini-RTS, Capture the Flag and Tower Defense). Mini-RTS, as a miniature version of StarCraft, captures key game dynamics and runs at 165K frame-per-second (FPS) on a laptop. When coupled with modern reinforcement learning methods, the system can train a full-game bot against built-in AIs end-to-end in one day with 6 CPUs and 1 GPU. In addition, our platform is flexible in terms of environment-agent communication topologies, choices of RL methods, changes in game parameters, and can host existing C/C++-based game environments like ALE. Using ELF, we thoroughly explore training parameters and show that a network with Leaky ReLU and Batch Normalization coupled with long-horizon training and progressive curriculum beats the rule-based built-in AI more than 70% of the time in the full game of Mini-RTS. Strong performance is also achieved on the other two games. In game replays, we show our agents learn interesting strategies. ELF, along with its RL platform, is open-sourced at https://github.com/facebookresearch/ELF.

************************************

Task-based End-to-end Model Learning in Stochastic Optimization

Priya Donti, Brandon Amos, J. Zico Kolter

With the increasing popularity of machine learning techniques, it has become common to see prediction algorithms operating within some larger process. However, the criteria by which we train these algorithms often differ from the ultimate criteria on which we evaluate them. This paper proposes an end-to-end approach for learning probabilistic machine learning models in a manner that directly captures the ultimate task-based objective for which they will be used, within the context of stochastic programming. We present three experimental evaluations of the proposed approach: a classical inventory stock problem, a real-world electrical grid scheduling task, and a real-world energy storage arbitrage task. We show that the proposed approach can outperform both traditional modeling and purely black-box policy optimization approaches in these applications.

************************************

Fader Networks:Manipulating Images by Sliding Attributes

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, Marc'Aurelio Ranzato

This paper introduces a new encoder-decoder architecture that is trained to reconstruct images by disentangling the salient information of the image and the val ues of attributes directly in the latent space. As a result, after training, our model can generate different realistic versions of an input image by varying the attribute values. By using continuous attribute values, we can choose how much a specific attribute is perceivable in the generated image. This property could allow for applications where users can modify an image using sliding knobs, like faders on a mixing console, to change the facial expression of a portrait, or to update the color of some objects. Compared to the state-of-the-art which mostly relies on training adversarial networks in pixel space by altering attribute values at train time, our approach results in much simpler training schemes and nicely scales to multiple attributes. We present evidence that our model can significantly change the perceived value of the attributes while preserving the naturalness of images.

************************************

VAE Learning via Stein Variational Gradient Descent

Yuchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, Lawrence Carin

A new method for learning variational autoencoders (VAEs) is developed, based on Stein variational gradient descent. A key advantage of this approach is that one need not make parametric assumptions about the form of the encoder distribution. Performance is further enhanced by integrating the proposed encoder with importance sampling. Excellent performance is demonstrated across multiple unsupervised and semi-supervised problems, including semi-supervised analysis of the ImageNet data, demonstrating the scalability of the model to large datasets.

************************************

Approximation and Convergence Properties of Generative Adversarial Learning

Shuang Liu, Olivier Bousquet, Kamalika Chaudhuri

Generative adversarial networks (GAN) approximate a target data distribution by jointly optimizing an objective function through a "two-player game" between a generator and a discriminator.  Despite their empirical success, however, two very basic questions on how well they can approximate the target distribution remain unanswered. First, it is not known how restricting the discriminator family affects the approximation quality. Second, while a number of different objective functions have been proposed, we do not understand when convergence to the global minima of the objective function leads to convergence to the target distribution under various notions of distributional convergence.   In this paper, we address these questions in a broad and unified setting by defining a notion of adversarial divergences that includes a number of recently proposed objective functions. We show that if the objective function is an adversarial divergence with some additional conditions, then using a restricted discriminator family has a moment-matching effect. Additionally, we show that for objective functions that are strict adversarial divergences, convergence in the objective function implies weak convergence, thus generalizing previous results.

************************************

VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, Charles Sutton

Deep generative models provide powerful tools for distributions over complicated manifolds, such as those of natural images.  But many of these methods, including generative adversarial networks (GANs), can be difficult to train, in part because they are prone to mode collapse, which means that they characterize only a few modes of the true distribution.  To address this, we introduce VEEGAN, which features a reconstructor network, reversing the action of the generator by mapping from data to noise.  Our training objective retains the original asymptotic consistency guarantee of GANs, and can be interpreted as a novel autoencoder loss over the noise.  In sharp contrast to a traditional autoencoder over data points, VEEGAN does not require specifying a loss function over the data, but rather only over the representations, which are standard normal by assumption.  On an extensive set of synthetic and real world image datasets, VEEGAN indeed resists mode collapsing to a far greater extent than other recent GAN variants, and produces more realistic samples.

************************************

Local Aggregative Games

Vikas Garg, Tommi Jaakkola

Aggregative games provide a rich abstraction to model strategic multi-agent interactions. We focus on learning local aggregative games, where the payoff of each player is a function of its own action and the aggregate behavior of its neighbors in a connected digraph. We show the existence of a pure strategy epsilon-Nash equilibrium in such games when the payoff functions are convex or sub-modular. We prove an information theoretic lower bound, in a value oracle model, on approximating the structure of the digraph with non-negative monotone sub-modular cost functions on the edge set cardinality. We also introduce gamma-aggregative games that generalize local aggregative games, and admit epsilon-Nash equilibrium that are stable with respect to small changes in some specified graph property. Moreover, we provide estimation algorithms for the game theoretic model that can meaningfully recover the underlying structure and payoff functions from real voting data.

***********************************

An Error Detection and Correction Framework for Connectomics

Jonathan Zung, Ignacio Tartavull, Kisuk Lee, H. Sebastian Seung

We define and study error detection and correction tasks that are useful for 3D reconstruction of neurons from electron microscopic imagery, and for image segmentation more generally. Both tasks take as input the raw image and a binary mask representing a candidate object. For the error detection task, the desired output is a map of split and merge errors in the object. For the error correction task, the desired output is the true object. We call this object mask pruning, because the candidate object mask is assumed to be a superset of the true object. We train multiscale 3D convolutional networks to perform both tasks. We find that the error-detecting net can achieve high accuracy. The accuracy of the error-correcting net is enhanced if its input object mask is ``advice'' (union of erroneous objects) from the error-detecting net.

***********************************

Hindsight Experience Replay

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, Wojciech Zaremba

Dealing with sparse rewards is one of the biggest challenges in Reinforcement Learning (RL). We present a novel technique called Hindsight Experience Replay which allows sample-efficient learning from rewards which are sparse and binary and therefore avoid the need for complicated reward engineering. It can be combined with an arbitrary off-policy RL algorithm and may be seen as a form of implicit curriculum. We demonstrate our approach on the task of manipulating objects with a robotic arm. In particular, we run experiments on three different tasks: pushing, sliding, and pick-and-place, in each case using only binary rewards indicating whether or not the task is completed. Our ablation studies show that Hindsight Experience Replay is a crucial ingredient which makes training possible in these challenging environments. We show that our policies trained on a physics simulation can be deployed on a physical robot and successfully complete the task. The video presenting our experiments is available at https://goo.gl/SMrQnI.

***********************************

Fixed-Rank Approximation of a Positive-Semidefinite Matrix from Streaming Data

Joel A. Tropp, Alp Yurtsever, Madeleine Udell, Volkan Cevher

Several important applications, such as streaming PCA and semidefinite programming, involve a large-scale positive-semidefinite (psd) matrix that is presented as a sequence of linear updates.  Because of storage limitations, it may only be possible to retain a sketch of the psd matrix. This paper develops a new algorithm for fixed-rank psd approximation from a sketch. The approach combines the Nyström approximation with a novel mechanism for rank truncation. Theoretical analysis establishes that the proposed method can achieve any prescribed relative error in the Schatten 1-norm and that it exploits the spectral decay of the input matrix.  Computer experiments show that the proposed method dominates alternative techniques for fixed-rank psd matrix approximation across a wide range of examples.

************************************

The Numerics of GANs

Lars Mescheder, Sebastian Nowozin, Andreas Geiger

In this paper, we analyze the numerics of common algorithms for training Generative Adversarial Networks (GANs). Using the formalism of smooth two-player games we analyze the associated gradient vector field of GAN training objectives. Our findings suggest that the convergence of current algorithms suffers due to two factors: i) presence of eigenvalues of the Jacobian of the gradient vector field with zero real-part, and ii) eigenvalues with big imaginary part. Using these findings, we design a new algorithm that overcomes some of these limitations and has better convergence properties. Experimentally, we demonstrate its superiority on training common GAN architectures and show convergence on GAN architectures that are known to be notoriously hard to train.

************************************

Cortical microcircuits as gated-recurrent neural networks

Rui Costa, Ioannis Alexandros Assael, Brendan Shillingford, Nando de Freitas, TIm Vogels

Cortical circuits exhibit intricate recurrent architectures that are remarkably similar across different brain areas. Such stereotyped structure suggests the existence of common computational principles. However, such principles have remained largely elusive. Inspired by gated-memory networks, namely long short-term memory networks (LSTMs), we introduce a recurrent neural network in which information is gated through inhibitory cells that are subtractive (subLSTM).  We propose a natural mapping of subLSTMs onto known canonical excitatory-inhibitory cortical microcircuits.   Our empirical evaluation across sequential image classification and language modelling tasks shows that subLSTM units can achieve similar performance to LSTM units. These results suggest that cortical circuits can be optimised to solve complex contextual problems and proposes a novel view on their computational function. Overall our work provides a step towards unifying recurrent networks as used in machine learning with their biological counterparts.

************************************

Deep Lattice Networks and Partial Monotonic Functions

Seungil You, David Ding, Kevin Canini, Jan Pfeifer, Maya Gupta

We propose learning deep models that are monotonic with respect to a user-specified set of inputs by alternating layers of linear embeddings, ensembles of lattices, and calibrators (piecewise linear functions), with appropriate constraints for monotonicity, and jointly training the resulting network. We implement the layers and projections with new computational graph nodes in TensorFlow and use the Adam optimizer and batched stochastic gradients. Experiments on benchmark and real-world datasets show that six-layer monotonic deep lattice networks achieve state-of-the art performance for classification and regression with monotonicity guarantees.

************************************

Zap Q-Learning

Adithya M Devraj, Sean Meyn

The  Zap Q-learning algorithm introduced in this paper is an improvement of Watkins' original algorithm and recent competitors in several respects. It is a matrix-gain algorithm designed so that its asymptotic variance is optimal. Moreover, an ODE analysis suggests that the transient behavior is a close match to a deterministic Newton-Raphson implementation. This is made possible by a  two time-scale update equation for the matrix gain sequence. The analysis suggests that the approach will lead to stable and efficient computation even for non-ideal parameterized settings. Numerical experiments confirm the quick convergence, even in such non-ideal cases.

************************************

Contrastive Learning for Image Captioning

Bo Dai, Dahua Lin

Image captioning, a popular topic in computer vision, has achieved substantial progress in recent years. However, the distinctiveness of natural descriptions is often overlooked in previous work. It is closely related to the quality of capt

ions, as distinctive captions are more likely to describe images with their unique aspects. In this work, we propose a new learning method, Contrastive Learning (CL), for image captioning. Specifically, via two constraints formulated on top of a reference model, the proposed method can encourage distinctiveness, while maintaining the overall quality of the generated captions. We tested our method on two challenging datasets, where it improves the baseline model by significant margins. We also showed in our studies that the proposed method is generic and can be used for models with various structures.
************************************
Variational Walkback: Learning a Transition Operator as a Stochastic Recurrent Net
Anirudh Goyal ALIAS PARTH GOYAL, Nan Rosemary Ke, Surya Ganguli, Yoshua Bengio
We propose a novel method to {\it directly} learn a stochastic transition operator whose repeated application provides generated samples. Traditional undirected graphical models approach this problem indirectly by learning a Markov chain model whose stationary distribution obeys detailed balance with respect to a parameterized energy function. The energy function is then modified so the model and data distributions match, with no guarantee on the number of steps required for the Markov chain to converge. Moreover, the detailed balance condition is highly restrictive: energy based models corresponding to neural networks must have symmetric weights, unlike biological neural circuits. In contrast, we develop a method for directly learning arbitrarily parameterized transition operators capable of expressing non-equilibrium stationary distributions that violate detailed balance, thereby enabling us to learn more biologically plausible asymmetric neural networks and more general non-energy based dynamical systems.   The proposed training objective, which we derive via principled variational methods, encourages the transition operator to "walk back" (prefer to revert its steps) in multi-step trajectories that start at data-points, as quickly as possible back to the original data points. We present a series of experimental results illustrating the soundness of the proposed approach, Variational Walkback (VW), on the MNIST, CIFAR-10, SVHN and CelebA datasets, demonstrating superior samples compared to earlier attempts to learn a transition operator. We also show that although each rapid training trajectory is limited to a finite but variable number of steps, our transition operator continues to generate good samples well past the length of such trajectories, thereby demonstrating the match of its non-equilibrium stationary distribution to the data distribution. Source Code:http://github.com/anirudh9119/walkback_nips17
************************************
Linear Time Computation of Moments in Sum-Product Networks
Han Zhao, Geoffrey J. Gordon
Bayesian online algorithms for Sum-Product Networks (SPNs) need to update their posterior distribution after seeing one single additional instance. To do so, they must compute moments of the model parameters under this distribution. The best existing method for computing such moments scales quadratically in the size of the SPN, although it scales linearly for trees. This unfortunate scaling makes Bayesian online algorithms prohibitively expensive, except for small or tree-structured SPNs. We propose an optimal linear-time algorithm that works even when the SPN is a general directed acyclic graph (DAG), which significantly broadens the applicability of Bayesian online algorithms for SPNs. There are three key ingredients in the design and analysis of our algorithm: 1). For each edge in the graph, we construct a linear time reduction from the moment computation problem to a joint inference problem in SPNs. 2). Using the property that each SPN computes a multilinear polynomial, we give an efficient procedure for polynomial evaluation by differentiation without expanding the network that may contain exponentially many monomials. 3). We propose a dynamic programming method to further reduce the computation of the moments of all the edges in the graph from quadratic to linear. We demonstrate the usefulness of our linear time algorithm by applying it to develop a linear time assume density filter (ADF) for SPNs.
************************************
SGD Learns the Conjugate Kernel Class of the Network

Amit Daniely

We show that the standard stochastic gradient decent (SGD) algorithm is guaranteed to learn, in polynomial time, a function that is competitive with the best function in the conjugate kernel space of the network, as defined in Daniely, Frostig and Singer. The result holds for log-depth networks from a rich family of architectures. To the best of our knowledge, it is the first polynomial-time guarantee for the standard neural network learning algorithm for networks of depth more that two. As corollaries, it follows that for neural networks of any depth between 2 and log(n), SGD is guaranteed to learn, in polynomial time, constant degree polynomials with polynomially bounded coefficients. Likewise, it follows that SGD on large enough networks can learn any continuous function (not in polynomial time), complementing classical expressivity results.

************************************

Learning to Pivot with Adversarial Networks

Gilles Louppe, Michael Kagan, Kyle Cranmer

Several techniques for domain adaptation have been proposed to account for differences in the distribution of the data used for training and testing. The majority of this work focuses on a binary domain label. Similar problems occur in a scientific context where there may be a continuous family of plausible data generation processes associated to the presence of systematic uncertainties. Robust inference is possible if it is based on a pivot -- a quantity whose distribution does not depend on the unknown values of the nuisance parameters that parametrize this family of data generation processes. In this work, we introduce and derive theoretical results for a training procedure based on adversarial networks for enforcing the pivotal property (or, equivalently, fairness with respect to continuous attributes) on a predictive model. The method includes a hyperparameter to control the trade-off between accuracy and robustness. We demonstrate the effectiveness of this approach with a toy example and examples from particle physics.

************************************

Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration

Jason Altschuler, Jonathan Niles-Weed, Philippe Rigollet

Computing optimal transport distances such as the earth mover's distance is a fundamental problem in machine learning, statistics, and computer vision. Despite the recent introduction of several algorithms with good empirical performance, it is unknown whether general optimal transport distances can be approximated in near-linear time. This paper demonstrates that this ambitious goal is in fact achieved by Cuturi's Sinkhorn Distances. This result relies on a new analysis of Sinkhorn iterations, which also directly suggests a new greedy coordinate descent algorithm Greenkhorn with the same theoretical guarantees. Numerical simulations illustrate that Greenkhorn significantly outperforms the classical Sinkhorn algorithm in practice.

************************************

Universal Style Transfer via Feature Transforms

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, Ming-Hsuan Yang

Universal style transfer aims to transfer arbitrary visual styles to content images. Existing feed-forward based methods, while enjoying the inference efficiency, are mainly limited by inability of generalizing to unseen styles or compromised visual quality. In this paper, we present a simple yet effective method that tackles these limitations without training on any pre-defined styles. The key ingredient of our method is a pair of feature transforms, whitening and coloring, that are embedded to an image reconstruction network. The whitening and coloring transforms reflect direct matching of feature covariance of the content image to a given style image, which shares similar spirits with the optimization of Gram matrix based cost in neural style transfer. We demonstrate the effectiveness of our algorithm by generating high-quality stylized images with comparisons to a number of recent methods. We also analyze our method by visualizing the whitened features and synthesizing textures by simple feature coloring.

************************************

Ensemble Sampling
Xiuyuan Lu, Benjamin Van Roy

Thompson sampling has emerged as an effective heuristic for a broad range of online decision problems. In its basic form, the algorithm requires computing and sampling from a posterior distribution over models, which is tractable only for simple special cases. This paper develops ensemble sampling, which aims to approximate Thompson sampling while maintaining tractability even in the face of complex models such as neural networks. Ensemble sampling dramatically expands on the range of applications for which Thompson sampling is viable. We establish a theoretical basis that supports the approach and present computational results that offer further insight.

************************************

Practical Data-Dependent Metric Compression with Provable Guarantees
Piotr Indyk, Ilya Razenshteyn, Tal Wagner

We introduce a new distance-preserving compact representation of multi-dimensional point-sets. Given n points in a d-dimensional space where each coordinate is represented using B bits (i.e., dB bits per point), it produces  a representation of size O( d log(d B/epsilon) +log n) bits per point from which one can approximate the distances up to a factor of 1 + epsilon. Our algorithm almost matches the recent bound of Indyk et al, 2017} while being much simpler. We compare our algorithm to Product Quantization (PQ) (Jegou et al, 2011) a state of the art heuristic metric compression method. We evaluate both algorithms on several data sets: SIFT, MNIST, New York City taxi time series and a synthetic one-dimensional data set embedded in a high-dimensional space. Our algorithm produces representations that are comparable to or better than those produced by PQ, while having provable guarantees on its performance.

************************************

Partial Hard Thresholding: Towards A Principled Analysis of Support Recovery
Jie Shen, Ping Li

In machine learning and compressed sensing, it is of central importance to understand when a tractable algorithm recovers the support of a sparse signal from its compressed measurements. In this paper, we present a principled analysis on the support recovery performance for a family of hard thresholding algorithms. To this end, we appeal to the partial hard thresholding (PHT) operator proposed recently by Jain et al. [IEEE Trans. Information Theory, 2017]. We show that under proper conditions, PHT recovers an arbitrary "s"-sparse signal within O(s kappa log kappa) iterations where "kappa" is an appropriate condition number. Specifying the PHT operator, we obtain the best known result for hard thresholding pursuit and orthogonal matching pursuit with replacement. Experiments on the simulated data complement our theoretical findings and also illustrate the effectiveness of PHT compared to other popular recovery methods.

************************************

Selective Classification for Deep Neural Networks
Yonatan Geifman, Ran El-Yaniv

Selective classification techniques (also known as reject option) have not yet been considered in the context of deep neural networks (DNNs). These techniques can potentially significantly improve DNNs prediction performance by trading-off coverage. In this paper we propose a method to construct a selective classifier given a trained neural network. Our method allows a user to set a desired risk level. At test time, the classifier rejects instances as needed, to grant the desired risk (with high probability). Empirical results over CIFAR and ImageNet convincingly demonstrate the viability of our method, which opens up possibilities to operate DNNs in mission-critical applications. For example, using our method an unprecedented 2% error in top-5 ImageNet classification can be guaranteed with probability 99.9%, with almost 60% test coverage.

************************************

Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space
Liwei Wang, Alexander Schwing, Svetlana Lazebnik

This paper explores image caption generation using conditional variational auto-

encoders (CVAEs). Standard CVAEs with a fixed Gaussian prior yield descriptions with too little variability. Instead, we propose two models that explicitly structure the latent space around K components corresponding to different types of image content, and combine components to create priors for images that contain multiple types of content simultaneously (e.g., several kinds of objects). Our first model uses a Gaussian Mixture model (GMM) prior, while the second one defines a novel Additive Gaussian (AG) prior that linearly combines component means. We show that both models produce captions that are more diverse and more accurate than a strong LSTM baseline or a “vanilla” CVAE with a fixed Gaussian prior, with AG-CVAE showing particular promise.
************************************

Deconvolutional Paragraph Representation Learning
Yizhe Zhang, Dinghan Shen, Guoyin Wang, Zhe Gan, Ricardo Henao, Lawrence Carin
Learning latent representations from long text sequences is an important first step in many natural language processing applications. Recurrent Neural Networks (RNNs) have become a cornerstone for this challenging task. However, the quality of sentences during RNN-based decoding (reconstruction) decreases with the length of the text. We propose a sequence-to-sequence, purely convolutional and deconvolutional autoencoding framework that is free of the above issue, while also being computationally efficient. The proposed method is simple, easy to implement and can be leveraged as a building block for many applications. We show empirically that compared to RNNs, our framework is better at reconstructing and correcting long paragraphs. Quantitative evaluation on semi-supervised text classification and summarization tasks demonstrate the potential for better utilization of long unlabeled text data.
************************************

Learning to See Physics via Visual De-animation
Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, Josh Tenenbaum
We introduce a paradigm for understanding physical scenes without human annotations. At the core of our system is a physical world representation that is first recovered by a perception module and then utilized by physics and graphics engines. During training, the perception module and the generative models learn by visual de-animation --- interpreting and reconstructing the visual information stream. During testing, the system first recovers the physical world state, and then uses the generative models for reasoning and future prediction.  Even more so than forward simulation, inverting a physics or graphics engine is a computationally hard problem; we overcome this challenge by using a convolutional inversion network. Our system quickly recognizes the physical world state from appearance and motion cues, and has the flexibility to incorporate both differentiable and non-differentiable physics and graphics engines. We evaluate our system on both synthetic and real datasets involving multiple physical scenes, and demonstrate that our system performs well on both physical state estimation and reasoning problems. We further show that the knowledge learned on the synthetic dataset generalizes to constrained real images.
************************************

Adversarial Symmetric Variational Autoencoder
Yuchen Pu, Weiyao Wang, Ricardo Henao, Liqun Chen, Zhe Gan, Chunyuan Li, Lawrence Carin
A new form of variational autoencoder (VAE) is developed, in which the joint distribution of data and codes is considered in two (symmetric) forms: (i) from observed data fed through the encoder to yield codes, and (ii) from latent codes drawn from a simple prior and propagated through the decoder to manifest data. Lower bounds are learned for marginal log-likelihood fits observed data and latent codes. When learning with the variational bound, one seeks to minimize the symmetric Kullback-Leibler divergence of joint density functions from (i) and (ii), while simultaneously seeking to maximize the two marginal log-likelihoods. To facilitate learning, a new form of adversarial training is developed. An extensive set of experiments is performed, in which we demonstrate state-of-the-art data reconstruction and generation on several image benchmarks datasets.
************************************

Model evidence from nonequilibrium simulations

Michael Habeck

The marginal likelihood, or model evidence, is a key quantity in Bayesian parameter estimation and model comparison. For many probabilistic models, computation of the marginal likelihood is challenging, because it involves a sum or integral over an enormous parameter space. Markov chain Monte Carlo (MCMC) is a powerful approach to compute marginal likelihoods. Various MCMC algorithms and evidence estimators have been proposed in the literature. Here we discuss the use of nonequilibrium techniques for estimating the marginal likelihood. Nonequilibrium estimators build on recent developments in statistical physics and are known as annealed importance sampling (AIS) and reverse AIS in probabilistic machine learning. We introduce estimators for the model evidence that combine forward and backward simulations and show for various challenging models that the evidence estimators outperform forward and reverse AIS.

************************************

Estimating High-dimensional Non-Gaussian Multiple Index Models via Stein’s Lemma

Zhuoran Yang, Krishnakumar Balasubramanian, Zhaoran Wang, Han Liu

We consider estimating the parametric components of semiparametric multi-index models in high dimensions. To bypass the requirements of Gaussianity or elliptical symmetry of covariates in existing methods, we propose to leverage a second-order Stein’s method with score function-based corrections. We prove that our estimator achieves a near-optimal statistical rate of convergence even when the score function or the response variable is heavy-tailed. To establish the key concentration results, we develop a data-driven truncation argument that may be of independent interest. We supplement our theoretical findings with simulations.

************************************

Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data

Stéphanie ALLASSONNIERE, Juliette Chevallier, Stephane Oudard

We introduce a hierarchical model which allows to estimate a group-average piecewise-geodesic trajectory in the Riemannian space of measurements and individual variability. This model falls into the well defined mixed-effect models. The subject-specific trajectories are defined through spatial and temporal transformations of the group-average piecewise-geodesic path, component by component. Thus we can apply our model to a wide variety of situations. Due to the non-linearity of the model, we use the Stochastic Approximation Expectation-Maximization algorithm to estimate the model parameters. Experiments on synthetic data validate this choice. The model is then applied to the metastatic renal cancer chemotherapy monitoring: we run estimations on RECIST scores of treated patients and estimate the time they escape from the treatment. Experiments highlight the role of the different parameters on the response to treatment.

************************************

SVD-Softmax: Fast Softmax Approximation on Large Vocabulary Neural Networks

Kyuhong Shim, Minjae Lee, Iksoo Choi, Yoonho Boo, Wonyong Sung

We propose a fast approximation method of a softmax function with a very large vocabulary using singular value decomposition (SVD). SVD-softmax targets fast and accurate probability estimation of the topmost probable words during inference of neural network language models. The proposed method transforms the weight matrix used in the calculation of the output vector by using SVD. The approximate probability of each word can be estimated with only a small part of the weight matrix by using a few large singular values and the corresponding elements for most of the words. We applied the technique to language modeling and neural machine translation and present a guideline for good approximation. The algorithm requires only approximately 20\% of arithmetic operations for an 800K vocabulary case and shows more than a three-fold speedup on a GPU.

************************************

Concentration of Multilinear Functions of the Ising Model with Applications to Network Data

Constantinos Daskalakis, Nishanth Dikkala, Gautam Kamath

We prove near-tight concentration of measure for polynomial functions of the Isi ng model, under high temperature, improving the radius of concentration guaranteed by known results by polynomial factors in the dimension (i.e.~the number of nodes in the Ising model). We show that our results are optimal up to logarithmic factors in the dimension. We obtain our results by extending and strengthening the exchangeable-pairs approach used to prove concentration of measure in this setting by Chatterjee. We demonstrate the efficacy of such functions as statistics for testing the strength  of interactions in social networks in both synthetic and real world data.

************************************

Rigorous Dynamics and Consistent Estimation in Arbitrarily Conditioned Linear Systems

Alyson K. Fletcher, Mojtaba Sahraee-Ardakan, Sundeep Rangan, Philip Schniter

The problem of estimating a random vector x from noisy linear measurements y=Ax+w with unknown parameters on the distributions of x and w, which must also be learned, arises in a wide range of statistical learning and linear inverse problems.  We show that a computationally simple iterative message-passing algorithm can provably obtain asymptotically consistent estimates in a certain high-dimensional large-system limit (LSL) under very general parameterizations.  Previous message passing techniques have required i.i.d. sub-Gaussian A matrices and often fail when the matrix is ill-conditioned. The proposed algorithm, called adaptive vector approximate message passing (Adaptive VAMP) with auto-tuning, applies to all right-rotationally random A.  Importantly, this class includes matrices with arbitrarily bad conditioning.  We show that the parameter estimates and mean squared error (MSE) of x in each iteration converge to deterministic limits that can be precisely predicted by a simple set of state evolution (SE) equations.  In addition, a simple testable condition is provided in which the MSE matches the Bayes-optimal value predicted by the replica method.  The paper thus provides a computationally simple method with provable guarantees of optimality and consistency over a large class of linear inverse problems.

************************************

OnACID: Online Analysis of Calcium Imaging Data in Real Time

Andrea Giovannucci, Johannes Friedrich, Matt Kaufman, Anne Churchland, Dmitri Chklovskii, Liam Paninski, Eftychios A. Pnevmatikakis

Optical imaging methods using calcium indicators are critical for monitoring the activity of large neuronal populations in vivo. Imaging experiments typically generate a large amount of data that needs to be processed to extract the activity of the imaged neuronal sources. While deriving such processing algorithms is an active area of research, most existing methods require the processing of large amounts of data at a time, rendering them vulnerable to the volume of the recorded data, and preventing real-time experimental interrogation. Here we introduce OnACID, an Online framework for the Analysis of streaming Calcium Imaging Data, including i) motion artifact correction, ii) neuronal source extraction, and iii) activity denoising and deconvolution. Our approach combines and extends previous work on online dictionary learning and calcium imaging data analysis, to deliver an automated pipeline that can discover and track the activity of hundreds of cells in real time, thereby enabling new types of closed-loop experiments. We apply our algorithm on two large scale experimental datasets, benchmark its performance on manually annotated data, and show that it outperforms a popular offline approach.

************************************

Action Centered Contextual Bandits

Kristjan Greenewald, Ambuj Tewari, Susan Murphy, Predag Klasnja

Contextual bandits have become popular as they offer a middle ground between very simple approaches based on multi-armed bandits and very complex approaches using the full power of reinforcement learning. They have demonstrated success in web applications and have a rich body of associated theoretical guarantees. Linear models are well understood theoretically and preferred by practitioners because they are not only easily interpretable but also simple to implement and debug. Furthermore, if the linear model is true, we get very strong performance guarantees. Unfortunately, in emerging applications in mobile health, the time-invaria nt linear model assumption is untenable. We provide an extension of the linear model for contextual bandits that has two parts: baseline reward and treatment effect. We allow the former to be complex but keep the latter simple. We argue that this model is plausible for mobile health applications. At the same time, it leads to algorithms with strong performance guarantees as in the linear model setting, while still allowing for complex nonlinear baseline modeling. Our theory is supported by experiments on data gathered in a recently concluded mobile health study.

************************************

Cost efficient gradient boosting

Sven Peter, Ferran Diego, Fred A. Hamprecht, Boaz Nadler

Many applications require learning classifiers or regressors that are both accurate and cheap to evaluate. Prediction cost can be drastically reduced if the learned predictor is constructed such that on the majority of the inputs, it uses cheap features and fast evaluations. The main challenge is to do so with little loss in accuracy. In this work we propose a budget-aware strategy based on deep boosted regression trees. In contrast to previous approaches to learning with cost penalties, our method can grow very deep trees that on average are nonetheless cheap to compute. We evaluate our method on a number of datasets and find that it outperforms the current state of the art by a large margin. Our algorithm is easy to implement and its learning time is comparable to that of the original gradient boosting. Source code is made available at http://github.com/svenpeter42/LightGBM-CEGB.

************************************

Eigenvalue Decay Implies Polynomial-Time Learnability for Neural Networks

Surbhi Goel, Adam Klivans

We consider the problem of learning function classes computed by neural networks with various activations (e.g. ReLU or Sigmoid), a task believed to be computationally intractable in the worst-case. A major open problem is to understand the minimal assumptions under which these classes admit provably efficient algorithms. In this work we show that a natural distributional assumption corresponding to {\em eigenvalue decay} of the Gram matrix yields polynomial-time algorithms in the non-realizable setting for expressive classes of networks (e.g. feed-forward networks of ReLUs). We make no assumptions on the structure of the network or the labels. Given sufficiently-strong eigenvalue decay, we obtain {\em fully}-polynomial time algorithms in {\em all} the relevant parameters with respect to square-loss. This is the first purely distributional assumption that leads to polynomial-time algorithms for networks of ReLUs. Further, unlike prior distributional assumptions (e.g., the marginal distribution is Gaussian), eigenvalue decay has been observed in practice on common data sets.

************************************

On Separability of Loss Functions, and Revisiting Discriminative Vs Generative Models

Adarsh Prasad, Alexandru Niculescu-Mizil, Pradeep K. Ravikumar

We revisit the classical analysis of generative vs discriminative models for general exponential families, and high-dimensional settings. Towards this, we develop novel technical machinery, including a notion of separability of general loss functions, which allow us to provide a general framework to obtain l∞ convergence rates for general M-estimators. We use this machinery to analyze l∞ and l2 convergence rates of generative and discriminative models, and provide insights into their nuanced behaviors in high-dimensions. Our results are also applicable to differential parameter estimation, where the quantity of interest is the difference between generative model parameters.

************************************

ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events

Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr. Prabhat, Chris Pal

Then detection and identification of extreme weather events in large-scale clima te simulations is an important problem for risk management, informing governmental policy decisions and advancing our basic understanding of the climate system. Recent work has shown that fully supervised convolutional neural networks (CNNs) can yield acceptable accuracy for classifying well-known types of extreme weather events when large amounts of labeled data are available. However, many different types of spatially localized climate patterns are of interest including hurricanes, extra-tropical cyclones, weather fronts, and blocking events among others. Existing labeled data for these patterns can be incomplete in various ways, such as covering only certain years or geographic areas and having false negatives. This type of climate data therefore poses a number of interesting machine learning challenges. We present a multichannel spatiotemporal CNN architecture for semi-supervised bounding box prediction and exploratory data analysis. We demonstrate that our approach is able to leverage temporal information and unlabeled data to improve the localization of extreme weather events. Further, we explore the representations learned by our model in order to better understand this important data. We present a dataset, ExtremeWeather, to encourage machine learning research in this area and to help facilitate further work in understanding and mitigating the effects of climate change. The dataset is available at extremeweatherdataset.github.io and the code is available at https://github.com/eracah/hur-detect.

************************************

A Meta-Learning Perspective on Cold-Start Recommendations for Items
Manasi Vartak, Arvind Thiagarajan, Conrado Miranda, Jeshua Bratman, Hugo Larochelle

Matrix factorization (MF) is one of the most popular techniques for product recommendation, but is known to suffer from serious cold-start problems. Item cold-start problems are particularly acute in settings such as Tweet recommendation where new items arrive continuously. In this paper, we present a meta-learning strategy to address item cold-start when new items arrive continuously. We propose two deep neural network architectures that implement our meta-learning strategy. The first architecture learns a linear classifier whose weights are determined by the item history while the second architecture learns a neural network whose biases are instead adjusted. We evaluate our techniques on the real-world problem of Tweet recommendation. On production data at Twitter, we demonstrate that our proposed techniques significantly beat the MF baseline and also outperform production models for Tweet recommendation.

************************************

Learning Unknown Markov Decision Processes: A Thompson Sampling Approach
Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, Rahul Jain

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.

************************************

Deep Hyperspherical Learning
Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, Le Song

Convolution as inner product has been the founding basis of convolutional neural networks (CNNs) and the key to end-to-end visual representation learning. Benefiting from deeper architectures, recent CNNs have demonstrated increasingly strong representation abilities. Despite such improvement, the increased depth and larger parameter space have also led to challenges in properly training a network. In light of such challenges, we propose hyperspherical convolution (SphereConv), a novel learning framework that gives angular representations on hyperspheres. We introduce SphereNet, deep hyperspherical convolution networks that are distinct from conventional inner product based convolutional networks. In particular, SphereNet adopts SphereConv as its basic convolution operator and is supervised by generalized angular softmax loss - a natural loss formulation under SphereConv. We show that SphereNet can effectively encode discriminative representation and alleviate training difficulty, leading to easier optimization, faster convergence and comparable (even better) classification accuracy over convolutional c ounterparts. We also provide some theoretical insights for the advantages of learning on hyperspheres. In addition, we introduce the learnable SphereConv, i.e., a natural improvement over prefixed SphereConv, and SphereNorm, i.e., hyperspherical learning as a normalization method. Experiments have verified our conclusions.

************************************

Interpretable and Globally Optimal Prediction for Textual Grounding using Image Concepts

Raymond Yeh, Jinjun Xiong, Wen-Mei Hwu, Minh Do, Alexander Schwing

Textual grounding is an important but challenging task for human-computer interaction, robotics and knowledge mining. Existing algorithms generally formulate the task as selection from a set of bounding box proposals obtained from deep net based systems. In this work, we demonstrate that we can cast the problem of textual grounding into a unified framework that permits efficient search over all possible bounding boxes. Hence, the method is able to consider significantly more proposals and doesn’t rely on a successful first stage hypothesizing bounding box proposals. Beyond, we demonstrate that the trained parameters of our model can be used as word-embeddings which capture spatial-image relationships and provide interpretability. Lastly, at the time of submission, our approach outperformed the current state-of-the-art methods on the Flickr 30k Entities and the ReferItGame dataset by 3.08% and 7.77% respectively.

************************************

Off-policy evaluation for slate recommendation

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, Imed Zitouni

This paper studies the evaluation of policies that recommend an ordered set of items (e.g., a ranking) based on some context---a common scenario in web search, ads, and recommendation. We build on techniques from combinatorial bandits to introduce a new practical estimator that uses logged data to estimate a policy's performance.  A thorough empirical evaluation on real-world data reveals that our estimator is accurate in a variety of settings, including as a subroutine in a learning-to-rank task, where it achieves competitive performance. We derive conditions under which our estimator is unbiased---these conditions are weaker than prior heuristics for slate evaluation---and experimentally demonstrate a smaller bias than parametric approaches, even when these conditions are violated. Finally, our theory and experiments also show exponential savings in the amount of required data compared with general unbiased estimators.

************************************

Unbiased estimates for linear regression via volume sampling

Michal Derezinski, Manfred K. K. Warmuth

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.

************************************

Revisiting Perceptron: Efficient and Label-Optimal Learning of Halfspaces

Songbai Yan, Chicheng Zhang

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.

************************************

Renyi Differential Privacy Mechanisms for Posterior Sampling

Joseph Geumlek, Shuang Song, Kamalika Chaudhuri

With the newly proposed privacy definition of Rényi Differential Privacy (RDP) in (Mironov, 2017), we re-examine the inherent privacy of releasing a single sample from a posterior distribution. We exploit the impact of the prior distribution in mitigating the influence of individual data points. In particular, we focus on sampling from an exponential family and specific generalized linear models, such as logistic regression. We propose novel RDP mechanisms as well as offering

a new RDP analysis for an existing method in order to add value to the RDP framework. Each method is capable of achieving arbitrary RDP privacy guarantees, and we offer experimental results of their efficacy.
************************************

Variable Importance Using Decision Trees
Jalil Kazemitabar, Arash Amini, Adam Bloniarz, Ameet S. Talwalkar
Decision trees and random forests are well established models that not only offer good predictive performance, but also provide rich feature importance information. While practitioners often employ variable importance methods that rely on this impurity-based information, these methods remain poorly characterized from a theoretical perspective. We provide novel insights into the performance of these methods by deriving finite sample performance guarantees in a high-dimensional setting under various modeling assumptions.  We further demonstrate the effectiveness of these impurity-based methods via an extensive set of simulations.
************************************

A simple model of recognition and recall memory
Nisheeth Srivastava, Edward Vul
We show that several striking differences in memory performance between recognition and recall tasks are explained by an ecological bias endemic in classic memory experiments - that such experiments universally involve more stimuli than retrieval cues. We show that while it is sensible to think of recall as simply retrieving items when probed with a cue - typically the item list itself -  it is better to think of recognition as retrieving cues when probed with items. To test this theory, by manipulating the number of items and cues in a memory experiment, we show a crossover effect in memory performance within subjects such that recognition performance is superior to recall performance when the number of items is greater than the number of cues and recall performance is better than recognition when the converse holds. We build a simple computational model around this theory, using sampling to approximate an ideal Bayesian observer encoding and retrieving situational co-occurrence frequencies of stimuli and retrieval cues. This model robustly reproduces a number of dissociations in recognition and recall previously used to argue for dual-process accounts of declarative memory.
************************************

Implicit Regularization in Matrix Factorization
Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, Nati Srebro
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.
************************************

Continuous DR-submodular  Maximization: Structure and Algorithms
An Bian, Kfir Levy, Andreas Krause, Joachim M. Buhmann
DR-submodular continuous functions are important objectives with wide real-world applications spanning MAP inference in  determinantal point processes (DPPs), and mean-field inference for probabilistic submodular models, amongst others. DR-submodularity captures a subclass of non-convex functions that enables both exact minimization and approximate maximization in polynomial time.  In this work we study the  problem of maximizing  non-monotone DR-submodular continuous functions under general down-closed convex constraints. We start by investigating geometric properties that underlie such objectives, e.g., a strong relation between (approximately) stationary points and global optimum is proved. These properties are then used to devise two optimization algorithms with provable guarantees. Concretely, we first devise a "two-phase'' algorithm with 1/4 approximation guarantee. This algorithm allows the use of existing methods for finding (approximately) stationary points as a subroutine, thus, harnessing recent progress in non-convex optimization. Then we present a non-monotone Frank-Wolfe variant with 1/e approximation guarantee and sublinear convergence rate. Finally, we extend our approach to a broader class of generalized DR-submodular continuous functions, which captures a wider spectrum of applications. Our theoretical findings are vali

dated on synthetic and real-world problem instances.

************************************

Decoupling "when to update" from "how to update"

Eran Malach, Shai Shalev-Shwartz

Deep learning requires data. A useful approach to obtain data is to  be creative and mine data from various sources, that were created for different purposes. Unfortunately, this approach often leads to noisy labels. In this paper, we propose a meta algorithm for tackling the noisy labels problem. The key idea is to decouple  when to update'' fromhow to update''. We demonstrate the effectiveness of our algorithm by mining data for gender classification by combining the Labeled Faces in the Wild (LFW) face recognition dataset with a textual genderizing service, which leads to a noisy dataset. While our approach is very simple to implement, it leads to state-of-the-art results. We analyze some convergence properties of the proposed algorithm.

************************************

Regret Analysis for Continuous Dueling Bandit

Wataru Kumagai

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.

************************************

One-Sided Unsupervised Domain Mapping

Sagie Benaim, Lior Wolf

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.

************************************

Poincaré Embeddings for Learning Hierarchical Representations

Maximillian Nickel, Douwe Kiela

Representation learning has become an invaluable approach for learning from symbolic data such as text and graphs. However, state-of-the-art embedding methods typically do not account for latent hierarchical structures which are characteristic for many complex symbolic datasets. In this work, we introduce a new approach for learning hierarchical representations of symbolic data by embedding them into hyperbolic space -- or more precisely into an n-dimensional Poincaré ball. Due to the underlying hyperbolic geometry, this allows us to learn parsimonious representations of symbolic data by simultaneously capturing hierarchy and similarity. We present an efficient algorithm to learn the embeddings based on Riemannian optimization and show experimentally that Poincaré embeddings can outperform Euclidean embeddings significantly on data with latent hierarchies, both in terms of representation capacity and in terms of generalization ability.

************************************

Variance-based Regularization with Convex Objectives

Hongseok Namkoong, John C. Duchi

We develop an approach to risk minimization and stochastic optimization that provides a convex surrogate for variance, allowing near-optimal and computationally efficient trading between approximation and estimation error. Our approach builds off of techniques for distributionally robust optimization and Owen's empirical likelihood, and we provide a number of finite-sample and asymptotic results characterizing the theoretical performance of the estimator. In particular, we show that our procedure comes with certificates of optimality, achieving (in some scenarios) faster rates of convergence than empirical risk minimization by virtue of automatically balancing bias and variance. We give corroborating empirical evidence showing that in practice, the estimator indeed trades between variance and absolute performance on a training sample, improving out-of-sample (test) performance over standard empirical risk minimization for a number of classification problems.

************************************

A Sharp Error Analysis for the Fused Lasso, with Application to Approximate Changepoint Screening

Kevin Lin, James L. Sharpnack, Alessandro Rinaldo, Ryan J. Tibshirani

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.

************************************

Cross-Spectral Factor Analysis

Neil Gallagher, Kyle R. Ulrich, Austin Talbot, Kafui Dzirasa, Lawrence Carin, David E. Carlson

In neuropsychiatric disorders such as schizophrenia or depression, there is often a disruption in the way that regions of the brain synchronize with one another. To facilitate understanding of network-level synchronization between brain regions, we introduce a novel model of multisite low-frequency neural recordings, such as local field potentials (LFPs) and electroencephalograms (EEGs). The proposed model, named Cross-Spectral Factor Analysis (CSFA), breaks the observed signal into factors defined by unique spatio-spectral properties. These properties are granted to the factors via a Gaussian process formulation in a multiple kernel learning framework. In this way, the LFP signals can be mapped to a lower dimensional space in a way that retains information of relevance to neuroscientists.  Critically, the factors are interpretable. The proposed approach empirically allows similar performance in classifying mouse genotype and behavioral context when compared to commonly used approaches that lack the interpretability of CSFA.  We also introduce a semi-supervised approach, termed discriminative CSFA (dCSFA). CSFA and dCSFA provide useful tools for understanding neural dynamics, particularly by aiding in the design of causal follow-up experiments.

************************************

Self-Normalizing Neural Networks

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, Sepp Hochreiter

Deep Learning has revolutionized vision via convolutional neural networks (CNNs) and natural language processing via recurrent neural networks (RNNs). However, success stories of Deep Learning with standard feed-forward neural networks (FNNs) are rare. FNNs that perform well are typically shallow and, therefore cannot exploit many levels of abstract representations. We introduce self-normalizing neural networks (SNNs) to enable high-level abstract representations. While batch normalization requires explicit normalization, neuron activations of SNNs automatically converge towards zero mean and unit variance. The activation function of SNNs are "scaled exponential linear units" (SELUs), which induce self-normalizing properties. Using the Banach fixed-point theorem, we prove that activations close to zero mean and unit variance that are propagated through many network layers will converge towards zero mean and unit variance -- even under the presence of noise and perturbations. This convergence property of SNNs allows to (1) train deep networks with many layers, (2) employ strong regularization, and (3) to make learning highly robust. Furthermore, for activations not close to unit variance, we prove an upper and lower bound on the variance, thus, vanishing and exploding gradients are impossible. We compared SNNs on (a) 121 tasks from the UCI machine learning repository, on (b) drug discovery benchmarks, and on (c) astronomy tasks with standard FNNs and other machine learning methods such as random forests and support vector machines. For FNNs we considered (i) ReLU networks without normalization, (ii) batch normalization, (iii) layer normalization, (iv) weight normalization, (v) highway networks, (vi) residual networks. SNNs significantly outperformed all competing FNN methods at 121 UCI tasks, outperformed all competing methods at the Tox21 dataset, and set a new record at an astronomy data set. The winning SNN architectures are often very deep.

************************************

Fast amortized inference of neural activity from calcium imaging data with variational autoencoders

Artur Speiser, Jinyao Yan, Evan W. Archer, Lars Buesing, Srinivas C. Turaga, Jakob H. Macke

Calcium imaging permits optical measurement of neural activity. Since intracellular calcium concentration is an indirect measurement of neural activity, computational tools are necessary to infer the true underlying spiking activity from fluorescence measurements. Bayesian model inversion can be used to solve this problem, but typically requires either computationally expensive MCMC sampling, or faster but approximate maximum-a-posteriori optimization.  Here, we introduce a flexible algorithmic framework for fast, efficient and accurate extraction of neural spikes from imaging data. Using the framework of variational autoencoders, we propose to amortize inference by training a deep neural network to perform model inversion efficiently. The recognition network is trained to produce samples from the posterior distribution over spike trains.  Once trained, performing inference amounts to a fast single forward pass through the network, without the need for iterative optimization or sampling. We show that amortization can be applied flexibly to a wide range of nonlinear generative models and significantly improves upon the state of the art in computation time, while achieving competitive accuracy.  Our framework is also able to represent posterior distributions over spike-trains. We demonstrate the generality of our method by proposing the first probabilistic approach for separating backpropagating action potentials from putative synaptic inputs in calcium imaging of dendritic spines.
************************************

Asynchronous Parallel Coordinate Minimization for MAP Inference
Ofer Meshi, Alexander Schwing

Finding the maximum a-posteriori (MAP) assignment is a central task in graphical models. Since modern applications give rise to very large problem instances, there is increasing need for efficient solvers. In this work we propose to improve the efficiency of coordinate-minimization-based dual-decomposition solvers by running their updates asynchronously in parallel. In this case message-passing inference is performed by multiple processing units simultaneously without coordination, all reading and writing to shared memory. We analyze the convergence properties of the resulting algorithms and identify settings where speedup gains can be expected. Our numerical evaluations show that this approach indeed achieves significant speedups in common computer vision tasks.
************************************

Inductive Representation Learning on Large Graphs
Will Hamilton, Zhitao Ying, Jure Leskovec

Low-dimensional embeddings of nodes in large graphs have proved extremely useful in a variety of prediction tasks, from content recommendation to identifying protein functions. However, most existing approaches require that all nodes in the graph are present during training of the embeddings; these previous approaches are inherently transductive and do not naturally generalize to unseen nodes. Here we present GraphSAGE, a general, inductive framework that leverages node feature information (e.g., text attributes) to efficiently generate node embeddings. Instead of training individual embeddings for each node, we learn a function that generates embeddings by sampling and aggregating features from a node's local neighborhood. Our algorithm outperforms strong baselines on three inductive node-classification benchmarks: we classify the category of unseen nodes in evolving information graphs based on citation and Reddit post data, and we show that our algorithm generalizes to completely unseen graphs using a multi-graph dataset of protein-protein interactions.
************************************

Data-Efficient Reinforcement Learning in Continuous State-Action Gaussian-POMDPs
Rowan McAllister, Carl Edward Rasmussen

We present a data-efficient reinforcement learning method for continuous state-action systems under significant observation noise. Data-efficient solutions under small noise exist, such as PILCO which learns the cartpole swing-up task in 30s. PILCO evaluates policies by planning state-trajectories using a dynamics model. However, PILCO applies policies to the observed state, therefore planning in observation space. We extend PILCO with filtering to instead plan in belief space, consistent with partially observable Markov decisions process (POMDP) planning. This enables data-efficient learning under significant observation noise, out performing more naive methods such as post-hoc application of a filter to policies optimised by the original (unfiltered) PILCO algorithm. We test our method on the cartpole swing-up task, which involves nonlinear dynamics and requires nonlinear control.

************************************

Coded Distributed Computing for Inverse Problems

Yaoqing Yang, Pulkit Grover, Soummya Kar

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.

************************************

Dykstra's Algorithm, ADMM, and Coordinate Descent: Connections, Insights, and Extensions

Ryan J. Tibshirani

We study connections between Dykstra's algorithm for projecting onto an intersection of convex sets, the augmented Lagrangian method of multipliers or ADMM, and block coordinate descent. We prove that coordinate descent for a regularized regression problem, in which the penalty is a separable sum of support functions, is exactly equivalent to Dykstra's algorithm applied to the dual problem. ADMM on the dual problem is also seen to be equivalent, in the special case of two sets, with one being a linear subspace. These connections, aside from being interesting in their own right, suggest new ways of analyzing and extending coordinate descent. For example, from existing convergence theory on Dykstra's algorithm over polyhedra, we discern that coordinate descent for the lasso problem converges at an (asymptotically) linear rate. We also develop two parallel versions of coordinate descent, based on the Dykstra and ADMM connections.

************************************

Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems

Ingmar Kanitscheider, Ila Fiete

Self-localization during navigation with noisy sensors in an ambiguous world is computationally challenging, yet animals and humans excel at it. In robotics, {\em Simultaneous Location and Mapping} (SLAM) algorithms solve this problem through joint sequential probabilistic inference of their own coordinates and those of external spatial landmarks. We generate the first neural solution to the SLAM problem by training recurrent LSTM networks to perform a set of hard 2D navigation tasks that require generalization to completely novel trajectories and environments. Our goal is to make sense of how the diverse phenomenology in the brain's spatial navigation circuits is related to their function. We show that the hidden unit representations exhibit several key properties of hippocampal place cells, including stable tuning curves that remap between environments. Our result is also a proof of concept for end-to-end-learning of a SLAM algorithm using recurrent networks, and a demonstration of why this approach may have some advantages for robotic SLAM.

************************************

SafetyNets: Verifiable Execution of Deep Neural Networks on an Untrusted Cloud

Zahra Ghodsi, Tianyu Gu, Siddharth Garg

Inference using deep neural networks is often outsourced to the cloud since it is a computationally demanding task.  However, this raises a fundamental issue of trust. How can a client be sure that the cloud has performed inference correctly? A lazy cloud provider might use a simpler but less accurate model to reduce its own computational load, or worse, maliciously modify the inference results sent to the client. We propose SafetyNets, a framework that enables an untrusted server (the cloud) to provide a client with a short mathematical proof of the correctness of inference tasks that they perform on behalf of the client. Specifically, SafetyNets develops and implements a specialized interactive proof (IP) protocol for verifiable execution of a class of deep neural networks, i.e., those that can be represented as arithmetic circuits. Our empirical results on three- and four-layer deep neural networks demonstrate the run-time costs of SafetyNets for both the client and server are low. SafetyNets detects any incorrect computations of the neural network by the untrusted server with high probability, while achieving state-of-the-art accuracy on the MNIST digit recognition (99.4%) and TIMIT speech recognition tasks (75.22%).

************************************

Improved Graph Laplacian via Geometric Self-Consistency

Dominique Joncas, Marina Meila, James McQueen

We address the problem of setting the kernel bandwidth, epps, used by Manifold Learning algorithms to construct the graph Laplacian. Exploiting the connection between manifold geometry, represented by the Riemannian metric, and the Laplace-Beltrami operator, we set epps by optimizing the Laplacian's ability to preserve the geometry of the data. Experiments show that this principled approach is effective and robust

************************************

Generalization Properties of Learning with Random Features

Alessandro Rudi, Lorenzo Rosasco

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.

************************************

Predictive-State Decoders: Encoding the Future into Recurrent Networks

Arun Venkatraman, Nicholas Rhinehart, Wen Sun, Lerrel Pinto, Martial Hebert, Byron Boots, Kris Kitani, J. Bagnell

Recurrent neural networks (RNNs) are a vital modeling technique that rely on internal states learned indirectly by optimization of a supervised, unsupervised, or reinforcement training loss. RNNs are used to model dynamic processes that are characterized by underlying latent states whose form is often unknown, precluding its analytic representation inside an RNN. In the Predictive-State Representation (PSR) literature, latent state processes are modeled by an internal state representation that directly models the distribution of future observations, and most recent work in this area has relied on explicitly representing and targeting sufficient statistics of this probability distribution. We seek to combine the advantages of RNNs and PSRs by augmenting existing state-of-the-art recurrent neural networks with Predictive-State Decoders (PSDs), which add supervision to the network's internal state representation to target predicting future observations. PSDs are simple to implement and easily incorporated into existing training pipelines via additional loss regularization. We demonstrate the effectiveness of PSDs with experimental results in three different domains:  probabilistic filtering, Imitation Learning, and Reinforcement Learning. In each, our method improves statistical performance of state-of-the-art recurrent baselines and does so with fewer iterations and less data.

************************************

Federated Multi-Task Learning

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, Ameet S. Talwalkar

Federated learning poses new statistical and systems challenges in training machine learning models over distributed networks of devices. In this work, we show that multi-task learning is naturally suited to handle the statistical challenges of this setting, and propose a novel systems-aware optimization method, MOCHA, that is robust to practical systems issues. Our method and theory for the first time consider issues of high communication cost, stragglers, and fault tolerance for distributed multi-task learning. The resulting method achieves significant speedups compared to alternatives in the federated setting, as we demonstrate through simulations on real-world federated datasets.

************************************

Learning Causal Structures Using Regression Invariance

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, Kun Zhang

We study causal discovery in a multi-environment setting, in which the functional relations for producing the variables from their direct causes remain the same across environments, while the distribution of exogenous noises may vary. We in troduce the idea of using the invariance of the functional relations of the variables to their causes across a set of environments for structure learning. We define a notion of completeness for a causal inference algorithm in this setting and prove the existence of such algorithm by proposing the baseline algorithm. Additionally, we present an alternate algorithm that has significantly improved computational and sample complexity compared to the baseline algorithm. Experiment results show that the proposed algorithm outperforms the other existing algorithms.

************************************

Practical Hash Functions for Similarity Estimation and Dimensionality Reduction
Søren Dahlgaard, Mathias Knudsen, Mikkel Thorup

Hashing is a basic tool for dimensionality reduction employed in several aspects of machine learning. However, the perfomance analysis is often carried out under the abstract assumption that a truly random unit cost hash function is used, without concern for which concrete hash function is employed. The concrete hash function may work fine on sufficiently random input. The question is if it can be trusted in the real world when faced with more structured input. In this paper we focus on two prominent applications of hashing, namely similarity estimation with the one permutation hashing (OPH) scheme of Li et al. [NIPS'12] and feature hashing (FH) of Weinberger et al. [ICML'09], both of which have found numerous applications, i.e. in approximate near-neighbour search with LSH and large-scale classification with SVM. We consider the recent mixed tabulation hash function of Dahlgaard et al. [FOCS'15] which was proved theoretically to perform like a truly random hash function in many applications, including the above OPH. Here we first show improved concentration bounds for FH with truly random hashing and then argue that mixed tabulation performs similar when the input vectors are sparse. Our main contribution, however, is an experimental comparison of different hashing schemes when used inside FH, OPH, and LSH. We find that mixed tabulation hashing is almost as fast as the classic multiply-mod-prime scheme ax+b mod p. Mutiply-mod-prime is guaranteed to work well on sufficiently random data, but we demonstrate that in the above applications, it can lead to bias and poor concentration on both real-world and synthetic data. We also compare with the very popular MurmurHash3, which has no proven guarantees. Mixed tabulation and MurmurHash3 both perform similar to truly random hashing in our experiments. However, mixed tabulation was 40% faster than MurmurHash3, and it has the proven guarantee of good performance on all possible input making it more reliable.

************************************

Gaussian Quadrature for Kernel Features
Tri Dao, Christopher M. De Sa, Christopher Ré

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.

************************************

Multi-Modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets
Karol Hausman, Yevgen Chebotar, Stefan Schaal, Gaurav Sukhatme, Joseph J. Lim

Imitation learning has traditionally been applied to learn a single task from demonstrations thereof. The requirement of structured and isolated demonstrations limits the scalability of imitation learning approaches as they are difficult to apply to real-world scenarios, where robots have to be able to execute a multitude of tasks. In this paper, we propose a multi-modal imitation learning framework that is able to segment and imitate skills from unlabelled and unstructured demonstrations by learning skill segmentation and imitation learning jointly. The extensive simulation results indicate that our method can efficiently separate the demonstrations into individual skills and learn to imitate them using a single multi-modal policy.

************************************

Greedy Algorithms for Cone Constrained Optimization with Convergence Guarantees
Francesco Locatello, Michael Tschannen, Gunnar Raetsch, Martin Jaggi

Greedy optimization methods such as Matching Pursuit (MP) and Frank-Wolfe (FW) algorithms regained popularity in recent years due to their simplicity, effectiveness and theoretical guarantees. MP and FW address optimization over the linear span and the convex hull of a set of atoms, respectively. In this paper, we consider the intermediate case of optimization over the convex cone, parametrized as the conic hull of a generic atom set, leading to the first principled definitions of non-negative MP algorithms for which we give explicit convergence rates and demonstrate excellent empirical performance. In particular, we derive sublinear (O(1/t)) convergence on general smooth and convex objectives, and linear convergence (O(e^{-t})) on strongly convex objectives, in both cases for general sets of atoms. Furthermore, we establish a clear correspondence of our algorithms to known algorithms from the MP and FW literature. Our novel algorithms and analyses target general atom sets and general objective functions, and hence are directly applicable to a large variety of learning settings.

************************************

On the Fine-Grained Complexity of Empirical Risk Minimization: Kernel Methods and Neural Networks

Arturs Backurs, Piotr Indyk, Ludwig Schmidt

Empirical risk minimization (ERM) is ubiquitous in machine learning and underlies most supervised learning methods. While there is a large body of work on algorithms for various ERM problems, the exact computational complexity of ERM is still not understood. We address this issue for multiple popular ERM problems including kernel SVMs, kernel ridge regression, and training the final layer of a neural network. In particular, we give conditional hardness results for these problems based on complexity-theoretic assumptions such as the Strong Exponential Time Hypothesis. Under these assumptions, we show that there are no algorithms that solve the aforementioned ERM problems to high accuracy in sub-quadratic time. We also give similar hardness results for computing the gradient of the empirical loss, which is the main computational burden in many non-convex learning tasks.

************************************

Acceleration and Averaging in Stochastic Descent Dynamics

Walid Krichene, Peter L. Bartlett

We formulate and study a general family of (continuous-time) stochastic dynamics for accelerated first-order minimization of smooth convex functions. Building on an averaging formulation of accelerated mirror descent, we propose a stochastic variant in which the gradient is contaminated by noise, and study the resulting stochastic differential equation. We prove a bound on the rate of change of an energy function associated with the problem, then use it to derive estimates of convergence rates of the function values (almost surely and in expectation), both for persistent and asymptotically vanishing noise. We discuss the interaction between the parameters of the dynamics (learning rate and averaging rates) and the covariation of the noise process. In particular, we show how the asymptotic rate of covariation affects the choice of parameters and, ultimately, the convergence rate.

************************************

LightGBM: A Highly Efficient Gradient Boosting Decision Tree

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu

Gradient Boosting Decision Tree (GBDT) is a popular machine learning algorithm, and has quite a few effective implementations such as XGBoost and pGBRT. Although many engineering optimizations have been adopted in these implementations, the efficiency and scalability are still unsatisfactory when the feature dimension is high and data size is large. A major reason is that for each feature, they need to scan all the data instances to estimate the information gain of all possible split points, which is very time consuming. To tackle this problem, we propose two novel techniques: \emph{Gradient-based One-Side Sampling} (GOSS) and \emph{Exclusive Feature Bundling} (EFB). With GOSS, we exclude a significant proportion of data instances with small gradients, and only use the rest to estimate the information gain. We prove that, since the data instances with larger gradients play a more important role in the computation of information gain, GOSS can obt

ain quite accurate estimation of the information gain with a much smaller data size. With EFB, we bundle mutually exclusive features (i.e., they rarely take nonzero values simultaneously), to reduce the number of features. We prove that finding the optimal bundling of exclusive features is NP-hard, but a greedy algorithm can achieve quite good approximation ratio (and thus can effectively reduce the number of features without hurting the accuracy of split point determination by much). We call our new GBDT implementation with GOSS and EFB \emph{LightGBM}. Our experiments on multiple public datasets show that, LightGBM speeds up the training process of conventional GBDT by up to over 20 times while achieving almost the same accuracy.
************************************

The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process
Hongyuan Mei, Jason M. Eisner

Many events occur in the world. Some event types are stochastically excited or inhibited—in the sense of having their probabilities elevated or decreased—by patterns in the sequence of previous events. Discovering such patterns can help us predict which type of event will happen next and when. We model streams of discrete events in continuous time, by constructing a neurally self-modulating multivariate point process in which the intensities of multiple event types evolve according to a novel continuous-time LSTM. This generative model allows past events to influence the future in complex and realistic ways, by conditioning future event intensities on the hidden state of a recurrent neural network that has consumed the stream of past events. Our model has desirable qualitative properties. It achieves competitive likelihood and predictive accuracy on real and synthetic datasets, including under missing-data conditions.
************************************

Bayesian Optimization with Gradients
Jian Wu, Matthias Poloczek, Andrew G. Wilson, Peter Frazier

Bayesian optimization has shown success in global optimization of expensive-to-evaluate multimodal objective functions. However, unlike most optimization methods, Bayesian optimization typically does not use derivative information. In this paper we show how Bayesian optimization can exploit derivative information to find good solutions with fewer objective function evaluations. In particular, we develop a novel Bayesian optimization algorithm, the derivative-enabled knowledge-gradient (dKG), which is one-step Bayes-optimal, asymptotically consistent, and provides greater one-step value of information than in the derivative-free setting. dKG accommodates noisy and incomplete derivative information, comes in both sequential and batch forms, and can optionally reduce the computational cost of inference through automatically selected retention of a single directional derivative. We also compute the dKG acquisition function and its gradient using a novel fast discretization-free technique. We show dKG provides state-of-the-art performance compared to a wide range of optimization procedures with and without gradients, on benchmarks including logistic regression, deep learning, kernel learning, and k-nearest neighbors.
************************************

Visual Reference Resolution using Attention Memory for Visual Dialog
Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, Leonid Sigal

Visual dialog is a task of answering a series of inter-dependent questions given an input image, and often requires to resolve visual references among the questions. This problem is different from visual question answering (VQA), which relies on spatial attention ({\em a.k.a. visual grounding}) estimated from an image and question pair. We propose a novel attention mechanism that exploits visual attentions in the past to resolve the current reference in the visual dialog scenario. The proposed model is equipped with an associative attention memory storing a sequence of previous (attention, key) pairs. From this memory, the model retrieves previous attention, taking into account recency, that is most relevant for the current question, in order to resolve potentially ambiguous reference(s). The model then merges the retrieved attention with the tentative one to obtain the final attention for the current question; specifically, we use dynamic parameter prediction to combine the two attentions conditioned on the question. Throug h extensive experiments on a new synthetic visual dialog dataset, we show that our model significantly outperforms the state-of-the-art (by ~16 % points) in the situation where the visual reference resolution plays an important role. Moreover, the proposed model presents superior performance (~2 % points improvement) in the Visual Dialog dataset, despite having significantly fewer parameters than the baselines.

************************************

Straggler Mitigation in Distributed Optimization Through Data Encoding

Can Karakus, Yifan Sun, Suhas Diggavi, Wotao Yin

Slow running or straggler tasks can significantly reduce computation speed in distributed computation. Recently, coding-theory-inspired approaches have been applied to mitigate the effect of straggling, through embedding redundancy in certain linear computational steps of the optimization algorithm, thus completing the computation without waiting for the stragglers. In this paper, we propose an alternate approach where we embed the redundancy directly in the data itself, and allow the computation to proceed completely oblivious to encoding. We propose several encoding schemes, and demonstrate that popular batch algorithms, such as gradient descent and L-BFGS, applied in a coding-oblivious manner, deterministically achieve sample path linear convergence to an approximate solution of the original problem, using an arbitrarily varying subset of the nodes at each iteration. Moreover, this approximation can be controlled by the amount of redundancy and the number of nodes used in each iteration. We provide experimental results demonstrating the advantage of the approach over uncoded and data replication strategies.

************************************

Using Options and Covariance Testing for Long Horizon Off-Policy Policy Evaluation

Zhaohan Guo, Philip S. Thomas, Emma Brunskill

Evaluating a policy by deploying it in the real world can be risky and costly. Off-policy policy evaluation (OPE) algorithms use historical data collected from running a previous policy to evaluate a new policy, which provides a means for evaluating a policy without requiring it to ever be deployed. Importance sampling is a popular OPE method because it is robust to partial observability and works with continuous states and actions. However, the amount of historical data required by importance sampling can scale exponentially with the horizon of the problem: the number of sequential decisions that are made. We propose using policies over temporally extended actions, called options, and show that combining these policies with importance sampling can significantly improve performance for long-horizon problems. In addition, we can take advantage of special cases that arise due to options-based policies to further improve the performance of importance sampling. We further generalize these special cases to a general covariance testing rule that can be used to decide which weights to drop in an IS estimate, and derive a new IS algorithm called Incremental Importance Sampling that can provide significantly more accurate estimates for a broad class of domains.

************************************

Attentional Pooling for Action Recognition

Rohit Girdhar, Deva Ramanan

We introduce a simple yet surprisingly powerful model to incorporate attention in action recognition and human object interaction tasks. Our proposed attention module can be trained with or without extra supervision, and gives a sizable boost in accuracy while keeping the network size and computational cost nearly the same. It leads to significant improvements over state of the art base architecture on three standard action recognition benchmarks across still images and videos, and establishes new state of the art on MPII dataset with 12.5% relative improvement. We also perform an extensive analysis of our attention module both empirically and analytically. In terms of the latter, we introduce a novel derivation of bottom-up and top-down attention as low-rank approximations of bilinear pooling methods (typically used for fine-grained classification). From this perspective, our attention formulation suggests a novel characterization of action recognition as a fine-grained recognition problem.

************************************

Testing and Learning on Distributions with Symmetric Noise Invariance

Ho Chung Law, Christopher Yau, Dino Sejdinovic

Kernel embeddings of distributions and the Maximum Mean Discrepancy (MMD), the resulting distance between distributions, are useful tools for fully nonparametric two-sample testing and learning on distributions. However, it is rarely that all possible differences between samples are of interest -- discovered differences can be due to different types of measurement noise, data collection artefacts or other irrelevant sources of variability. We propose distances between distributions which encode invariance to additive symmetric noise, aimed at testing whether the assumed true underlying processes differ. Moreover, we construct invariant features of distributions, leading to learning algorithms robust to the impairment of the input distributions with symmetric additive noise.

************************************

Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results

Antti Tarvainen, Harri Valpola

The recently proposed Temporal Ensembling has achieved state-of-the-art results in several semi-supervised learning benchmarks. It maintains an exponential moving average of label predictions on each training example, and penalizes predictions that are inconsistent with this target. However, because the targets change only once per epoch, Temporal Ensembling becomes unwieldy when learning large datasets. To overcome this problem, we propose Mean Teacher, a method that averages model weights instead of label predictions. As an additional benefit, Mean Teacher improves test accuracy and enables training with fewer labels than Temporal Ensembling. Without changing the network architecture, Mean Teacher achieves an error rate of 4.35% on SVHN with 250 labels, outperforming Temporal Ensembling trained with 1000 labels. We also show that a good network architecture is crucial to performance. Combining Mean Teacher and Residual Networks, we improve the state of the art on CIFAR-10 with 4000 labels from 10.55% to 6.28%, and on ImageNet 2012 with 10% of the labels from 35.24% to 9.11%.

************************************

Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments

Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, Igor Mordatch

We explore deep reinforcement learning methods for multi-agent domains. We begin by analyzing the difficulty of traditional algorithms in the multi-agent case: Q-learning is challenged by an inherent non-stationarity of the environment, while policy gradient suffers from a variance that increases as the number of agents grows. We then present an adaptation of actor-critic methods that considers action policies of other agents and is able to successfully learn policies that require complex multi-agent coordination. Additionally, we introduce a training regimen utilizing an ensemble of policies for each agent that leads to more robust multi-agent policies. We show the strength of our approach compared to existing methods in cooperative as well as competitive scenarios, where agent populations are able to discover various physical and informational coordination strategies.

************************************

Log-normality and Skewness of Estimated State/Action Values in Reinforcement Learning

Liangpeng Zhang, Ke Tang, Xin Yao

Under/overestimation of state/action values are harmful for reinforcement learning agents. In this paper, we show that a state/action value estimated using the Bellman equation can be decomposed to a weighted sum of path-wise values that follow log-normal distributions. Since log-normal distributions are skewed, the distribution of estimated state/action values can also be skewed, leading to an imbalanced likelihood of under/overestimation. The degree of such imbalance can vary greatly among actions and policies within a single problem instance, making the agent prone to select actions/policies that have inferior expected return and higher likelihood of overestimation. We present a comprehensive analysis to such skewness, examine its factors and impacts through both theoretical and empiric al results, and discuss the possible ways to reduce its undesirable effects.

************************************

Bayesian Compression for Deep Learning

Christos Louizos, Karen Ullrich, Max Welling

Compression and computational efficiency in deep learning have become a problem of great significance. In this work, we argue that the most principled and effective way to attack this problem is by adopting a Bayesian point of view, where through sparsity inducing priors we prune large parts of the network. We introduce two novelties in this paper: 1) we use hierarchical priors to prune nodes instead of individual weights, and 2) we use the posterior uncertainties to determine the optimal fixed point precision to encode the weights. Both factors significantly contribute to achieving the state of the art in terms of compression rates, while still staying competitive with methods designed to optimize for speed or energy efficiency.

************************************

Is Input Sparsity Time Possible for Kernel Low-Rank Approximation?

Cameron Musco, David Woodruff

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.

************************************

Convergent Block Coordinate Descent for Training Tikhonov Regularized Deep Neural Networks

Ziming Zhang, Matthew Brand

By lifting the ReLU function into a higher dimensional space, we develop a smooth multi-convex formulation for training feed-forward deep neural networks (DNNs). This allows us to develop a block coordinate descent (BCD) training algorithm consisting of a sequence of numerically well-behaved convex optimizations. Using ideas from proximal point methods in convex analysis, we prove that this BCD algorithm will converge globally to a stationary point with R-linear convergence rate of order one. In experiments with the MNIST database, DNNs trained with this BCD algorithm consistently yielded better test-set error rates than identical DNN architectures trained via all the stochastic gradient descent (SGD) variants in the Caffe toolbox.

************************************

Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes

Ahmed M. Alaa, Mihaela van der Schaar

Predicated on the increasing abundance of electronic health records, we investigate the problem of inferring individualized treatment effects using observational data. Stemming from the potential outcomes model, we propose a novel multi-task learning framework in which factual and counterfactual outcomes are modeled as the outputs of a function in a vector-valued reproducing kernel Hilbert space (vvRKHS). We develop a nonparametric Bayesian method for learning the treatment effects using a multi-task Gaussian process (GP) with a linear coregionalization kernel as a prior over the vvRKHS. The Bayesian approach allows us to compute individualized measures of confidence in our estimates via pointwise credible intervals, which are crucial for realizing the full potential of precision medicine. The impact of selection bias is alleviated via a risk-based empirical Bayes method for adapting the multi-task GP prior, which jointly minimizes the empirical error in factual outcomes and the uncertainty in (unobserved) counterfactual outcomes. We conduct experiments on observational datasets for an interventional social program applied to premature infants, and a left ventricular assist device applied to cardiac patients wait-listed for a heart transplant. In both experiments, we show that our method significantly outperforms the state-of-the-art.

************************************

Learning Overcomplete HMMs

Vatsal Sharan, Sham M. Kakade, Percy S. Liang, Gregory Valiant

We study the basic problem of learning overcomplete HMMs---those that have many hidden states but a small output alphabet. Despite having significant practical importance, such HMMs are poorly understood with no known positive or negative results for efficient learning. In this paper, we present several new results---both positive and negative---which help define the boundaries between the tractable-learning setting and the intractable setting. We show positive results for a large subclass of HMMs whose transition matrices are sparse, well-conditioned and have small probability mass on short cycles. We also show that learning is impossible given only a polynomial number of samples for HMMs with a small output alphabet and whose transition matrices are random regular graphs with large degree. We also discuss these results in the context of learning HMMs which can capture long-term dependencies.

************************************

Convolutional Phase Retrieval
Qing Qu, Yuqian Zhang, Yonina Eldar, John Wright

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.

************************************

Stochastic and Adversarial Online Learning without Hyperparameters
Ashok Cutkosky, Kwabena A. Boahen

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.

************************************

Masked Autoregressive Flow for Density Estimation
George Papamakarios, Theo Pavlakou, Iain Murray

Autoregressive models are among the best performing neural density estimators. We describe an approach for increasing the flexibility of an autoregressive model, based on modelling the random numbers that the model uses internally when generating data. By constructing a stack of autoregressive models, each modelling the random numbers of the next model in the stack, we obtain a type of normalizing flow suitable for density estimation, which we call Masked Autoregressive Flow. This type of flow is closely related to Inverse Autoregressive Flow and is a generalization of Real NVP. Masked Autoregressive Flow achieves state-of-the-art performance in a range of general-purpose density estimation tasks.

************************************

QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding
Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, Milan Vojnovic

Parallel implementations of stochastic gradient descent (SGD) have received significant research attention, thanks to its excellent scalability properties. A fundamental barrier when parallelizing SGD is the high bandwidth cost of communicating gradient updates between nodes; consequently, several lossy compresion heuristics have been proposed, by which nodes only communicate quantized gradients. Although effective in practice, these heuristics do not always guarantee convergence, and it is not clear whether they can be improved.  In this paper, we propose Quantized SGD (QSGD), a family of compression schemes for gradient updates which provides convergence guarantees. QSGD allows the user to smoothly trade off \emph{communication bandwidth} and \emph{convergence time}: nodes can adjust the number of bits sent per iteration, at the cost of possibly higher variance. We show that this trade-off is inherent, in the sense that improving it past some threshold would violate  information-theoretic lower bounds. QSGD guarantees convergence for convex and non-convex objectives,  under asynchrony, and can be extended to stochastic variance-reduced techniques.   When applied to  training deep neural networks for image classification and  automated speech recognition, QSGD leads to significant reductions in  end-to-end training time. For example, on 16GPUs, we can train the ResNet152  network to full accuracy on ImageNet 1.8x faster than the full-precision  variant.

************************************

Learning Hierarchical Information Flow with Recurrent Neural Modules

Danijar Hafner, Alexander Irpan, James Davidson, Nicolas Heess

We propose ThalNet, a deep learning model inspired by neocortical communication via the thalamus. Our model consists of recurrent neural modules that send features through a routing center, endowing the modules with the flexibility to share features over multiple time steps. We show that our model learns to route information hierarchically, processing input data by a chain of modules. We observe common architectures, such as feed forward neural networks and skip connections, emerging as special cases of our architecture, while novel connectivity patterns are learned for the text8 compression task. Our model outperforms standard recurrent neural networks on several sequential benchmarks.

************************************

Deanonymization in the Bitcoin P2P Network

Giulia Fanti, Pramod Viswanath

Recent attacks on Bitcoin's peer-to-peer (P2P) network demonstrated that its transaction-flooding protocols, which are used to ensure network consistency, may enable user deanonymization---the linkage of a user's IP address with her pseudonym in the Bitcoin network. In 2015, the Bitcoin community responded to these attacks by changing the network's flooding mechanism to a different protocol, known as diffusion. However, it is unclear if diffusion actually improves the system's anonymity. In this paper, we model the Bitcoin networking stack and analyze its anonymity properties, both pre- and post-2015. The core problem is one of epidemic source inference over graphs, where the observational model and spreading mechanisms are informed by Bitcoin's implementation; notably, these models have not been studied in the epidemic source detection literature before. We identify and analyze near-optimal source estimators. This analysis suggests that Bitcoin's networking protocols (both pre- and post-2015) offer poor anonymity properties on networks with a regular-tree topology. We confirm this claim in simulation on a 2015 snapshot of the real Bitcoin P2P network topology.

************************************

Learning with Average Top-k Loss

Yanbo Fan, Siwei Lyu, Yiming Ying, Baogang Hu

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.

************************************

MaskRNN: Instance Level Video Object Segmentation

Yuan-Ting Hu, Jia-Bin Huang, Alexander Schwing

Instance level video object segmentation is an important technique for video editing and compression. To capture the temporal coherence, in this paper, we develop MaskRNN, a recurrent neural net approach which fuses in each frame the output of two deep nets for each object instance - a binary segmentation net providing a mask and a localization net providing a bounding box. Due to the recurrent component and the localization component, our method is able to take advantage of long-term temporal structures of the video data as well as rejecting outliers. We validate the proposed algorithm on three challenging benchmark datasets, the DAVIS-2016 dataset, the DAVIS-2017 dataset, and the Segtrack v2 dataset, achieving state-of-the-art performance on all of them.

************************************

Max-Margin Invariant Features from Transformed Unlabelled Data

Dipan Pal, Ashwin Kannan, Gautam Arakalgud, Marios Savvides

The study of representations invariant to common transformations of the data is important to learning. Most techniques have focused on local approximate invariance implemented within expensive optimization frameworks lacking explicit theoretical guarantees. In this paper, we study kernels that are invariant to a unitary group while having theoretical guarantees in addressing the important practical issue of unavailability of transformed versions of labelled data. A problem we call the Unlabeled Transformation Problem which is a special form of semi-supervised learning and one-shot learning. We present a theoretically motivated alter nate approach to the invariant kernel SVM based on which we propose Max-Margin Invariant Features (MMIF) to solve this problem. As an illustration, we design an framework for face recognition and demonstrate the efficacy of our approach on a large scale semi-synthetic dataset with 153,000 images and a new challenging protocol on Labelled Faces in the Wild (LFW) while out-performing strong baselines.

************************************

Sparse Approximate Conic Hulls

Greg Van Buskirk, Benjamin Raichel, Nicholas Ruozzi

We consider the problem of computing a restricted nonnegative matrix factorization (NMF) of an m\times n matrix X.  Specifically, we seek a factorization X\approx BC, where the k columns of B are a subset of those from X and C\in\Re_{\geq 0}^{k\times n}.  Equivalently, given the matrix X, consider the problem of finding a small subset, S, of the columns of X such that the conic hull of S \eps-approximates the conic hull of the columns of X, i.e., the distance of every column of X to the conic hull of the columns of S should be at most an \eps-fraction of the angular diameter of X.   If k is the size of the smallest \eps-approximation, then we produce an O(k/\eps^{2/3}) sized O(\eps^{1/3})-approximation, yielding the first provable, polynomial time \eps-approximation for this class of NMF problems, where also desirably the approximation is independent of n and m. Furthermore, we prove an approximate conic Carathéodory theorem, a general sparsity result, that shows that any column of X can be \eps-approximated with an O(1/\eps^2) sparse combination from S.   Our results are facilitated by a reduction to the problem of approximating convex hulls,  and we prove that both the convex and conic hull variants are d-sum-hard, resolving an open problem.  Finally, we provide experimental results for the convex and conic algorithms on a variety of feature selection tasks.

************************************

Label Distribution Learning Forests

Wei Shen, KAI ZHAO, Yilu Guo, Alan L. Yuille

Label distribution learning (LDL) is a general learning framework, which assigns to an instance a distribution over a set of labels rather than a single label or multiple labels. Current LDL methods have either restricted assumptions on the expression form of the label distribution or limitations in representation learning, e.g., to learn deep features in an end-to-end manner. This paper presents label distribution learning forests (LDLFs) - a novel label distribution learning algorithm based on differentiable decision trees, which have several advantages: 1) Decision trees have the potential to model any general form of label distributions by a mixture of leaf node predictions. 2) The learning of differentiable decision trees can be combined with representation learning. We define a distribution-based loss function for a forest, enabling all the trees to be learned jointly, and show that an update function for leaf node predictions, which guarantees a strict decrease of the loss function, can be derived by variational bounding. The effectiveness of the proposed LDLFs is verified on several LDL tasks and a computer vision application, showing significant improvements to the state-of-the-art LDL methods.

************************************

Efficient Sublinear-Regret Algorithms for Online Sparse Linear Regression with Limited Observation

Shinji Ito, Daisuke Hatano, Hanna Sumita, Akihiro Yabe, Takuro Fukunaga, Naonori Kakimura, Ken-Ichi Kawarabayashi

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.

************************************

Accelerated First-order Methods for Geodesically Convex Optimization on Riemannian Manifolds

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, Licheng Jiao

In this paper, we propose an accelerated first-order method for geodesically con vex optimization, which is the generalization of the standard Nesterov's accelerated method from Euclidean space to nonlinear Riemannian space. We first derive two equations and obtain two nonlinear operators for geodesically convex optimization instead of the linear extrapolation step in Euclidean space. In particular, we analyze the global convergence properties of our accelerated method for geodesically strongly-convex problems, which show that our method improves the convergence rate from $O((1-\mu/L)^{k})$ to $O((1-\sqrt{\mu/L})^{k})$. Moreover, our method also improves the global convergence rate on geodesically general convex problems from $O(1/k)$ to $O(1/k^{2})$. Finally, we give a specific iterative scheme for matrix Karcher mean problems, and validate our theoretical results with experiments.

************************************

Hierarchical Implicit Models and Likelihood-Free Variational Inference

Dustin Tran, Rajesh Ranganath, David Blei

Implicit probabilistic models are a flexible class of models defined by a simulation process for data. They form the basis for models which encompass our understanding of the physical word. Despite this fundamental nature, the use of implicit models remains limited due to challenge in positing complex latent structure in them, and the ability to inference in such models with large data sets. In this paper, we first introduce the hierarchical implicit models (HIMs). HIMs combine the idea of implicit densities with hierarchical Bayesian modeling thereby defining models via simulators of data with rich hidden structure. Next, we develop likelihood-free variational inference (LFVI), a scalable variational inference algorithm for HIMs. Key to LFVI is specifying a variational family that is also implicit. This matches the model's flexibility and allows for accurate approximation of the posterior. We demonstrate diverse applications: a large-scale physical simulator for predator-prey populations in ecology; a Bayesian generative adversarial network for discrete data; and a deep implicit model for symbol generation.

************************************

Learning the Morphology of Brain Signals Using Alpha-Stable Convolutional Sparse Coding

Mainak Jas, Tom Dupré la Tour, Umut Simsekli, Alexandre Gramfort

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.

************************************

Modulating early visual processing by language

Harm de Vries, Florian Strub, Jeremie Mary, Hugo Larochelle, Olivier Pietquin, Aaron C. Courville

It is commonly assumed that language refers to high-level visual concepts while leaving low-level visual processing unaffected. This view dominates the current literature in computational models for language-vision tasks, where visual and linguistic inputs are mostly processed independently before being fused into a single representation. In this paper, we deviate from this classic pipeline and propose to modulate the \emph{entire visual processing} by a linguistic input. Specifically, we introduce Conditional Batch Normalization (CBN) as an efficient mechanism to modulate convolutional feature maps by a linguistic embedding. We apply CBN to a pre-trained Residual Network (ResNet), leading to the MODulatEd ResNet (\MRN) architecture, and show that this significantly improves strong baselines on two visual question answering tasks. Our ablation study confirms that modulating from the early stages of the visual processing is beneficial.

************************************

Discriminative State Space Models

Vitaly Kuznetsov, Mehryar Mohri

In this paper, we introduce and analyze Discriminative State-Space Models for forecasting non-stationary time series. We provide data-dependent generalization guarantees for learning these models based on the recently introduced notion of discrepancy. We provide an in-depth analysis of the complexity of such models. Fi

nally, we also study the generalization guarantees for several structural risk minimization approaches to this problem and provide an efficient implementation for one of them which is based on a convex objective.
************************************

Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols

Serhii Havrylov, Ivan Titov

Learning to communicate through interaction, rather than relying on explicit supervision, is often considered a prerequisite for developing a general AI. We study a setting where two agents engage in playing a referential game and, from scratch, develop a communication protocol necessary to succeed in this game. Unlike previous work, we require that messages they exchange, both at train and test time, are in the form of a language (i.e. sequences of discrete symbols). We compare a reinforcement learning approach and one using a differentiable relaxation (straight-through Gumbel-softmax estimator) and observe that the latter is much faster to converge and it results in more effective protocols. Interestingly, we also observe that the protocol we induce by optimizing the communication success exhibits a degree of compositionality and variability (i.e. the same information can be phrased in different ways), both properties characteristic of natural languages. As the ultimate goal is to ensure that communication is accomplished in natural language, we also perform experiments where we inject prior information about natural language into our model and study properties of the resulting protocol.
************************************

Wider and Deeper, Cheaper and Faster: Tensorized LSTMs for Sequence Learning

Zhen He, Shaobing Gao, Liang Xiao, Daxue Liu, Hangen He, David Barber

Long Short-Term Memory (LSTM) is a popular approach to boosting the ability of Recurrent Neural Networks to store longer term temporal information. The capacity of an LSTM network can be increased by widening and adding layers. However, usually the former introduces additional parameters, while the latter increases the runtime. As an alternative we propose the Tensorized LSTM in which the hidden states are represented by tensors and updated via a cross-layer convolution. By increasing the tensor size, the network can be widened efficiently without additional parameters since the parameters are shared across different locations in the tensor; by delaying the output, the network can be deepened implicitly with little additional runtime since deep computations for each timestep are merged into temporal computations of the sequence. Experiments conducted on five challenging sequence learning tasks show the potential of the proposed model.
************************************

Online Influence Maximization under Independent Cascade Model with Semi-Bandit Feedback

Zheng Wen, Branislav Kveton, Michal Valko, Sharan Vaswani

We study the online influence maximization problem in social networks under the independent cascade model. Specifically, we aim to learn the set of "best influencers" in a social network online while repeatedly interacting with it. We address the challenges of (i) combinatorial action space, since the number of feasible influencer sets grows exponentially with the maximum number of influencers, and (ii) limited feedback, since only the influenced portion of the network is observed. Under a stochastic semi-bandit feedback, we propose and analyze IMLinUCB, a computationally efficient UCB-based algorithm. Our bounds on the cumulative regret are polynomial in all quantities of interest, achieve near-optimal dependence on the number of interactions and reflect the topology of the network and the activation probabilities of its edges, thereby giving insights on the problem complexity. To the best of our knowledge, these are the first such results. Our experiments show that in several representative graph topologies, the regret of IMLinUCB scales as suggested by our upper bounds. IMLinUCB permits linear generalization and thus is both statistically and computationally suitable for large-scale problems. Our experiments also show that IMLinUCB with linear generalization can lead to low regret in real-world online influence maximization.
************************************

Smooth Primal-Dual Coordinate Descent Algorithms for Nonsmooth Convex Optimization
Ahmet Alacaoglu, Quoc Tran Dinh, Olivier Fercoq, Volkan Cevher
We propose a new randomized coordinate descent method for a convex optimization template  with broad applications. Our analysis relies on a novel combination of four ideas applied to the primal-dual gap function: smoothing, acceleration, homotopy, and coordinate descent with non-uniform sampling. As a result, our method features the first convergence rate guarantees among the coordinate descent methods, that are the best-known under a variety of common structure assumptions on the template. We provide numerical evidence to support the theoretical results with a comparison to state-of-the-art algorithms.
************************************
Linearly constrained Gaussian processes
Carl Jidling, Niklas Wahlström, Adrian Wills, Thomas B. Schön
We consider a modification of the covariance function in Gaussian processes to correctly account for known linear constraints. By modelling the target function as a transformation of an underlying function, the constraints are explicitly incorporated in the model such that they are guaranteed to be fulfilled by any sample drawn or prediction made. We also propose a constructive procedure for designing the transformation operator and illustrate the result on both simulated and real-data examples.
************************************
Solid Harmonic Wavelet Scattering: Predicting Quantum Molecular Energy from Invariant Descriptors of 3D  Electronic Densities
Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, Stephane Mallat
We introduce a solid harmonic wavelet scattering representation, invariant  to rigid motion and stable to deformations, for regression and classification  of 2D and 3D signals. Solid harmonic wavelets are computed by multiplying solid  harmonic functions with Gaussian windows dilated at different scales. Invariant  scattering coefficients are obtained by cascading such wavelet transforms with  the complex modulus nonlinearity. We study an application of solid harmonic  scattering invariants to the estimation of quantum molecular energies, which  are also invariant to rigid motion and stable with respect to deformations. A multilinear regression  over scattering invariants provides close to state of the art results over  small and large databases of organic molecules.
************************************
On Frank-Wolfe and Equilibrium Computation
Jacob D. Abernethy, Jun-Kun Wang
We consider the Frank-Wolfe (FW) method for constrained convex optimization, and we show that this classical technique can be interpreted from a different perspective: FW emerges as the computation of an equilibrium (saddle point) of a special convex-concave zero sum game. This saddle-point trick relies on the existence of no-regret online learning to both generate a sequence of iterates but also to provide a proof of convergence through vanishing regret. We show that our stated equivalence has several nice properties, as it exhibits a modularity that gives rise to various old and new algorithms. We explore a few such resulting methods, and provide experimental results to demonstrate correctness and efficiency.
************************************
Generalizing GANs: A Turing Perspective
Roderich Gross, Yue Gu, Wei Li, Melvin Gauci
Recently, a new class of machine learning algorithms has emerged, where models and discriminators are generated in a competitive setting. The most prominent example is Generative Adversarial Networks (GANs). In this paper we examine how these algorithms relate to the Turing test, and derive what - from a Turing perspective - can be considered their defining features. Based on these features, we outline directions for generalizing GANs - resulting in the family of algorithms referred to as Turing Learning. One such direction is to allow the discriminators to interact with the processes from which the data samples are obtained, making them "interrogators", as in the Turing test. We validate this idea using two case studies. In the first case study, a computer infers the behavior of an agent

while controlling its environment. In the second case study, a robot infers its own sensor configuration while controlling its movements. The results confirm that by allowing discriminators to interrogate, the accuracy of models is improved.

************************************

Predicting Scene Parsing and Motion Dynamics in the Future

Xiaojie Jin, Huaxin Xiao, Xiaohui Shen, Jimei Yang, Zhe Lin, Yunpeng Chen, Zequn Jie, Jiashi Feng, Shuicheng Yan

It is important for intelligent systems, e.g. autonomous vehicles and robotics to anticipate the future in order to plan early and make decisions accordingly. Predicting the future scene parsing and motion dynamics helps the agents better understand the visual environment better as the former provides dense semantic segmentations, i.e. what objects will be present and where they will appear, while the latter provides dense motion information, i.e. how the objects move in the future. In this paper, we propose a novel model to predict the scene parsing and motion dynamics in unobserved future video frames simultaneously. Using history information (preceding frames and corresponding scene parsing results) as input, our model is able to predict the scene parsing and motion for arbitrary time steps ahead. More importantly, our model is superior compared to other methods that predict parsing and motion separately, as the complementary relationship between the two tasks are fully utilized in our model through joint learning. To our best knowledge, this is the first attempt in jointly predicting scene parsing and motion dynamics in the future frames. On the large-scale Cityscapes dataset, it is demonstrated that our model produces significantly better parsing and motion prediction results compared to well established baselines. In addition, we also show our model can be used to predict the steering angle of the vehicles, which further verifies the ability of our model to learn underlying latent parameters.

************************************

A Screening Rule for l1-Regularized Ising Model Estimation

Zhaobin Kuang, Sinong Geng, David Page

We discover a screening rule for l1-regularized Ising model estimation. The simple closed-form screening rule is a necessary and sufficient condition for exactly recovering the blockwise structure of a solution under any given regularization parameters. With enough sparsity, the screening rule can be combined with various optimization procedures to deliver solutions efficiently in practice. The screening rule is especially suitable for large-scale exploratory data analysis, where the number of variables in the dataset can be thousands while we are only interested in the relationship among a handful of variables within moderate-size clusters for interpretability. Experimental results on various datasets demonstrate the efficiency and insights gained from the introduction of the screening rule.

************************************

A Minimax Optimal Algorithm for Crowdsourcing

Thomas Bonald, Richard Combes

We consider the problem of accurately estimating the reliability of workers based on noisy labels they provide, which is a fundamental question in crowdsourcing. We propose a novel lower bound on the minimax estimation error which applies to any estimation procedure. We further propose Triangular Estimation (TE), an algorithm for estimating the reliability of workers. TE has low complexity, may be implemented in a streaming setting when labels are provided by workers in real time, and does not rely on an iterative procedure. We prove that TE is minimax optimal and matches our lower bound. We conclude by assessing the performance of TE and other state-of-the-art algorithms on both synthetic and real-world data.

************************************

Communication-Efficient Distributed Learning of Discrete Distributions

Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, Ludwig Schmidt

We initiate a systematic investigation of distribution learning (density estimation) when the data is distributed across multiple servers. The servers must comm unicate with a referee and the goal is to estimate the underlying distribution with as few bits of communication as possible. We focus on non-parametric density estimation of discrete distributions with respect to the l1 and l2 norms. We provide the first non-trivial upper and lower bounds on the communication complexity of this basic estimation task in various settings of interest. Specifically, our results include the following:   1. When the unknown discrete distribution is unstructured and each server has only one sample, we show that any blackboard protocol (i.e., any protocol in which servers interact arbitrarily using public messages) that learns the distribution must essentially communicate the entire sample.  2. For the case of structured distributions, such as k-histograms and monotone distributions, we design distributed learning algorithms that achieve significantly better communication guarantees than the naive ones, and obtain tight upper and lower bounds in several regimes. Our distributed learning algorithms run in near-linear time and are robust to model misspecification.  Our results provide insights on the interplay between structure and communication efficiency for a range of fundamental distribution estimation tasks.

************************************

### VAIN: Attentional Multi-agent Predictive Modeling

Yedid Hoshen

Multi-agent predictive modeling is an essential step for understanding physical, social and team-play systems. Recently, Interaction Networks (INs) were proposed for the task of modeling multi-agent physical systems. One of the drawbacks of INs is scaling with the number of interactions in the system (typically quadratic or higher order in the number of agents). In this paper we introduce VAIN, a novel attentional architecture for multi-agent predictive modeling that scales linearly with the number of agents. We show that VAIN is effective for multi-agent predictive modeling. Our method is evaluated on tasks from challenging multi-agent prediction domains: chess and soccer, and outperforms competing multi-agent approaches.

************************************

### Hierarchical Attentive Recurrent Tracking

Adam Kosiorek, Alex Bewley, Ingmar Posner

Class-agnostic object tracking is particularly difficult in cluttered environments as target specific discriminative models cannot be learned a priori. Inspired by how the human visual cortex employs spatial attention and separate where'' andwhat'' processing pathways to actively suppress irrelevant visual features, this work develops a hierarchical attentive recurrent model for single object tracking in videos. The first layer of attention discards the majority of background by selecting a region containing the object of interest, while the subsequent layers tune in on visual features particular to the tracked object.    This framework is fully differentiable and can be trained in a purely data driven fashion by gradient methods. To improve training convergence, we augment the loss function with terms for auxiliary tasks relevant for tracking. Evaluation of the proposed model is performed on two datasets: pedestrian tracking on the KTH activity recognition dataset and the more difficult KITTI object tracking dataset.

************************************

### Sobolev Training for Neural Networks

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, Razvan Pascanu

At the heart of deep learning we aim to use neural networks as function approximators -  training them to produce outputs from inputs in emulation of a ground truth function or data creation process. In many cases we only have access to input-output pairs from the ground truth, however it is becoming more common to have access to derivatives of the target output with respect to the input -- for example when the ground truth function is itself a neural network such as in network compression or distillation.  Generally these target derivatives are not computed, or are ignored. This paper introduces Sobolev Training for neural networks, which is a method for incorporating these target derivatives in addition the to target values while training. By optimising neural networks to not only approximate the function’s outputs but also the function’s derivatives we encode addit ional information about the target function within the parameters of the neural network. Thereby we can improve the quality of our predictors, as well as the data-efficiency and generalization capabilities of our learned function approximation. We provide theoretical justifications for such an approach as well as examples of empirical evidence on three distinct domains: regression on classical optimisation datasets, distilling policies of an agent playing Atari, and on large-scale applications of synthetic gradients.  In all three domains the use of Sobolev Training, employing target derivatives in addition to target values, results in models with higher accuracy and stronger generalisation.

************************************

Doubly Accelerated Stochastic Variance Reduced Dual Averaging Method for Regularized Empirical Risk Minimization

Tomoya Murata, Taiji Suzuki

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.

************************************

Learning with Feature Evolvable Streams

Bo-Jian Hou, Lijun Zhang, Zhi-Hua Zhou

Learning with streaming data has attracted much attention during the past few years.Though most studies consider data stream with fixed features, in real practice the features may be evolvable. For example, features of data gathered by limited lifespan sensors will change when these sensors are substituted by new ones. In this paper, we propose a novel learning paradigm: Feature Evolvable Streaming Learning where old features would vanish and new features would occur. Rather than relying on only the current features, we attempt to recover the vanished features and exploit it to improve performance. Specifically, we learn two models from the recovered features and the current features, respectively. To benefit from the recovered features, we develop two ensemble methods. In the first method, we combine the predictions from two models and theoretically show that with the assistance of old features, the performance on new features can be improved. In the second approach, we dynamically select the best single prediction and establish a better performance guarantee when the best model switches. Experiments on both synthetic and real data validate the effectiveness of our proposal.

************************************

Safe Model-based Reinforcement Learning with Stability Guarantees

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, Andreas Krause

Reinforcement learning is a powerful paradigm for learning optimal policies from experimental data. However, to find optimal policies, most reinforcement learning algorithms explore all possible actions, which may be harmful for real-world systems. As a consequence, learning algorithms are rarely applied on safety-critical systems in the real world. In this paper, we present a learning algorithm that explicitly considers safety, defined in terms of stability guarantees. Specifically, we extend control-theoretic results on Lyapunov stability verification and show how to use statistical models of the dynamics to obtain high-performance control policies with provable stability certificates. Moreover, under additional regularity assumptions in terms of a Gaussian process prior, we prove that one can effectively and safely collect data in order to learn about the dynamics and thus both improve control performance and expand the safe region of the state space. In our experiments, we show how the resulting algorithm can safely optimize a neural network policy on a simulated inverted pendulum, without the pendulum ever falling down.

************************************

Time-dependent spatially varying graphical models, with application to brain fMRI data analysis

Kristjan Greenewald, Seyoung Park, Shuheng Zhou, Alexander Giessing

In this work, we present an additive model for space-time data that splits the data into a temporally correlated component and a spatially correlated component. We model the spatially correlated portion using a time-varying Gaussian graphic al model. Under assumptions on the smoothness of changes in covariance matrices, we derive strong single sample convergence results, confirming our ability to estimate meaningful graphical structures as they evolve over time. We apply our methodology to the discovery of time-varying spatial structures in human brain fMRI signals.

************************************

Clone MCMC: Parallel High-Dimensional Gaussian Gibbs Sampling
Andrei-Cristian Barbos, Francois Caron, Jean-François Giovannelli, Arnaud Doucet

We propose a generalized Gibbs sampler algorithm for obtaining samples approximately distributed from a high-dimensional Gaussian distribution. Similarly to Hogwild methods, our approach does not target the original Gaussian distribution of interest, but an approximation to it. Contrary to Hogwild methods, a single parameter allows us to trade bias for variance. We show empirically that our method is very flexible and performs well compared to Hogwild-type algorithms.

************************************

Context Selection for Embedding Models
Liping Liu, Francisco Ruiz, Susan Athey, David Blei

Word embeddings are an effective tool to analyze language. They have been recently extended to model other types of data beyond text, such as items in recommendation systems. Embedding models consider the probability of a target observation (a word or an item) conditioned on the elements in the context (other words or items). In this paper, we show that conditioning on all the elements in the context is not optimal. Instead, we model the probability of the target conditioned on a learned subset of the elements in the context. We use amortized variational inference to automatically choose this subset. Compared to standard embedding models, this method improves predictions and the quality of the embeddings.

************************************

Union of Intersections (UoI) for Interpretable Data Driven Discovery and Prediction
Kristofer Bouchard, Alejandro Bujan, Fred Roosta, Shashanka Ubaru, Mr. Prabhat, Antoine Snijders, Jian-Hua Mao, Edward Chang, Michael W. Mahoney, Sharmodeep Bhattacharya

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.

************************************

Good Semi-supervised Learning That Requires a Bad GAN
Zihang Dai, Zhilin Yang, Fan Yang, William W. Cohen, Russ R. Salakhutdinov

Semi-supervised learning methods based on generative adversarial networks (GANs) obtained strong empirical results, but it is not clear 1) how the discriminator benefits from joint training with a generator, and 2) why good semi-supervised classification performance and a good generator cannot be obtained at the same time. Theoretically we show that given the discriminator objective, good semi-supervised learning indeed requires a bad generator, and propose the definition of a preferred generator. Empirically, we derive a novel formulation based on our analysis that substantially improves over feature matching GANs, obtaining state-of-the-art results on multiple benchmark datasets.

************************************

Targeting EEG/LFP Synchrony with Neural Nets
Yitong Li, michael Murias, samantha Major, geraldine Dawson, Kafui Dzirasa, Lawrence Carin, David E. Carlson

We consider the analysis of Electroencephalography (EEG) and Local Field Potential (LFP) datasets, which are “big” in terms of the size of recorded data but rarely have sufficient labels required to train complex models (e.g., conventional deep learning methods).  Furthermore, in many scientific applications, the goal is to be able to understand the underlying features related to the classification, which prohibits the blind application of deep networks. This motivates the development of a new model based on {\em parameterized} convolutional filters guided by previous neuroscience research; the filters learn relevant frequency bands

while targeting synchrony, which are frequency-specific power and phase correlations between electrodes. This results in a highly expressive convolutional neural network with only a few hundred parameters, applicable to smaller datasets. The proposed approach is demonstrated to yield competitive (often state-of-the-art) predictive performance during our empirical tests while yielding interpretable features.  Furthermore, a Gaussian process adapter is developed to combine analysis over distinct electrode layouts, allowing the joint processing of multiple datasets to address overfitting and improve generalizability.  Finally, it is demonstrated that the proposed framework effectively tracks neural dynamics on children in a clinical trial on Autism Spectrum Disorder.
************************************
Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes
Anton Mallasto, Aasa Feragen
We introduce a novel framework for statistical analysis of populations of non-degenerate Gaussian processes (GPs), which are natural representations of uncertain curves. This allows inherent variation or uncertainty in function-valued data to be properly incorporated in the population analysis. Using the 2-Wasserstein metric we geometrize the space of GPs with L2 mean and covariance functions over compact index spaces. We prove uniqueness of the barycenter of a population of GPs, as well as convergence of the metric and the barycenter of their finite-dimensional counterparts. This justifies practical computations. Finally, we demonstrate our framework through experimental validation on GP datasets representing brain connectivity and climate development. A Matlab library for relevant computations will be published at https://sites.google.com/view/antonmallasto/software.
************************************
Online Dynamic Programming
Holakou Rahmanian, Manfred K. K. Warmuth
We consider the problem of repeatedly solving a variant of the same dynamic programming problem in successive trials. An instance of the type of problems we consider is to find a good binary search tree in a changing environment. At the beginning of each trial, the learner probabilistically chooses a tree with the n keys at the internal nodes and the n + 1 gaps between keys at the leaves. The learner is then told the frequencies of the keys and gaps and is charged by the average search cost for the chosen tree. The problem is online because the frequencies can change between trials. The goal is to develop algorithms with the property that their total average search cost (loss) in all trials is close to the total loss of the best tree chosen in hindsight for all trials. The challenge, of course, is that the algorithm has to deal with exponential number of trees. We develop a general methodology for tackling such problems for a wide class of dynamic programming algorithms. Our framework allows us to extend online learning algorithms like Hedge and Component Hedge to a significantly wider class of combinatorial objects than was possible before.
************************************
Neural Discrete Representation Learning
Aaron van den Oord, Oriol Vinyals, koray kavukcuoglu
Learning useful representations without supervision remains a key challenge in machine learning. In this paper, we propose a simple yet powerful generative model that learns such discrete representations. Our model, the Vector Quantised-Variational AutoEncoder (VQ-VAE), differs from VAEs in two key ways: the encoder network outputs discrete, rather than continuous, codes; and the prior is learnt rather than static. In order to learn a discrete latent representation, we incorporate ideas from vector quantisation (VQ). Using the VQ method allows the model to circumvent issues of ``posterior collapse'' -— where the latents are ignored when they are paired with a powerful autoregressive decoder -— typically observed in the VAE framework. Pairing these representations with an autoregressive prior, the model can generate high quality images, videos, and speech as well as doing high quality speaker conversion and unsupervised learning of phonemes, providing further evidence of the utility of the learnt representations.
************************************

Probabilistic Rule Realization and Selection

Haizi Yu, Tianxi Li, Lav R. Varshney

Abstraction and realization are bilateral processes that are key in deriving intelligence and creativity. In many domains, the two processes are approached through \emph{rules}: high-level principles that reveal invariances within similar yet diverse examples. Under a probabilistic setting for discrete input spaces, we focus on the rule realization problem which generates input sample distributions that follow the given rules. More ambitiously, we go beyond a mechanical realization that takes whatever is given, but instead ask for proactively selecting reasonable rules to realize. This goal is demanding in practice, since the initial rule set may not always be consistent and thus intelligent compromises are needed. We formulate both rule realization and selection as two strongly connected components within a single and symmetric bi-convex problem, and derive an efficient algorithm that works at large scale. Taking music compositional rules as the main example throughout the paper, we demonstrate our model's efficiency in not only music realization (composition) but also music interpretation and understanding (analysis).

************************************

A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, Ole Winther

This paper takes a step towards temporal reasoning in a dynamically changing video, not in the pixel space that constitutes its frames, but in a latent space that describes the non-linear dynamics of the objects in its world. We introduce the Kalman variational auto-encoder, a framework for unsupervised learning of sequential data that disentangles two latent representations: an object's representation, coming from a recognition model, and a latent state describing its dynamics. As a result, the evolution of the world can be imagined and missing data imputed, both without the need to generate high dimensional frames at each time step. The model is trained end-to-end on videos of a variety of simulated physical systems, and outperforms competing methods in generative and missing data imputation tasks.

************************************

Stabilizing Training of Generative Adversarial Networks through Regularization

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, Thomas Hofmann

Deep generative models based on Generative Adversarial Networks (GANs) have demonstrated impressive sample quality but in order to work they require a careful choice of architecture, parameter initialization, and selection of hyper-parameters. This fragility is in part due to a dimensional mismatch or non-overlapping support between the model distribution and the data distribution, causing their density ratio and the associated f -divergence to be undefined. We overcome this fundamental limitation and propose a new regularization approach with low computational cost that yields a stable GAN training procedure. We demonstrate the effectiveness of this regularizer accross several architectures trained on common benchmark image generation tasks. Our regularization turns GAN models into reliable building blocks for deep learning.

************************************

Training Deep Networks without Learning Rates Through Coin Betting

Francesco Orabona, Tatiana Tommasi

Deep learning methods achieve state-of-the-art performance in many application scenarios. Yet, these methods require a significant amount of hyperparameters tuning in order to achieve the best results. In particular, tuning the learning rates in the stochastic optimization process is still one of the main bottlenecks. In this paper, we propose a new stochastic gradient descent procedure for deep networks that does not require any learning rate setting. Contrary to previous methods, we do not adapt the learning rates nor we make use of the assumed curvature of the objective function. Instead, we reduce the optimization process to a game of betting on a coin and propose a learning rate free optimal algorithm for this scenario. Theoretical convergence is proven for convex and quasi-convex functions and empirical evidence shows the advantage of our algorithm over p opular stochastic gradient algorithms.

************************************

Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis

Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, Jiashi Feng

Synthesizing realistic profile faces is promising for more efficiently training deep pose-invariant models for large-scale unconstrained face recognition, by populating samples with extreme poses and avoiding tedious annotations. However, learning from synthetic faces may not achieve the desired performance due to the discrepancy between distributions of the synthetic and real face images. To narrow this gap, we propose a Dual-Agent Generative Adversarial Network (DA-GAN) model, which can improve the realism of a face simulator's output using unlabeled real faces, while preserving the identity information during the realism refinement. The dual agents are specifically designed for distinguishing real v.s. fake and identities simultaneously. In particular, we employ an off-the-shelf 3D face model as a simulator to generate profile face images with varying poses. DA-GAN leverages a fully convolutional network as the generator to generate high-resolution images and an auto-encoder as the discriminator with the dual agents. Besides the novel architecture, we make several key modifications to the standard GAN to preserve pose and texture, preserve identity and stabilize training process: (i) a pose perception loss; (ii) an identity perception loss; (iii) an adversarial loss with a boundary equilibrium regularization term. Experimental results show that DA-GAN not only presents compelling perceptual results but also significantly outperforms state-of-the-arts on the large-scale and challenging NIST IJB-A unconstrained face recognition benchmark. In addition, the proposed DA-GAN is also promising as a new approach for solving generic transfer learning problems more effectively.

************************************

Thy Friend is My Friend: Iterative Collaborative Filtering for Sparse Matrix Estimation

Christian Borgs, Jennifer Chayes, Christina E. Lee, Devavrat Shah

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.

************************************

Positive-Unlabeled Learning with Non-Negative Risk Estimator

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, Masashi Sugiyama

From only positive (P) and unlabeled (U) data, a binary classifier could be trained with PU learning, in which the state of the art is unbiased PU learning. However, if its model is very flexible, empirical risks on training data will go negative, and we will suffer from serious overfitting. In this paper, we propose a non-negative risk estimator for PU learning: when getting minimized, it is more robust against overfitting, and thus we are able to use very flexible models (such as deep neural networks) given limited P data. Moreover, we analyze the bias, consistency, and mean-squared-error reduction of the proposed risk estimator, and bound the estimation error of the resulting empirical risk minimizer. Experiments demonstrate that our risk estimator fixes the overfitting problem of its unbiased counterparts.

************************************

Gradient descent GAN optimization is locally stable

Vaishnavh Nagarajan, J. Zico Kolter

Despite the growing prominence of generative adversarial networks (GANs), optimization in GANs is still a poorly understood topic. In this paper, we analyze the ``gradient descent'' form of GAN optimization, i.e., the natural setting where we simultaneously take small gradient steps in both generator and discriminator parameters. We show that even though GAN optimization does \emph{not} correspond to a convex-concave game (even for simple parameterizations), un der   proper conditions, equilibrium points of this  optimization    procedure are still \emph{locally asymptotically stable} for the traditional   GAN formulation. On the other hand, we show that the recently proposed   Wasserstein GAN can have non-convergent limit cycles near equilibrium.   Motivated by this stability analysis, we    propose an additional regularization term for gradient descent GAN updates,   which \emph{is} able to guarantee local stability for both the WGAN and the   traditional GAN, and also shows practical promise in speeding up convergence and addressing mode collapse.

************************************

Faster and Non-ergodic O(1/K) Stochastic Alternating Direction Method of Multipliers

Cong Fang, Feng Cheng, Zhouchen Lin

We study stochastic convex optimization subjected to linear equality constraints. Traditional Stochastic Alternating Direction Method of Multipliers and its Nesterov's acceleration scheme can only achieve ergodic O(1/\sqrt{K}) convergence rates, where K is the number of iteration. By introducing Variance Reduction (VR) techniques, the convergence rates improve to ergodic O(1/K). In this paper, we propose a new stochastic ADMM which elaborately integrates Nesterov's extrapolation and VR techniques. With Nesterov’s extrapolation, our algorithm can achieve a non-ergodic O(1/K) convergence rate which is optimal for separable linearly constrained non-smooth convex problems, while the convergence rates of VR based ADMM methods are actually tight O(1/\sqrt{K}) in non-ergodic sense. To the best of our knowledge, this is the first work that achieves a truly accelerated, stochastic convergence rate for constrained convex problems. The experimental results demonstrate that our algorithm is significantly faster than the existing state-of-the-art stochastic ADMM methods.

************************************

Group Sparse Additive Machine

Hong Chen, Xiaoqian Wang, Cheng Deng, Heng Huang

A family of learning algorithms generated from additive models have attracted much attention recently for their flexibility and interpretability in high dimensional data analysis. Among them, learning models with grouped variables have shown competitive performance for prediction and variable selection. However, the previous works mainly focus on the least squares regression problem, not the classification task. Thus, it is desired to design the new additive classification model with variable selection capability for many real-world applications which focus on high-dimensional data classification. To address this challenging problem, in this paper, we investigate the classification with group sparse additive models in reproducing kernel Hilbert spaces. A novel classification method, called as \emph{group sparse additive machine} (GroupSAM), is proposed to explore and utilize the structure information among the input variables. Generalization error bound is derived and proved by integrating the sample error analysis with empirical covering numbers and the hypothesis error estimate with the stepping stone technique. Our new bound shows that GroupSAM can achieve a satisfactory learning rate with polynomial decay. Experimental results on synthetic data and seven benchmark datasets consistently show the effectiveness of our new approach.

************************************

PixelGAN Autoencoders

Alireza Makhzani, Brendan J. Frey

In this paper, we describe the "PixelGAN autoencoder", a generative autoencoder in which the generative path is a convolutional autoregressive neural network on pixels (PixelCNN) that is conditioned on a latent code, and the recognition path uses a generative adversarial network (GAN) to impose a prior distribution on the latent code. We show that different priors result in different decompositions of information between the latent code and the autoregressive decoder. For example, by imposing a Gaussian distribution as the prior, we can achieve a global vs. local decomposition, or by imposing a categorical distribution as the prior, we can disentangle the style and content information of images in an unsupervised fashion. We further show how the PixelGAN autoencoder with a categorical prior can be directly used in semi-supervised settings and achieve competitive semi-supervised classification results on the MNIST, SVHN and NORB datasets.

************************************

Excess Risk Bounds for the Bayes Risk using Variational Inference in Latent Gaussian Models

Rishit Sheth, Roni Khardon

Bayesian models are established as one of the main successful paradigms for complex problems in machine learning. To handle intractable inference, research in this area has developed new approximation methods that are fast and effective. However, theoretical analysis of the performance of such approximations is not well developed. The paper furthers such analysis by providing bounds on the excess risk of variational inference algorithms and related regularized loss minimization algorithms for a large class of latent variable models with Gaussian latent variables. We strengthen previous results for variational algorithms by showing they are competitive with any point-estimate predictor. Unlike previous work, we also provide bounds on the risk of the \emph{Bayesian} predictor and not just the risk of the Gibbs predictor for the same approximate posterior. The bounds are applied in complex models including sparse Gaussian processes and correlated topic models. Theoretical results are complemented by identifying novel approximations to the Bayesian objective that attempt to minimize the risk directly. An empirical evaluation compares the variational and new algorithms shedding further light on their performance.

************************************

Online control of the false discovery rate with decaying memory

Aaditya Ramdas, Fanny Yang, Martin J. Wainwright, Michael I. Jordan

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.

************************************

Safe and Nested Subgame Solving for Imperfect-Information Games

Noam Brown, Tuomas Sandholm

In imperfect-information games, the optimal strategy in a subgame may depend on the strategy in other, unreached subgames. Thus a subgame cannot be solved in isolation and must instead consider the strategy for the entire game as a whole, unlike perfect-information games. Nevertheless, it is possible to first approximate a solution for the whole game and then improve it in individual subgames. This is referred to as subgame solving. We introduce subgame-solving techniques that outperform prior methods both in theory and practice. We also show how to adapt them, and past subgame-solving techniques, to respond to opponent actions that are outside the original action abstraction; this significantly outperforms the prior state-of-the-art approach, action translation. Finally, we show that subgame solving can be repeated as the game progresses down the game tree, leading to far lower exploitability. These techniques were a key component of Libratus, the first AI to defeat top humans in heads-up no-limit Texas hold'em poker.

************************************

A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent

Ben London

We study the generalization error of randomized learning algorithms -- focusing on stochastic gradient descent (SGD) -- using a novel combination of PAC-Bayes and algorithmic stability. Importantly, our generalization bounds hold for all posterior distributions on an algorithm's random hyperparameters, including distributions that depend on the training data. This inspires an adaptive sampling algorithm for SGD that optimizes the posterior at runtime. We analyze this algorithm in the context of our generalization bounds and evaluate it on a benchmark dataset. Our experiments demonstrate that adaptive sampling can reduce empirical risk faster than uniform sampling while also improving out-of-sample accuracy.

************************************

Dynamic Safe Interruptibility for Decentralized Multi-Agent Reinforcement Learning

El Mahdi El Mhamdi, Rachid Guerraoui, Hadrien Hendrikx, Alexandre Maurer

In reinforcement learning, agents learn by performing actions and observing their outcomes. Sometimes, it is desirable for a human operator to interrupt an agent in order to prevent dangerous situations from happening. Yet, as part of their learning process, agents may link these interruptions, that impact their reward, to specific states and deliberately avoid them. The situation is particularly challenging in a multi-agent context because agents might not only learn from their own past interruptions, but also from those of other agents. Orseau and Armstrong defined safe interruptibility for one learner, but their work does not naturally extend to multi-agent systems. This paper introduces dynamic safe interruptibility, an alternative definition more suited to decentralized learning problems, and studies this notion in two learning frameworks: joint action learners and independent learners. We give realistic sufficient conditions on the learning algorithm to enable dynamic safe interruptibility in the case of joint action learners, yet show that these conditions are not sufficient for independent learners. We show however that if agents can detect interruptions, it is possible to prune the observations to ensure dynamic safe interruptibility even for independent learners.

************************************

Toward Multimodal Image-to-Image Translation

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, Eli Shechtman

Many image-to-image translation problems are ambiguous, as a single input image may correspond to multiple possible outputs. In this work, we aim to model a distribution of possible outputs in a conditional generative modeling setting. The ambiguity of the mapping is distilled in a low-dimensional latent vector, which can be randomly sampled at test time. A generator learns to map the given input, combined with this latent code, to the output. We explicitly encourage the connection between output and the latent code to be invertible. This helps prevent a many-to-one mapping from the latent code to the output during training, also known as the problem of mode collapse, and produces more diverse results. We explore several variants of this approach by employing different training objectives, network architectures, and methods of injecting the latent code. Our proposed method encourages bijective consistency between the latent encoding and output modes. We present a systematic comparison of our method and other variants on both perceptual realism and diversity.

************************************

The Marginal Value of Adaptive Gradient Methods in Machine Learning

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, Benjamin Recht

Adaptive optimization methods, which perform local optimization with a metric constructed from the history of iterates, are becoming increasingly popular for training deep neural networks.  Examples include AdaGrad, RMSProp, and Adam. We show that for simple overparameterized problems, adaptive methods often find drastically different solutions than gradient descent (GD) or stochastic gradient descent (SGD).  We construct an illustrative binary classification problem where the data is linearly separable, GD and SGD achieve zero test error, and AdaGrad, Adam, and RMSProp attain test errors arbitrarily close to half.  We additionally study the empirical generalization capability of adaptive methods on several state-of-the-art deep learning models. We observe that the solutions found by adaptive methods generalize worse (often significantly worse) than SGD, even when these solutions have better training performance. These results suggest that practitioners should reconsider the use of adaptive methods to train neural networks.

************************************

Mean Field Residual Networks: On the Edge of Chaos

Ge Yang, Samuel Schoenholz

We study randomly initialized residual networks using mean field theory and the theory of difference equations. Classical feedforward neural networks, such as those with tanh activations, exhibit exponential behavior on the average when propagating inputs forward or gradients backward. The exponential forward dynamics causes rapid collapsing of the input space geometry, while the exponential backw ard dynamics causes drastic vanishing or exploding gradients. We show, in contrast, that by adding skip connections, the network will, depending on the nonlinearity, adopt subexponential forward and backward dynamics, and in many cases in fact polynomial. The exponents of these polynomials are obtained through analytic methods and proved and verified empirically to be correct. In terms of the "edge of chaos" hypothesis, these subexponential and polynomial laws allow residual networks to "hover over the boundary between stability and chaos," thus preserving the geometry of the input space and the gradient information flow. In our experiments, for each activation function we study here, we initialize residual networks with different hyperparameters and train them on MNIST. Remarkably, our initialization time theory can accurately predict test time performance of these networks, by tracking either the expected amount of gradient explosion or the expected squared distance between the images of two input vectors. Importantly, we show, theoretically as well as empirically, that common initializations such as the Xavier or the He schemes are not optimal for residual networks, because the optimal initialization variances depend on the depth. Finally, we have made mathematical contributions by deriving several new identities for the kernels of powers of ReLU functions by relating them to the zeroth Bessel function of the second kind.

************************************

Non-convex Finite-Sum Optimization Via SCSG Methods

Lihua Lei, Cheng Ju, Jianbo Chen, Michael I. Jordan

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.

************************************

First-Order Adaptive Sample Size Methods to Reduce Complexity of Empirical Risk Minimization

Aryan Mokhtari, Alejandro Ribeiro

This paper studies empirical risk minimization (ERM) problems for large-scale datasets and incorporates the idea of adaptive sample size methods to improve the guaranteed convergence bounds for first-order stochastic and deterministic methods. In contrast to traditional methods that attempt to solve the ERM problem corresponding to the full dataset directly, adaptive sample size schemes start with a small number of samples and solve the corresponding ERM problem to its statistical accuracy. The sample size is then grown geometrically -- e.g., scaling by a factor of two -- and use the solution of the previous ERM as a warm start for the new ERM. Theoretical analyses show that the use of adaptive sample size methods reduces the overall computational cost of achieving the statistical accuracy of the whole dataset for a broad range of deterministic and stochastic first-order methods. The gains are specific to the choice of method. When particularized to, e.g., accelerated gradient descent and stochastic variance reduce gradient, the computational cost advantage is a logarithm of the number of training samples. Numerical experiments on various datasets confirm theoretical claims and showcase the gains of using the proposed adaptive sample size scheme.

************************************

Doubly Stochastic Variational Inference for Deep Gaussian Processes

Hugh Salimbeni, Marc Deisenroth

Deep Gaussian processes (DGPs) are multi-layer generalizations of GPs, but inference in these models has proved challenging. Existing approaches to inference in DGP models assume approximate posteriors that force independence between the layers, and do not work well in practice. We present a doubly stochastic variational inference algorithm, which does not force independence between layers. With our method of inference we demonstrate that a DGP model can be used effectively on data ranging in size from hundreds to a billion points. We provide strong empirical evidence that our inference scheme for DGPs works well in practice in both classification and regression.

************************************

From Parity to Preference-based Notions of Fairness in Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, Adrian Weller

The adoption of automated, data-driven decision making in an ever expanding range of applications has raised concerns about its potential unfairness towards certain social groups. In this context, a number of recent studies have focused on defining, detecting, and removing unfairness from data-driven decision systems. However, the existing notions of fairness, based on parity (equality) in treatment or outcomes for different social groups, tend to be quite stringent, limiting the overall decision making accuracy. In this paper, we draw inspiration from the fair-division and envy-freeness literature in economics and game theory and propose preference-based notions of fairness -- given the choice between various sets of decision treatments or outcomes, any group of users would collectively prefer its treatment or outcomes, regardless of the (dis)parity as compared to the other groups. Then, we introduce tractable proxies to design margin-based classifiers that satisfy these preference-based notions of fairness. Finally, we experiment with a variety of synthetic and real-world datasets and show that preference-based fairness allows for greater decision accuracy than parity-based fairness.

************************************

Nonparametric Online Regression while Learning the Metric

Ilja Kuzborskij, Nicolò Cesa-Bianchi

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.

************************************

Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite Sum Structure

Alberto Bietti, Julien Mairal

Stochastic optimization algorithms with variance reduction have proven successful for minimizing large finite sums of functions. Unfortunately, these techniques are unable to deal with stochastic perturbations of input data, induced for example by data augmentation. In such cases, the objective is no longer a finite sum, and the main candidate for optimization is the stochastic gradient descent method (SGD). In this paper, we introduce a variance reduction approach for these settings when the objective is composite and strongly convex. The convergence rate outperforms SGD with a typically much smaller constant factor, which depends on the variance of gradient estimates only due to perturbations on a single example.

************************************

Working hard to know your neighbor's margins: Local descriptor learning loss

Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, Jiri Matas

We introduce a loss for metric learning, which is inspired by the Lowe's matching criterion for SIFT. We show that the proposed loss, that maximizes the distance between the closest positive and closest negative example in the batch, is better than complex regularization methods; it works well for both shallow and deep convolution network architectures. Applying the novel loss to the L2Net CNN architecture results in a compact descriptor named HardNet. It has the same dimensionality as SIFT (128) and shows state-of-art performance in wide baseline stereo, patch verification and instance retrieval benchmarks.

************************************

Hiding Images in Plain Sight: Deep Steganography

Shumeet Baluja

Steganography is the practice of concealing a secret message within another, ordinary, message.  Commonly, steganography is used to unobtrusively hide a small message within the noisy regions of a larger image.  In this study, we attempt to place a full size color image within another image of the same size.  Deep neural networks are simultaneously trained to create the hiding and revealing processes and are designed to specifically work as a pair.  The system is trained on images drawn randomly from the ImageNet database, and works well on natural image s from a wide variety of sources.  Beyond demonstrating the successful application of deep learning to hiding images, we carefully examine how the result is achieved and explore extensions.  Unlike many popular steganographic methods that encode the secret message within the least significant bits of the carrier image, our approach compresses and distributes the secret image's representation across all of the available bits.

************************************

Lookahead  Bayesian Optimization with Inequality Constraints
Remi Lam, Karen Willcox

We consider the task of optimizing an objective function subject to inequality constraints when both the objective and the constraints are expensive to evaluate. Bayesian optimization (BO) is a popular way to tackle optimization problems with expensive objective function evaluations, but has mostly been applied to unconstrained problems. Several BO approaches have been proposed to address expensive constraints but are limited to greedy strategies maximizing immediate reward. To address this limitation, we propose a lookahead approach that selects the next evaluation in order to maximize the long-term feasible reduction of the objective function. We present numerical experiments demonstrating the performance improvements of such a lookahead approach compared to several greedy BO algorithms, including constrained expected improvement (EIC) and predictive entropy search with constraint (PESC).

************************************

Online Learning with Transductive Regret
Mehryar Mohri, Scott Yang

We study online learning with the general notion of transductive regret, that is regret with modification rules applying to expert sequences (as opposed to single experts) that are representable by weighted finite-state transducers. We show how transductive regret generalizes existing notions of regret, including: (1) external regret; (2) internal regret; (3) swap regret; and (4) conditional swap regret. We present a general and efficient online learning algorithm for minimizing transductive regret. We further extend that to design efficient algorithms for the time-selection and sleeping expert settings. A by-product of our study is an algorithm for swap regret, which, under mild assumptions, is more efficient than existing ones, and a substantially more efficient algorithm for time selection swap regret.

************************************

Pixels to Graphs by Associative Embedding
Alejandro Newell, Jia Deng

Graphs are a useful abstraction of image content. Not only can graphs represent details about individual objects in a scene but they can capture the interactions between pairs of objects. We present a method for training a convolutional neural network such that it takes in an input image and produces a full graph definition. This is done end-to-end in a single stage with the use of associative embeddings. The network learns to simultaneously identify all of the elements that make up a graph and piece them together. We benchmark on the Visual Genome dataset, and demonstrate state-of-the-art performance on the challenging task of scene graph generation.

************************************

Accelerated Stochastic Greedy Coordinate Descent by Soft Thresholding Projection onto Simplex
Chaobing Song, Shaobo Cui, Yong Jiang, Shu-Tao Xia

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.

************************************

Reinforcement Learning under Model Mismatch
Aurko Roy, Huan Xu, Sebastian Pokutta

We study reinforcement learning under model misspecification,   where we do not have access to the true environment but only to a   reasonably close approximati on to it. We address this problem by extending the framework of robust MDPs to the model-free Reinforcement Learning setting, where we do not have access to the model parameters, but can only sample states from it. We define robust versions of Q-learning, Sarsa, and TD-learning and prove convergence to an approximately optimal robust policy and approximate value function respectively. We scale up the robust algorithms to large MDPs via function approximation and prove convergence under two different settings. We prove convergence of robust approximate policy iteration and robust approximate value iteration for linear architectures (under mild assumptions). We also define a robust loss function, the mean squared robust projected Bellman error and give stochastic gradient descent algorithms that are guaranteed to converge to a local minimum.

************************************

Concrete Dropout

Yarin Gal, Jiri Hron, Alex Kendall

Dropout is used as a practical tool to obtain uncertainty estimates in large vision models and reinforcement learning (RL) tasks. But to obtain well-calibrated uncertainty estimates, a grid-search over the dropout probabilities is necessary—a prohibitive operation with large models, and an impossible one with RL. We propose a new dropout variant which gives improved performance and better calibrated uncertainties. Relying on recent developments in Bayesian deep learning, we use a continuous relaxation of dropout’s discrete masks. Together with a principled optimisation objective, this allows for automatic tuning of the dropout probability in large models, and as a result faster experimentation cycles. In RL this allows the agent to adapt its uncertainty dynamically as more data is observed. We analyse the proposed variant extensively on a range of tasks, and give insights into common practice in the field where larger dropout probabilities are often used in deeper model layers.

************************************

Multiresolution Kernel Approximation for Gaussian Process Regression

Yi Ding, Risi Kondor, Jonathan Eskreis-Winkler

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.

************************************

Near Minimax Optimal Players for the Finite-Time 3-Expert Prediction Problem

Yasin Abbasi Yadkori, Peter L. Bartlett, Victor Gabillon

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.

************************************

Learned D-AMP: Principled Neural Network based Compressive Image Recovery

Chris Metzler, Ali Mousavi, Richard Baraniuk

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.

************************************

Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks

Designing optimal treatment plans for patients with comorbidities requires accurate cause-specific mortality prognosis. Motivated by the recent availability of linked electronic health records, we develop a nonparametric Bayesian model for survival analysis with competing risks, which can be used for jointly assessing a patient's risk of multiple (competing) adverse outcomes. The model views a patient's survival times with respect to the competing risks as the outputs of a deep multi-task Gaussian process (DMGP), the inputs to which are the patients' covariates. Unlike parametric survival analysis methods based on Cox and Weibull mo dels, our model uses DMGPs to capture complex non-linear interactions between the patients' covariates and cause-specific survival times, thereby learning flexible patient-specific and cause-specific survival curves, all in a data-driven fashion without explicit parametric assumptions on the hazard rates. We propose a variational inference algorithm that is capable of learning the model parameters from time-to-event data while handling right censoring. Experiments on synthetic and real data show that our model outperforms the state-of-the-art survival models.

************************************

Unsupervised Transformation Learning via Convex Relaxations

Tatsunori B. Hashimoto, Percy S. Liang, John C. Duchi

Our goal is to extract meaningful transformations from raw images, such as varying the thickness of lines in handwriting or the lighting in a portrait. We propose an unsupervised approach to learn such transformations by attempting to reconstruct an image from a linear combination of transformations of its nearest neighbors.  On handwritten digits and celebrity portraits, we show that even with linear transformations, our method generates visually high-quality modified images.  Moreover, since our method is semiparametric and does not model the data distribution, the learned transformations extrapolate off the training data and can be applied to new types of images.

************************************

Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, Luc V. Gool

We present a new approach to learn compressible representations in deep architectures with an end-to-end training strategy.  Our method is based on a soft (continuous) relaxation of quantization and entropy, which we anneal to their discrete counterparts throughout training.  We showcase this method for two challenging applications: Image compression and neural network compression.  While these tasks have typically been approached with different methods, our soft-to-hard quantization approach gives results competitive with the state-of-the-art for both.

************************************

Accuracy First: Selecting a Differential Privacy Level for Accuracy Constrained ERM

Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, Steven Z. Wu

Traditional approaches to differential privacy assume a fixed privacy requirement ε for a computation, and attempt to maximize the accuracy of the computation subject to the privacy constraint. As differential privacy is increasingly deployed in practical settings, it may often be that there is instead a fixed accuracy requirement for a given computation and the data analyst would like to maximize the privacy of the computation subject to the accuracy constraint. This raises the question of how to find and run a maximally private empirical risk minimizer subject to a given accuracy requirement. We propose a general “noise reduction” framework that can apply to a variety of private empirical risk minimization (ERM) algorithms, using them to “search” the space of privacy levels to find the empirically strongest one that meets the accuracy constraint, and incurring only logarithmic overhead in the number of privacy levels searched. The privacy analysis of our algorithm leads naturally to a version of differential privacy where the privacy parameters are dependent on the data, which we term ex-post privacy, and which is related to the recently introduced notion of privacy odometers. We also give an ex-post privacy analysis of the classical AboveThreshold privacy tool, modifying it to allow for queries chosen depending on the database. Finally, we apply our approach to two common objective functions, regularized linear and logistic regression, and empirically compare our noise reduction methods to (i) inverting the theoretical utility guarantees of standard private ERM algorithms and (ii) a stronger empirical baseline based on binary search.

************************************

Triple Generative Adversarial Nets

Chongxuan LI, Taufik Xu, Jun Zhu, Bo Zhang

Generative Adversarial Nets (GANs) have shown promise in image generation and semi-supervised learning (SSL). However, existing GANs in SSL have two problems: (1) the generator and the discriminator (i.e. the classifier) may not be optimal at the same time; and (2) the generator cannot control the semantics of the generated samples. The problems essentially arise from the two-player formulation, where a single discriminator shares incompatible roles of identifying fake samples and predicting labels and it only estimates the data without considering the labels. To address the problems, we present triple generative adversarial net (Triple-GAN), which consists of three players---a generator, a discriminator and a classifier. The generator and the classifier characterize the conditional distributions between images and labels, and the discriminator solely focuses on identifying fake image-label pairs. We design compatible utilities to ensure that the distributions characterized by the classifier and the generator both converge to the data distribution. Our results on various datasets demonstrate that Triple-GAN as a unified model can simultaneously (1) achieve the state-of-the-art classification results among deep generative models, and (2) disentangle the classes and styles of the input and transfer smoothly in the data space via interpolation in the latent space class-conditionally.

************************************

Deep Learning with Topological Signatures

Christoph Hofer, Roland Kwitt, Marc Niethammer, Andreas Uhl

Inferring topological and geometrical information from data can offer an alternative perspective in machine learning problems. Methods from topological data analysis, e.g., persistent homology, enable us to obtain such information, typically in the form of summary representations of topological features. However, such topological signatures often  come with an unusual structure (e.g., multisets of intervals) that is highly impractical for most machine learning techniques. While many strategies have been proposed to map these topological signatures into machine learning compatible representations, they suffer from being agnostic to the target learning task. In contrast, we propose a technique that enables us to input topological signatures to deep neural networks and learn a task-optimal representation during training. Our approach is realized as a novel input layer with favorable theoretical properties. Classification experiments on 2D object shapes and social network graphs demonstrate the versatility of the approach and, in case of the latter, we even outperform the state-of-the-art by a large margin.

************************************

Revenue Optimization with Approximate Bid Predictions

Andres Munoz, Sergei Vassilvitskii

In the context of advertising auctions, finding good reserve prices is a notoriously challenging learning problem. This is due to the heterogeneity of ad opportunity types, and the non-convexity of the objective function. In this work, we show how to reduce reserve price optimization to the standard setting of prediction under squared loss, a well understood problem in the learning community.  We further bound the gap between the expected bid and revenue in terms of the average loss of the predictor. This is the first result that formally relates the revenue gained to the quality of a standard machine learned model.

************************************

Mapping distinct timescales of functional interactions among brain networks

Mali Sundaresan, Arshed Nabeel, Devarajan Sridharan

Brain processes occur at various timescales, ranging from milliseconds (neurons) to minutes and hours (behavior). Characterizing functional coupling among brain regions at these diverse timescales is key to understanding how the brain produces behavior. Here, we apply instantaneous and lag-based measures of conditional linear dependence, based on Granger-Geweke causality (GC), to infer network connections at distinct timescales from functional magnetic resonance imaging (fMRI) data. Due to the slow sampling rate of fMRI, it is widely held that GC produces spurious and unreliable estimates of functional connectivity when applied to fMRI data. We challenge this claim with simulations and a novel machine learning approach. First, we show, with simulated fMRI data, that instantaneous and lag-based GC identify distinct timescales and complementary patterns of functional co nnectivity. Next, we analyze fMRI scans from 500 subjects and show that a linear classifier trained on either instantaneous or lag-based GC connectivity reliably distinguishes task versus rest brain states, with ~80-85% cross-validation accuracy. Importantly, instantaneous and lag-based GC exploit markedly different spatial and temporal patterns of connectivity to achieve robust classification. Our approach enables identifying functionally connected networks that operate at distinct timescales in the brain.

************************************

Improved Training of Wasserstein GANs

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron C. Courville

Generative Adversarial Networks (GANs) are powerful generative models, but suffer from training instability. The recently proposed Wasserstein GAN (WGAN) makes progress toward stable training of GANs, but sometimes can still generate only poor samples or fail to converge. We find that these problems are often due to the use of weight clipping in WGAN to enforce a Lipschitz constraint on the critic, which can lead to undesired behavior. We propose an alternative to clipping weights: penalize the norm of gradient of the critic with respect to its input. Our proposed method performs better than standard WGAN and enables stable training of a wide variety of GAN architectures with almost no hyperparameter tuning, including 101-layer ResNets and language models with continuous generators. We also achieve high quality generations on CIFAR-10 and LSUN bedrooms.

************************************

Adaptive stimulus selection for optimizing neural population responses

Benjamin Cowley, Ryan Williamson, Katerina Clemens, Matthew Smith, Byron M. Yu

Adaptive stimulus selection methods in neuroscience have primarily focused on maximizing the firing rate of a single recorded neuron. When recording from a population of neurons, it is usually not possible to find a single stimulus that maximizes the firing rates of all neurons. This motivates optimizing an objective function that takes into account the responses of all recorded neurons together. We propose “Adept,” an adaptive stimulus selection method that can optimize population objective functions. In simulations, we first confirmed that population objective functions elicited more diverse stimulus responses than single-neuron objective functions. Then, we tested Adept in a closed-loop electrophysiological experiment in which population activity was recorded from macaque V4, a cortical area known for mid-level visual processing. To predict neural responses, we used the outputs of a deep convolutional neural network model as feature embeddings. Images chosen by Adept elicited mean neural responses that were 20% larger than those for randomly-chosen natural images, and also evoked a larger diversity of neural responses. Such adaptive stimulus selection methods can facilitate experiments that involve neurons far from the sensory periphery, for which it is often unclear which stimuli to present.

************************************

Matrix Norm Estimation from a Few Entries

Ashish Khetan, Sewoong Oh

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.

************************************

On the Power of Truncated SVD for General High-rank Matrix Estimation Problems

Simon S. Du, Yining Wang, Aarti Singh

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.

************************************

TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, Hai Li

High network communication cost for synchronizing gradients and parameters is th e well-known bottleneck of distributed training. In this work, we propose TernGrad that uses ternary gradients to accelerate distributed deep learning in data parallelism. Our approach requires only three numerical levels {-1,0,1}, which can aggressively reduce the communication time. We mathematically prove the convergence of TernGrad under the assumption of a bound on gradients. Guided by the bound, we propose layer-wise ternarizing and gradient clipping to improve its convergence. Our experiments show that applying TernGrad on AlexNet does not incur any accuracy loss and can even improve accuracy. The accuracy loss of GoogLeNet induced by TernGrad is less than 2% on average. Finally, a performance model is proposed to study the scalability of TernGrad. Experiments show significant speed gains for various deep neural networks. Our source code is available.

************************************

GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter

Generative Adversarial Networks (GANs) excel at creating realistic images with complex models for which maximum likelihood is infeasible. However, the convergence of GAN training has still not been proved. We propose a two time-scale update rule (TTUR) for training GANs with stochastic gradient descent on arbitrary GAN loss functions. TTUR has an individual learning rate for both the discriminator and the generator. Using the theory of stochastic approximation, we prove that the TTUR converges under mild assumptions to a stationary local Nash equilibrium. The convergence carries over to the popular Adam optimization, for which we prove that it follows the dynamics of a heavy ball with friction and thus prefers flat minima in the objective landscape. For the evaluation of the performance of GANs at image generation, we introduce the `Fréchet Inception Distance'' (FID) which captures the similarity of generated images to real ones better than the Inception Score. In experiments, TTUR improves learning for DCGANs and Improved Wasserstein GANs (WGAN-GP) outperforming conventional GAN training on CelebA, CIFAR-10, SVHN, LSUN Bedrooms, and the One Billion Word Benchmark.

************************************

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg, Su-In Lee

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between accuracy and interpretability. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

************************************

Nonbacktracking Bounds on the Influence in Independent Cascade Models

Emmanuel Abbe, Sanjeev Kulkarni, Eun Jee Lee

This paper develops upper and lower bounds on the influence measure in a network, more precisely, the expected number of nodes that a seed set can influence in the independent cascade model. In particular, our bounds exploit nonbacktracking walks, Fortuin-Kasteleyn-Ginibre type inequalities, and are computed by message passing algorithms. Nonbacktracking walks have recently allowed for headways in community detection, and this paper shows that their use can also impact the in fluence computation. Further, we provide parameterized versions of the bounds that control the trade-off between the efficiency and the accuracy. Finally, the tightness of the bounds is illustrated with simulations on various network models.

************************************

Linear Convergence of a Frank-Wolfe Type Algorithm over Trace-Norm Balls
Zeyuan Allen-Zhu, Elad Hazan, Wei Hu, Yuanzhi Li
We propose a rank-k variant of the classical Frank-Wolfe algorithm to solve convex optimization over a trace-norm ball. Our algorithm replaces the top singular-vector computation (1-SVD) in Frank-Wolfe with a top-k singular-vector computation (k-SVD), which can be done by repeatedly applying 1-SVD k times. Alternatively, our algorithm can be viewed as a rank-k restricted version of projected gradient descent. We show that our algorithm has a linear convergence rate when the objective function is smooth and strongly convex, and the optimal solution has rank at most k. This improves the convergence rate and the total time complexity of the Frank-Wolfe method and its variants.

************************************

Fully Decentralized Policies for Multi-Agent Systems: An Information Theoretic Approach
Roel Dobbe, David Fridovich-Keil, Claire Tomlin
Learning cooperative policies for multi-agent systems is often challenged by partial observability and a lack of coordination. In some settings, the structure of a problem allows a distributed solution with limited communication. Here, we consider a scenario where no communication is available, and instead we learn local policies for all agents that collectively mimic the solution to a centralized multi-agent static optimization problem. Our main contribution is an information theoretic framework based on rate distortion theory which facilitates analysis of how well the resulting fully decentralized policies are able to reconstruct the optimal solution. Moreover, this framework provides a natural extension that addresses which nodes an agent should communicate with to improve the performance of its individual policy.

************************************

Neural system identification for large populations separating “what” and “where”
David Klindt, Alexander S. Ecker, Thomas Euler, Matthias Bethge
Neuroscientists classify neurons into different types that perform similar computations at different locations in the visual field. Traditional methods for neural system identification do not capitalize on this separation of “what” and “where”. Learning deep convolutional feature spaces that are shared among many neurons provides an exciting path forward, but the architectural design needs to account for data limitations: While new experimental techniques enable recordings from thousands of neurons, experimental time is limited so that one can sample only a small fraction of each neuron's response space. Here, we show that a major bottleneck for fitting convolutional neural networks (CNNs) to neural data is the estimation of the individual receptive field locations – a problem that has been scratched only at the surface thus far. We propose a CNN architecture with a sparse readout layer factorizing the spatial (where) and feature (what) dimensions. Our network scales well to thousands of neurons and short recordings and can be trained end-to-end. We evaluate this architecture on ground-truth data to explore the challenges and limitations of CNN-based system identification. Moreover, we show that our network model outperforms current state-of-the art system identification models of mouse primary visual cortex.

************************************

Learning Active Learning from Data
Ksenia Konyushkova, Raphael Sznitman, Pascal Fua
In this paper, we suggest a novel data-driven approach to active learning (AL). The key idea is to train a regressor that predicts the expected error reduction for a candidate sample in a particular learning state. By formulating the query selection procedure as a regression problem we are not restricted to working with existing AL heuristics; instead, we learn strategies based on experience from previous AL outcomes. We show that a strategy can be learnt either from simple

synthetic 2D datasets or  from a subset of domain-specific data. Our method yields strategies that work well on real data from a wide range of domains.

************************************

Controllable Invariance through Adversarial Feature Learning

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig

Learning meaningful representations that maintain the content necessary for a particular task while filtering away detrimental variations is a problem of great interest in machine learning. In this paper, we tackle the problem of learning representations invariant to a specific factor or trait of data. The representation learning process is formulated as an adversarial minimax game. We analyze the optimal equilibrium of such a game and find that it amounts to maximizing the uncertainty of inferring the detrimental factor given the representation while maximizing the certainty of making task-specific predictions. On three benchmark tasks, namely fair and bias-free classification, language-independent generation, and lighting-independent image classification, we show that the proposed framework induces an invariant representation, and leads to better generalization evidenced by the improved performance.

************************************

Visual Interaction Networks: Learning a Physics Simulator from Video

Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, Andrea Tacchetti

From just a glance, humans can make rich predictions about the future of a wide range of physical systems.  On the other hand, modern approaches from engineering, robotics, and graphics are often restricted to narrow domains or require information about the underlying state. We introduce the Visual Interaction Network, a general-purpose model for learning the dynamics of a physical system from raw visual observations. Our model consists of a perceptual front-end based on convolutional neural networks and a dynamics predictor based on interaction networks. Through joint training, the perceptual front-end learns to parse a dynamic visual scene into a set of factored latent object representations. The dynamics predictor learns to roll these states forward in time by computing their interactions, producing a predicted physical trajectory of arbitrary length. We found that from just six input video frames the Visual Interaction Network can generate accurate future trajectories of hundreds of time steps on a wide range of physical systems. Our model can also be applied to scenes with invisible objects, inferring their future states from their effects on the visible objects, and can implicitly infer the unknown mass of objects. This work opens new opportunities for model-based decision-making and planning from raw sensory observations in complex physical environments.

************************************

Repeated Inverse Reinforcement Learning

Kareem Amin, Nan Jiang, Satinder Singh

We introduce a novel repeated Inverse Reinforcement Learning problem: the agent has to act on behalf of a human in a sequence of tasks and wishes to minimize the number of tasks that it surprises the human by acting suboptimally with respect to how the human would have acted. Each time the human is surprised, the agent is provided a demonstration of the desired behavior by the human. We formalize this problem, including how the sequence of tasks is chosen, in a few different ways and provide some foundational results.

************************************

Inference in Graphical Models via Semidefinite Programming Hierarchies

Murat A. Erdogdu, Yash Deshpande, Andrea Montanari

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.

************************************

Gauging Variational Inference

Sung-Soo Ahn, Michael Chertkov, Jinwoo Shin

Computing partition function is the most important statistical inference task ar ising in applications of Graphical Models (GM). Since it is computationally intractable, approximate methods have been used in practice, where mean-field (MF) and belief propagation (BP) are arguably the most popular and successful approaches of a variational type. In this paper, we propose two new variational schemes, coined Gauged-MF (G-MF) and Gauged-BP (G-BP), improving MF and BP, respectively. Both provide lower bounds for the partition function by utilizing the so-called gauge transformation which modifies factors of GM while keeping the partition function invariant. Moreover, we prove that both G-MF and G-BP are exact for GMs with a single loop of a special structure, even though the bare MF and BP perform badly in this case. Our extensive experiments indeed confirm that the proposed algorithms outperform and generalize MF and BP.

************************************

Teaching Machines to Describe Images with Natural Language Feedback

huan ling, Sanja Fidler

Robots will eventually be part of every household. It is thus critical to enable algorithms to learn from and be guided by non-expert users. In this paper, we bring a human in the loop, and enable a human teacher to give feedback to a learning agent in the form of natural language. A descriptive sentence can provide a stronger learning signal than a numeric reward in that it can easily point to where the mistakes are and how to correct them. We focus on the problem of image captioning in which the quality of the output can easily be judged by non-experts. We propose a phrase-based captioning model trained with policy gradients, and design a critic that provides reward to the learner by conditioning on the human-provided feedback. We show that by exploiting descriptive feedback our model learns to perform better than when given independently written human captions.

************************************

Associative Embedding: End-to-End Learning for Joint Detection and Grouping

Alejandro Newell, Zhiao Huang, Jia Deng

We introduce associative embedding, a novel method for supervising convolutional neural networks for the task of detection and grouping. A number of computer vision problems can be framed in this manner including multi-person pose estimation, instance segmentation, and multi-object tracking. Usually the grouping of detections is achieved with multi-stage pipelines, instead we propose an approach that teaches a network to simultaneously output detections and group assignments. This technique can be easily integrated into any state-of-the-art network architecture that produces pixel-wise predictions. We show how to apply this method to multi-person pose estimation and report state-of-the-art performance on the MPII and MS-COCO datasets.

************************************

Information Theoretic Properties of Markov Random Fields, and their Algorithmic Applications

Linus Hamilton, Frederic Koehler, Ankur Moitra

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.

************************************

Subset Selection and Summarization in Sequential Data

Ehsan Elhamifar, M. Clara De Paolis Kaluza

Subset selection, which is the task of finding a small subset of representative items from a large ground set, finds numerous applications in different areas. Sequential data, including time-series and ordered data, contain important structural relationships among items, imposed by underlying dynamic models of data, that should play a vital role in the selection of representatives. However, nearly all existing subset selection techniques ignore underlying dynamics of data and treat items independently, leading to incompatible sets of representatives. In this paper, we develop a new framework for sequential subset selection that finds a set of representatives compatible with the dynamic models of data. To do so, we equip items with transition dynamic models and pose the problem as an integer binary optimization over assignments of sequential items to representatives, t hat leads to high encoding, diversity and transition potentials. Our formulation generalizes the well-known facility location objective to deal with sequential data, incorporating transition dynamics among facilities. As the proposed formulation is non-convex, we derive a max-sum message passing algorithm to solve the problem efficiently. Experiments on synthetic and real data, including instructional video summarization, show that our sequential subset selection framework not only achieves better encoding and diversity than the state of the art, but also successfully incorporates dynamics of data, leading to compatible representatives.

************************************

Z-Forcing: Training Stochastic Recurrent Networks

Anirudh Goyal ALIAS PARTH GOYAL, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, Yoshua Bengio

Many efforts have been devoted to training generative latent variable models with autoregressive decoders, such as recurrent neural networks (RNN). Stochastic recurrent models have been successful in capturing the variability observed in natural sequential data such as speech. We unify successful ideas from recently proposed architectures into a stochastic recurrent model: each step in the sequence is associated with a latent variable that is used to condition the recurrent dynamics for future steps. Training is performed with amortised variational inference where the approximate posterior is augmented with a RNN that runs backward through the sequence. In addition to maximizing the variational lower bound, we ease training of the latent variables by adding an auxiliary cost which forces them to reconstruct the state of the backward recurrent network. This provides the latent variables with a task-independent objective that enhances the performance of the overall model. We found this strategy to perform better than alternative approaches such as KL annealing. Although being conceptually simple, our model achieves state-of-the-art results on standard speech benchmarks such as TIMIT and Blizzard and competitive performance on sequential MNIST. Finally, we apply our model to language modeling on the IMDB dataset where the auxiliary cost helps in learning interpretable latent variables.

************************************

Regret Minimization in MDPs with Options without Prior Knowledge

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, Emma Brunskill

The option framework integrates temporal abstraction into the reinforcement learning model through the introduction of macro-actions (i.e., options). Recent works leveraged on the mapping of Markov decision processes (MDPs) with options to semi-MDPs (SMDPs) and introduced SMDP-versions of exploration-exploitation algorithms (e.g., RMAX-SMDP and UCRL-SMDP) to analyze the impact of options on the learning performance. Nonetheless, the PAC-SMDP sample complexity of RMAX-SMDP can hardly be translated into equivalent PAC-MDP theoretical guarantees, while UCRL-SMDP requires prior knowledge of the parameters characterizing the distributions of the cumulative reward and duration of each option, which are hardly available in practice. In this paper, we remove this limitation by combining the SMDP view together with the inner Markov structure of options into a novel algorithm whose regret performance matches UCRL-SMDP's up to an additive regret term. We show scenarios where this term is negligible and the advantage of temporal abstraction is preserved. We also report preliminary empirical result supporting the theoretical findings.

************************************

Learning Identifiable Gaussian Bayesian Networks in Polynomial Time and Sample Complexity

Asish Ghoshal, Jean Honorio

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.

************************************

Learning Neural Representations of Human Cognition across Many fMRI Studies

Arthur Mensch, Julien Mairal, Danilo Bzdok, Bertrand Thirion, Gael Varoquaux

Cognitive neuroscience is enjoying rapid increase in extensive public brain-imaging datasets. It opens the door to large-scale statistical models. Finding a unified perspective for all available data calls for scalable and automated solutions to an old challenge: how to aggregate heterogeneous information on brain function into a universal cognitive system that relates mental operations/cognitive processes/psychological tasks to brain networks? We cast this challenge in a machine-learning approach to predict conditions from statistical brain maps across different studies. For this, we leverage multi-task learning and multi-scale dimension reduction to learn low-dimensional representations of brain images that carry cognitive information and can be robustly associated with psychological stimuli. Our multi-dataset classification model achieves the best prediction performance on several large reference datasets, compared to models without cognitive-aware low-dimension representations; it brings a substantial performance boost to the analysis of small datasets, and can be introspected to identify universal template cognitive concepts.

************************************

Conic Scan-and-Cover algorithms for nonparametric topic modeling

Mikhail Yurochkin, Aritra Guha, XuanLong Nguyen

We propose new algorithms for topic modeling when the number of topics is unknown. Our approach relies on an analysis of the concentration of mass and angular geometry of the topic simplex, a convex polytope constructed by taking the convex hull of vertices representing the latent topics. Our algorithms are shown in practice to have accuracy comparable to a Gibbs sampler in terms of topic estimation, which requires the number of topics be given. Moreover, they are one of the fastest among several state of the art parametric techniques. Statistical consistency of our estimator is established under some conditions.

************************************

Online Learning for Multivariate Hawkes Processes

Yingxiang Yang, Jalal Etesami, Niao He, Negar Kiyavash

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.

************************************

An Empirical Study on The Properties of Random Bases for Kernel Methods

Maximilian Alber, Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, Fei Sha

Kernel machines as well as neural networks possess universal function approximation properties. Nevertheless in practice their ways of choosing the appropriate function class differ. Specifically neural networks learn a representation by adapting their basis functions to the data and the task at hand, while kernel methods typically use a basis that is not adapted during training. In this work, we contrast random features of approximated kernel machines with learned features of neural networks. Our analysis reveals how these random and adaptive basis functions affect the quality of learning. Furthermore, we present basis adaptation schemes that allow for a more compact representation, while retaining the generalization properties of kernel machines.

************************************

Nearest-Neighbor Sample Compression: Efficiency, Consistency, Infinite Dimensions

Aryeh Kontorovich, Sivan Sabato, Roi Weiss

We examine the Bayes-consistency of a recently proposed 1-nearest-neighbor-based multiclass learning algorithm. This algorithm is derived from sample compression bounds and enjoys the statistical advantages of tight, fully empirical generalization bounds, as well as the algorithmic advantages of a faster runtime and memory savings. We prove that this algorithm is strongly Bayes-consistent in metric spaces with finite doubling dimension --- the first consistency result for an efficient nearest-neighbor sample compression scheme. Rather surprisingly, we discover that this algorithm continues to be Bayes-consistent even in a certain infinite-dimensional setting, in which the basic measure-theoretic conditions on w

hich classic consistency proofs hinge are violated. This is all the more surprising, since it is known that k-NN is not Bayes-consistent in this setting. We pose several challenging open problems for future research.
************************************

Causal Effect Inference with Deep Latent-Variable Models
Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, Max Welling
Learning individual-level causal effects from observational data, such as inferring the most effective medication for a specific patient, is a problem of growing importance for policy makers. The most important aspect of inferring causal effects from observational data is the handling of confounders, factors that affect both an intervention and its outcome. A carefully designed observational study attempts to measure all important confounders. However, even if one does not have direct access to all confounders, there may exist noisy and uncertain measurement of proxies for confounders. We build on recent advances in latent variable modeling to simultaneously estimate the unknown latent space summarizing the confounders and the causal effect. Our method is based on Variational Autoencoders (VAE) which follow the causal structure of inference with proxies. We show our method is significantly more robust than existing methods, and matches the state-of-the-art on previous benchmarks focused on individual treatment effects.
************************************

Estimating Accuracy from Unlabeled Data: A Probabilistic Logic Approach
Emmanouil Platanios, Hoifung Poon, Tom M. Mitchell, Eric J. Horvitz
We propose an efficient method to estimate the accuracy of classifiers using only unlabeled data. We consider a setting with multiple classification problems where the target classes may be tied together through logical constraints. For example, a set of classes may be mutually exclusive, meaning that a data instance can belong to at most one of them. The proposed method is based on the intuition that: (i) when classifiers agree, they are more likely to be correct, and (ii) when the classifiers make a prediction that violates the constraints, at least one classifier must be making an error. Experiments on four real-world data sets produce accuracy estimates within a few percent of the true accuracy, using solely unlabeled data. Our models also outperform existing state-of-the-art solutions in both estimating accuracies, and combining multiple classifier outputs. The results emphasize the utility of logical constraints in estimating accuracy, thus validating our intuition.
************************************

A Decomposition of Forecast Error in Prediction Markets
Miro Dudik, Sebastien Lahaie, Ryan M. Rogers, Jennifer Wortman Vaughan
We analyze sources of error in prediction market forecasts in order to bound the difference between a security's price and the ground truth it estimates. We consider cost-function-based prediction markets in which an automated market maker adjusts security prices according to the history of trade. We decompose the forecasting error into three components: sampling error, arising because traders only possess noisy estimates of ground truth; market-maker bias, resulting from the use of a particular market maker (i.e., cost function) to facilitate trade; and convergence error, arising because, at any point in time, market prices may still be in flux. Our goal is to make explicit the tradeoffs between these error components, influenced by design decisions such as the functional form of the cost function and the amount of liquidity in the market. We consider a specific model in which traders have exponential utility and exponential-family beliefs representing noisy estimates of ground truth. In this setting, sampling error vanishes as the number of traders grows, but there is a tradeoff between the other two components. We provide both upper and lower bounds on market-maker bias and convergence error, and demonstrate via numerical simulations that these bounds are tight. Our results yield new insights into the question of how to set the market's liquidity parameter and into the forecasting benefits of enforcing coherent prices across securities.
************************************

Ranking Data with Continuous Labels through Oriented Recursive Partitions

Stéphan Clémençon, Mastane Achab

We formulate a supervised learning problem, referred to as continuous ranking, where a continuous real-valued label Y is assigned to an observable r.v. X taking its values in a feature space X and the goal is to order all possible observations x in X by means of a scoring function s : X → R so that s(X) and Y tend to increase or decrease together with highest probability. This problem generalizes bi/multi-partite ranking to a certain extent and the task of finding optimal scoring functions s(x) can be naturally cast as optimization of a dedicated functional cri- terion, called the IROC curve here, or as maximization of the Kendall τ related to the pair (s(X), Y ). From the theoretical side, we describe the optimal elements of this problem and provide statistical guarantees for empirical Kendall τ maximiza- tion under appropriate conditions for the class of scoring function candidates. We also propose a recursive statistical learning algorithm tailored to empirical IROC curve optimization and producing a piecewise constant scoring function that is fully described by an oriented binary tree. Preliminary numerical experiments highlight the difference in nature between regression and continuous ranking and provide strong empirical evidence of the performance of empirical optimizers of the criteria proposed.

************************************

Scalable Log Determinants for Gaussian Process Kernel Learning

Kun Dong, David Eriksson, Hannes Nickisch, David Bindel, Andrew G. Wilson

For applications as varied as Bayesian neural networks, determinantal point processes, elliptical graphical models, and kernel learning for Gaussian processes (GPs), one must compute a log determinant of an n by n positive definite matrix, and its derivatives---leading to prohibitive O(n^3) computations.  We propose novel O(n) approaches to estimating these quantities from only fast matrix vector multiplications (MVMs). These stochastic approximations are based on Chebyshev, Lanczos, and surrogate models, and converge quickly even for kernel matrices that have challenging spectra.  We leverage these approximations to develop a scalable Gaussian process approach to kernel learning. We find that Lanczos is generally superior to Chebyshev for kernel learning, and that a surrogate approach can be highly efficient and accurate with popular kernels.

************************************

Fair Clustering Through Fairlets

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Sergei Vassilvitskii

We study the question of fair clustering under the {\em disparate impact} doctrine, where each protected class must have approximately equal representation in every cluster. We formulate the fair clustering problem under both the k-center and the k-median objectives, and show that even with two protected classes the problem is challenging, as the optimum solution can violate common conventions---for instance a point may no longer be assigned to its nearest cluster center!  En route we introduce the concept of fairlets, which are minimal sets that satisfy fair representation while approximately preserving the clustering objective.  We show that any fair clustering problem can be decomposed into first finding good fairlets, and then using existing machinery for traditional clustering algorithms.  While finding good fairlets can be NP-hard, we proceed to obtain efficient approximation algorithms based on minimum cost flow.  We empirically demonstrate the \emph{price of fairness} by quantifying the value of fair clustering on real-world datasets with sensitive attributes.

************************************

A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, Arthur Gretton

We propose a novel adaptive test of goodness-of-fit, with computational cost linear in the number of samples. We learn the test features that best indicate the differences between observed samples and a reference model, by minimizing the false negative rate. These features are constructed via Stein's method, meaning that it is not necessary to compute the normalising constant of the model. We analyse the asymptotic Bahadur efficiency of the new test, and prove that under a mean-shift alternative, our test always has greater relative efficiency than a previous linear-time kernel test, regardless of the choice of parameters for that t

est. In experiments, the performance of our method exceeds that of the earlier linear-time test, and matches or exceeds the power of a quadratic-time kernel test. In high dimensions and where model structure may be exploited, our goodness of fit test performs far better than a quadratic-time two-sample test based on the Maximum Mean Discrepancy, with samples drawn from the model.
************************************
Rotting Bandits
Nir Levine, Koby Crammer, Shie Mannor
The Multi-Armed Bandits (MAB) framework highlights the trade-off between acquiring new knowledge (Exploration) and leveraging available knowledge (Exploitation). In the classical MAB problem, a decision maker must choose an arm at each time step, upon which she receives a reward. The decision maker's objective is to maximize her cumulative expected reward over the time horizon. The MAB problem has been studied extensively, specifically under the assumption of the arms' rewards distributions being stationary, or quasi-stationary, over time. We consider a variant of the MAB framework, which we termed Rotting Bandits, where each arm's expected reward decays as a function of the number of times it has been pulled. We are motivated by many real-world scenarios such as online advertising, content recommendation, crowdsourcing, and more. We present algorithms, accompanied by simulations, and derive theoretical guarantees.
************************************
Scalable Planning with Tensorflow for Hybrid Nonlinear Domains
Ga Wu, Buser Say, Scott Sanner
Given recent deep learning results that demonstrate the ability to effectively optimize high-dimensional non-convex functions with gradient descent optimization on GPUs, we ask in this paper whether symbolic gradient optimization tools such as Tensorflow can be effective for planning in hybrid (mixed discrete and continuous) nonlinear domains with high dimensional state and action spaces?  To this end, we demonstrate that hybrid planning with Tensorflow and RMSProp gradient descent is competitive with mixed integer linear program (MILP) based optimization on piecewise linear planning domains (where we can compute optimal solutions) and substantially outperforms state-of-the-art interior point methods for nonlinear planning domains.  Furthermore, we remark that Tensorflow is highly scalable, converging to a strong plan on a large-scale concurrent domain with a total of 576,000 continuous action parameters distributed over a horizon of 96 time steps and 100 parallel instances in only 4 minutes.  We provide a number of insights that clarify such strong performance including observations that despite long horizons, RMSProp avoids both the vanishing and exploding gradient problems. Together these results suggest a new frontier for highly scalable planning in nonlinear hybrid domains by leveraging GPUs and the power of recent advances in gradient descent with highly optimized toolkits like Tensorflow.
************************************
Probabilistic Models for Integration Error in the Assessment of Functional Cardiac Models
Chris Oates, Steven Niederer, Angela Lee, François-Xavier Briol, Mark Girolami
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.
************************************
Bandits Dueling on Partially Ordered Sets
Julien Audiffren, Liva Ralaivola
We address the problem of dueling bandits defined on partially ordered sets, or posets.  In this setting, arms may not be comparable, and there may be several (incomparable) optimal arms.  We propose an algorithm, UnchainedBandits, that efficiently finds the set of optimal arms, or Pareto front, of any poset  even when pairs of comparable arms cannot be a priori distinguished from pairs of incomparable arms,   with a set of minimal assumptions. This means that UnchainedBandits does not require information about comparability and can be used with limited knowledge of the poset. To achieve this, the algorithm relies on the concept of

decoys, which stems from social psychology. We also provide theoretical guarantees on both the regret incurred and the number of comparison required by UnchainedBandits, and we report compelling empirical results.

************************************

Decomposition-Invariant Conditional Gradient for General Polytopes with Line Search

Mohammad Ali Bashiri, Xinhua Zhang

Frank-Wolfe (FW) algorithms with linear convergence rates have recently achieved great efficiency in many applications. Garber and Meshi (2016) designed a new decomposition-invariant pairwise FW variant with favorable dependency on the domain geometry. Unfortunately, it applies only to a restricted class of polytopes and cannot achieve theoretical and practical efficiency at the same time. In this paper, we show that by employing an away-step update, similar rates can be generalized to arbitrary polytopes with strong empirical performance. A new "condition number" of the domain is introduced which allows leveraging the sparsity of the solution. We applied the method to a reformulation of SVM, and the linear convergence rate depends, for the first time, on the number of support vectors.

************************************

Multiscale Semi-Markov Dynamics for Intracortical Brain-Computer Interfaces

Daniel Milstein, Jason Pacheco, Leigh Hochberg, John D. Simeral, Beata Jarosiewicz, Erik Sudderth

Intracortical brain-computer interfaces (iBCIs) have allowed people with tetraplegia to control a computer cursor by imagining the movement of their paralyzed arm or hand. State-of-the-art decoders deployed in human iBCIs are derived from a Kalman filter that assumes Markov dynamics on the angle of intended movement, and a unimodal dependence on intended angle for each channel of neural activity. Due to errors made in the decoding of noisy neural data, as a user attempts to move the cursor to a goal, the angle between cursor and goal positions may change rapidly. We propose a dynamic Bayesian network that includes the on-screen goal position as part of its latent state, and thus allows the person’s intended angle of movement to be aggregated over a much longer history of neural activity. This multiscale model explicitly captures the relationship between instantaneous angles of motion and long-term goals, and incorporates semi-Markov dynamics for motion trajectories. We also introduce a multimodal likelihood model for recordings of neural populations which can be rapidly calibrated for clinical applications. In offline experiments with recorded neural data, we demonstrate significantly improved prediction of motion directions compared to the Kalman filter. We derive an efficient online inference algorithm, enabling a clinical trial participant with tetraplegia to control a computer cursor with neural activity in real time. The observed kinematics of cursor movement are objectively straighter and smoother than prior iBCI decoding models without loss of responsiveness.

************************************

Fast Black-box Variational Inference through Stochastic Trust-Region Optimization

Jeffrey Regier, Michael I. Jordan, Jon McAuliffe

We introduce TrustVI, a fast second-order algorithm for black-box variational inference based on trust-region optimization and the reparameterization trick. At each iteration, TrustVI proposes and assesses a step based on minibatches of draws from the variational distribution. The algorithm provably converges to a stationary point. We implemented TrustVI in the Stan framework and compared it to two alternatives: Automatic Differentiation Variational Inference (ADVI) and Hessian-free Stochastic Gradient Variational Inference (HFSGVI). The former is based on stochastic first-order optimization. The latter uses second-order information, but lacks convergence guarantees. TrustVI typically converged at least one order of magnitude faster than ADVI, demonstrating the value of stochastic second-order information. TrustVI often found substantially better variational distributions than HFSGVI, demonstrating that our convergence theory can matter in practice.

************************************

Revisit Fuzzy Neural Network: Demystifying Batch Normalization and ReLU with Gen

### eralized Hamming Network

Lixin Fan

We revisit fuzzy neural network with a cornerstone notion of generalized hamming distance, which provides a novel and theoretically justified framework to re-interpret many useful neural network techniques in terms of fuzzy logic. In particular, we conjecture and empirically illustrate that, the celebrated batch normalization (BN) technique actually adapts the “normalized” bias such that it approximates the rightful bias induced by the generalized hamming distance. Once the due bias is enforced analytically, neither the optimization of bias terms nor the sophisticated batch normalization is needed. Also in the light of generalized hamming distance, the popular rectified linear units (ReLU) can be treated as setting a minimal hamming distance threshold between network inputs and weights. This thresholding scheme, on the one hand, can be improved by introducing double-thresholding on both positive and negative extremes of neuron outputs. On the other hand, ReLUs turn out to be non-essential and can be removed from networks trained for simple tasks like MNIST classification. The proposed generalized hamming network (GHN) as such not only lends itself to rigorous analysis and interpretation within the fuzzy logic theory but also demonstrates fast learning speed, well-controlled behaviour and state-of-the-art performances on a variety of learning tasks.

************************************

### Optimized Pre-Processing for Discrimination Prevention

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, Kush R. Varshney

Non-discrimination is a recognized objective in algorithmic decision making. In this paper, we introduce a novel probabilistic formulation of data pre-processing for reducing discrimination. We propose a convex optimization for learning a data transformation with three goals: controlling discrimination, limiting distortion in individual data samples, and preserving utility. We characterize the impact of limited sample size in accomplishing this objective. Two instances of the proposed optimization are applied to datasets, including one on real-world criminal recidivism. Results show that discrimination can be greatly reduced at a small cost in classification accuracy.

************************************

### Scalable Demand-Aware Recommendation

Jinfeng Yi, Cho-Jui Hsieh, Kush R. Varshney, Lijun Zhang, Yao Li

Recommendation for e-commerce with a mix of durable and nondurable goods has characteristics that distinguish it from the well-studied media recommendation problem. The demand for items is a combined effect of form utility and time utility, i.e., a product must both be intrinsically appealing to a consumer and the time must be right for purchase. In particular for durable goods, time utility is a function of inter-purchase duration within product category because consumers are unlikely to purchase two items in the same category in close temporal succession. Moreover, purchase data, in contrast to rating data, is implicit with non-purchases not necessarily indicating dislike. Together, these issues give rise to the positive-unlabeled demand-aware recommendation problem that we pose via joint low-rank tensor completion and product category inter-purchase duration vector estimation. We further relax this problem and propose a highly scalable alternating minimization approach with which we can solve problems with millions of users and millions of items in a single thread. We also show superior prediction accuracies on multiple real-world datasets.

************************************

### Learning a Multi-View Stereo Machine

Abhishek Kar, Christian Häne, Jitendra Malik

We present a learnt system for multi-view stereopsis. In contrast to recent learning based methods for 3D reconstruction, we leverage the underlying 3D geometry of the problem through feature projection and unprojection along viewing rays. By formulating these operations in a differentiable manner, we are able to learn the system end-to-end for the task of metric 3D reconstruction. End-to-end learning allows us to jointly reason about shape priors while conforming to geometri

c constraints, enabling reconstruction from much fewer images (even a single image) than required by classical approaches as well as completion of unseen surfaces. We thoroughly evaluate our approach on the ShapeNet dataset and demonstrate the benefits over classical approaches and recent learning based methods.
************************************

On Blackbox Backpropagation and Jacobian Sensing
Krzysztof M. Choromanski, Vikas Sindhwani
From a small number of calls to a given “blackbox" on random input perturbations, we show how to efficiently recover its unknown Jacobian, or estimate the left action of its Jacobian on a given vector. Our methods are based on a novel combination of compressed sensing and graph coloring techniques, and provably exploit structural prior knowledge about the Jacobian such as sparsity and symmetry while being noise robust. We demonstrate efficient backpropagation through noisy blackbox layers in a deep neural net, improved data-efficiency in the task of linearizing the dynamics of a rigid body system, and the generic ability to handle a rich class of input-output dependency structures in Jacobian estimation problems.
************************************

Learning Disentangled Representations with Semi-Supervised Deep Generative Models
Siddharth N, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, Philip Torr
Variational autoencoders (VAEs) learn representations of data by jointly training a probabilistic encoder and decoder network. Typically these models encode all features of the data into a single variable. Here we are interested in learning disentangled representations that encode distinct aspects of the data into separate variables. We propose to learn such representations using model architectures that generalise from standard VAEs, employing a general graphical model structure in the encoder and decoder. This allows us to train partially-specified models that make relatively strong assumptions about a subset of interpretable variables and rely on the flexibility of neural networks to learn representations for the remaining variables. We further define a general objective for semi-supervised learning in this model class, which can be approximated using an importance sampling procedure. We evaluate our framework's ability to learn disentangled representations, both by qualitative exploration of its generative capacity, and quantitative evaluation of its discriminative ability on a variety of models and datasets.
************************************

GP CaKe: Effective brain connectivity with causal kernels
Luca Ambrogioni, Max Hinne, Marcel Van Gerven, Eric Maris
A fundamental goal in network neuroscience is to understand how activity in one brain region drives activity elsewhere, a process referred to as effective connectivity. Here we propose to model this causal interaction using integro-differential equations and causal kernels that allow for a rich analysis of effective connectivity. The approach combines the tractability and flexibility of autoregressive modeling with the biophysical interpretability of dynamic causal modeling. The causal kernels are learned nonparametrically using Gaussian process regression, yielding an efficient framework for causal inference. We construct a novel class of causal covariance functions that enforce the desired properties of the causal kernels, an approach which we call GP CaKe. By construction, the model and its hyperparameters have biophysical meaning and are therefore easily interpretable. We demonstrate the efficacy of GP CaKe on a number of simulations and give an example of a realistic application on magnetoencephalography (MEG) data.
************************************

Certified Defenses for Data Poisoning Attacks
Jacob Steinhardt, Pang Wei W. Koh, Percy S. Liang
Machine learning systems trained on user-provided data are susceptible to data poisoning attacks, whereby malicious users inject false training data with the aim of corrupting the learned model. While recent work has proposed a number of attacks and defenses, little is understood about the worst-case loss of a defense

in the face of a determined attacker. We address this by constructing approximate upper bounds on the loss across a broad family  of attacks, for defenders that first perform outlier removal followed by empirical risk minimization. Our approximation relies on two assumptions: (1) that the dataset is large enough for  statistical concentration between train and test error to hold, and (2) that outliers  within the clean (non-poisoned) data do not have a strong effect on the model. Our bound comes paired with a candidate attack that often nearly matches the upper bound, giving us a powerful tool for quickly assessing defenses on a given dataset. Empirically, we find that even under a simple defense, the MNIST-1-7 and Dogfish datasets are resilient to attack, while in contrast the IMDB sentiment dataset can be driven from 12% to 23% test error by adding only 3% poisoned data.

************************************

Towards Generalization and Simplicity in Continuous Control
Aravind Rajeswaran, Kendall Lowrey, Emanuel V. Todorov, Sham M. Kakade

The remarkable successes of deep learning in speech recognition and computer vision have motivated efforts to adapt similar techniques to other problem domains, including reinforcement learning (RL). Consequently, RL methods have produced rich motor behaviors on simulated robot tasks, with their success largely attributed to the use of multi-layer neural networks. This work is among the first to carefully study what might be responsible for these recent advancements. Our main result calls this emerging narrative into question by showing that much simpler architectures -- based on linear and RBF parameterizations  -- achieve comparable performance to state of the art results.  We not only study different policy representations with regard to performance measures at hand, but also towards robustness to external perturbations. We again find that the learned neural network policies --- under the standard training scenarios ---  are no more robust than linear (or RBF) policies; in fact, all three are remarkably brittle. Finally, we then directly modify the training scenarios in order to favor more robust policies, and we again do not find a compelling case to favor multi-layer architectures.  Overall, this study suggests that multi-layer architectures should not be the default choice, unless a side-by-side comparison to simpler architectures shows otherwise. More generally, we hope that these results lead to more interest in carefully studying the architectural choices, and associated trade-offs, for training generalizable and robust policies.

************************************

Imagination-Augmented Agents for Deep Reinforcement Learning
Sébastien Racanière, Theophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, Razvan Pascanu, Peter Battaglia, Demis Hassabis, David Silver, Daan Wierstra

We introduce Imagination-Augmented Agents (I2As), a novel architecture for deep reinforcement learning combining model-free and model-based aspects.   In contrast to most existing model-based reinforcement learning and planning methods, which prescribe how a model should be used to arrive at a policy, I2As learn to interpret predictions from a trained environment model to construct implicit plans in arbitrary ways, by using the predictions as additional context in deep policy networks. I2As show improved data efficiency, performance, and robustness to model misspecification compared to several strong baselines.

************************************

Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles
Balaji Lakshminarayanan, Alexander Pritzel, Charles Blundell

Deep neural networks (NNs) are powerful black box predictors that have recently achieved impressive performance on a wide spectrum of tasks. Quantifying predictive uncertainty in NNs is a challenging and yet unsolved problem.  Bayesian NNs, which learn a distribution over weights, are currently the state-of-the-art for estimating predictive uncertainty; however these require significant modifications to the training procedure and are computationally expensive compared to standard (non-Bayesian)  NNs.  We propose an alternative to Bayesian NNs that is simple to implement, readily parallelizable, requires very little hyperparameter tu ning, and yields high quality predictive uncertainty estimates. Through a series of experiments on classification and regression benchmarks, we demonstrate that our method produces well-calibrated uncertainty estimates which are as good or better than approximate Bayesian NNs. To assess robustness to dataset shift, we evaluate the predictive uncertainty on test examples from known and unknown distributions, and show that our method is able to express higher uncertainty on out-of-distribution examples. We demonstrate the scalability of our method by evaluating predictive uncertainty estimates on ImageNet.

************************************

Adaptive Active Hypothesis Testing under Limited Information

Fabio Cecchi, Nidhi Hegde

We consider the problem of active sequential hypothesis testing where a Bayesian decision maker must infer the true hypothesis from a set of hypotheses. The decision maker may choose for a set of actions, where the outcome of an action is corrupted by independent noise. In this paper we consider a special case where the decision maker has limited knowledge about the distribution of observations for each action, in that only a binary value is observed. Our objective is to infer the true hypothesis with low error, while minimizing the number of action sampled. Our main results include the derivation of a lower bound on sample size for our system under limited knowledge and the design of an active learning policy that matches this lower bound and outperforms similar known algorithms.

************************************

Translation Synchronization via Truncated Least Squares

Xiangru Huang, Zhenxiao Liang, Chandrajit Bajaj, Qixing Huang

In this paper, we introduce a robust algorithm, \textsl{TranSync}, for the 1D translation synchronization problem, in which the aim is to recover the global coordinates of a set of nodes from noisy measurements of relative coordinates along an observation graph. The basic idea of TranSync is to apply truncated least squares, where the solution at each step is used to gradually prune out noisy measurements. We analyze TranSync under both deterministic and randomized noisy models, demonstrating its robustness and stability. Experimental results on synthetic and real datasets show that TranSync is superior to state-of-the-art convex formulations in terms of both efficiency and accuracy.

************************************

Limitations on Variance-Reduction and Acceleration Schemes for Finite Sums Optimization

Yossi Arjevani

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.

************************************

Flexpoint: An Adaptive Numerical Format for Efficient Training of Deep Neural Networks

Urs Köster, Tristan Webb, Xin Wang, Marcel Nassar, Arjun K. Bansal, William Constable, Oguz Elibol, Scott Gray, Stewart Hall, Luke Hornof, Amir Khosrowshahi, Carey Kloss, Ruby J. Pai, Naveen Rao

Deep neural networks are commonly developed and trained in 32-bit floating point format. Significant gains in performance and energy efficiency could be realized by training and inference in numerical formats optimized for deep learning. Despite advances in limited precision inference in recent years, training of neural networks in low bit-width remains a challenging problem. Here we present the Flexpoint data format, aiming at a complete replacement of 32-bit floating point format training and inference, designed to support modern deep network topologies without modifications. Flexpoint tensors have a shared exponent that is dynamically adjusted to minimize overflows and maximize available dynamic range. We validate Flexpoint by training AlexNet, a deep residual network and a generative adversarial network, using a simulator implemented with the \emph{neon} deep learning framework. We demonstrate that 16-bit Flexpoint closely matches 32-bit floating point in training all three models, without any need for tuning of model hy perparameters. Our results suggest Flexpoint as a promising numerical format for future hardware for training and inference.

************************************

Recursive Sampling for the Nystrom Method

Cameron Musco, Christopher Musco

We give the first algorithm for kernel Nystrom approximation that runs in linear time in the number of training points and is provably accurate for all kernel matrices, without dependence on regularity or incoherence conditions. The algorithm projects the kernel onto a set of s landmark points sampled by their ridge leverage scores, requiring just O(ns) kernel evaluations and O(ns^2) additional runtime. While leverage score sampling has long been known to give strong theoretical guarantees for Nystrom approximation, by employing a fast recursive sampling scheme, our algorithm is the first to make the approach scalable. Empirically we show that it finds more accurate kernel approximations in less time than popular techniques such as classic Nystrom approximation and the random Fourier features method.

************************************

Early stopping for kernel boosting algorithms: A general analysis with localized complexities

Yuting Wei, Fanny Yang, Martin J. Wainwright

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.

************************************

Interpolated Policy Gradient: Merging On-Policy and Off-Policy Gradient Estimation for Deep Reinforcement Learning

Shixiang (Shane) Gu, Timothy Lillicrap, Richard E. Turner, Zoubin Ghahramani, Bernhard Schölkopf, Sergey Levine

Off-policy model-free deep reinforcement learning methods using previously collected data can improve sample efficiency over on-policy policy gradient techniques. On the other hand, on-policy algorithms are often more stable and easier to use. This paper examines, both theoretically and empirically, approaches to merging on- and off-policy updates for deep reinforcement learning.  Theoretical results show that off-policy updates with a value function estimator can be interpolated with on-policy policy gradient updates whilst still satisfying performance bounds. Our analysis uses control variate methods to produce a family of policy gradient algorithms, with several recently proposed algorithms being special cases of this family. We then provide an empirical comparison of these techniques with the remaining algorithmic details fixed, and show how different mixing of off-policy gradient estimates with on-policy samples contribute to improvements in empirical performance. The final algorithm provides a generalization and unification of existing deep policy gradient techniques, has theoretical guarantees on the bias introduced by off-policy updates, and improves on the state-of-the-art model-free deep RL methods on a number of OpenAI Gym continuous control benchmarks.

************************************

Parameter-Free Online Learning via Model Selection

Dylan J. Foster, Satyen Kale, Mehryar Mohri, Karthik Sridharan

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.

************************************

Predicting User Activity Level In Point Processes With Mass Transport Equation

Yichen Wang, Xiaojing Ye, Hongyuan Zha, Le Song

Point processes are powerful tools to model user activities and have a plethora of applications in social sciences. Predicting user activities based on point processes is a central problem. However, existing works are mostly problem specific, use heuristics, or simplify the stochastic nature of point processes. In this paper, we propose a framework that provides an unbiased estimator of the probability mass function of point processes. In particular, we design a key reformulation of the prediction problem, and further derive a differential-difference equation to compute a conditional probability mass function. Our framework is applicable to general point processes and prediction tasks, and achieves superb predictive and efficiency performance in diverse real-world applications compared to state-of-arts.

************************************

The Importance of Communities for Learning to Influence

Eric Balkanski, Nicole Immorlica, Yaron Singer

We consider the canonical problem of influence maximization in social networks. Since the seminal work of Kempe, Kleinberg, and Tardos there have been two, largely disjoint efforts on this problem. The first studies the problem associated with learning the generative model that produces cascades, and the second focuses on the algorithmic challenge of identifying a set of influencers, assuming the generative model is known. Recent results on learning and optimization imply that in general, if the generative model is not known but rather learned from training data, no algorithm for influence maximization can yield a constant factor approximation guarantee using polynomially-many samples, drawn from any distribution.  In this paper we describe a simple algorithm for maximizing influence from training data. The main idea behind the algorithm is to leverage the strong community structure of social networks and identify a set of individuals who are influentials but whose communities have little overlap. Although in general, the approximation guarantee of such an algorithm is unbounded, we show that this algorithm performs well experimentally. To analyze its performance, we prove this algorithm obtains a constant factor approximation guarantee on graphs generated through the stochastic block model, traditionally used to model networks with community structure.

************************************

Gradients of Generative Models for Improved Discriminative Analysis of Tandem Mass Spectra

John T. Halloran, David M. Rocke

Tandem mass spectrometry (MS/MS) is a high-throughput technology used to identify the proteins in a complex biological sample, such as a drop of blood. A collection of spectra is generated at the output of the process, each spectrum of which is representative of a peptide (protein subsequence) present in the original complex sample. In this work, we leverage the log-likelihood gradients of generative models to improve the identification of such spectra. In particular, we show that the gradient of a recently proposed dynamic Bayesian network (DBN) may be naturally employed by a kernel-based discriminative classifier. The resulting Fisher kernel substantially improves upon recent attempts to combine generative and discriminative models for post-processing analysis, outperforming all other methods on the evaluated datasets. We extend the improved accuracy offered by the Fisher kernel framework to other search algorithms by introducing Theseus, a DBN representating a large number of widely used MS/MS scoring functions. Furthermore, with gradient ascent and max-product inference at hand, we use Theseus to learn model parameters without any supervision.

************************************

On the Optimization Landscape of Tensor Decompositions

Rong Ge, Tengyu Ma

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.

************************************

Counterfactual Fairness

Matt J. Kusner, Joshua Loftus, Chris Russell, Ricardo Silva

Machine learning can impact people with legal or ethical consequences when it is used to automate decisions in areas such as insurance, lending, hiring, and predictive policing.  In many of these scenarios, previous decisions have been made that are unfairly biased against certain subpopulations, for example those of a particular race, gender, or sexual orientation.  Since this past data may be biased, machine learning predictors must account for this to avoid perpetuating or creating discriminatory practices. In this paper, we develop a framework for modeling fairness using tools from causal inference. Our definition of counterfactual fairness captures the intuition that a decision is fair towards an individual if it  the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. We demonstrate our framework on a real-world problem of fair prediction of success in law school.

************************************

Efficient Online Linear Optimization with Approximation Algorithms

Dan Garber

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.

************************************

Inhomogeneous Hypergraph Clustering with Applications

Pan Li, Olgica Milenkovic

Hypergraph partitioning is an important problem in machine learning, computer vision and network analytics. A widely used method for hypergraph partitioning relies on minimizing a normalized sum of the costs of partitioning hyperedges across clusters. Algorithmic solutions based on this approach assume that different partitions of a hyperedge incur the same cost. However, this assumption fails to leverage the fact that different subsets of vertices within the same hyperedge may have different structural importance. We hence propose a new hypergraph clustering technique, termed inhomogeneous hypergraph partitioning, which assigns different costs to different hyperedge cuts. We prove that inhomogeneous partitioning produces a quadratic approximation to the optimal solution if the inhomogeneous costs satisfy submodularity constraints. Moreover, we demonstrate that inhomogenous partitioning offers significant performance improvements in applications such as structure learning of rankings, subspace segmentation and motif clustering.

************************************

Runtime Neural Pruning

Ji Lin, Yongming Rao, Jiwen Lu, Jie Zhou

In this paper, we propose a Runtime Neural Pruning (RNP) framework which prunes the deep neural network dynamically at the runtime. Unlike existing neural pruning methods which produce a fixed pruned model for deployment, our method preserves the full ability of the original network and conducts pruning according to the input image and current feature maps adaptively. The pruning is performed in a bottom-up, layer-by-layer manner, which we model as a Markov decision process and use reinforcement learning for training. The agent judges the importance of each convolutional kernel and conducts channel-wise pruning conditioned on different samples, where the network is pruned more when the image is easier for the task. Since the ability of network is fully preserved, the balance point is easily adjustable according to the available resources. Our method can be applied to off-the-shelf network structures and reach a better tradeoff between speed and accuracy, especially with a large pruning rate.

************************************

Train longer, generalize better: closing the generalization gap in large batch training of neural networks

Elad Hoffer, Itay Hubara, Daniel Soudry

Background: Deep learning models are typically trained using stochastic gradient descent or one of its variants. These methods update the weights using their gradient, estimated from a small fraction of the training data. It has been observed that when using large batch sizes there is a persistent degradation in generalization performance -  known as the "generalization gap" phenomenon. Identifying the origin of this gap and closing it had remained an open problem.  Contributions: We examine the initial high learning rate training phase. We find that the weight distance from its initialization grows logarithmically with the number of weight updates. We therefore propose a "random walk on a random landscape" statistical model which is known to exhibit similar "ultra-slow" diffusion behavior. Following this hypothesis we conducted experiments to show empirically that the "generalization gap" stems from the relatively small number of updates rather than the batch size, and can be completely eliminated by adapting the training regime used. We further investigate different techniques to train models in the large-batch regime and present a novel algorithm named "Ghost Batch Normalization" which enables significant decrease in the generalization gap without increasing the number of updates. To validate our findings we conduct several additional experiments on MNIST, CIFAR-10, CIFAR-100 and ImageNet. Finally, we reassess common practices and beliefs concerning training of deep models and suggest they may not be optimal to achieve good generalization.

************************************

Monte-Carlo Tree Search by Best Arm Identification

Emilie Kaufmann, Wouter M. Koolen

Recent advances in bandit tools and techniques for sequential learning are steadily enabling new applications and are promising the resolution of a range of challenging related problems. We study the game tree search problem, where the goal is to quickly identify the optimal move in a given game tree by sequentially sampling its stochastic payoffs. We develop new algorithms for trees of arbitrary depth, that operate by summarizing all deeper levels of the tree into confidence intervals at depth one, and applying a best arm identification procedure at the root. We prove new sample complexity guarantees with a refined dependence on the problem instance. We show experimentally that our algorithms outperform existing elimination-based algorithms and match  previous special-purpose methods for depth-two trees.

************************************

Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, Wang-chun WOO

With the goal of making high-resolution forecasts of regional rainfall, precipitation nowcasting has become an important and fundamental technology underlying various public services ranging from rainstorm warnings to flight safety. Recently, the Convolutional LSTM (ConvLSTM) model has been shown to outperform traditional optical flow based methods for precipitation nowcasting, suggesting that deep learning models have a huge potential for solving the problem. However, the convolutional recurrence structure in ConvLSTM-based models is location-invariant while natural motion and transformation (e.g., rotation) are location-variant in general. Furthermore, since deep-learning-based precipitation nowcasting is a newly emerging area, clear evaluation protocols have not yet been established. To address these problems, we propose both a new model and a benchmark for precipitation nowcasting. Specifically, we go beyond ConvLSTM and propose the Trajectory GRU (TrajGRU) model that can actively learn the location-variant structure for recurrent connections. Besides, we provide a benchmark that includes a real-world large-scale dataset from the Hong Kong Observatory, a new training loss, and a comprehensive evaluation protocol to facilitate future research and gauge the state of the art.

************************************

Scalable Model Selection for Belief Networks

Zhao Song, Yusuke Muraoka, Ryohei Fujimaki, Lawrence Carin

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.

************************************

Collaborative Deep Learning in Fixed Topology Networks

Zhanhong Jiang, Aditya Balu, Chinmay Hegde, Soumik Sarkar

There is significant recent interest to parallelize deep learning algorithms in order to handle the enormous growth in data and model sizes. While most advances focus on model parallelization and engaging multiple computing agents via using a central parameter server, aspect of data parallelization along with decentralized computation has not been explored sufficiently. In this context, this paper presents a new consensus-based distributed SGD (CDSGD) (and its momentum variant, CDMSGD) algorithm for collaborative deep learning over fixed topology networks that enables data parallelization as well as decentralized computation. Such a framework can be extremely useful for learning agents with access to only local/private data in a communication constrained environment. We analyze the convergence properties of the proposed algorithm with strongly convex and nonconvex objective functions with fixed and diminishing step sizes using concepts of Lyapunov function construction. We demonstrate the efficacy of our algorithms in comparison with the baseline centralized SGD and the recently proposed federated averaging algorithm (that also enables data parallelism) based on benchmark datasets such as MNIST, CIFAR-10 and CIFAR-100.

************************************

On the Complexity of Learning Neural Networks

Le Song, Santosh Vempala, John Wilmes, Bo Xie

The stunning empirical successes of neural networks currently lack rigorous theoretical explanation. What form would such an explanation take, in the face of existing complexity-theoretic lower bounds? A first step might be to show that data generated by neural networks with a single hidden layer, smooth activation functions and benign input distributions can be learned efficiently. We demonstrate here a comprehensive lower bound ruling out this possibility: for a wide class of activation functions (including all currently used), and inputs drawn from any logconcave distribution, there is a family of one-hidden-layer functions whose output is a sum gate that are hard to learn in a precise sense: any statistical query algorithm (which includes all known variants of stochastic gradient descent with any loss function) needs an exponential number of queries even using tolerance inversely proportional to the input dimensionality. Moreover, this hard family of functions is realizable with a small (sublinear in dimension) number of activation units in the single hidden layer. The lower bound is also robust to small perturbations of the true weights. Systematic experiments illustrate a phase transition in the training error as predicted by the analysis.

************************************

A Sample Complexity Measure with Applications to Learning Optimal Auctions

Vasilis Syrgkanis

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.

************************************

On Optimal Generalizability in Parametric Learning

Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, Vahid Tarokh

We consider the parametric learning problem, where the objective of the learner is determined by a parametric loss function. Employing empirical risk minimization with possibly regularization, the inferred parameter vector will be biased toward the training samples. Such bias is measured by the cross validation procedure in practice where the data set is partitioned into a training set used for training and a validation set, which is not used in training and is left to measure the out-of-sample performance. A classical cross validation strategy is the leave-one-out cross validation (LOOCV) where one sample is left out for validation and training is done on the rest of the samples that are presented to the learner, and this process is repeated on all  of the samples. LOOCV is rarely used in practice due to the high computational complexity. In this paper, we first develop a computationally efficient approximate LOOCV (ALOOCV) and provide theoretical guarantees for its performance. Then we use ALOOCV to provide an optimization algorithm for finding the regularizer in the empirical risk minimization framework. In our numerical experiments, we illustrate the accuracy and efficiency of ALOOCV  as well as our proposed framework for the optimization of the regularizer.

************************************

K-Medoids For K-Means Seeding

James Newling, François Fleuret

We show experimentally that the algorithm CLARANS of Ng and Han (1994) finds better K-medoids solutions than the Voronoi iteration algorithm of Hastie et al. (2001). This finding, along with the similarity between the Voronoi iteration algorithm and Lloyd's K-means algorithm, motivates us to use CLARANS as a K-means initializer. We show that CLARANS outperforms other algorithms on 23/23 datasets with a mean decrease over k-means++ of 30% for initialization mean squared error (MSE) and 3% for final MSE. We introduce algorithmic improvements to CLARANS which improve its complexity and runtime, making it a viable initialization scheme for large datasets.

************************************

Learning Deep Structured Multi-Scale Features using Attention-Gated CRFs for Contour Prediction

Dan Xu, Wanli Ouyang, Xavier Alameda-Pineda, Elisa Ricci, Xiaogang Wang, Nicu Sebe

Recent works have shown that exploiting multi-scale representations deeply learned via convolutional neural networks (CNN) is of tremendous importance for accurate contour detection. This paper presents a novel approach for predicting contours which advances the state of the art in two fundamental aspects, i.e. multi-scale feature generation and fusion. Different from previous works directly considering multi-scale feature maps obtained from the inner layers of a primary CNN architecture, we introduce a hierarchical deep model which produces more rich and complementary representations. Furthermore, to refine and robustly fuse the representations learned at different scales, the novel Attention-Gated Conditional Random Fields (AG-CRFs) are proposed. The experiments ran on two publicly available datasets (BSDS500 and NYUDv2) demonstrate the effectiveness of the latent AG-CRF model and of the overall hierarchical framework.

************************************

Geometric Descent Method for Convex Composite Minimization

Shixiang Chen, Shiqian Ma, Wei Liu

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.

************************************

Label Efficient Learning of Transferable Representations acrosss Domains and Tasks

Zelun Luo, Yuliang Zou, Judy Hoffman, Li F. Fei-Fei

We propose a framework that learns a representation transferable across different domains and tasks in a data efficient manner. Our approach battles domain shift with a domain adversarial loss, and generalizes the embedding to novel task using a metric learning-based approach. Our model is simultaneously optimized on labeled source data and unlabeled or sparsely labeled data in the target domain. Our method shows compelling results on novel classes within a new domain even when only a few labeled examples per class are available, outperforming the prevalent fine-tuning approach. In addition, we demonstrate the effectiveness of our framework on the transfer learning task from image object recognition to video action recognition.

************************************

Improving Regret Bounds for Combinatorial Semi-Bandits with Probabilistically Triggered Arms and Its Applications

Qinshi Wang, Wei Chen

We study combinatorial multi-armed bandit with probabilistically triggered arms (CMAB-T) and semi-bandit feedback. We resolve a serious issue in the prior CMAB-T studies where the regret bounds contain a possibly exponentially large factor of 1/p, where p is the minimum positive probability that an arm is triggered by any action. We address this issue by introducing a triggering probability modulated (TPM) bounded smoothness condition into the influence maximization bandit an d combinatorial cascading bandit satisfy this TPM condition. As a result, we completely remove the factor of 1/p* from the regret bounds, achieving significantly better regret bounds for influence maximization and cascading bandits than before. Finally, we provide lower bound results showing that the factor 1/p* is unavoidable for general CMAB-T problems, suggesting that the TPM condition is crucial in removing this factor.

************************************

Matching neural paths: transfer from recognition to correspondence search
Nikolay Savinov, Lubor Ladicky, Marc Pollefeys

Many machine learning tasks require finding per-part correspondences between objects. In this work we focus on low-level correspondences --- a highly ambiguous matching problem. We propose to use a hierarchical semantic representation of the objects, coming from a convolutional neural network, to solve this ambiguity. Training it for low-level correspondence prediction directly might not be an option in some domains where the ground-truth correspondences are hard to obtain. We show how transfer from recognition can be used to avoid such training. Our idea is to mark parts as "matching" if their features are close to each other at all the levels of convolutional feature hierarchy (neural paths). Although the overall number of such paths is exponential in the number of layers, we propose a polynomial algorithm for aggregating all of them in a single backward pass. The empirical validation is done on the task of stereo correspondence and demonstrates that we achieve competitive results among the methods which do not use labeled target domain data.

************************************

Convergence Analysis of Two-layer Neural Networks with ReLU Activation
Yuanzhi Li, Yang Yuan

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.

************************************

Quantifying how much sensory information in a neural code is relevant for behavior
Giuseppe Pica, Eugenio Piasini, Houman Safaai, Caroline Runyan, Christopher Harvey, Mathew Diamond, Christoph Kayser, Tommaso Fellin, Stefano Panzeri

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.

************************************

Self-supervised Learning of Motion Capture
Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, Katerina Fragkiadaki

Current state-of-the-art solutions for motion capture from a single camera are optimization driven: they optimize the parameters of a 3D human model so that its re-projection matches measurements in the video (e.g. person segmentation, optical flow, keypoint detections etc.). Optimization models are susceptible to local minima. This has been the bottleneck that forced using clean green-screen like backgrounds at capture time, manual initialization, or switching to multiple cameras as input resource. In this work, we propose a learning based motion capture model for single camera input. Instead of optimizing mesh and skeleton parameters directly, our model optimizes neural network weights that predict 3D shape and skeleton configurations given a monocular RGB video. Our model is trained using a combination of strong supervision from synthetic data, and self-supervision from differentiable rendering of (a) skeletal keypoints, (b) dense 3D mesh motion, and (c) human-background segmentation, in an end-to-end framework. Empirically we show our model combines the best of both worlds of supervised learning and test-time optimization: supervised learning initializes the model parameters in the right regime, ensuring good pose and surface initialization at test time, without manual effort. Self-supervision by back-propagating through differentiable rendering allows (unsupervised) adaptation of the model to the test data, and offers much tighter fit than a pretrained fixed model. We show that the proposed model improves with experience and converges to low-error solutions where previous optimization methods fail.

************************************

Toward Goal-Driven Neural Network Models for the Rodent Whisker-Trigeminal System

Chengxu Zhuang, Jonas Kubilius, Mitra JZ Hartmann, Daniel L. Yamins

In large part, rodents “see” the world through their whiskers, a powerful tactile sense enabled by a series of brain areas that form the whisker-trigeminal system. Raw sensory data arrives in the form of mechanical input to the exquisitely sensitive, actively-controllable whisker array, and is processed through a sequence of neural circuits, eventually arriving in cortical regions that communicate with decision making and memory areas. Although a long history of experimental studies has characterized many aspects of these processing stages, the computational operations of the whisker-trigeminal system remain largely unknown. In the present work, we take a goal-driven deep neural network (DNN) approach to modeling these computations. First, we construct a biophysically-realistic model of the rat whisker array. We then generate a large dataset of whisker sweeps across a wide variety of 3D objects in highly-varying poses, angles, and speeds. Next, we train DNNs from several distinct architectural families to solve a shape recognition task in this dataset. Each architectural family represents a structurally-distinct hypothesis for processing in the whisker-trigeminal system, corresponding to different ways in which spatial and temporal information can be integrated. We find that most networks perform poorly on the challenging shape recognition task, but that specific architectures from several families can achieve reasonable performance levels. Finally, we show that Representational Dissimilarity Matrices (RDMs), a tool for comparing population codes between neural systems, can separate these higher performing networks with data of a type that could plausibly be collected in a neurophysiological or imaging experiment. Our results are a proof-of-concept that DNN models of the whisker-trigeminal system are potentially within reach.

************************************

Clustering Billions of Reads for DNA Data Storage

Cyrus Rashtchian, Konstantin Makarychev, Miklos Racz, Siena Ang, Djordje Jevdjic, Sergey Yekhanin, Luis Ceze, Karin Strauss

Storing data in synthetic DNA offers the possibility of improving information density and durability  by several orders of magnitude compared to current storage technologies. However, DNA data storage requires a computationally intensive process to retrieve the data. In particular, a crucial step in the data retrieval pipeline involves clustering billions of strings with respect to edit distance. Datasets in this domain have many notable properties, such as containing a very large number of small clusters that are well-separated in the edit distance metric space. In this regime, existing algorithms are unsuitable because of either their long running time or low accuracy. To address this issue, we present a novel distributed algorithm for approximately computing the underlying clusters. Our algorithm converges efficiently on any dataset that satisfies certain separability properties, such as those coming from DNA data storage systems. We also prove that, under these assumptions, our algorithm is robust to outliers and high levels of noise. We provide empirical justification of the accuracy, scalability, and convergence of our algorithm on real and synthetic data. Compared to the state-of-the-art algorithm for clustering DNA sequences, our algorithm simultaneously achieves higher accuracy and a 1000x speedup on three real datasets.

************************************

AIDE: An algorithm for measuring the accuracy of probabilistic inference algorithms

Marco Cusumano-Towner, Vikash K. Mansinghka

Approximate probabilistic inference algorithms are central to many fields. Examples include sequential Monte Carlo inference in robotics, variational inference in machine learning, and Markov chain Monte Carlo inference in statistics. A key problem faced by practitioners is measuring the accuracy of an approximate infe rence algorithm on a specific data set. This paper introduces the auxiliary inference divergence estimator (AIDE), an algorithm for measuring the accuracy of approximate inference algorithms. AIDE is based on the observation that inference algorithms can be treated as probabilistic models and the random variables used within the inference algorithm can be viewed as auxiliary variables. This view leads to a new estimator for the symmetric KL divergence between the approximating distributions of two inference algorithms. The paper illustrates application of AIDE to algorithms for inference in regression, hidden Markov, and Dirichlet process mixture models. The experiments show that AIDE captures the qualitative behavior of a broad class of inference algorithms and can detect failure modes of inference algorithms that are missed by standard heuristics.

************************************

Information-theoretic analysis of generalization capability of learning algorithms

Aolin Xu, Maxim Raginsky

We derive upper bounds on the generalization error of a learning algorithm in terms of the mutual information between its input and output. The bounds provide an information-theoretic understanding of generalization in learning problems, and give theoretical guidelines for striking the right balance between data fit and generalization by controlling the input-output mutual information.  We propose a number of methods for this purpose, among which are algorithms that regularize the ERM algorithm with relative entropy or with random noise. Our work extends and leads to nontrivial improvements on the recent results of Russo and Zou.

************************************

MarrNet: 3D Shape Reconstruction via 2.5D Sketches

Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, Josh Tenenbaum

3D object reconstruction from a single image is a highly under-determined problem, requiring strong prior knowledge of plausible 3D shapes. This introduces challenge for learning-based approaches, as 3D object annotations in real images are scarce. Previous work chose to train on synthetic data with ground truth 3D information, but suffered from the domain adaptation issue when tested on real data.  In this work, we propose an end-to-end trainable framework, sequentially estimating 2.5D sketches and 3D object shapes. Our disentangled, two-step formulation has three advantages. First, compared to full 3D shape, 2.5D sketches are much easier to be recovered from a 2D image, and to transfer from synthetic to real data. Second, for 3D reconstruction from the 2.5D sketches, we can easily transfer the learned model on synthetic data to real images, as rendered 2.5D sketches are invariant to object appearance variations in real images, including lighting, texture, etc. This further relieves the domain adaptation problem. Third, we derive differentiable projective functions from 3D shape to 2.5D sketches, making the framework end-to-end trainable on real images, requiring no real-image annotations. Our framework achieves state-of-the-art performance on 3D shape reconstruction.

************************************

Flexible statistical inference for mechanistic models of neural dynamics

Jan-Matthis Lueckmann, Pedro J. Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, Jakob H. Macke

Mechanistic models of single-neuron dynamics have been extensively studied in computational neuroscience. However, identifying which models can quantitatively reproduce empirically measured data has been challenging. We propose to overcome this limitation by using likelihood-free inference approaches (also known as Approximate Bayesian Computation, ABC) to perform full Bayesian inference on single-neuron models. Our approach builds on recent advances in ABC by learning a neural network which maps features of the observed data to the posterior distribution over parameters. We learn a Bayesian mixture-density network approximating the posterior over multiple rounds of adaptively chosen simulations. Furthermore, we propose an efficient approach for handling missing features and parameter settings for which the simulator fails, as well as a strategy for automatically learning relevant features using recurrent neural networks. On synthetic data, our approach efficiently estimates posterior distributions and recovers ground-truth parameters. On in-vitro recordings of membrane voltages, we recover multivariate posteriors over biophysical parameters, which yield model-predicted voltage traces that accurately match empirical data. Our approach will enable neuroscientists to perform Bayesian inference on complex neuron models without having to design model-specific algorithms, closing the gap between mechanistic and statistical approaches to single-neuron modelling.

************************************

ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching

Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, Lawrence Carin

We investigate the non-identifiability issues associated with bidirectional adversarial training for joint distribution matching. Within a framework of conditional entropy, we propose both adversarial and non-adversarial approaches to learn desirable matched joint distributions for unsupervised and supervised tasks. We unify a broad family of adversarial models as joint distribution matching problems. Our approach stabilizes learning of unsupervised bidirectional adversarial learning methods. Further, we introduce an extension for semi-supervised learning tasks. Theoretical results are validated in synthetic data and real-world applications.

************************************

Speeding Up Latent Variable Gaussian Graphical Model Estimation via Nonconvex Optimization

Pan Xu, Jian Ma, Quanquan Gu

We study the estimation of the latent variable Gaussian graphical model (LVGGM), where the precision matrix is the superposition of a sparse matrix and a low-rank matrix.  In order to speed up the estimation of the sparse plus low-rank components, we propose a sparsity constrained maximum likelihood estimator based on matrix factorization and an efficient alternating gradient descent algorithm with hard thresholding to solve it. Our algorithm is orders of magnitude faster than the convex relaxation based methods for LVGGM. In addition, we prove that our algorithm is guaranteed to linearly converge to the unknown sparse and low-rank components up to the optimal statistical precision. Experiments on both synthetic and genomic data demonstrate the superiority of our algorithm over the state-of-the-art algorithms and corroborate our theory.

************************************

Sparse convolutional coding for neuronal assembly detection

Sven Peter, Elke Kirschbaum, Martin Both, Lee Campbell, Brandon Harvey, Conor Heins, Daniel Durstewitz, Ferran Diego, Fred A. Hamprecht

Cell assemblies, originally proposed by Donald Hebb (1949), are subsets of neurons firing in a temporally coordinated way that gives rise to repeated motifs supposed to underly neural representations and information processing. Although Hebb's original proposal dates back many decades, the detection of assemblies and their role in coding is still an open and current research topic, partly because simultaneous recordings from large populations of neurons became feasible only relatively recently. Most current and easy-to-apply computational techniques focus on the identification of strictly synchronously spiking neurons. In this paper we propose a new algorithm, based on sparse convolutional coding, for detecting recurrent motifs of arbitrary structure up to a given length. Testing of our algorithm on synthetically generated datasets shows that it outperforms established methods and accurately identifies the temporal structure of embedded assemblies, even when these contain overlapping neurons or when strong background noise is present. Moreover, exploratory analysis of experimental datasets from hippocampal slices and cortical neuron cultures have provided promising results.

************************************

Neural Networks for Efficient Bayesian Decoding of Natural Images from Retinal Neurons

Nikhil Parthasarathy, Eleanor Batty, William Falcon, Thomas Rutten, Mohit Rajpal, E.J. Chichilnisky, Liam Paninski

Decoding sensory stimuli from neural signals can be used to reveal how we sense our physical environment, and is valuable for the design of brain-machine interfaces. However, existing linear techniques for neural decoding may not fully reveal or exploit the fidelity of the neural signal. Here we develop a new approximate Bayesian method for decoding natural images from the spiking activity of populations of retinal ganglion cells (RGCs). We sidestep known computational challenges with Bayesian inference by exploiting artificial neural networks developed for computer vision, enabling fast nonlinear decoding that incorporates natural scene statistics implicitly. We use a decoder architecture that first linearly reconstructs an image from RGC spikes, then applies a convolutional autoencoder to enhance the image. The resulting decoder, trained on natural images and simulated neural responses, significantly outperforms linear decoding, as well as simple point-wise nonlinear decoding. These results provide a tool for the assessment and optimization of retinal prosthesis technologies, and reveal that the retina may provide a more accurate representation of the visual scene than previously appreciated.

************************************

Plan, Attend, Generate: Planning for Sequence-to-Sequence Models

Caglar Gulcehre, Francis Dutil, Adam Trischler, Yoshua Bengio

We investigate the integration of a planning mechanism into sequence-to-sequence models using attention. We develop a model which can plan ahead in the future when it computes its alignments between input and output sequences, constructing a matrix of proposed future alignments and a commitment vector that governs whether to follow or recompute the plan. This mechanism is inspired by the recently proposed strategic attentive reader and writer (STRAW) model for Reinforcement Learning. Our proposed model is end-to-end trainable using primarily differentiable operations. We show that it outperforms a strong baseline on character-level translation tasks from WMT'15, the algorithmic task of finding Eulerian circuits of graphs, and question generation from the text. Our analysis demonstrates that the model computes qualitatively intuitive alignments, converges faster than the baselines, and achieves superior performance with fewer parameters.

************************************

Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems

Yonatan Belinkov, James Glass

Neural networks have become ubiquitous in automatic speech recognition systems. While neural networks are typically used as acoustic models in more complex systems, recent studies have explored end-to-end speech recognition systems based on neural networks, which can be trained to directly predict text from input acoustic features. Although such systems are conceptually elegant and simpler than traditional systems, it is less obvious how to interpret the trained models. In this work, we analyze the speech representations learned by a deep end-to-end model that is based on convolutional and recurrent layers, and trained with a connectionist temporal classification (CTC) loss. We use a pre-trained model to generate frame-level features which are given to a classifier that is trained on frame classification into phones. We evaluate representations from different layers of the deep model and compare their quality for predicting phone labels. Our experiments shed light on important aspects of the end-to-end model such as layer depth, model complexity, and other design choices.

************************************

Multi-Task Learning for Contextual Bandits

Aniket Anand Deshmukh, Urun Dogan, Clay Scott

Contextual bandits are a form of multi-armed bandit in which the agent has access to predictive side information (known as the context) for each arm at each time step, and have been used to model personalized news recommendation, ad placement, and other applications. In this work, we propose a multi-task learning framework for contextual bandit problems. Like multi-task learning in the batch setting, the goal is to leverage similarities in contexts for different arms so as to improve the agent's ability to predict rewards from contexts. We propose an upper confidence bound-based multi-task learning algorithm for contextual bandits, establish a corresponding regret bound, and interpret this bound to quantify the advantages of learning in the presence of high task (arm) similarity. We also describe an effective scheme for estimating task similarity from data, and demonstrate our algorithm's performance on several data sets.

************************************

Temporal Coherency based Criteria for Predicting Video Frames using Deep Multi-stage Generative Adversarial Networks

Prateep Bhattacharjee, Sukhendu Das

Predicting the future from a sequence of video frames has been recently a sought after yet challenging task in the field of computer vision and machine learning. Although there have been efforts for tracking using motion trajectories and flow features, the complex problem of generating unseen frames has not been studied extensively. In this paper, we deal with this problem using convolutional models within a multi-stage Generative Adversarial Networks (GAN) framework. The proposed method uses two stages of GANs to generate a crisp and clear set of future frames. Although GANs have been used in the past for predicting the future, none of the works consider the relation between subsequent frames in the temporal dimension. Our main contribution lies in formulating two objective functions based on the Normalized Cross Correlation (NCC) and the Pairwise Contrastive Divergence (PCD) for solving this problem. This method, coupled with the traditional L1 loss, has been experimented with three real-world video datasets, viz. Sports-1M, UCF-101 and the KITTI. Performance analysis reveals superior results over the recent state-of-the-art methods.

************************************

Improving the Expected Improvement Algorithm

Chao Qin, Diego Klabjan, Daniel Russo

The expected improvement (EI) algorithm is a popular strategy for information collection in optimization under uncertainty. The algorithm is widely known to be too greedy, but nevertheless enjoys wide use due to its simplicity and ability to handle uncertainty and noise in a coherent decision theoretic framework. To provide rigorous insight into EI, we study its properties in a simple setting of Bayesian optimization where the domain consists of a finite grid of points. This is the so-called best-arm identification problem, where the goal is to allocate measurement effort wisely to confidently identify the best arm using a small number of measurements. In this framework, one can show formally that EI is far from optimal. To overcome this shortcoming, we introduce a simple modification of the expected improvement algorithm. Surprisingly, this simple change results in an algorithm that is asymptotically optimal for Gaussian best-arm identification problems, and provably outperforms standard EI by an order of magnitude.

************************************

Towards Accurate Binary Convolutional Neural Network

Xiaofan Lin, Cong Zhao, Wei Pan

We introduce a novel scheme to train binary convolutional neural networks (CNNs) -- CNNs with weights and activations constrained to {-1,+1} at run-time. It has been known that using binary weights and activations drastically reduce memory size and accesses, and can replace arithmetic operations with more efficient bitwise operations, leading to much faster test-time inference and lower power consumption. However, previous works on binarizing CNNs usually result in severe prediction accuracy degradation. In this paper, we address this issue with two major innovations: (1) approximating full-precision weights with the linear combination of multiple binary weight bases; (2) employing multiple binary activations to alleviate information loss. The implementation of the resulting binary CNN, denoted as ABC-Net, is shown to achieve much closer performance to its full-precision counterpart, and even reach the comparable prediction accuracy on ImageNet and forest trail datasets, given adequate binary weight bases and activations.

************************************

Spectrally-normalized margin bounds for neural networks

Peter L. Bartlett, Dylan J. Foster, Matus J. Telgarsky

This paper presents a margin-based multiclass generalization bound for neural networks that scales with their margin-normalized "spectral complexity": their Lipschitz constant, meaning the product of the spectral norms of the weight matrice s, times a certain correction factor. This bound is empirically investigated for a standard AlexNet network trained with SGD on the MNIST and CIFAR10 datasets, with both original and random labels;  the bound, the Lipschitz constants, and the excess risks are all in direct correlation, suggesting both that SGD selects predictors whose complexity scales with the difficulty of the learning task, and secondly that the presented bound is sensitive to this complexity.

************************************

Consistent Multitask Learning with Nonlinear Output Relations
Carlo Ciliberto, Alessandro Rudi, Lorenzo Rosasco, Massimiliano Pontil

Key to multitask learning is exploiting the relationships between different tasks to improve prediction performance. Most previous methods have focused on the case where tasks relations can be modeled as linear operators and regularization approaches can be used successfully. However, in practice assuming the tasks to be linearly related is often restrictive, and allowing for nonlinear structures is a challenge. In this paper, we tackle this issue by casting the problem within the framework of structured prediction. Our main contribution is a novel algorithm for learning multiple tasks which are related by a system of nonlinear equations that their joint outputs need to satisfy. We show that our algorithm can be efficiently implemented and study its generalization properties, proving universal consistency and learning rates. Our theoretical analysis highlights the benefits of non-linear multitask learning over learning the tasks independently. Encouraging experimental results show the benefits of the proposed method in practice.

************************************

Deep Recurrent Neural Network-Based Identification of Precursor microRNAs
Seunghyun Park, Seonwoo Min, Hyun-Soo Choi, Sungroh Yoon

MicroRNAs (miRNAs) are small non-coding ribonucleic acids (RNAs) which play key roles in post-transcriptional gene regulation. Direct identification of mature miRNAs is infeasible due to their short lengths, and researchers instead aim at identifying precursor miRNAs (pre-miRNAs). Many of the known pre-miRNAs have distinctive stem-loop secondary structure, and structure-based filtering is usually the first step to predict the possibility of a given sequence being a pre-miRNA. To identify new pre-miRNAs that often have non-canonical structure, however, we need to consider additional features other than structure. To obtain such additional characteristics, existing computational methods rely on manual feature extraction, which inevitably limits the efficiency, robustness, and generalization of computational identification. To address the limitations of existing approaches, we propose a pre-miRNA identification method that incorporates (1) a deep recurrent neural network (RNN) for automated feature learning and classification, (2) multimodal architecture for seamless integration of prior knowledge (secondary structure), (3) an attention mechanism for improving long-term dependence modeling, and (4) an RNN-based class activation mapping for highlighting the learned representations that can contrast pre-miRNAs and non-pre-miRNAs. In our experiments with recent benchmarks, the proposed approach outperformed the compared state-of-the-art alternatives in terms of various performance metrics.

************************************

Boltzmann Exploration Done Right
Nicolò Cesa-Bianchi, Claudio Gentile, Gabor Lugosi, Gergely Neu

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.

************************************

End-to-end Differentiable Proving
Tim Rocktäschel, Sebastian Riedel

We introduce deep neural networks for end-to-end differentiable theorem proving that operate on dense vector representations of symbols.  These neural networks are recursively constructed by following the backward chaining algorithm as used in Prolog.  Specifically, we replace symbolic unification with a differentiable computation on vector representations of symbols using a radial basis function kernel, thereby combining symbolic reasoning with learning subsymbolic vector representations.  The resulting neural network can be trained to infer facts from a given incomplete knowledge base using gradient descent.  By doing so, it learns to (i) place representations of similar symbols in close proximity in a vector space, (ii) make use of such similarities to prove facts, (iii) induce logical rules, and (iv) it can use provided and induced logical rules for complex multi-hop reasoning.  On four benchmark knowledge bases we demonstrate that this architecture outperforms ComplEx, a state-of-the-art neural link prediction model, while at the same time inducing interpretable function-free first-order logic rules.

************************************

Matching on Balanced Nonlinear Representations for Treatment Effects Estimation
Sheng Li, Yun Fu

Estimating treatment effects from observational data is challenging due to the missing counterfactuals. Matching is an effective strategy to tackle this problem. The widely used matching estimators such as nearest neighbor matching (NNM) pair the treated units with the most similar control units in terms of covariates, and then estimate treatment effects accordingly. However, the existing matching estimators have poor performance when the distributions of control and treatment groups are unbalanced. Moreover, theoretical analysis suggests that the bias of causal effect estimation would increase with the dimension of covariates. In this paper, we aim to address these problems by learning low-dimensional balanced and nonlinear representations (BNR) for observational data. In particular, we convert counterfactual prediction as a classification problem, develop a kernel learning model with domain adaptation constraint, and design a novel matching estimator. The dimension of covariates will be significantly reduced after projecting data to a low-dimensional subspace. Experiments on several synthetic and real-world datasets demonstrate the effectiveness of our approach.

************************************

Tomography of the London Underground: a Scalable Model for Origin-Destination Data
Nicolò Colombo, Ricardo Silva, Soong Moon Kang

The paper addresses the classical network tomography problem of inferring local traffic given origin-destination observations. Focussing on large complex public transportation systems, we build a scalable model that exploits input-output information to estimate the unobserved link/station loads and the users path preferences. Based on the reconstruction of the users' travel time distribution, the model is flexible enough to capture possible different path-choice strategies and correlations between users travelling on similar paths at similar times. The corresponding likelihood function is intractable for medium or large-scale networks and we propose two distinct strategies, namely the exact maximum-likelihood inference of an approximate but tractable model and the variational inference of the original intractable model. As an application of our approach, we consider the emblematic case of the London Underground network, where a tap-in/tap-out system tracks the start/exit time and location of all journeys in a day. A set of synthetic simulations and real data provided by Transport For London are used to validate and test the model on the predictions of observable and unobservable quantities.

************************************

Gaussian process based nonlinear latent structure discovery in multivariate spike train data
Anqi Wu, Nicholas A. Roy, Stephen Keeley, Jonathan W. Pillow

A large body of recent work focuses on methods for extracting low-dimensional latent structure from multi-neuron spike train data. Most such methods employ either linear latent dynamics or linear mappings from latent space to log spike rates. Here we propose a doubly nonlinear latent variable model that can identify low-dimensional structure underlying apparently high-dimensional spike train data. We introduce the Poisson Gaussian-Process Latent Variable Model (P-GPLVM), which consists of Poisson spiking observations and two underlying Gaussian processes—one governing a temporal latent variable and another governing a set of nonline

ar tuning curves. The use of nonlinear tuning curves enables discovery of low-dimensional latent structure even when spike responses exhibit high linear dimensionality (e.g., as found in hippocampal place cell codes). To learn the model from data, we introduce the decoupled Laplace approximation, a fast approximate inference method that allows us to efficiently optimize the latent path while marginalizing over tuning curves. We show that this method outperforms previous Laplace-approximation-based inference methods in both the speed of convergence and accuracy. We apply the model to spike trains recorded from hippocampal place cells and show that it compares favorably to a variety of previous methods for latent structure discovery, including variational auto-encoder (VAE) based methods that parametrize the nonlinear mapping from latent space to spike rates with a deep neural network.
************************************
Multi-Objective Non-parametric Sequential Prediction
Guy Uziel, Ran El-Yaniv
Online-learning research has mainly been focusing on minimizing one objective function. In many real-world applications, however, several objective functions have to be considered simultaneously. Recently, an algorithm for dealing with several objective functions in the i.i.d. case has been presented.  In this paper, we extend the multi-objective framework to the case of stationary and ergodic processes, thus  allowing dependencies among observations. We first identify an asymptomatic lower bound for any prediction strategy and then present an algorithm whose predictions achieve the optimal solution while  fulfilling  any continuous and convex constraining criterion.
************************************
Optimal Sample Complexity of M-wise Data for Top-K Ranking
Minje Jang, Sunghyun Kim, Changho Suh, Sewoong Oh
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.
************************************
From which world is your graph
Cheng Li, Felix MF Wong, Zhenming Liu, Varun Kanade
Discovering statistical structure from links is a fundamental problem in the analysis of social networks. Choosing a misspecified model, or equivalently, an incorrect inference algorithm will result in an invalid analysis or even falsely uncover patterns that are in fact artifacts of the model. This work focuses on unifying two of the most widely used link-formation models: the stochastic block model (SBM) and the small world (or latent space) model (SWM). Integrating techniques from kernel learning, spectral graph theory, and nonlinear dimensionality reduction, we develop the first statistically sound polynomial-time algorithm to discover latent patterns in sparse graphs for both models. When the network comes from an SBM, the algorithm outputs a block structure. When it is from an SWM, the algorithm outputs estimates of each node's latent position.
************************************
An Empirical Bayes Approach to Optimizing Machine Learning Algorithms
James McInerney
There is rapidly growing interest in using Bayesian optimization to tune model and inference hyperparameters for machine learning algorithms that take a long time to run. For example, Spearmint is a popular software package for selecting the optimal number of layers and learning rate in neural networks. But given that there is uncertainty about which hyperparameters give the best predictive performance, and given that fitting a model for each choice of hyperparameters is costly, it is arguably wasteful to "throw away" all but the best result, as per Bayesian optimization. A related issue is the danger of overfitting the validation data when optimizing many hyperparameters. In this paper, we consider an alternative approach that uses more samples from the hyperparameter selection procedure to average over the uncertainty in model hyperparameters. The resulting approach, empirical Bayes for hyperparameter averaging (EB-Hyp) predicts held-out data b

etter than Bayesian optimization in two experiments on latent Dirichlet allocation and deep latent Gaussian models. EB-Hyp suggests a simpler approach to evaluating and deploying machine learning algorithms that does not require a separate validation data set and hyperparameter selection procedure.

************************************

Multiscale Quantization for Fast Similarity Search

Xiang Wu, Ruiqi Guo, Ananda Theertha Suresh, Sanjiv Kumar, Daniel N. Holtmann-Rice, David Simcha, Felix Yu

We propose a multiscale quantization approach for fast similarity search on large, high-dimensional datasets. The key insight of the approach is that quantization methods, in particular product quantization, perform poorly when there is large variance in the norms of the data points. This is a common scenario for real-world datasets, especially when doing product quantization of residuals obtained from coarse vector quantization. To address this issue, we propose a multiscale formulation where we learn a separate scalar quantizer of the residual norm scales. All parameters are learned jointly in a stochastic gradient descent framework to minimize the overall quantization error. We provide theoretical motivation for the proposed technique and conduct comprehensive experiments on two large-scale public datasets, demonstrating substantial improvements in recall over existing state-of-the-art methods.

************************************

Bregman Divergence for Stochastic Variance Reduction: Saddle-Point and Adversarial Prediction

Zhan Shi, Xinhua Zhang, Yaoliang Yu

Adversarial machines, where a learner competes against an adversary, have regained much recent interest in machine learning. They are naturally in the form of saddle-point optimization, often with separable structure but sometimes also with unmanageably large dimension. In this work we show that adversarial prediction under multivariate losses can be solved much faster than they used to be. We first reduce the problem size exponentially by using appropriate sufficient statistics, and then we adapt the new stochastic variance-reduced algorithm of Balamurugan & Bach (2016) to allow any Bregman divergence. We prove that the same linear rate of convergence is retained and we show that for adversarial prediction using KL-divergence we can further achieve a speedup of #example times compared with the Euclidean alternative. We verify the theoretical findings through extensive experiments on two example applications: adversarial prediction and LPboosting.

************************************

Perturbative Black Box Variational Inference

Robert Bamler, Cheng Zhang, Manfred Opper, Stephan Mandt

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.

************************************

Kernel Feature Selection via Conditional Covariance Minimization

Jianbo Chen, Mitchell Stern, Martin J. Wainwright, Michael I. Jordan

We propose a method for feature selection that employs kernel-based measures of independence to find a subset of covariates that is maximally predictive of the response. Building on past work in kernel dimension reduction, we show how to perform feature selection via a constrained optimization problem involving the trace of the conditional covariance operator. We prove various consistency results for this procedure, and also demonstrate that our method compares favorably with other state-of-the-art algorithms on a variety of synthetic and real data sets.

************************************

Active Learning from Peers

Keerthiram Murugesan, Jaime Carbonell

This paper addresses the challenge of learning from peers in an online multitask setting. Instead of always requesting a label from a human oracle, the proposed method first determines if the learner for each task can acquire that label wit h sufficient confidence from its peers either as a task-similarity weighted sum, or from the single most similar task.  If so, it saves the oracle query for later use in more difficult cases, and if not it queries the human oracle.  The paper develops the new algorithm to exhibit this behavior and proves a theoretical mistake bound for the method compared to the best linear predictor in hindsight.  Experiments over three multitask learning benchmark datasets show clearly superior performance over baselines such as assuming task independence, learning only from the oracle and not learning from peer tasks.

************************************

On Fairness and Calibration

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, Kilian Q. Weinberger

The machine learning community has become increasingly concerned with the potential for bias and discrimination in predictive models. This has motivated a growing line of work on what it means for a classification procedure to be "fair." In this paper, we investigate the tension between minimizing error disparity across different population groups while maintaining calibrated probability estimates. We show that calibration is compatible only with a single error constraint (i.e. equal false-negatives rates across groups), and show that any algorithm that satisfies this relaxation is no better than randomizing a percentage of predictions for an existing classifier. These unsettling findings, which extend and generalize existing results, are empirically confirmed on several datasets.

************************************

One-Shot Imitation Learning

Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, Wojciech Zaremba

Imitation learning has been commonly applied to solve different tasks in isolation. This usually requires either careful feature engineering, or a significant number of samples. This is far from what we desire: ideally, robots should be able to learn from very few demonstrations of any given task, and instantly generalize to new situations of the same task, without requiring task-specific engineering. In this paper, we propose a meta-learning framework for achieving such capability, which we call one-shot imitation learning.  Specifically, we consider the setting where there is a very large (maybe infinite) set of tasks, and each task has many instantiations.  For example, a task could be to stack all blocks on a table into a single tower, another task could be to place all blocks on a table into two-block towers, etc. In each case, different instances of the task would consist of different sets of blocks with different initial states.  At training time, our algorithm is presented with pairs of demonstrations for a subset of all tasks.  A neural net is trained that takes as input one demonstration and the current state (which initially is the initial state of the other demonstration of the pair), and outputs an action with the goal that the resulting sequence of states and actions matches as closely as possible with the second demonstration. At test time, a demonstration of a single instance of a new task is presented, and the neural net is expected to perform well on new instances of this new task. Our experiments show that the use of soft attention allows the model to generalize to conditions and tasks unseen in the training data. We anticipate that by training this model on a much greater variety of tasks and settings, we will obtain a general system that can turn any demonstrations into robust policies that can accomplish an overwhelming variety of tasks.

************************************

Triangle Generative Adversarial Networks

Zhe Gan, Liqun Chen, Weiyao Wang, Yuchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, Lawrence Carin

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.

************************************

Learning Populations of Parameters

Kevin Tian, Weihao Kong, Gregory Valiant

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.

************************************

Multi-Armed Bandits with Metric Movement Costs

Tomer Koren, Roi Livni, Yishay Mansour

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.

************************************

Structured Embedding Models for Grouped Data

Maja Rudolph, Francisco Ruiz, Susan Athey, David Blei

Word embeddings are a powerful approach for analyzing language, and exponential family embeddings (EFE) extend them to other types of data. Here we develop structured exponential family embeddings (S-EFE), a method for discovering embeddings that vary across related groups of data. We study how the word usage of U.S. Congressional speeches varies across states and party affiliation, how words are used differently across sections of the ArXiv, and how the co-purchase patterns of groceries can vary across seasons. Key to the success of our method is that the groups share statistical information. We develop two sharing strategies: hierarchical modeling and amortization. We demonstrate the benefits of this approach in empirical studies of speeches, abstracts, and shopping baskets. We show how SEFE enables group-specific interpretation of word usage, and outperforms EFE in predicting held-out data.

************************************

Conservative Contextual Linear Bandits

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, Benjamin Van Roy

Safety is a desirable property that can immensely increase the applicability of learning algorithms in real-world decision-making problems. It is much easier for a company to deploy an algorithm that is safe, i.e., guaranteed to perform at least as well as a baseline. In this paper, we study the issue of safety in contextual linear bandits that have application in many different fields including personalized ad recommendation in online marketing. We formulate a notion of safety for this class of algorithms. We develop a safe contextual linear bandit algorithm, called conservative linear UCB (CLUCB), that simultaneously minimizes its regret and satisfies the safety constraint, i.e., maintains its performance above a fixed percentage of the performance of a baseline strategy, uniformly over time. We prove an upper-bound on the regret of CLUCB and show that it can be decomposed into two terms: 1) an upper-bound for the regret of the standard linear UCB algorithm that grows with the time horizon and 2) a constant term that accounts for the loss of being conservative in order to satisfy the safety constraint. We empirically show that our algorithm is safe and validate our theoretical analysis.

************************************

Regularized Modal Regression with Applications in Cognitive Impairment Prediction

Xiaoqian Wang, Hong Chen, Weidong Cai, Dinggang Shen, Heng Huang

Linear regression models have been successfully used to function estimation and model selection in high-dimensional data analysis. However, most existing methods are built on least squares with the mean square error (MSE) criterion, which are sensitive to outliers and their performance may be degraded for heavy-tailed noise. In this paper, we go beyond this criterion by investigating the regularized modal regression from a statistical learning viewpoint. A new regularized modal regression model is proposed for estimation and variable selection, which is robust to outliers, heavy-tailed noise, and skewed noise. On the theoretical side, we establish the approximation estimate for learning the conditional mode function, the sparsity analysis for variable selection, and the robustness characterization. On the application side, we applied our model to successfully improve the cognitive impairment prediction using the Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort data.

************************************

Adversarial Ranking for Language Generation

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, Ming-ting Sun

Generative adversarial networks (GANs) have great successes on synthesizing data. However, the existing GANs restrict the discriminator to be a binary classifier, and thus limit their learning capacity for tasks that need to synthesize output with rich structures such as natural language descriptions. In this paper, we propose a novel generative adversarial network, RankGAN, for generating high-quality language descriptions. Rather than training the discriminator to learn and assign absolute binary predicate for individual data sample, the proposed RankGAN is able to analyze and rank a collection of human-written and machine-written sentences by giving a reference group. By viewing a set of data samples collectively and evaluating their quality through relative ranking scores, the discriminator is able to make better assessment which in turn helps to learn a better generator. The proposed RankGAN is optimized through the policy gradient technique. Experimental results on multiple public datasets clearly demonstrate the effectiveness of the proposed approach.

************************************

Diving into the shallows: a computational perspective on large-scale shallow learning

SIYUAN MA, Mikhail Belkin

Remarkable recent success of deep neural networks has not been easy to analyze theoretically. It has been  particularly hard to disentangle relative significance of architecture and optimization in achieving accurate classification on large datasets. On the flip side, shallow methods (such as kernel methods) have  encountered obstacles in scaling to large data, despite excellent performance on smaller datasets, and extensive theoretical analysis. Practical methods, such as variants of gradient descent used so successfully in deep learning, seem to perform below par when applied to kernel methods. This difficulty has sometimes been attributed to the limitations of shallow  architecture.   In this paper we  identify a basic limitation in gradient descent-based optimization methods when used in conjunctions with smooth kernels. Our analysis demonstrates that only a vanishingly small fraction of the function space is reachable after a polynomial number of gradient descent iterations. That drastically limits the approximating power of gradient descent leading to over-regularization. The issue is purely algorithmic, persisting even in the limit of infinite data.  To address this shortcoming in practice, we introduce EigenPro iteration, a simple and direct preconditioning scheme using a small number of approximately computed eigenvectors. It can also be viewed as learning a kernel optimized for gradient descent. Injecting this small, computationally inexpensive and SGD-compatible, amount of approximate second-order information leads to major improvements in convergence. For large data, this leads to a  significant performance  boost over the state-of-the-art kernel methods. In particular, we are able to match or improve the results reported in the literature at a small fraction of their computational budget. For complete version of this paper see https://arxiv.org/abs/1703.10622.

************************************

Integration Methods and Optimization Algorithms

Damien Scieur, Vincent Roulet, Francis Bach, Alexandre d'Aspremont

We show that accelerated optimization methods can be seen as particular instances of multi-step integration schemes from numerical analysis, applied to the gradient flow equation. Compared with recent advances in this vein, the differential equation considered here is the basic gradient flow, and we derive a class of multi-step schemes which includes accelerated algorithms, using classical conditions from numerical analysis. Multi-step schemes integrate the differential equation using larger step sizes, which intuitively explains the acceleration phenomenon.

************************************

The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings

Krzysztof M. Choromanski, Mark Rowland, Adrian Weller

We examine a class of embeddings based on structured random matrices with orthogonal rows which can be applied in many machine learning applications including dimensionality reduction and kernel approximation. For both the Johnson-Lindenstrauss transform and the angular kernel, we show that we can select matrices yielding guaranteed improved performance in accuracy and/or speed compared to earlier methods. We introduce matrices with complex entries which give significant further accuracy improvement. We provide geometric and Markov chain-based perspectives to help understand the benefits, and empirical results which suggest that the approach is helpful in a wider range of applications.

************************************

A KL-LUCB algorithm for Large-Scale Crowdsourcing

Ervin Tanczos, Robert Nowak, Bob Mankoff

This paper focuses on best-arm identification in multi-armed bandits with bounded rewards. We develop an algorithm that is a fusion of lil-UCB and KL-LUCB, offering the best qualities of the two algorithms in one method. This is achieved by proving a novel anytime confidence bound for the mean of bounded distributions, which is the analogue of the LIL-type bounds recently developed for sub-Gaussian distributions. We corroborate our theoretical results with numerical experiments based on the New Yorker Cartoon Caption Contest.

************************************

Q-LDA: Uncovering Latent Patterns in Text-based Sequential Decision Processes

Jianshu Chen, Chong Wang, Lin Xiao, Ji He, Lihong Li, Li Deng

In sequential decision making, it is often important and useful for end users to understand the underlying patterns or causes that lead to the corresponding decisions. However, typical deep reinforcement learning algorithms seldom provide such information due to their black-box nature. In this paper, we present a probabilistic model, Q-LDA, to uncover latent patterns in text-based sequential decision processes. The model can be understood as a variant of latent topic models that are tailored to maximize total rewards; we further draw an interesting connection between an approximate maximum-likelihood estimation of Q-LDA and the celebrated Q-learning algorithm. We demonstrate in the text-game domain that our proposed method not only provides a viable mechanism to uncover latent patterns in decision processes, but also obtains state-of-the-art rewards in these games.

************************************

Streaming Weak Submodularity: Interpreting Neural Networks on the Fly

Ethan Elenberg, Alexandros G. Dimakis, Moran Feldman, Amin Karbasi

In many machine learning applications, it is important to explain the predictions of a black-box classifier. For example, why does a deep neural network assign an image to a particular class? We cast interpretability of black-box classifiers as a combinatorial maximization problem and propose an efficient streaming algorithm to solve it subject to cardinality constraints. By extending ideas from Badanidiyuru et al. [2014], we provide a constant factor approximation guarantee for our algorithm in the case of random stream order and a weakly submodular objective function. This is the first such theoretical guarantee for this general class of functions, and we also show that no such algorithm exists for a worst case stream order. Our algorithm obtains similar explanations of Inception V3 predictions 10 times faster than the state-of-the-art LIME framework of Ribeiro et al. [2016].

************************************

Decomposable Submodular Function Minimization: Discrete and Continuous

Alina Ene, Huy Nguyen, László A. Végh

This paper investigates connections between discrete and continuous approaches for decomposable submodular function minimization. We provide improved running time estimates for the state-of-the-art continuous algorithms for the problem using combinatorial arguments. We also provide a systematic experimental comparison of the two types of methods, based on a clear distinction between level-0 and level-1 algorithms.

************************************

Learning Affinity via Spatial Propagation Networks

Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, Jan Kautz

In this paper, we propose a spatial propagation networks for learning affinity matrix. We show that by constructing a row/column linear propagation model, the spatially variant transformation matrix constitutes an affinity matrix that models dense, global pairwise similarities of an image. Specifically, we develop a three-way connection for the linear propagation model, which (a) formulates a sparse transformation matrix where all elements can be the output from a deep CNN, but (b) results in a dense affinity matrix that is effective to model any task-specific pairwise similarity. Instead of designing the similarity kernels according to image features of two points, we can directly output all similarities in a pure data-driven manner. The spatial propagation network is a generic framework that can be applied to numerous tasks, which traditionally benefit from designed affinity, e.g., image matting, colorization, and guided filtering, to name a few. Furthermore, the model can also learn semantic-aware affinity for high-level vision tasks due to the learning capability of the deep model. We validate the proposed framework by refinement of object segmentation. Experiments on the HELEN face parsing and PASCAL VOC-2012 semantic segmentation tasks show that the spatial propagation network provides general, effective and efficient solutions for generating high-quality segmentation results.

************************************

Gated Recurrent Convolution Neural Network for OCR

Jianfeng Wang, Xiaolin Hu

Optical Character Recognition (OCR) aims to recognize text in natural images. Inspired by a recently proposed model for general image classification, Recurrent Convolution Neural Network (RCNN), we propose a new architecture named Gated RCNN (GRCNN) for solving this problem. Its critical component, Gated Recurrent Convolution Layer (GRCL), is constructed by adding a gate to the Recurrent Convolution Layer (RCL), the critical component of RCNN. The gate controls the context modulation in RCL and balances the feed-forward information and the recurrent information. In addition, an efficient Bidirectional Long Short-Term Memory (BLSTM) is built for sequence modeling. The GRCNN is combined with BLSTM to recognize text in natural images. The entire GRCNN-BLSTM model can be trained end-to-end. Experiments show that the proposed model outperforms existing methods on several benchmark datasets including the IIIT-5K, Street View Text (SVT) and ICDAR.

************************************

Multi-view Matrix Factorization for Linear Dynamical System Estimation

Mahdi Karami, Martha White, Dale Schuurmans, Csaba Szepesvari

We consider maximum likelihood estimation of linear dynamical systems with generalized-linear observation models. Maximum likelihood is typically considered to be hard in this setting since latent states and transition parameters must be inferred jointly. Given that expectation-maximization does not scale and is prone to local minima, moment-matching approaches from the subspace identification literature have become standard, despite known statistical efficiency issues. In this paper, we instead reconsider likelihood maximization and develop an optimization based strategy for recovering the latent states and transition parameters. Key to the approach is a two-view reformulation of maximum likelihood estimation for linear dynamical systems that enables the use of global optimization algorithms for matrix factorization. We show that the proposed estimation strategy outperforms widely-used identification algorithms such as subspace identification methods, both in terms of accuracy and runtime.

************************************

Policy Gradient With Value Function Approximation For Collective Multiagent Planning

Duc Thien Nguyen, Akshat Kumar, Hoong Chuin Lau

Decentralized (PO)MDPs provide an expressive framework for sequential decision making in a multiagent system. Given their computational complexity, recent research has focused on tractable yet practical subclasses of Dec-POMDPs. We address such a subclass called CDec-POMDP where the collective behavior of a population of agents affects the joint-reward and environment dynamics. Our main contributi on is an actor-critic (AC) reinforcement learning method for optimizing CDec-POMDP policies. Vanilla AC has slow convergence for larger problems. To address this, we show how a particular decomposition of the approximate action-value function over agents leads to effective updates, and also derive a new way to train the critic based on local reward signals. Comparisons on a synthetic benchmark and a real world taxi fleet optimization problem show that our new AC approach provides better quality solutions than previous best approaches.

************************************

Stochastic Submodular Maximization: The Case of Coverage Functions
Mohammad Karimi, Mario Lucic, Hamed Hassani, Andreas Krause

Stochastic optimization of continuous objectives is at the heart of modern machine learning. However, many important problems are of discrete nature and often involve submodular objectives. We seek to unleash the power of stochastic continuous optimization, namely stochastic gradient descent and its variants, to such discrete problems. We first introduce the problem of stochastic submodular optimization, where one needs to optimize a submodular objective which is given as an expectation. Our model captures situations where the discrete objective arises as an empirical risk (e.g., in the case of exemplar-based clustering), or is given as an explicit stochastic model (e.g., in the case of influence maximization in social networks). By exploiting that common extensions act linearly on the class of submodular functions, we employ projected stochastic gradient ascent and its variants in the continuous domain, and perform rounding to obtain discrete solutions. We focus on the rich and widely used family of weighted coverage functions. We show that our approach yields solutions that are guaranteed to match the optimal approximation guarantees, while reducing the computational cost by several orders of magnitude, as we demonstrate empirically.

************************************

Stochastic Approximation for Canonical Correlation Analysis
Raman Arora, Teodor Vanislavov Marinov, Poorya Mianjy, Nati Srebro

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.

************************************

Linear regression without correspondence
Daniel J. Hsu, Kevin Shi, Xiaorui Sun

This article considers algorithmic and statistical aspects of linear regression when the correspondence between the covariates and the responses is unknown. First, a fully polynomial-time approximation scheme is given for the natural least squares optimization problem in any constant dimension. Next, in an average-case and noise-free setting where the responses exactly correspond to a linear function of i.i.d. draws from a standard multivariate normal distribution, an efficient algorithm based on lattice basis reduction is shown to exactly recover the unknown linear function in arbitrary dimension. Finally, lower bounds on the signal-to-noise ratio are established for approximate recovery of the unknown linear function by any estimator.

************************************

Structured Generative Adversarial Networks
Zhijie Deng, Hao Zhang, Xiaodan Liang, Luona Yang, Shizhen Xu, Jun Zhu, Eric P. Xing

We study the problem of conditional generative modeling based on designated semantics or structures. Existing models that build conditional generators either require massive labeled instances as supervision or are unable to accurately control the semantics of generated samples. We propose structured generative adversarial networks (SGANs) for semi-supervised conditional generative modeling. SGAN assumes the data x is generated conditioned on two independent latent variables: y that encodes the designated semantics, and z that contains other factors of variation. To ensure disentangled semantics in y and z, SGAN builds two collaborative games in the hidden space to minimize the reconstruction error of y and z, respectively. Training SGAN also involves solving two adversarial games that have their equilibrium concentrating at the true joint data distributions p(x, z) and p(x, y), avoiding distributing the probability mass diffusely over data space that MLE-based methods may suffer. We assess SGAN by evaluating its trained networks, and its performance on downstream tasks. We show that SGAN delivers a highly controllable generator, and disentangled representations; it also establishes start-of-the-art results across multiple datasets when applied for semi-supervised image classification (1.27%, 5.73%, 17.26% error rates on MNIST, SVHN and CIFAR-10 using 50, 1000 and 4000 labels, respectively). Benefiting from the separate modeling of y and z, SGAN can generate images with high visual quality and strictly following the designated semantic, and can be extended to a wide spectrum of applications, such as style transfer.

************************************

### Dynamic-Depth Context Tree Weighting

Joao V. Messias, Shimon Whiteson

Reinforcement learning (RL) in partially observable settings is challenging because the agent’s observations are not Markov. Recently proposed methods can learn variable-order Markov models of the underlying process but have steep memory requirements and are sensitive to aliasing between observation histories due to sensor noise. This paper proposes dynamic-depth context tree weighting (D2-CTW), a model-learning method that addresses these limitations. D2-CTW dynamically expands a suffix tree while ensuring that the size of the model, but not its depth, remains bounded. We show that D2-CTW approximately matches the performance of state-of-the-art alternatives at stochastic time-series prediction while using at least an order of magnitude less memory. We also apply D2-CTW to model-based RL, showing that, on tasks that require memory of past observations, D2-CTW can learn without prior knowledge of a good state representation, or even the length of history upon which such a representation should depend.

************************************

### Fast, Sample-Efficient Algorithms for Structured Phase Retrieval

Gauri Jagatap, Chinmay Hegde

We consider the problem of recovering a signal x in $R^n$, from magnitude-only measurements, $y_i = |a_i^T x|$ for $i=\{1,2...m\}$. Also known as the phase retrieval problem, it is a fundamental challenge in nano-, bio- and astronomical imaging systems, astronomical imaging, and speech processing. The problem is ill-posed, and therefore additional assumptions on the signal and/or the measurements are necessary. In this paper, we first study the case where the underlying signal x is s-sparse. We develop a novel recovery algorithm that we call Compressive Phase Retrieval with Alternating Minimization, or CoPRAM. Our algorithm is simple and can be obtained via a natural combination of the classical alternating minimization approach for phase retrieval, with the CoSaMP algorithm for sparse recovery. Despite its simplicity, we prove that our algorithm achieves a sample complexity of $O(s^2 \log n)$ with Gaussian samples, which matches the best known existing results. It also demonstrates linear convergence in theory and practice and requires no extra tuning parameters other than the signal sparsity level s. We then consider the case where the underlying signal x arises from to structured sparsity models. We specifically examine the case of block-sparse signals with uniform block size of b and block sparsity $k=s/b$. For this problem, we design a recovery algorithm that we call Block CoPRAM that further reduces the sample complexity to $O(ks \log n)$. For sufficiently large block lengths of $b=\Theta(s)$, this bound equates to $O(s \log n)$. To our knowledge, this constitutes the first end-to-end linearly convergent family of algorithms for phase retrieval where the Gaussian sample complexity has a sub-quadratic dependence on the sparsity level of the signal.

************************************

### Hierarchical Methods of Moments

Matteo Ruffini, Guillaume Rabusseau, Borja Balle

Spectral methods of moments provide a powerful tool for learning the parameters of latent variable models. Despite their theoretical appeal, the applicability of these methods to real data is still limited due to a lack of robustness to model misspecification. In this paper we present a hierarchical approach to methods

of moments to circumvent such limitations. Our method is based on replacing the tensor decomposition step used in previous algorithms with approximate joint diagonalization. Experiments on topic modeling show that our method outperforms previous tensor decomposition methods in terms of speed and model quality.
************************************

A New Alternating Direction Method for Linear Programming
Sinong Wang, Ness Shroff
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.
************************************

Near Optimal Sketching of Low-Rank Tensor Regression
Xingguo Li, Jarvis Haupt, David Woodruff
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.
************************************

Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models
Sergey Ioffe
Batch Normalization is quite effective at accelerating and improving the training of deep models. However, its effectiveness diminishes when the training minibatches are small, or do not consist of independent samples. We hypothesize that this is due to the dependence of model layer inputs on all the examples in the minibatch, and different activations being produced between training and inference. We propose Batch Renormalization, a simple and effective extension to ensure that the training and inference models generate the same outputs that depend on individual examples rather than the entire minibatch. Models trained with Batch Renormalization perform substantially better than batchnorm when training with small  or non-i.i.d. minibatches. At the same time, Batch Renormalization retains the benefits of batchnorm such as insensitivity to initialization and training efficiency.
************************************

Position-based Multiple-play Bandit Problem with Unknown Position Bias
Junpei Komiyama, Junya Honda, Akiko Takeda
Motivated by online advertising, we study a multiple-play multi-armed bandit problem with position bias that involves several slots and the latter slots yield fewer rewards. We characterize the hardness of the problem by deriving an asymptotic regret bound. We propose the Permutation Minimum Empirical Divergence (PMED) algorithm and derive its asymptotically optimal regret bound. Because of the uncertainty of the position bias, the optimal algorithm for such a problem requires non-convex optimizations that are different from usual partial monitoring and semi-bandit problems. We propose a cutting-plane method and related bi-convex relaxation for these optimizations by using auxiliary variables.
************************************

Deep Voice 2: Multi-Speaker Neural Text-to-Speech
Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, Yanqi Zhou
We introduce a technique for augmenting neural text-to-speech (TTS) with low-dimensional trainable speaker embeddings to generate different voices from a single model. As a starting point, we show improvements over the two state-of-the-art approaches for single-speaker neural TTS: Deep Voice 1 and Tacotron. We introduce Deep Voice 2, which is based on a similar pipeline with Deep Voice 1, but constructed with higher performance building blocks and demonstrates a significant audio quality improvement over Deep Voice 1. We improve Tacotron by introducing a post-processing neural vocoder, and demonstrate a significant audio quality improvement. We then demonstrate our technique for multi-speaker speech synthesis for both Deep Voice 2 and Tacotron on two multi-speaker TTS datasets. We show tha

t a single neural TTS system can learn hundreds of unique voices from less than half an hour of data per speaker, while achieving high audio quality synthesis and preserving the speaker identities almost perfectly.
************************************
Eigen-Distortions of Hierarchical Representations
Alexander Berardino, Valero Laparra, Johannes Ballé, Eero Simoncelli
We develop a method for comparing hierarchical image representations in terms of their ability to explain perceptual sensitivity in humans. Specifically, we utilize Fisher information to establish a model-derived prediction of sensitivity to local perturbations of an image. For a given image, we compute the eigenvectors of the Fisher information matrix with largest and smallest eigenvalues, corresponding to the model-predicted most- and least-noticeable image distortions, respectively. For human subjects, we then measure the amount of each distortion that can be reliably detected when added to the image. We use this method to test the ability of a variety of representations to mimic human perceptual sensitivity. We find that the early layers of VGG16, a deep neural network optimized for object recognition, provide a better match to human perception than later layers, and a better match than a 4-stage convolutional neural network (CNN) trained on a database of human ratings of distorted image quality. On the other hand, we find that simple models of early visual processing, incorporating one or more stages of local gain control, trained on the same database of distortion ratings, provide substantially better predictions of human sensitivity than either the CNN, or any combination of layers of VGG16.
************************************
Learning to Prune Deep Neural Networks via Layer-wise Optimal Brain Surgeon
Xin Dong, Shangyu Chen, Sinno Pan
How to develop slim and accurate deep neural networks has become crucial for real- world applications, especially for those employed in embedded systems. Though previous work along this research line has shown some promising results, most existing methods either fail to significantly compress a well-trained deep network or require a heavy retraining process for the pruned deep network to re-boost its prediction performance. In this paper, we propose a new layer-wise pruning method for deep neural networks. In our proposed method, parameters of each individual layer are pruned independently based on second order derivatives of a layer-wise error function with respect to the corresponding parameters. We prove that the final prediction performance drop after pruning is bounded by a linear combination of the reconstructed errors caused at each layer. By controlling layer-wise errors properly, one only needs to perform a light retraining process on the pruned network to resume its original prediction performance. We conduct extensive experiments on benchmark datasets to demonstrate the effectiveness of our pruning method compared with several state-of-the-art baseline methods. Codes of our work are released at: https://github.com/csyhhu/L-OBS.
************************************
Deliberation Networks: Sequence Generation Beyond One-Pass Decoding
Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, Tie-Yan Liu
The encoder-decoder framework has achieved promising progress for many sequence generation tasks, including machine translation, text summarization, dialog system, image captioning, etc. Such a framework adopts an one-pass forward process while decoding and generating a sequence, but lacks the deliberation process: A generated sequence is directly used as final output without further polishing. However, deliberation is a common behavior in human's daily life like reading news and writing papers/articles/books. In this work, we introduce the deliberation process into the encoder-decoder framework and propose deliberation networks for sequence generation. A deliberation network has two levels of decoders, where the first-pass decoder generates a raw sequence and the second-pass decoder polishes and refines the raw sentence with deliberation. Since the second-pass deliberation decoder has global information about what the sequence to be generated might be, it has the potential to generate a better sequence by looking into future words in the raw sentence. Experiments on neural machine translation and text summarization demonstrate the effectiveness of the proposed deliberation network

s. On the WMT 2014 English-to-French translation task, our model establishes a new state-of-the-art BLEU score of 41.5.

************************************

Do Deep Neural Networks Suffer from Crowding?
Anna Volokitin, Gemma Roig, Tomaso A. Poggio

Crowding is a visual effect suffered by humans, in which an object that can be recognized in isolation can no longer be recognized when other objects, called flankers, are placed close to it. In this work, we study the effect of crowding in artificial Deep Neural Networks (DNNs) for object recognition. We analyze both deep convolutional neural networks  (DCNNs) as well as an extension of DCNNs that are multi-scale and that change the receptive field size of the convolution filters with their position in the image.  The latter  networks, that we call eccentricity-dependent, have been proposed for modeling the feedforward path of the primate visual cortex. Our results reveal that the eccentricity-dependent model, trained on target objects in isolation, can recognize such targets in the presence of flankers, if the targets are near the center of the image, whereas DCNNs cannot. Also, for all tested networks, when trained on targets in isolation, we find that recognition accuracy of the networks decreases the closer the flankers are to the target and the more flankers there are. We find that visual similarity between the target and flankers also plays a role and that pooling in early layers of the network leads to more crowding. Additionally, we show that incorporating flankers  into the images of the training set for learning the DNNs does not lead to robustness against configurations not seen at training.

************************************

Non-Stationary Spectral Kernels
Sami Remes, Markus Heinonen, Samuel Kaski

We propose non-stationary spectral kernels for Gaussian process regression by modelling the spectral density of a non-stationary kernel function as a mixture of input-dependent Gaussian process frequency density surfaces. We solve the generalised Fourier transform with such a model, and present a family of non-stationary and non-monotonic kernels that can learn input-dependent and potentially long-range, non-monotonic covariances between inputs. We derive efficient inference using model whitening and marginalized posterior, and show with case studies that these kernels are necessary when modelling even rather simple time series, image or geospatial data with non-stationary characteristics.

************************************

Extracting low-dimensional dynamics from multiple large-scale neural population recordings by learning to predict correlations
Marcel Nonnenmacher, Srinivas C. Turaga, Jakob H. Macke

A powerful approach for understanding neural population dynamics is to extract low-dimensional trajectories from population recordings using dimensionality reduction methods. Current approaches for dimensionality reduction on neural data are limited to single population recordings, and can not identify dynamics embedded across multiple measurements. We propose an approach for extracting low-dimensional dynamics from multiple, sequential recordings. Our algorithm scales to data comprising millions of observed dimensions, making it possible to access dynamics distributed across large populations or multiple brain areas. Building on subspace-identification approaches for dynamical systems, we perform parameter estimation by minimizing a moment-matching objective using a scalable stochastic gradient descent algorithm: The model is optimized to predict temporal covariations across neurons and across time. We show how this approach naturally handles missing data and multiple partial recordings, and can identify dynamics and predict correlations even in the presence of severe subsampling and small overlap between recordings. We demonstrate the effectiveness of the approach both on simulated data and a whole-brain larval zebrafish imaging dataset.

************************************

Minimizing a Submodular Function from Samples
Eric Balkanski, Yaron Singer

In this paper we consider the problem of minimizing a submodular function from training data. Submodular functions can be efficiently minimized and are conse- q uently heavily applied in machine learning. There are many cases, however, in which we do not know the function we aim to optimize, but rather have access to training data that is used to learn the function. In this paper we consider the question of whether submodular functions can be minimized in such cases. We show that even learnable submodular functions cannot be minimized within any non-trivial approximation when given access to polynomially-many samples. Specifically, we show that there is a class of submodular functions with range in [0, 1] such that, despite being PAC-learnable and minimizable in polynomial-time, no algorithm can obtain an approximation strictly better than 1/2 − o(1) using polynomially-many samples drawn from any distribution. Furthermore, we show that this bound is tight using a trivial algorithm that obtains an approximation of 1/2.

************************************

A graph-theoretic approach to multitasking

Noga Alon, Daniel Reichman, Igor Shinkar, Tal Wagner, Sebastian Musslick, Jonathan D. Cohen, Tom Griffiths, Biswadip dey, Kayhan Ozcimder

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.

************************************

Adversarial Surrogate Losses for Ordinal Regression

Rizal Fathony, Mohammad Ali Bashiri, Brian Ziebart

Ordinal regression seeks class label predictions when the penalty incurred for mistakes increases according to an ordering over the labels. The absolute error is a canonical example. Many existing methods for this task reduce to binary classification problems and employ surrogate losses, such as the hinge loss. We instead derive uniquely defined surrogate ordinal regression loss functions by seeking the predictor that is robust to the worst-case approximations of training data labels, subject to matching certain provided training data statistics. We demonstrate the advantages of our approach over other surrogate losses based on hinge loss approximations using UCI ordinal prediction tasks.

************************************

Self-Supervised Intrinsic Image Decomposition

Michael Janner, Jiajun Wu, Tejas D. Kulkarni, Ilker Yildirim, Josh Tenenbaum

Intrinsic decomposition from a single image is a highly challenging task, due to its inherent ambiguity and the scarcity of training data. In contrast to traditional fully supervised learning approaches, in this paper we propose learning intrinsic image decomposition by explaining the input image. Our model, the Rendered Intrinsics Network (RIN), joins together an image decomposition pipeline, which predicts reflectance, shape, and lighting conditions given a single image, with a recombination function, a learned shading model used to recompose the original input based off of intrinsic image predictions. Our network can then use unsupervised reconstruction error as an additional signal to improve its intermediate representations. This allows large-scale unlabeled data to be useful during training, and also enables transferring learned knowledge to images of unseen object categories, lighting conditions, and shapes. Extensive experiments demonstrate that our method performs well on both intrinsic image decomposition and knowledge transfer.

************************************

On-the-fly Operation Batching in Dynamic Computation Graphs

Graham Neubig, Yoav Goldberg, Chris Dyer

Dynamic neural networks toolkits such as PyTorch, DyNet, and Chainer offer more flexibility for implementing models that cope with data of varying dimensions and structure, relative to toolkits that operate on statically declared computations (e.g., TensorFlow, CNTK, and Theano). However, existing toolkits - both static and dynamic - require that the developer organize the computations into the batches necessary for exploiting high-performance data-parallel algorithms and hardware. This batching task is generally difficult, but it becomes a major hurdle as architectures become complex. In this paper, we present an algorithm, and its implementation in the DyNet toolkit, for automatically batching operations. Dev

elopers simply write minibatch computations as aggregations of single instance computations, and the batching algorithm seamlessly executes them, on the fly, in computationally efficient batches. On a variety of tasks, we obtain throughput similar to manual batches, as well as comparable speedups over single-instance learning on architectures that are impractical to batch manually.
************************************
Fitting Low-Rank Tensors in Constant Time
Kohei Hayashi, Yuichi Yoshida
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.
************************************
Random Projection Filter Bank for Time Series Data
Amir-massoud Farahmand, Sepideh Pourazarm, Daniel Nikovski
We propose Random Projection Filter Bank (RPFB) as a generic and simple approach to extract features from time series data. RPFB is a set of randomly generated stable autoregressive filters that are convolved with the input time series to generate the features. These features can be used by any conventional machine learning algorithm for solving tasks such as time series prediction, classification with time series data, etc. Different filters in RPFB extract different aspects of the time series, and together they provide a reasonably good summary of the time series. RPFB is easy to implement, fast to compute, and parallelizable. We provide an error upper bound indicating that RPFB provides a reasonable approximation to a class of dynamical systems. The empirical results in a series of synthetic and real-world problems show that RPFB is an effective method to extract features from time series.
************************************
Dynamic Revenue Sharing
Santiago Balseiro, Max Lin, Vahab Mirrokni, Renato Leme, IIIS Song Zuo
Many online platforms act as intermediaries between a seller and a set of buyers. Examples of such settings include online retailers (such as Ebay) selling items on behalf of sellers to buyers, or advertising exchanges (such as AdX) selling pageviews on behalf of publishers to advertisers. In such settings, revenue sharing is a central part of running such a marketplace for the intermediary, and fixed-percentage revenue sharing schemes are often used to split the revenue among the platform and the sellers. In particular, such revenue sharing schemes require the platform to (i) take at most a constant fraction \alpha of the revenue from auctions and (ii) pay the seller at least the seller declared opportunity cost c for each item sold. A straightforward way to satisfy the constraints is to set a reserve price at c / (1 - \alpha) for each item, but it is not the optimal solution on maximizing the profit of the intermediary.  While previous studies (by Mirrokni and Gomes, and by Niazadeh et al) focused on revenue-sharing schemes in static double auctions, in this paper, we take advantage of the repeated nature of the auctions. In particular, we introduce dynamic revenue sharing schemes where we balance the two constraints over different auctions to achieve higher profit and seller revenue. This is directly motivated by the practice of advertising exchanges where the fixed-percentage revenue-share should be met across all auctions and not in each auction. In this paper, we characterize the optimal revenue sharing scheme that satisfies both constraints in expectation. Finally, we empirically evaluate our revenue sharing scheme on real data.
************************************
Prototypical Networks for Few-shot Learning
Jake Snell, Kevin Swersky, Richard Zemel
We propose Prototypical Networks for the problem of few-shot classification, where a classifier must generalize to new classes not seen in the training set, given only a small number of examples of each new class. Prototypical Networks learn a metric space in which classification can be performed by computing distances to prototype representations of each class. Compared to recent approaches for few-shot learning, they reflect a simpler inductive bias that is beneficial in th

is limited-data regime, and achieve excellent results. We provide an analysis showing that some simple design decisions can yield substantial improvements over recent approaches involving complicated architectural choices and meta-learning. We further extend Prototypical Networks to zero-shot learning and achieve state-of-the-art results on the CU-Birds dataset.

************************************

Unsupervised learning of object frames by dense equivariant image labelling

James Thewlis, Hakan Bilen, Andrea Vedaldi

One of the key challenges of visual perception is to extract abstract models of 3D objects and object categories from visual measurements, which are affected by complex nuisance factors such as viewpoint, occlusion, motion, and deformations. Starting from the recent idea of viewpoint factorization, we propose a new approach that, given a large number of images of an object and no other supervision, can extract a dense object-centric coordinate frame. This coordinate frame is invariant to deformations of the images and comes with a dense equivariant labelling neural network that can map image pixels to their corresponding object coordinates. We demonstrate the applicability of this method to simple articulated objects and deformable objects such as human faces, learning embeddings from random synthetic transformations or optical flow correspondences, all without any manual supervision.

************************************

Unified representation of tractography and diffusion-weighted MRI data using sparse multidimensional arrays

Cesar F. Caiafa, Olaf Sporns, Andrew Saykin, Franco Pestilli

Recently, linear formulations and convex optimization methods have been proposed to predict diffusion-weighted Magnetic Resonance Imaging (dMRI) data given estimates of brain connections generated using tractography algorithms. The size of the linear models comprising such methods grows with both dMRI data and connectome resolution, and can become very large when applied to modern data. In this paper, we introduce a method to encode dMRI signals and large connectomes, i.e., those that range from hundreds of thousands to millions of fascicles (bundles of neuronal axons), by using a sparse tensor decomposition. We show that this tensor decomposition accurately approximates the Linear Fascicle Evaluation (LiFE) model, one of the recently developed linear models. We provide a theoretical analysis of the accuracy of the sparse decomposed model, LiFESD, and demonstrate that it can reduce the size of the model significantly. Also, we develop algorithms to implement the optimisation solver using the tensor representation in an efficient way.

************************************

Random Permutation Online Isotonic Regression

Wojciech Kotlowski, Wouter M. Koolen, Alan Malek

We revisit isotonic regression on linear orders, the problem of fitting monotonic functions to best explain the data, in an online setting. It was previously shown that online isotonic regression is unlearnable in a fully adversarial model, which lead to its study in the fixed design model. Here, we instead develop the more practical random permutation model. We show that the regret is bounded above by the excess leave-one-out loss for which we develop efficient algorithms and matching lower bounds. We also analyze the class of simple and popular forward algorithms and recommend where to look for  algorithms for online isotonic regression on partial orders.

************************************

PRUNE: Preserving Proximity and Global Ranking for Network Embedding

Yi-An Lai, Chin-Chi Hsu, Wen Hao Chen, Mi-Yen Yeh, Shou-De Lin

We investigate an unsupervised generative approach for network embedding. A multi-task Siamese neural network structure is formulated to connect embedding vectors and our objective to preserve the global node ranking and local proximity of nodes. We provide deeper analysis to connect the proposed proximity objective to link prediction and community detection in the network. We show our model can satisfy the following design properties: scalability, asymmetry, unity and simplicity. Experiment results not only verify the above design properties but also de monstrate the superior performance in learning-to-rank, classification, regression, and link prediction tasks.

************************************

Online to Offline Conversions, Universality and Adaptive Minibatch Sizes
Kfir Levy

We present an approach towards convex optimization that relies on a novel scheme which converts adaptive online algorithms into offline methods. In the offline optimization setting, our derived methods are shown to obtain favourable adaptive guarantees which depend on the harmonic sum of the queried gradients. We further show that our methods implicitly adapt to the objective's structure: in the smooth case fast convergence rates are ensured without any prior knowledge of the smoothness parameter, while still maintaining guarantees in the non-smooth setting. Our approach has a natural extension to the stochastic setting, resulting in a lazy version of SGD (stochastic GD), where minibathces are chosen adaptively depending on the magnitude of the gradients. Thus providing a principled approach towards choosing minibatch sizes.

************************************

Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network
Wengong Jin, Connor Coley, Regina Barzilay, Tommi Jaakkola

The prediction of organic reaction outcomes is a fundamental problem in computational chemistry. Since a reaction may involve hundreds of atoms, fully exploring the space of possible transformations is intractable. The current solution utilizes reaction templates to limit the space, but it suffers from coverage and efficiency issues. In this paper, we propose a template-free approach to efficiently explore the space of product molecules by first pinpointing the reaction center -- the set of nodes and edges where graph edits occur. Since only a small number of atoms contribute to reaction center, we can directly enumerate candidate products. The generated candidates are scored by a Weisfeiler-Lehman Difference Network that models high-order interactions between changes occurring at nodes across the molecule. Our framework outperforms the top-performing template-based approach with a 10% margin, while running orders of magnitude faster. Finally, we demonstrate that the model accuracy rivals the performance of domain experts.

************************************

Inferring Generative Model Structure with Static Analysis
Paroma Varma, Bryan D. He, Payal Bajaj, Nishith Khandwala, Imon Banerjee, Daniel Rubin, Christopher Ré

Obtaining enough labeled data to robustly train complex discriminative models is a major bottleneck in the machine learning pipeline. A popular solution is combining multiple sources of weak supervision using generative models. The structure of these models affects the quality of the training labels, but is difficult to learn without any ground truth labels. We instead rely on weak supervision sources having some structure by virtue of being encoded programmatically. We present Coral, a paradigm that infers generative model structure by statically analyzing the code for these heuristics, thus significantly reducing the amount of data required to learn structure. We prove that Coral's sample complexity scales quasilinearly with the number of heuristics and number of relations identified, improving over the standard sample complexity, which is exponential in n for learning n-th degree relations. Empirically, Coral matches or outperforms traditional structure learning approaches by up to 3.81 F1 points. Using Coral to model dependencies instead of assuming independence results in better performance than a fully supervised model by 3.07 accuracy points when heuristics are used to label radiology data without ground truth labels.

************************************

Influence Maximization with $\varepsilon$-Almost Submodular Threshold Functions
Qiang Li, Wei Chen, Institute of Computing Xiaoming Sun, Institute of Computing Jialin Zhang

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.

************************************

### Improved Dynamic Regret for Non-degenerate Functions

Lijun Zhang, Tianbao Yang, Jinfeng Yi, Rong Jin, Zhi-Hua Zhou

Recently, there has been a growing research interest in the analysis of dynamic regret, which measures the performance of an online learner against a sequence of local minimizers. By exploiting the strong convexity, previous studies have shown that the dynamic regret can be upper bounded by the path-length of the comparator sequence. In this paper, we illustrate that the dynamic regret can be further improved by allowing the learner to query the gradient of the function multiple times, and meanwhile the strong convexity can be weakened to other non-degenerate conditions. Specifically, we introduce the squared path-length, which could be much smaller than the path-length, as a new regularity of the comparator sequence. When multiple gradients are accessible to the learner, we first demonstrate that the dynamic regret of strongly convex functions can be upper bounded by the minimum of the path-length and the squared path-length. We then extend our theoretical guarantee to functions that are semi-strongly convex or self-concordant. To the best of our knowledge, this is the first time that semi-strong convexity and self-concordance are utilized to tighten the dynamic regret.

************************************

### AdaGAN: Boosting Generative Models

Ilya O. Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann SIMON-GABRIEL, Bernhard Schölkopf

Generative Adversarial Networks (GAN) are an effective method for training generative models of complex data such as natural images. However, they are notoriously hard to train and can suffer from the problem of missing modes where the model is not able to produce examples in certain regions of the space. We propose an iterative procedure, called AdaGAN, where at every step we add a new component into a mixture model by running a GAN algorithm on a re-weighted sample. This is inspired by boosting algorithms, where many potentially weak individual predictors are greedily aggregated to form a strong composite predictor. We prove analytically that such an incremental procedure leads to convergence to the true distribution in a finite number of steps if each step is optimal, and convergence at an exponential rate otherwise. We also illustrate experimentally that this procedure addresses the problem of missing modes.

************************************

### Large-Scale Quadratically Constrained Quadratic Program via Low-Discrepancy Sequences

Kinjal Basu, Ankan Saha, Shaunak Chatterjee

We consider the problem of solving a large-scale Quadratically Constrained Quadratic Program. Such problems occur naturally in many scientific and web applications. Although there are efficient methods which tackle this problem, they are mostly not scalable. In this paper, we develop a method that transforms the quadratic constraint into a linear form by a sampling a set of low-discrepancy points. The transformed problem can then be solved by applying any state-of-the-art large-scale solvers. We show the convergence of our approximate solution to the true solution as well as some finite sample error bounds. Experimental results are also shown to prove scalability in practice.

************************************

### Graph Matching via Multiplicative Update Algorithm

Bo Jiang, Jin Tang, Chris Ding, Yihong Gong, Bin Luo

As a fundamental problem in computer vision, graph matching problem can usually be formulated as a Quadratic Programming (QP) problem with doubly stochastic and discrete (integer) constraints. Since it is NP-hard, approximate algorithms are required. In this paper, we present a new algorithm, called Multiplicative Update Graph Matching (MPGM), that develops a multiplicative update technique to solve the QP matching problem. MPGM has three main benefits: (1) theoretically, MPGM solves the general QP problem with doubly stochastic constraint naturally whose convergence and KKT optimality are guaranteed. (2) Em- pirically, MPGM generally returns a sparse solution and thus can also incorporate the discrete constraint approximately. (3) It is efficient and simple to implement. Experimental resu lts show the benefits of MPGM algorithm.

************************************

Neural Expectation Maximization

Klaus Greff, Sjoerd van Steenkiste, Jürgen Schmidhuber

Many real world tasks such as reasoning and physical interaction require identification and manipulation of conceptual entities. A first step towards solving these tasks is the automated discovery of distributed symbol-like representations. In this paper, we explicitly formalize this problem as inference in a spatial mixture model where each component is parametrized by a neural network.  Based on the Expectation Maximization framework we then derive a differentiable clustering method that simultaneously learns how to group and represent individual entities.  We evaluate our method on the (sequential) perceptual grouping task and find that it is able to accurately recover the constituent objects.  We demonstrate that the learned representations are useful for next-step prediction.

************************************

Hunt For The Unique, Stable, Sparse And Fast Feature Learning On Graphs

Saurabh Verma, Zhi-Li Zhang

For the purpose of learning on graphs, we hunt for a graph feature representation that exhibit certain uniqueness, stability and sparsity properties while also being amenable to fast computation. This leads to the discovery of family of graph spectral distances (denoted as FGSD) and their based graph feature representations, which we prove to possess most of these desired properties. To both evaluate the quality of graph features produced by FGSD and demonstrate their utility, we apply them to the graph classification problem. Through extensive experiments, we show that a simple SVM based classification algorithm, driven with our powerful FGSD based graph features, significantly outperforms all the more sophisticated state-of-art algorithms on the unlabeled node datasets in terms of both accuracy and speed; it also yields very competitive results on the labeled datasets - despite the fact it does not utilize any node label information.

************************************

Expectation Propagation with Stochastic Kinetic Model in Complex Interaction Systems

Le Fang, Fan Yang, Wen Dong, Tong Guan, Chunming Qiao

Technological breakthroughs allow us to collect data with increasing spatio-temporal resolution from complex interaction systems. The combination of high-resolution observations, expressive dynamic models, and efficient machine learning algorithms can lead to crucial insights into complex interaction dynamics and the functions of these systems. In this paper, we formulate the dynamics of a complex interacting network as a stochastic process driven by a sequence of events, and develop expectation propagation algorithms to make inferences from noisy observations. To avoid getting stuck at a local optimum, we formulate the problem of minimizing Bethe free energy as a constrained primal problem and take advantage of the concavity of dual problem in the feasible domain of dual variables guaranteed by duality theorem. Our expectation propagation algorithms demonstrate better performance in inferring the interaction dynamics in complex transportation networks than competing models such as particle filter, extended Kalman filter, and deep neural networks.

************************************

Welfare Guarantees from Data

Darrell Hoy, Denis Nekipelov, Vasilis Syrgkanis

Analysis of efficiency of outcomes in game theoretic settings has been a main item of study at the intersection of economics and computer science. The notion of the price of anarchy takes a worst-case stance to efficiency analysis, considering instance independent guarantees of efficiency. We propose a data-dependent analog of the price of anarchy that refines this worst-case assuming access to samples of strategic behavior. We focus on auction settings, where the latter is non-trivial due to the private information held by participants. Our approach to bounding the efficiency from data is robust to statistical errors and mis-specification. Unlike traditional econometrics, which seek to learn the private information of players from observed behavior and then analyze properties of the outco me, we directly quantify the inefficiency without going through the private information. We apply our approach to datasets from a sponsored search auction system and find empirical results that are a significant improvement over bounds from worst-case analysis.

************************************

Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference
Abhishek Kumar, Prasanna Sattigeri, Tom Fletcher

Semi-supervised learning methods using Generative adversarial networks (GANs) have shown promising empirical success recently. Most of these methods use a shared discriminator/classifier which discriminates real examples from fake while also predicting the class label. Motivated by the ability of the GANs generator to capture the data manifold well, we propose to estimate the tangent space to the data manifold using GANs and employ it to inject invariances into the classifier. In the process, we propose enhancements over existing methods for learning the inverse mapping (i.e., the encoder)  which greatly improves in terms of semantic similarity of the reconstructed sample with the input sample. We observe considerable empirical gains in semi-supervised learning over baselines, particularly in the cases when the number of labeled examples is low. We also provide insights into how fake examples influence the semi-supervised learning procedure.

************************************

Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples
Moustapha M. Cisse, Yossi Adi, Natalia Neverova, Joseph Keshet

Generating adversarial examples is a critical step for evaluating and improving the robustness of learning machines. So far, most existing methods only work for classification and are not designed to alter the true performance measure of the problem at hand. We introduce a novel flexible approach named Houdini for generating adversarial examples specifically tailored for the final performance measure of the task considered, be it combinatorial and non-decomposable. We successfully apply Houdini to a range of applications such as speech recognition, pose estimation and semantic segmentation. In all cases, the attacks based on Houdini achieve higher success rate than those based on the traditional surrogates used to train the models while using a less perceptible adversarial perturbation.

************************************

Clustering Stable Instances of Euclidean k-means.
Aravindan Vijayaraghavan, Abhratanu Dutta, Alex Wang

The Euclidean k-means problem is arguably the most widely-studied clustering problem in machine learning. While the k-means objective is NP-hard in the worst-case, practitioners have enjoyed remarkable success in applying heuristics like Lloyd's algorithm for this problem. To address this disconnect, we study the following question: what properties of real-world instances will enable us to design efficient algorithms and prove guarantees for finding the optimal clustering?

We consider a natural notion called additive perturbation stability that we believe captures many practical instances of Euclidean k-means clustering. Stable instances have unique optimal k-means solutions that does not change even when each point is perturbed a little (in Euclidean distance). This captures the property that k-means optimal solution should be tolerant to measurement errors and uncertainty in the points. We design efficient algorithms that provably recover the optimal clustering for instances that are additive perturbation stable. When the instance has some additional separation, we can design a simple, efficient algorithm with provable guarantees that is also robust to outliers. We also complement these results by studying the amount of stability in real datasets, and demonstrating that our algorithm performs well on these benchmark datasets.

************************************

Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin
Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi

The past decade has seen a revolution in genomic technologies that enabled a flood of genome-wide profiling of chromatin marks.  Recent literature tried to understand gene regulation by predicting gene expression from large-scale chromatin

measurements.  Two fundamental challenges exist for such learning tasks: (1) genome-wide chromatin signals are spatially structured, high-dimensional and highly modular; and (2) the core aim is to understand what are the relevant factors and how they work together.  Previous studies either failed to model complex dependencies among input signals or relied on separate feature analysis to explain the decisions. This paper presents an attention-based deep learning approach; AttentiveChrome, that uses a unified architecture to model and to interpret dependencies among chromatin factors for controlling gene regulation. AttentiveChrome uses a hierarchy of multiple Long Short-Term Memory (LSTM) modules to encode the input signals and to model how various chromatin marks cooperate automatically. AttentiveChrome trains two levels of attention jointly with the target prediction, enabling it to attend differentially to relevant marks and to locate important positions per mark. We evaluate the model across 56 different cell types (tasks) in human. Not only is the proposed architecture more accurate, but its attention scores also provide a better interpretation than state-of-the-art feature visualization methods such as saliency map.
************************************

Deep Reinforcement Learning from Human Preferences

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, Dario Amodei

For sophisticated reinforcement learning (RL) systems to interact usefully with real-world environments, we need to communicate complex goals to these systems. In this work, we explore goals defined in terms of (non-expert) human preferences between pairs of trajectory segments. Our approach separates learning the goal from learning the behavior to achieve it. We show that this approach can effectively solve complex RL tasks without access to the reward function, including Atari games and simulated robot locomotion, while providing feedback on about 0.1% of our agent's interactions with the environment. This reduces the cost of human oversight far enough that it can be practically applied to state-of-the-art RL systems. To demonstrate the flexibility of our approach, we show that we can successfully train complex novel behaviors with about an hour of human time. These behaviors and environments are considerably more complex than any which have been previously learned from human feedback.
************************************

Subset Selection under Noise

Chao Qian, Jing-Cheng Shi, Yang Yu, Ke Tang, Zhi-Hua Zhou

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.
************************************

PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space

Charles Ruizhongtai Qi, Li Yi, Hao Su, Leonidas J. Guibas

Few prior works study deep learning on point sets. PointNet is a pioneer in this direction. However, by design PointNet does not capture local structures induced by the metric space points live in, limiting its ability to recognize fine-grained patterns and generalizability to complex scenes. In this work, we introduce a hierarchical neural network that applies PointNet recursively on a nested partitioning of the input point set. By exploiting metric space distances, our network is able to learn local features with increasing contextual scales. With further observation that point sets are usually sampled with varying densities, which results in greatly decreased performance for networks trained on uniform densities, we propose novel set learning layers to adaptively combine features from multiple scales. Experiments show that our network called PointNet++ is able to learn deep point set features efficiently and robustly. In particular, results significantly better than state-of-the-art have been obtained on challenging benchmarks of 3D point clouds.
************************************

Approximation Bounds for Hierarchical Clustering: Average Linkage, Bisecting K-means, and Local Search

Benjamin Moseley, Joshua Wang

Hierarchical clustering is a data analysis method that has been used for decades. Despite its widespread use, the method has an underdeveloped analytical foundation. Having a well understood foundation would both support the currently used methods and help guide future improvements. The goal of this paper is to give an analytic framework to better understand observations seen in practice. This paper considers the dual of a problem framework for hierarchical clustering introduced by Dasgupta. The main result is that one of the most popular algorithms used in practice, average linkage agglomerative clustering, has a small constant approximation ratio for this objective. Furthermore, this paper establishes that using bisecting k-means divisive clustering has a very poor lower bound on its approximation ratio for the same objective.  However, we show that there are divisive algorithms that perform well with respect to this objective by giving two constant approximation algorithms. This paper is some of the first work to establish guarantees on widely used hierarchical algorithms for a natural objective function.  This objective and analysis give insight into what these popular algorithms are optimizing and when they will perform well.

************************************

Thinking Fast and Slow with Deep Learning and Tree Search

Thomas Anthony, Zheng Tian, David Barber

Sequential decision making problems, such as structured prediction, robotic control, and game playing, require a combination of planning policies and generalisation of those plans. In this paper, we present Expert Iteration (ExIt), a novel reinforcement learning algorithm which decomposes the problem into separate planning and generalisation tasks. Planning new policies is performed by tree search, while a deep neural network generalises those plans. Subsequently, tree search is improved by using the neural network policy to guide search, increasing the strength of new plans. In contrast, standard deep Reinforcement Learning algorithms rely on a neural network not only to generalise plans, but to discover them too. We show that ExIt outperforms REINFORCE for training a neural network to play the board game Hex, and our final tree search agent, trained tabula rasa, defeats MoHex1.0, the most recent Olympiad Champion player to be publicly released.

************************************

Learning Combinatorial Optimization Algorithms over Graphs

Elias Khalil, Hanjun Dai, Yuyu Zhang, Bistra Dilkina, Le Song

The design of good heuristics or approximation algorithms for NP-hard combinatorial optimization problems often requires significant specialized knowledge and trial-and-error. Can we automate this challenging, tedious process, and learn the algorithms instead? In many real-world applications, it is typically the case that the same optimization problem is solved again and again on a regular basis, maintaining the same problem structure but differing in the data. This provides an opportunity for learning heuristic algorithms that exploit the structure of such recurring problems.  In this paper, we propose a unique combination of reinforcement learning and graph embedding to address this challenge. The learned greedy policy behaves like a meta-algorithm that incrementally constructs a solution, and the action is determined by the output of a graph embedding network capturing the current state of the solution. We show that our framework can be applied to a diverse range of optimization problems over graphs, and learns effective algorithms for the Minimum Vertex Cover, Maximum Cut and Traveling Salesman problems.

************************************

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice

Jeffrey Pennington, Samuel Schoenholz, Surya Ganguli

It is well known that weight initialization in deep networks can have a dramatic impact on learning speed. For example, ensuring the mean squared singular value of a network's input-output Jacobian is O(1) is essential for avoiding exponentially vanishing or exploding gradients. Moreover, in deep linear networks, ensuring that all singular values of the Jacobian are concentrated near 1 can yield a dramatic additional speed-up in learning; this is a property known as dynamical

isometry. However, it is unclear how to achieve dynamical isometry in nonlinear deep networks.  We address this question by employing powerful tools from free probability theory to analytically compute the {\it entire} singular value distribution of a deep network's input-output Jacobian. We explore the dependence of the singular value distribution on the depth of the network, the weight initialization, and the choice of nonlinearity. Intriguingly, we find that ReLU networks are incapable of dynamical isometry. On the other hand, sigmoidal networks can achieve isometry, but only with orthogonal weight initialization. Moreover, we demonstrate empirically that deep nonlinear networks achieving dynamical isometry learn orders of magnitude faster than networks that do not. Indeed, we show that properly-initialized deep sigmoidal networks consistently outperform deep ReLU networks. Overall, our analysis reveals that controlling the entire distribution of Jacobian singular values is an important design consideration in deep learning.
************************************
Adaptive Classification for Prediction Under a Budget
Feng Nan, Venkatesh Saligrama
We propose a novel adaptive approximation approach for test-time resource-constrained prediction motivated by Mobile, IoT, health, security and other applications, where constraints in the form of computation, communication, latency and feature acquisition costs arise. We learn an adaptive low-cost system by training a gating and prediction model that limits utilization of a high-cost model to hard input instances and gates easy-to-handle input instances to a low-cost model. Our method is based on adaptively approximating the high-cost model in regions where low-cost models suffice for making highly accurate predictions. We pose an empirical loss minimization problem with cost constraints to jointly train gating and prediction models. On a number of benchmark datasets our method outperforms state-of-the-art achieving higher accuracy for the same cost.
************************************
Online Convex Optimization with Stochastic Constraints
Hao Yu, Michael Neely, Xiaohan Wei
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.
************************************
Structured Bayesian Pruning via Log-Normal Multiplicative Noise
Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, Dmitry P. Vetrov
Dropout-based regularization methods can be regarded as injecting random noise with pre-defined magnitude to different parts of the neural network during training. It was recently shown that Bayesian dropout procedure not only improves gener- alization but also leads to extremely sparse neural architectures by automatically setting the individual noise magnitude per weight. However, this sparsity can hardly be used for acceleration since it is unstructured. In the paper, we propose a new Bayesian model that takes into account the computational structure of neural net- works and provides structured sparsity, e.g. removes neurons and/or convolutional channels in CNNs. To do this we inject noise to the neurons outputs while keeping the weights unregularized. We establish the probabilistic model with a proper truncated log-uniform prior over the noise and truncated log-normal variational approximation that ensures that the KL-term in the evidence lower bound is com- puted in closed-form. The model leads to structured sparsity by removing elements with a low SNR from the computation graph and provides significant acceleration on a number of deep neural architectures. The model is easy to implement as it can be formulated as a separate dropout-like layer.
************************************
Clustering with Noisy Queries
Arya Mazumdar, Barna Saha
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-auth

ors prior to requesting a name change in the electronic proceedings.

************************************

Compression-aware Training of Deep Networks

Jose M. Alvarez, Mathieu Salzmann

In recent years, great progress has been made in a variety of application domains thanks to the development of increasingly deeper neural networks. Unfortunately, the huge number of units of these networks makes them expensive both computationally and memory-wise. To overcome this, exploiting the fact that deep networks are over-parametrized, several compression strategies have been proposed. These methods, however, typically start from a network that has been trained in a standard manner, without considering such a future compression. In this paper, we propose to explicitly account for compression in the training process. To this end, we introduce a regularizer that encourages the parameter matrix of each layer to have low rank during training. We show that accounting for compression during training allows us to learn much more compact, yet at least as effective, models than state-of-the-art compression techniques.

************************************

Maxing and Ranking with Few Assumptions

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, Vaishakh Ravindrakumar

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.

************************************

Subspace Clustering via Tangent Cones

Amin Jalali, Rebecca Willett

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.

************************************

DropoutNet: Addressing Cold Start in Recommender Systems

Maksims Volkovs, Guangwei Yu, Tomi Poutanen

Latent models have become the default choice for recommender systems due to their performance and scalability. However, research in this area has primarily focused on modeling user-item interactions,  and few latent models have been developed for cold start. Deep learning has recently achieved remarkable success showing excellent results for diverse input types. Inspired by these results we propose a neural network based latent model called DropoutNet to address the cold start problem in recommender systems. Unlike existing approaches that incorporate additional content-based objective terms, we instead focus on the optimization and show that neural network models can be explicitly trained for cold start through dropout. Our model can  be applied on top of any existing latent model effectively providing cold start capabilities, and full power of deep architectures. Empirically we demonstrate  state-of-the-art accuracy on publicly available benchmarks. Code is available at  https://github.com/layer6ai-labs/DropoutNet.

************************************

Unsupervised Image-to-Image Translation Networks

Ming-Yu Liu, Thomas Breuel, Jan Kautz

Unsupervised image-to-image translation aims at learning a joint distribution of images in different domains by using images from the marginal distributions in individual domains. Since there exists an infinite set of joint distributions that can arrive the given marginal distributions, one could infer nothing about the joint distribution from the marginal distributions without additional assumptions. To address the problem, we make a shared-latent space assumption and propose an unsupervised image-to-image translation framework based on Coupled GANs. We compare the proposed framework with competing approaches and present high quality image translation results on various challenging unsupervised image translation tasks, including street scene image translation, animal image translation, an

d face image translation. We also apply the proposed framework to domain adaptation and achieve state-of-the-art performance on benchmark datasets. Code and additional results are available in https://github.com/mingyuliutw/unit.
************************************

SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability
Maithra Raghu, Justin Gilmer, Jason Yosinski, Jascha Sohl-Dickstein
We propose a new technique, Singular Vector Canonical Correlation Analysis (SVCCA), a tool for quickly comparing two representations in a way that is both invariant to affine transform (allowing comparison between different layers and networks) and fast to compute (allowing more comparisons to be calculated than with previous methods).  We deploy this tool to measure the intrinsic dimensionality of layers, showing in some cases needless over-parameterization; to probe learning dynamics throughout training, finding that networks converge to final representations from the bottom up; to show where class-specific information in networks is formed; and to suggest new training regimes that simultaneously save computation and overfit less.
************************************

Fast Rates for Bandit Optimization with Upper-Confidence Frank-Wolfe
Quentin Berthet, Vianney Perchet
We consider the problem of bandit optimization, inspired by stochastic optimization and online learning problems with bandit feedback. In this problem, the objective is to minimize a global loss function of all the actions, not necessarily a cumulative loss. This framework allows us to study a very general class of problems, with applications in statistics, machine learning, and other fields. To solve this problem, we analyze the Upper-Confidence Frank-Wolfe algorithm, inspired by techniques for bandits and convex optimization. We give theoretical guarantees for the performance of this algorithm over various classes of functions, and discuss the optimality of these results.
************************************

Identifying Outlier Arms in Multi-Armed Bandit
Honglei Zhuang, Chi Wang, Yifan Wang
We study a novel problem lying at the intersection of two areas: multi-armed bandit and outlier detection. Multi-armed bandit is a useful tool to model the process of incrementally collecting data for multiple objects in a decision space. Outlier detection is a powerful method to narrow down the attention to a few objects after the data for them are collected. However, no one has studied how to detect outlier objects while incrementally collecting data for them, which is necessary when data collection is expensive. We formalize this problem as identifying outlier arms in a multi-armed bandit. We propose two sampling strategies with theoretical guarantee, and analyze their sampling efficiency. Our experimental results on both synthetic and real data show that our solution saves 70-99% of data collection cost from baseline while having nearly perfect accuracy.
************************************

Discovering Potential Correlations via Hypercontractivity
Hyeji Kim, Weihao Gao, Sreeram Kannan, Sewoong Oh, Pramod Viswanath
Discovering a correlation from one variable to another variable is of fundamental scientific and practical interest. While existing correlation measures are suitable for discovering average correlation, they fail to discover hidden or potential correlations. To bridge this gap, (i) we postulate a set of natural axioms that we expect a measure of potential correlation to satisfy; (ii) we show that the rate of information bottleneck, i.e., the hypercontractivity coefficient, satisfies all the proposed axioms; (iii) we provide a novel estimator to estimate the hypercontractivity coefficient from samples; and (iv) we provide numerical experiments demonstrating that this proposed estimator discovers potential correlations among various indicators of WHO datasets, is robust in discovering gene interactions from gene expression time series data, and is statistically more powerful than the estimators for other correlation measures in binary hypothesis testing of canonical examples of potential correlations.
************************************

A Dirichlet Mixture Model of Hawkes Processes for Event Sequence Clustering

Hongteng Xu, Hongyuan Zha

How to cluster event sequences generated via different point processes is an interesting and important problem in statistical machine learning. To solve this problem, we propose and discuss an effective model-based clustering method based on a novel Dirichlet mixture model of a special but significant type of point processes --- Hawkes process. The proposed model generates the event sequences with different clusters from the Hawkes processes with different parameters, and uses a Dirichlet process as the prior distribution of the clusters. We prove the identifiability of our mixture model and propose an effective variational Bayesian inference algorithm to learn our model. An adaptive inner iteration allocation strategy is designed to accelerate the convergence of our algorithm. Moreover, we investigate the sample complexity and the computational complexity of our learning algorithm in depth. Experiments on both synthetic and real-world data show that the clustering method based on our model can learn structural triggering patterns hidden in asynchronous event sequences robustly and achieve superior performance on clustering purity and consistency compared to existing methods.

************************************

Efficient Approximation Algorithms for Strings Kernel Based Sequence Classification

Muhammad Farhan, Juvaria Tariq, Arif Zaman, Mudassir Shabbir, Imdad Ullah Khan

Sequence classification algorithms, such as SVM, require a definition of distance (similarity) measure between two sequences. A commonly used notion of similarity is the number of matches between k-mers (k-length subsequences) in the two sequences. Extending this definition, by considering two k-mers to match if their distance is at most m, yields better classification performance. This, however, makes the problem computationally much more complex. Known algorithms to compute this similarity have computational complexity that render them applicable only for small values of k and m. In this work, we develop novel techniques to efficiently and accurately estimate the pairwise similarity score, which enables us to use much larger values of k and m, and get higher predictive accuracy. This opens up a broad avenue of applying this classification approach to audio, images, and text sequences. Our algorithm achieves excellent approximation performance with theoretical guarantees. In the process we solve an open combinatorial problem, which was posed as a major hindrance to the scalability of existing solutions. We give analytical bounds on quality and runtime of our algorithm and report its empirical performance on real world biological and music sequences datasets.

************************************

Multi-output Polynomial Networks and Factorization Machines

Mathieu Blondel, Vlad Niculae, Takuma Otsuka, Naonori Ueda

Factorization machines and polynomial networks are supervised polynomial models based on an efficient low-rank decomposition. We extend these models to the multi-output setting, i.e., for learning vector-valued functions, with application to multi-class or multi-task problems. We cast this as the problem of learning a 3-way tensor whose slices share a common basis and propose a convex formulation of that problem. We then develop an efficient conditional gradient algorithm and prove its global convergence, despite the fact that it involves a non-convex basis selection step. On classification tasks, we show that our algorithm achieves excellent accuracy with much sparser models than existing methods. On recommendation system tasks, we show how to combine our algorithm with a reduction from ordinal regression to multi-output classification and show that the resulting algorithm outperforms simple baselines in terms of ranking accuracy.

************************************

Tractability in Structured Probability Spaces

Arthur Choi, Yujia Shen, Adnan Darwiche

Recently, the Probabilistic Sentential Decision Diagram (PSDD) has been proposed as a framework for systematically inducing and learning distributions over structured objects, including combinatorial objects such as permutations and rankings, paths and matchings on a graph, etc. In this paper, we study the scalability of such models in the context of representing and learning distributions over ro

utes on a map. In particular, we introduce the notion of a hierarchical route distribution and show how they can be leveraged to construct tractable PSDDs over route distributions, allowing them to scale to larger maps. We illustrate the utility of our model empirically, in a route prediction task, showing how accuracy can be increased significantly compared to Markov models.
************************************
Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search
Luigi Acerbi, Wei Ji Ma
Computational models in fields such as computational neuroscience  are often evaluated via stochastic simulation or numerical approximation. Fitting these models implies a difficult optimization problem over complex, possibly noisy parameter landscapes. Bayesian optimization (BO) has been successfully applied to solving expensive black-box problems in engineering and machine learning.  Here we explore whether BO can be applied as a general tool for model fitting. First, we present a novel hybrid BO algorithm, Bayesian adaptive direct search (BADS), that achieves competitive performance with an affordable computational overhead for the running time of typical models. We then perform an extensive benchmark of BADS vs. many common and state-of-the-art nonconvex, derivative-free optimizers, on a set of model-fitting problems with real data and models from six studies in behavioral, cognitive, and computational neuroscience. With default settings, BADS consistently finds comparable or better solutions than other methods, including `vanilla' BO, showing great promise for advanced BO techniques, and BADS in particular, as a general model-fitting tool.
************************************
Multi-Information Source Optimization
Matthias Poloczek, Jialei Wang, Peter Frazier
We consider Bayesian methods for multi-information source optimization (MISO), in which we seek to optimize an expensive-to-evaluate black-box objective function while also accessing cheaper but biased and noisy approximations ("information sources"). We present a novel algorithm that outperforms the state of the art for this problem by using a Gaussian process covariance kernel better suited to MISO than those used by previous approaches, and an acquisition function based on a one-step optimality analysis supported by efficient parallelization. We also provide a novel technique to guarantee the asymptotic quality of the solution provided by this algorithm. Experimental evaluations demonstrate that this algorithm consistently finds designs of higher value at less cost than previous approaches.
************************************
Differentially private Bayesian learning on distributed data
Mikko Heikkilä, Eemil Lagerspetz, Samuel Kaski, Kana Shimizu, Sasu Tarkoma, Antti Honkela
Many applications of machine learning, for example in health care, would benefit from methods that can guarantee privacy of data subjects. Differential privacy (DP) has become established as a standard for protecting learning results. The standard DP algorithms require a single trusted party to have access to the entire data, which is a clear weakness, or add prohibitive amounts of noise. We consider DP Bayesian learning in a distributed setting, where each party only holds a single sample or a few samples of the data. We propose a learning strategy based on a secure multi-party sum function for aggregating summaries from data holders and the Gaussian mechanism for DP. Our method builds on an asymptotically optimal and practically efficient DP Bayesian inference with rapidly diminishing extra cost.
************************************
MMD GAN: Towards Deeper Understanding of Moment Matching Network
Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabas Poczos
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.

************************************

Convergence of Gradient EM on Multi-component Mixture of Gaussians

Bowei Yan, Mingzhang Yin, Purnamrita Sarkar

In this paper, we study convergence properties of the gradient variant of Expectation-Maximization algorithm~\cite{lange1995gradient} for Gaussian Mixture Models for arbitrary number of clusters and mixing coefficients. We derive the convergence rate depending on the mixing coefficients, minimum and maximum pairwise distances between the true centers, dimensionality and number of components; and obtain a near-optimal local contraction radius. While there have been some recent notable works that derive local convergence rates for EM in the two symmetric mixture of Gaussians, in the more general case, the derivations need structurally different and non-trivial arguments. We use recent tools from learning theory and empirical processes to achieve our theoretical results.

************************************

Bayesian Dyadic Trees and Histograms for  Regression

Stéphanie van der Pas, Veronika Ro■ková

Many machine learning  tools for regression are based on recursive partitioning of the covariate space into smaller regions, where the regression function can be estimated locally. Among these, regression trees and their ensembles have demonstrated impressive empirical performance.    In this work,  we shed light on the machinery behind Bayesian variants of these methods.  In particular, we study Bayesian regression histograms, such as Bayesian dyadic trees, in the simple regression case with just one predictor.   We focus on the reconstruction of regression surfaces that are piecewise constant, where the number of jumps is unknown. We show that with suitably designed priors, posterior distributions concentrate around the true step regression function at a near-minimax rate. These results {\sl do not} require the knowledge of the true number of steps, nor the width of the true partitioning cells. Thus, Bayesian dyadic regression trees are fully adaptive and can recover the true piecewise regression function nearly as well as if we knew the exact number and location  of jumps. Our results constitute the first step towards  understanding why Bayesian trees and their ensembles have worked so well in practice.  As an aside, we discuss prior distributions  on balanced interval partitions and how they relate to an old  problem in geometric probability. Namely, we relate the probability of covering the circumference of a circle with random arcs whose endpoints are confined to a grid, a new variant of the original problem.

************************************

Efficient and Flexible Inference for Stochastic Systems

Stefan Bauer, Nico S. Gorbach, Djordje Miladinovic, Joachim M. Buhmann

Many real world dynamical systems are described by stochastic differential equations. Thus parameter inference is a challenging and important problem in many disciplines. We provide a grid free and flexible algorithm offering parameter and state inference for stochastic systems and compare our approch based on variational approximations to state of the art methods showing significant advantages both in runtime and accuracy.

************************************

Learning ReLUs via Gradient Descent

Mahdi Soltanolkotabi

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.

************************************

Learning Graph Representations with Embedding Propagation

Alberto Garcia Duran, Mathias Niepert

We propose EP, Embedding Propagation, an unsupervised learning framework for graph-structured data. EP learns vector representations of graphs by passing two types of messages between neighboring nodes. Forward messages consist of label representations such as representations of words and other attributes associated with the nodes. Backward messages consist of gradients that result from aggregatin g the label representations and applying a reconstruction loss. Node representations are finally computed from the representation of their labels. With significantly fewer parameters and hyperparameters, an instance of EP is competitive with and often outperforms state of the art unsupervised and semi-supervised learning methods on a range of benchmark data sets.

************************************

Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation

Matthias Hein, Maksym Andriushchenko

Recent work has shown that state-of-the-art classifiers are quite brittle, in the sense that a small adversarial change of an originally with high confidence correctly classified  input leads to a wrong classification again with high confidence. This raises concerns that such classifiers are vulnerable to attacks and calls into question their usage in safety-critical systems. We show in this paper for the first time formal guarantees on the robustness of a classifier by giving instance-specific \emph{lower bounds} on the norm of the input manipulation required to change the classifier decision. Based on this analysis we propose the Cross-Lipschitz regularization functional. We show that using this form of regularization in kernel methods resp. neural networks improves the robustness of the classifier without any loss in prediction performance.

************************************

Collapsed variational Bayes for Markov jump processes

Boqian Zhang, Jiangwei Pan, Vinayak A. Rao

Markov jump processes are continuous-time stochastic processes widely used in statistical applications in the natural sciences, and more recently in machine learning. Inference for these models typically proceeds via Markov chain Monte Carlo, and can suffer from various computational challenges. In this work, we propose a novel collapsed variational inference algorithm to address this issue. Our work leverages ideas from discrete-time Markov chains, and exploits a connection between these two through an idea called uniformization. Our algorithm proceeds by marginalizing out the parameters of the Markov jump process, and then approximating the distribution over the trajectory with a factored distribution over segments of a piecewise-constant function. Unlike MCMC schemes that marginalize out transition times of a piecewise-constant process, our scheme optimizes the discretization of time, resulting in significant computational savings. We apply our ideas to synthetic data as well as a dataset of check-in recordings, where we demonstrate superior performance over state-of-the-art MCMC methods.

************************************

Is the Bellman residual a bad proxy?

Matthieu Geist, Bilal Piot, Olivier Pietquin

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.

************************************

Efficient Use of Limited-Memory Accelerators for Linear Learning on Heterogeneous Systems

Celestine Dünner, Thomas Parnell, Martin Jaggi

We propose a generic algorithmic building block to accelerate training of  machine learning models on heterogeneous compute systems. Our scheme allows to efficiently employ compute accelerators such as GPUs and FPGAs for the training of large-scale machine learning models, when the training data exceeds their memory capacity. Also, it provides adaptivity to any system's memory hierarchy in terms of size and processing speed. Our technique is built upon novel theoretical insights regarding primal-dual coordinate methods, and uses duality gap information to dynamically decide which part of the data should be made available for fast processing. To illustrate the power of our approach we demonstrate its performance for training of generalized linear models on a large-scale dataset exceeding the memory size of a modern GPU, showing an order-of-magnitude speedup over existing approaches.

************************************

Noise-Tolerant Interactive Learning Using Pairwise Comparisons
Yichong Xu, Hongyang Zhang, Kyle Miller, Aarti Singh, Artur Dubrawski

We study the problem of interactively learning a binary classifier using noisy labeling and pairwise comparison oracles, where the comparison oracle answers which one in the given two instances is more likely to be positive. Learning from such oracles has multiple applications where obtaining direct labels is harder but pairwise comparisons are easier, and the algorithm can leverage both types of oracles. In this paper, we attempt to characterize how the access to an easier comparison oracle helps in improving the label and total query complexity. We show that the comparison oracle reduces the learning problem to that of learning a threshold function. We then present an algorithm that interactively queries the label and comparison oracles and we characterize its query complexity under Tsybakov and adversarial noise conditions for the comparison and labeling oracles. Our lower bounds show that our label and total query complexity is almost optimal.

************************************

Near-Optimal Edge Evaluation in Explicit Generalized Binomial Graphs
Sanjiban Choudhury, Shervin Javdani, Siddhartha Srinivasa, Sebastian Scherer

Robotic motion-planning problems, such as a UAV flying fast in a partially-known environment or a robot arm moving around cluttered objects, require finding collision-free paths quickly. Typically, this is solved by constructing a graph, where vertices represent robot configurations and edges represent potentially valid movements of the robot between theses configurations. The main computational bottlenecks are expensive edge evaluations to check for collisions. State of the art planning methods do not reason about the optimal sequence of edges to evaluate in order to find a collision free path quickly. In this paper, we do so by drawing a novel equivalence between motion planning and the Bayesian active learning paradigm of decision region determination (DRD). Unfortunately, a straight application of ex- isting methods requires computation exponential in the number of edges in a graph. We present BISECT, an efficient and near-optimal algorithm to solve the DRD problem when edges are independent Bernoulli random variables. By leveraging this property, we are able to significantly reduce computational complexity from exponential to linear in the number of edges. We show that BISECT outperforms several state of the art algorithms on a spectrum of planning problems for mobile robots, manipulators, and real flight data collected from a full scale helicopter. Open-source code and details can be found here: https://github.com/sanjibac/matlablearningcollision_checking

************************************

Minimal Exploration in Structured Stochastic Bandits
Richard Combes, Stefan Magureanu, Alexandre Proutiere

This paper introduces and addresses a wide class of stochastic bandit problems where the function mapping the arm to the corresponding reward exhibits some known structural properties. Most existing structures (e.g. linear, lipschitz, unimodal, combinatorial, dueling,...) are covered by our framework. We derive an asymptotic instance-specific regret lower bound for these problems, and develop OSSB, an algorithm whose regret matches this fundamental limit. OSSB is not based on the classical principle of ``optimism in the face of uncertainty'' or on Thompson sampling, and rather aims at matching the minimal exploration rates of sub-optimal arms as characterized in the derivation of the regret lower bound. We illustrate the efficiency of OSSB using numerical experiments in the case of the linear bandit problem and show that OSSB outperforms existing algorithms, including Thompson sampling

************************************

Learning Efficient Object Detection Models with Knowledge Distillation
Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, Manmohan Chandraker

Despite significant accuracy improvement in convolutional neural networks (CNN) based object detectors, they often require prohibitive runtimes to process an image for real-time applications. State-of-the-art models often use very deep networks with a large number of floating point operations. Efforts such as model com pression learn compact models with fewer number of parameters, but with much reduced accuracy. In this work, we propose a new framework to learn compact and fast ob- ject detection networks with improved accuracy using knowledge distillation [20] and hint learning [34]. Although knowledge distillation has demonstrated excellent improvements for simpler classification setups, the complexity of detection poses new challenges in the form of regression, region proposals and less voluminous la- bels. We address this through several innovations such as a weighted cross-entropy loss to address class imbalance, a teacher bounded loss to handle the regression component and adaptation layers to better learn from intermediate teacher distribu- tions. We conduct comprehensive empirical evaluation with different distillation configurations over multiple datasets including PASCAL, KITTI, ILSVRC and MS-COCO. Our results show consistent improvement in accuracy-speed trade-offs for modern multi-class detection models.

************************************

Learning Chordal Markov Networks via Branch and Bound
Kari Rantanen, Antti Hyttinen, Matti Järvisalo
We present a new algorithmic approach for the task of finding a chordal Markov network structure that maximizes a given scoring function. The algorithm is based on branch and bound and integrates dynamic programming for both domain pruning and for obtaining strong bounds for search-space pruning. Empirically, we show that the approach dominates in terms of running times a recent integer programming approach (and thereby also a recent constraint optimization approach) for the problem. Furthermore, our algorithm scales at times further with respect to the number of variables than a state-of-the-art dynamic programming algorithm for the problem, with the potential of reaching 20 variables and at the same time circumventing the tight exponential lower bounds on memory consumption of the pure dynamic programming approach.

************************************

Efficient Optimization for Linear Dynamical Systems with Applications to Clustering and Sparse Coding
Wenbing Huang, Mehrtash Harandi, Tong Zhang, Lijie Fan, Fuchun Sun, Junzhou Huang
Linear Dynamical Systems (LDSs) are fundamental tools for modeling spatio-temporal data in various disciplines. Though rich in modeling, analyzing LDSs is not free of difficulty, mainly because LDSs do not comply with Euclidean geometry and hence conventional learning techniques can not be applied directly. In this paper, we propose an efficient projected gradient descent method to minimize a general form of a loss function and demonstrate how clustering and sparse coding with LDSs can be solved by the proposed method efficiently. To this end, we first derive a novel canonical form for representing the parameters of an LDS, and then show how gradient-descent updates through the projection on the space of LDSs can be achieved dexterously. In contrast to previous studies, our solution avoids any approximation in LDS modeling or during the optimization process. Extensive experiments reveal the superior performance of the proposed method in terms of the convergence and classification accuracy over state-of-the-art techniques.

************************************

Deep Subspace Clustering Networks
Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, Ian Reid
We present a novel deep neural network architecture for unsupervised subspace clustering. This architecture is built upon deep auto-encoders, which non-linearly map the input data into a latent space. Our key idea is to introduce a novel self-expressive layer between the encoder and the decoder to mimic the "self-expressiveness" property that has proven effective in traditional subspace clustering. Being differentiable, our new self-expressive layer provides a simple but effective way to learn pairwise affinities between all data points through a standard back-propagation procedure. Being nonlinear, our neural-network based method is able to cluster data points having complex (often nonlinear) structures. We further propose pre-training and fine-tuning strategies that let us effectively learn the parameters of our subspace clustering networks. Our experiments show that the proposed method significantly outperforms the state-of-the-art unsupervise

d subspace clustering methods.

************************************

Robust Estimation of Neural Signals in Calcium Imaging

Hakan Inan, Murat A. Erdogdu, Mark Schnitzer

Calcium imaging is a prominent technology in neuroscience research which allows for simultaneous recording of large numbers of neurons in awake animals. Automated extraction of neurons and their temporal activity from imaging datasets is an important step in the path to producing neuroscience results. However, nearly all imaging datasets contain gross contaminating sources which could originate from the technology used, or the underlying biological tissue. Although past work has considered the effects of contamination under limited circumstances, there has not been a general framework treating contamination and its effects on the statistical estimation of calcium signals. In this work, we proceed in a new direction and propose to extract cells and their activity using robust statistical estimation. Using the theory of M-estimation, we derive a minimax optimal robust loss, and also find a simple and practical optimization routine for this loss with provably fast convergence. We use our proposed robust loss in a matrix factorization framework to extract the neurons and their temporal activity in calcium imaging datasets. We demonstrate the superiority of our robust estimation approach over existing methods on both simulated and real datasets.

************************************

Fast-Slow Recurrent Neural Networks

Asier Mujika, Florian Meier, Angelika Steger

Processing sequential data of variable length is a major challenge in a wide range of applications, such as speech recognition, language modeling, generative image modeling and machine translation. Here, we address this challenge by proposing a novel recurrent neural network (RNN) architecture, the Fast-Slow RNN (FS-RNN). The FS-RNN incorporates the strengths of both multiscale RNNs and deep transition RNNs as it processes sequential data on different timescales and learns complex transition functions from one time step to the next. We evaluate the FS-RNN on two character based language modeling data sets, Penn Treebank and Hutter Prize Wikipedia, where we improve state of the art results to 1.19 and 1.25 bits-per-character (BPC), respectively. In addition, an ensemble of two FS-RNNs achieves 1.20 BPC on Hutter Prize Wikipedia outperforming the best known compression algorithm with respect to the BPC measure. We also present an empirical investigation of the learning and network dynamics of the FS-RNN, which explains the improved performance compared to other RNN architectures. Our approach is general as any kind of RNN cell is a possible building block for the FS-RNN architecture, and thus can be flexibly applied to different tasks.

************************************

PredRNN: Recurrent Neural Networks for Predictive Learning using Spatiotemporal LSTMs

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, Philip S. Yu

The predictive learning of spatiotemporal sequences aims to generate future images by learning from the historical frames, where spatial appearances and temporal variations are two crucial structures. This paper models these structures by presenting a predictive recurrent neural network (PredRNN). This architecture is enlightened by the idea that spatiotemporal predictive learning should memorize both spatial appearances and temporal variations in a unified memory pool. Concretely, memory states are no longer constrained inside each LSTM unit. Instead, they are allowed to zigzag in two directions: across stacked RNN layers vertically and through all RNN states horizontally. The core of this network is a new Spatiotemporal LSTM (ST-LSTM) unit that extracts and memorizes spatial and temporal representations simultaneously. PredRNN achieves the state-of-the-art prediction performance on three video prediction datasets and is a more general framework, that can be easily extended to other predictive learning tasks by integrating with other architectures.

************************************

Dual Discriminator Generative Adversarial Nets

Tu Nguyen, Trung Le, Hung Vu, Dinh Phung

We propose in this paper a novel approach to tackle the problem of mode collapse encountered in generative adversarial network (GAN). Our idea is intuitive but proven to be very effective, especially in addressing some key limitations of GAN. In essence, it combines the Kullback-Leibler (KL) and reverse KL divergences into a unified objective function, thus it exploits the complementary statistical properties from these divergences to effectively diversify the estimated density in capturing multi-modes. We term our method dual discriminator generative adversarial nets (D2GAN) which, unlike GAN, has two discriminators; and together with a generator, it also has the analogy of a minimax game, wherein a discriminator rewards high scores for samples from data distribution whilst another discriminator, conversely, favoring data from the generator, and the generator produces data to fool both two discriminators. We develop theoretical analysis to show that, given the maximal discriminators, optimizing the generator of D2GAN reduces to minimizing both KL and reverse KL divergences between data distribution and the distribution induced from the data generated by the generator, hence effectively avoiding the mode collapsing problem. We conduct extensive experiments on synthetic and real-world large-scale datasets (MNIST, CIFAR-10, STL-10, ImageNet), where we have made our best effort to compare our D2GAN with the latest state-of-the-art GAN's variants in comprehensive qualitative and quantitative evaluations. The experimental results demonstrate the competitive and superior performance of our approach in generating good quality and diverse samples over baselines, and the capability of our method to scale up to ImageNet database.

************************************

Beyond Parity: Fairness Objectives for Collaborative Filtering

Sirui Yao, Bert Huang

We study fairness in collaborative-filtering recommender systems, which are sensitive to discrimination that exists in historical data. Biased data can lead collaborative-filtering methods to make unfair predictions for users from minority groups. We identify the insufficiency of existing fairness metrics and propose four new metrics that address different forms of unfairness. These fairness metrics can be optimized by adding fairness terms to the learning objective. Experiments on synthetic and real data show that our new metrics can better measure fairness than the baseline, and that the fairness objectives effectively help reduce unfairness.

************************************

Multitask Spectral Learning of Weighted Automata

Guillaume Rabusseau, Borja Balle, Joelle Pineau

We consider the problem of estimating multiple related functions computed by weighted  automata~(WFA). We first present a natural notion of relatedness between WFAs by considering to which extent several WFAs can share a common underlying representation. We then introduce the model of vector-valued WFA which conveniently helps us formalize this notion of relatedness. Finally, we propose a spectral learning algorithm for vector-valued WFAs to tackle the multitask learning problem. By jointly learning multiple tasks in the form of a vector-valued WFA, our algorithm enforces the discovery of a representation space shared between tasks. The benefits of the proposed multitask approach are theoretically motivated and showcased through experiments on both synthetic and real world datasets.

************************************

A simple neural network module for relational reasoning

Adam Santoro, David Raposo, David G. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, Timothy Lillicrap

Relational reasoning is a central component of generally intelligent behavior, but has proven difficult for neural networks to learn. In this paper we describe how to use Relation Networks (RNs) as a simple plug-and-play module to solve problems that fundamentally hinge on relational reasoning. We tested RN-augmented networks on three tasks: visual question answering using a challenging dataset called CLEVR, on which we achieve state-of-the-art, super-human performance; text-based question answering using the bAbI suite of tasks; and complex reasoning about dynamical physical systems. Then, using a curated dataset called Sort-of-CLEVR we show that powerful convolutional networks do not have a general capacity t o solve relational questions, but can gain this capacity when augmented with RNs. Thus, by simply augmenting convolutions, LSTMs, and MLPs with RNs, we can remove computational burden from network components that are not well-suited to handle relational reasoning, reduce overall network complexity, and gain a general ability to reason about the relations between entities and their properties.

************************************

Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks

Arash Vahdat

Collecting large training datasets, annotated with high-quality labels, is costly and time-consuming. This paper proposes a novel framework for training deep convolutional neural networks from noisy labeled datasets that can be obtained cheaply. The problem is formulated using an undirected graphical model that represents the relationship between noisy and clean labels, trained in a semi-supervised setting. In our formulation, the inference over latent clean labels is tractable and is regularized during training using auxiliary sources of information. The proposed model is applied to the image labeling problem and is shown to be effective in labeling unseen images as well as reducing label noise in training on CIFAR-10 and MS COCO datasets.

************************************

Stochastic Mirror Descent in Variationally Coherent Optimization Problems

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, Peter W. Glynn

In this paper, we examine a class of non-convex stochastic optimization problems which we call variationally coherent, and which properly includes pseudo-/quasiconvex and star-convex optimization problems. To solve such problems, we focus on the widely used stochastic mirror descent (SMD) family of algorithms (which contains stochastic gradient descent as a special case), and we show that the last iterate of SMD converges to the problem’s solution set with probability 1. This result contributes to the landscape of non-convex stochastic optimization by clarifying that neither pseudo-/quasi-convexity nor star-convexity is essential for (almost sure) global convergence; rather, variational coherence, a much weaker requirement, suffices. Characterization of convergence rates for the subclass of strongly variationally coherent optimization problems as well as simulation results are also presented.

************************************

Polynomial Codes: an Optimal Design for High-Dimensional Coded Matrix Multiplication

Qian Yu, Mohammad Maddah-Ali, Salman Avestimehr

We consider a large-scale matrix multiplication problem where the computation is carried out using a distributed system with a master node and multiple worker nodes, where each worker can store parts of the input matrices. We propose a computation strategy that leverages ideas from coding theory to design intermediate computations at the worker nodes, in order to optimally deal with straggling workers. The proposed strategy, named as \emph{polynomial codes}, achieves the optimum recovery threshold, defined as the minimum number of workers that the master needs to wait for in order to compute the output. This is the first code that achieves the optimal utilization of redundancy for tolerating stragglers or failures in distributed matrix multiplication. Furthermore, by leveraging the algebraic structure of polynomial codes, we can map the reconstruction problem of the final output to a polynomial interpolation problem, which can be solved efficiently.   Polynomial codes provide order-wise improvement over the state of the art in terms of recovery threshold, and are also optimal in terms of several other metrics including computation latency and communication load.  Moreover, we extend this code to distributed convolution and show its order-wise optimality.

************************************

From Bayesian Sparsity to Gated Recurrent Nets

Hao He, Bo Xin, Satoshi Ikehata, David Wipf

The iterations of many first-order algorithms, when applied to minimizing common regularized regression functions, often resemble neural network layers with pre

-specified weights.  This observation has prompted the development of learning-based approaches that purport to replace these iterations with enhanced surrogates forged as DNN models from available training data.  For example, important NP-hard sparse estimation problems have recently benefitted from this genre of upgrade, with simple feedforward or recurrent networks ousting proximal gradient-based iterations.  Analogously, this paper demonstrates that more powerful Bayesian algorithms for promoting sparsity, which rely on complex multi-loop majorization-minimization techniques, mirror the structure of more sophisticated long short-term memory (LSTM) networks, or alternative gated feedback networks previously designed for sequence prediction.  As part of this development, we examine the parallels between latent variable trajectories operating across multiple time-scales during optimization, and the activations within deep network structures designed to adaptively model such characteristic sequences.  The resulting insights lead to a novel sparse estimation system that, when granted training data, can estimate optimal solutions efficiently in regimes where other algorithms fail, including practical direction-of-arrival (DOA) and 3D geometry recovery problems.  The underlying principles we expose are also suggestive of a learning process for a richer class of multi-loop algorithms in other domains.

************************************

Compatible Reward Inverse Reinforcement Learning

Alberto Maria Metelli, Matteo Pirotta, Marcello Restelli

Inverse Reinforcement Learning (IRL) is an effective approach to recover a reward function that explains the behavior of an expert by observing a set of demonstrations.  This paper is about a novel model-free IRL approach that, differently from most of the existing IRL algorithms, does not require to specify a function space where to search for the expert's reward function. Leveraging on the fact that the policy gradient needs to be zero for any optimal policy, the algorithm generates a set of basis functions that span the subspace of reward functions that make the policy gradient vanish. Within this subspace, using a second-order criterion, we search for the reward function that penalizes the most a deviation from the expert's policy. After introducing our approach for finite domains, we extend it to continuous ones. The proposed approach is empirically compared to other IRL methods both in the (finite) Taxi domain and in the (continuous) Linear Quadratic Gaussian (LQG) and Car on the Hill environments.

************************************

Consistent Robust Regression

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, Purushottam Kar

We present the first efficient and provably consistent estimator for the robust regression problem. The area of robust learning and optimization has generated a significant amount of interest in the learning and statistics communities in recent years owing to its applicability in scenarios with corrupted data, as well as in handling model mis-specifications. In particular, special interest has been devoted to the fundamental problem of robust linear regression where estimators that can tolerate corruption in up to a constant fraction of the response variables are widely studied. Surprisingly however, to this date, we are not aware of a polynomial time estimator that offers a consistent estimate in the presence of dense, unbounded corruptions. In this work we present such an estimator, called CRR. This solves an open problem put forward in the work of (Bhatia et al, 2015). Our consistency analysis requires a novel two-stage proof technique involving a careful analysis of the stability of ordered lists which may be of independent interest. We show that CRR not only offers consistent estimates, but is empirically far superior to several other recently proposed algorithms for the robust regression problem, including extended Lasso and the TORRENT algorithm. In comparison, CRR offers comparable or better model recovery but with runtimes that are faster by an order of magnitude.

************************************

Scalable Variational Inference for Dynamical Systems

Nico S. Gorbach, Stefan Bauer, Joachim M. Buhmann

Gradient matching is a promising tool for learning parameters and state dynamics of ordinary differential equations. It is a grid free inference approach, which

, for fully observable systems is at times competitive with numerical integration. However, for many real-world applications, only sparse observations are available or even unobserved variables are included in the model description. In these cases most gradient matching methods are difficult to apply or simply do not provide satisfactory results. That is why, despite the high computational cost, numerical integration is still the gold standard in many applications. Using an existing gradient matching approach, we propose a scalable variational inference framework which can infer states and parameters simultaneously, offers computational speedups, improved accuracy and works well even under model misspecifications in a partially observable system.

************************************

### Learning multiple visual domains with residual adapters

Sylvestre-Alvise Rebuffi, Hakan Bilen, Andrea Vedaldi

There is a growing interest in learning data representations that work well for many different types of problems and data. In this paper, we look in particular at the task of learning a single visual representation that can be successfully utilized in the analysis of very different types of images, from dog breeds to stop signs and digits. Inspired by recent work on learning networks that predict the parameters of another, we develop a tunable deep network architecture that, by means of adapter residual modules, can be steered on the fly to diverse visual domains. Our method achieves a high degree of parameter sharing while maintaining or even improving the accuracy of domain-specific representations. We also introduce the Visual Decathlon Challenge, a benchmark that evaluates the ability of  representations to capture simultaneously ten very different visual domains and measures their ability to recognize well uniformly.

************************************

### Incorporating Side Information by Adaptive Convolution

Di Kang, Debarun Dhar, Antoni Chan

Computer vision tasks often have side information available that is helpful to solve the task. For example, for crowd counting, the camera perspective (e.g., camera angle and height) gives a clue about the appearance and scale of people in the scene. While side information has been shown to be useful for counting systems using traditional hand-crafted features, it has not been fully utilized in counting systems based on deep learning. In order to incorporate the available side information, we propose an adaptive convolutional neural network (ACNN), where the convolution filter weights adapt to the current scene context via the side information. In particular, we model the filter weights as a low-dimensional manifold within the high-dimensional space of filter weights. The filter weights are generated using a learned ``filter manifold'' sub-network, whose input is the side information. With the help of side information and adaptive weights, the ACNN can disentangle the variations related to the side information, and extract discriminative features related to the current context (e.g. camera perspective, noise level, blur kernel parameters). We demonstrate the effectiveness of  ACNN incorporating side information on 3 tasks: crowd counting, corrupted digit recognition, and image deblurring. Our experiments show that ACNN improves the performance compared to a plain CNN with a similar number of parameters. Since existing crowd counting datasets do not contain ground-truth side information, we collect a new dataset with the ground-truth camera angle and height as the side information.

************************************

### Hierarchical Clustering Beyond the Worst-Case

Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn

Hiererachical clustering, that is computing a recursive partitioning of a dataset to obtain clusters at increasingly finer granularity is a fundamental problem in data analysis. Although hierarchical clustering has mostly been studied through procedures such as linkage algorithms, or top-down heuristics, rather than as optimization problems, recently Dasgupta [1] proposed an objective function for hierarchical clustering and initiated a line of work developing algorithms that explicitly optimize an objective (see also [2, 3, 4]). In this paper, we consider a fairly general random graph model for hierarchical clustering, called the h ierarchical stochastic blockmodel (HSBM), and show that in certain regimes the SVD approach of McSherry [5] combined with specific linkage methods results in a clustering that give an O(1)-approximation to Dasgupta’s cost function. We also show that an approach based on SDP relaxations for balanced cuts based on the work of Makarychev et al. [6], combined with the recursive sparsest cut algorithm of Dasgupta, yields an O(1) approximation in slightly larger regimes and also in the semi-random setting, where an adversary may remove edges from the random graph generated according to an HSBM. Finally, we report empirical evaluation on synthetic and real-world data showing that our proposed SVD-based method does indeed achieve a better cost than other widely-used heurstics and also results in a better classification accuracy when the underlying problem was that of multi-class classification.

************************************

Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference

Geoffrey Roeder, Yuhuai Wu, David K. Duvenaud

We propose a simple and general variant of the standard reparameterized gradient estimator for the variational evidence lower bound. Specifically, we remove a part of the total derivative with respect to the variational parameters that corresponds to the score function. Removing this term produces an unbiased gradient estimator whose variance approaches zero as the approximate posterior approaches the exact posterior. We analyze the behavior of this gradient estimator theoretically and empirically, and generalize it to more complex variational distributions such as mixtures and importance-weighted posteriors.

************************************

Multiplicative Weights Update with Constant Step-Size in Congestion Games: Convergence, Limit Cycles and Chaos

Gerasimos Palaiopanos, Ioannis Panageas, Georgios Piliouras

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.

************************************

QMDP-Net: Deep Learning for Planning under Partial Observability

Peter Karkus, David Hsu, Wee Sun Lee

This paper introduces the QMDP-net, a neural network architecture for planning under partial observability. The QMDP-net combines the strengths of model-free learning and model-based planning. It is a recurrent policy network, but it represents a policy for a parameterized set of tasks by connecting a model with a planning algorithm that solves the model, thus embedding the solution structure of planning in a network learning architecture. The QMDP-net is fully differentiable and allows for end-to-end training. We train a QMDP-net on different tasks so that it can generalize to new ones in the parameterized task set and “transfer” to other similar tasks beyond the set. In preliminary experiments, QMDP-net showed strong performance on several robotic tasks in simulation. Interestingly, while QMDP-net encodes the QMDP algorithm, it sometimes outperforms the QMDP algorithm in the experiments, as a result of end-to-end learning.

************************************

Deep Supervised Discrete Hashing

Qi Li, Zhenan Sun, Ran He, Tieniu Tan

With the rapid growth of image and video data on the web, hashing has been extensively studied for image or video search in recent years. Benefiting from recent advances in deep learning, deep hashing methods have achieved promising results for image retrieval. However, there are some limitations of previous deep hashing methods (e.g., the semantic information is not fully exploited). In this paper, we develop a deep supervised discrete hashing algorithm based on the assumption that the learned binary codes should be ideal for classification. Both the pairwise label information and the classification information are used to learn the hash codes within one stream framework. We constrain the outputs of the last layer to be binary codes directly, which is rarely investigated in deep hashing a lgorithm. Because of the discrete nature of hash codes, an alternating minimization method is used to optimize the objective function. Experimental results have shown that our method outperforms current state-of-the-art methods on benchmark datasets.

************************************

Approximation Algorithms for $\ell_0$-Low Rank Approximation

Karl Bringmann, Pavel Kolev, David Woodruff

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.

************************************

ADMM without a Fixed Penalty Parameter: Faster Convergence with New Adaptive Penalization

Yi Xu, Mingrui Liu, Qihang Lin, Tianbao Yang

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.

************************************

A Learning Error Analysis for Structured Prediction with Approximate Inference

Yuanbin Wu, Man Lan, Shiliang Sun, Qi Zhang, Xuanjing Huang

In this work, we try to understand the differences between exact and approximate inference algorithms in structured prediction. We compare the estimation and approximation error of both underestimate and overestimate models. The result shows that, from the perspective of learning errors, performances of approximate inference could be as good as exact inference. The error analyses also suggest a new margin for existing learning algorithms. Empirical evaluations on text classification, sequential labelling and dependency parsing witness the success of approximate inference and the benefit of the proposed margin.

************************************

Simple strategies for recovering inner products from coarsely quantized random projections

Ping Li, Martin Slawski

Random projections have been increasingly adopted for a diverse set of tasks in machine learning involving dimensionality reduction. One specific line of research on this topic has investigated the use of quantization subsequent to projection with the aim of additional data compression. Motivated by applications in nearest neighbor search and linear learning, we revisit the problem of recovering inner products (respectively cosine similarities) in such setting. We show that even under coarse scalar quantization with 3 to 5 bits per projection, the loss in accuracy tends to range from negligible'' tomoderate''. One implication is that in most scenarios of practical interest, there is no need for a sophisticated recovery approach like maximum likelihood estimation as considered in previous work on the subject. What we propose herein also yields considerable improvements in terms of accuracy over the Hamming distance-based approach in Li et al. (ICML 2014) which is comparable in terms of simplicity

************************************

Trimmed Density Ratio Estimation

Song Liu, Akiko Takeda, Taiji Suzuki, Kenji Fukumizu

Density ratio estimation is a vital tool in both machine learning and statistical community. However, due to the unbounded nature of density ratio, the estimation proceudre can be vulnerable to corrupted data points, which often pushes the estimated ratio toward infinity. In this paper, we present a robust estimator which automatically identifies and trims outliers. The proposed estimator has a convex formulation, and the global optimum can be obtained via subgradient descent. We analyze the parameter estimation error of this estimator under high-dimensional settings. Experiments are conducted to verify the effectiveness of the estimator.

************************************

Adaptive Batch Size for Safe Policy Gradients

Matteo Papini, Matteo Pirotta, Marcello Restelli

Policy gradient methods are among the best Reinforcement Learning (RL) techniques to solve complex control problems. In real-world RL applications, it is common to have a good initial policy whose performance needs to be improved and it may not be acceptable to try bad policies during the learning process. Although several methods for choosing the step size exist, research paid less attention to determine the batch size, that is the number of samples used to estimate the gradient direction for each update of the policy parameters. In this paper, we propose a set of methods to jointly optimize the step and the batch sizes that guarantee (with high probability) to improve the policy performance after each update. Besides providing theoretical guarantees, we show numerical simulations to analyse the behaviour of our methods.

************************************

Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting

Rebecca Morrison, Ricardo Baptista, Youssef Marzouk

We present an algorithm to identify sparse dependence structure in continuous and non-Gaussian probability distributions, given a corresponding set of data. The conditional independence structure of an arbitrary distribution can be represented as an undirected graph (or Markov random field), but most algorithms for learning this structure are restricted to the discrete or Gaussian cases. Our new approach allows for more realistic and accurate descriptions of the distribution in question, and in turn better estimates of its sparse Markov structure. Sparsity in the graph is of interest as it can accelerate inference, improve sampling methods, and reveal important dependencies between variables. The algorithm relies on exploiting the connection between the sparsity of the graph and the sparsity of transport maps, which deterministically couple one probability measure to another.

************************************

REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models

George Tucker, Andriy Mnih, Chris J. Maddison, John Lawson, Jascha Sohl-Dickstein

Learning in models with discrete latent variables is challenging due to high variance gradient estimators. Generally, approaches have relied on control variates to reduce the variance of the REINFORCE estimator. Recent work \citep{jang2016categorical, maddison2016concrete} has taken a different approach, introducing a continuous relaxation of discrete variables to produce low-variance, but biased, gradient estimates. In this work, we combine the two approaches through a novel control variate that produces low-variance, \emph{unbiased} gradient estimates. Then, we introduce a modification to the continuous relaxation and show that the tightness of the relaxation can be adapted online, removing it as a hyperparameter. We show state-of-the-art variance reduction on several benchmark generative modeling tasks, generally leading to faster convergence to a better final log-likelihood.

************************************

Submultiplicative Glivenko-Cantelli and Uniform Convergence of Revenues

Noga Alon, Moshe Babaioff, Yannai A. Gonczarowski, Yishay Mansour, Shay Moran, Amir Yehudayoff

In this work we derive a variant of the classic Glivenko-Cantelli Theorem, which asserts uniform convergence of the empirical Cumulative Distribution Function (CDF) to the CDF of the underlying distribution. Our variant allows for tighter convergence bounds for extreme values of the CDF.  We apply our bound in the context of revenue learning, which is a well-studied problem in economics and algorithmic game theory. We derive sample-complexity bounds on the uniform convergence rate of the empirical revenues to the true revenues, assuming a bound on the k'th moment of the valuations, for any (possibly fractional) k > 1.  For uniform convergence in the limit, we give a complete characterization and a zero-one law: if the first moment of the valuations is finite, then uniform convergence almos t surely occurs; conversely, if the first moment is infinite, then uniform convergence almost never occurs.

************************************

Tensor Biclustering

Soheil Feizi, Hamid Javadi, David Tse

Consider a dataset where data is collected on multiple features of multiple individuals over multiple times. This type of data can be represented as a three dimensional individual/feature/time tensor and has become increasingly prominent in various areas of science. The tensor biclustering problem computes a subset of individuals and a subset of features whose signal trajectories over time lie in a low-dimensional subspace, modeling similarity among the signal trajectories while allowing different scalings across different individuals or different features. We study the information-theoretic limit of this problem under a generative model. Moreover, we propose an efficient spectral algorithm to solve the tensor biclustering problem and analyze its achievability bound in an asymptotic regime. Finally, we show the efficiency of our proposed method in several synthetic and real datasets.

************************************

On the Model Shrinkage Effect of Gamma Process Edge Partition Models

Iku Ohama, Issei Sato, Takuya Kida, Hiroki Arimura

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.

************************************

Estimating Mutual Information for Discrete-Continuous Mixtures

Weihao Gao, Sreeram Kannan, Sewoong Oh, Pramod Viswanath

Estimation of mutual information from observed samples is a basic primitive in machine learning, useful in several learning tasks including correlation mining, information bottleneck, Chow-Liu tree, and conditional independence testing in (causal) graphical models. While mutual information is a quantity well-defined for general probability spaces, estimators have been developed only in the special case of discrete or continuous pairs of random variables. Most of these estimators operate using the 3H -principle, i.e., by calculating the three (differential) entropies of X, Y and the pair (X,Y). However, in general mixture spaces, such individual entropies are not well defined, even though mutual information is. In this paper, we develop a novel estimator for estimating mutual information in discrete-continuous mixtures. We prove the consistency of this estimator theoretically as well as demonstrate its excellent empirical performance. This problem is relevant in a wide-array of applications, where some variables are discrete, some continuous, and others are a mixture between continuous and discrete components.

************************************

Reconstructing perceived faces from brain activations with deep adversarial neural decoding

Ya■mur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, Marcel A. J. van Gerven

Here, we present a novel approach to solve the problem of reconstructing perceived stimuli from brain responses by combining probabilistic inference with deep learning. Our approach first inverts the linear transformation from latent features to brain responses with maximum a posteriori estimation and then inverts the nonlinear transformation from perceived stimuli to latent features with adversarial training of convolutional neural networks. We test our approach with a functional magnetic resonance imaging experiment and show that it can generate state-of-the-art reconstructions of perceived faces from brain activations.

************************************

An inner-loop free solution to inverse problems using deep neural networks

Kai Fan, Qi Wei, Lawrence Carin, Katherine A. Heller

We propose a new method that uses deep learning techniques to accelerate the popular alternating direction method of multipliers (ADMM) solution for inverse pro blems. The ADMM updates consist of a proximity operator, a least squares regression that includes a big matrix inversion, and an explicit solution for updating the dual variables. Typically, inner loops are required to solve the first two sub-minimization problems due to the intractability of the prior and the matrix inversion.  To avoid such drawbacks or limitations, we propose an inner-loop free update rule with two pre-trained deep convolutional architectures. More specifically, we learn a conditional denoising auto-encoder which imposes an implicit data-dependent prior/regularization on ground-truth in the first sub-minimization problem. This design follows an empirical Bayesian strategy, leading to so-called amortized inference. For matrix inversion in the second sub-problem, we learn a convolutional neural network to approximate the matrix inversion, i.e., the inverse mapping is learned by feeding the input through the learned forward network. Note that training this neural network does not require ground-truth or measurements, i.e., data-independent. Extensive experiments on both synthetic data and real datasets demonstrate the efficiency and accuracy of the proposed method compared with the conventional ADMM solution using inner loops for solving inverse problems.

************************************

A framework for Multi-A(rmed)/B(andit) Testing with Online FDR Control
Fanny Yang, Aaditya Ramdas, Kevin G. Jamieson, Martin J. Wainwright

We propose an alternative framework to existing setups for controlling false alarms when multiple A/B tests are run over time. This setup arises in many practical applications, e.g. when pharmaceutical companies test new treatment options against control pills for different diseases, or when internet companies test their default webpages versus various alternatives over time. Our framework proposes to replace a sequence of A/B tests by a sequence of best-arm MAB instances, which can be continuously monitored by the data scientist. When interleaving the MAB tests with an online false discovery rate (FDR) algorithm, we can obtain the best of both worlds: low sample complexity and any time online FDR control. Our main contributions are: (i) to propose reasonable definitions of a null hypothesis for MAB instances; (ii) to demonstrate how one can derive an always-valid sequential p-value that allows continuous monitoring of each MAB test; and (iii) to show that using rejection thresholds of online-FDR algorithms as the confidence levels for the MAB algorithms results in both sample-optimality, high power and low FDR at any point in time. We run extensive simulations to verify our claims, and also report results on real data collected from the New Yorker Cartoon Caption contest.

************************************

Interactive Submodular Bandit
Lin Chen, Andreas Krause, Amin Karbasi

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.

************************************

Hash Embeddings for Efficient Word Representations
Dan Tito Svenstrup, Jonas Hansen, Ole Winther

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.

************************************

Learning Low-Dimensional Metrics
Blake Mason, Lalit Jain, Robert Nowak

This paper investigates the theoretical foundations of metric learning, focused on three key questions that are not fully addressed in prior work:  1) we consider learning general low-dimensional (low-rank) metrics as well as sparse metrics;2) we develop upper and lower (minimax) bounds on the generalization error; 3)we quantify the sample complexity of metric learning in terms of the dimension of the feature space and the dimension/rank of the underlying metric; 4) we also b ound the accuracy of the learned metric relative to the underlying true generative metric. All the results involve novel mathematical approaches to the metric learning problem, and also shed new light on the special case of ordinal embedding (aka non-metric multidimensional scaling).

************************************

Unsupervised Sequence Classification using Sequential Output Statistics
Yu Liu, Jianshu Chen, Li Deng

We consider learning a sequence classifier without labeled data by using sequential output statistics. The problem is highly valuable since obtaining labels in training data is often costly, while the sequential output statistics (e.g., language models) could be obtained independently of input data and thus with low or no cost. To address the problem, we propose an unsupervised learning cost function and study its properties. We show that, compared to earlier works, it is less inclined to be stuck in trivial solutions and avoids the need for a strong generative model. Although it is harder to optimize in its functional form, a stochastic primal-dual gradient method is developed to effectively solve the problem. Experiment results on real-world datasets demonstrate that the new unsupervised learning method gives drastically lower errors than other baseline methods. Specifically, it reaches test errors about twice of those obtained by fully supervised learning.

************************************

Deep Sets
Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R. Salakhutdinov, Alexander J. Smola

We study the problem of designing models for machine learning tasks defined on sets. In contrast to the traditional approach of operating on fixed dimensional vectors, we consider objective functions defined on sets and are invariant to permutations. Such problems are widespread, ranging from the estimation of population statistics, to anomaly detection in piezometer data of embankment dams, to cosmology. Our main theorem characterizes the permutation invariant objective functions and provides a family of functions to which any permutation invariant objective function must belong. This family of functions has a special structure which enables us to design a deep network architecture that can operate on sets and which can be deployed on a variety of scenarios including both unsupervised and supervised learning tasks. We demonstrate the applicability of our method on population statistic estimation, point cloud classification, set expansion, and outlier detection.

************************************

Optimal Shrinkage of Singular Values Under Random Data Contamination
Danny Barash, Matan Gavish

A low rank matrix X has been contaminated by uniformly distributed noise, missing values, outliers and corrupt entries. Reconstruction of X from the singular values and singular vectors of the contaminated matrix Y is a key problem in machine learning, computer vision and data science. In this paper we show that common contamination models (including arbitrary combinations of uniform noise, missing values, outliers and corrupt entries) can be described efficiently using a single framework. We develop an asymptotically optimal algorithm that estimates X by manipulation of the singular values of Y, which applies to any of the contamination models considered. Finally, we find an explicit signal-to-noise cutoff, below which estimation of X from the singular value decomposition of Y must fail, in a well-defined sense.

************************************

Learning Mixture of Gaussians with Streaming Data
Aditi Raghunathan, Prateek Jain, Ravishankar Krishnawamy

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.

************************************

Learning to Compose Domain-Specific Transformations for Data Augmentation

Alexander J. Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, Christopher Ré

Data augmentation is a ubiquitous technique for increasing the size of labeled training sets by leveraging task-specific data transformations that preserve class labels. While it is often easy for domain experts to specify individual transformations, constructing and tuning the more sophisticated compositions typically needed to achieve state-of-the-art results is a time-consuming manual task in practice. We propose a method for automating this process by learning a generative sequence model over user-specified transformation functions using a generative adversarial approach. Our method can make use of arbitrary, non-deterministic transformation functions, is robust to misspecified user input, and is trained on unlabeled data. The learned transformation model can then be used to perform data augmentation for any end discriminative model. In our experiments, we show the efficacy of our approach on both image and text datasets, achieving improvements of 4.0 accuracy points on CIFAR-10, 1.4 F1 points on the ACE relation extraction task, and 3.4 accuracy points when using domain-specific transformation operations on a medical imaging dataset as compared to standard heuristic augmentation approaches.

************************************

Preventing Gradient Explosions in Gated Recurrent Units

Sekitoshi Kanai, Yasuhiro Fujiwara, Sotetsu Iwamura

A gated recurrent unit (GRU) is a successful recurrent neural network architecture for time-series data. The GRU is typically trained using a gradient-based method, which is subject to the exploding gradient problem in which the gradient increases significantly. This problem is caused by an abrupt change in the dynamics of the GRU due to a small variation in the parameters. In this paper, we find a condition under which the dynamics of the GRU changes drastically and propose a learning method to address the exploding gradient problem. Our method constrains the dynamics of the GRU so that it does not drastically change. We evaluated our method in experiments on language modeling and polyphonic music modeling. Our experiments showed that our method can prevent the exploding gradient problem and improve modeling accuracy.

************************************

Streaming Sparse Gaussian Process Approximations

Thang D. Bui, Cuong Nguyen, Richard E. Turner

Sparse pseudo-point approximations for Gaussian process (GP) models provide a suite of methods that support deployment of GPs in the large data regime and enable analytic intractabilities to be sidestepped. However, the field lacks a principled method to handle streaming data in which both the posterior distribution over function values and the hyperparameter estimates are updated in an online fashion. The small number of existing approaches either use suboptimal hand-crafted heuristics for hyperparameter learning, or suffer from catastrophic forgetting or slow updating when new data arrive. This paper develops a new principled framework for deploying Gaussian process probabilistic models in the streaming setting, providing  methods for learning hyperparameters and optimising pseudo-input locations. The proposed framework is assessed using synthetic and real-world datasets.

************************************

Differentially Private Empirical Risk Minimization Revisited: Faster and More General

Di Wang, Minwei Ye, Jinhui Xu

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.

************************************

Unbounded cache model for online language modeling with open vocabulary

Edouard Grave, Moustapha M. Cisse, Armand Joulin

Recently, continuous cache models were proposed as extensions to recurrent neural network language models, to adapt their predictions to local changes in the da ta distribution. These models only capture the local context, of up to a few thousands tokens. In this paper, we propose an extension of continuous cache models, which can scale to larger contexts. In particular, we use a large scale non-parametric memory component that stores all the hidden activations seen in the past. We leverage recent advances in approximate nearest neighbor search and quantization algorithms to store millions of representations while searching them efficiently. We conduct extensive experiments showing that our approach significantly improves the perplexity of pre-trained language models on new distributions, and can scale efficiently to much larger contexts than previously proposed local cache models.

************************************

Shape and Material from Sound

Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Josh Tenenbaum, Bill Freeman

Hearing an object falling onto the ground, humans can recover rich information including its rough shape, material, and falling height. In this paper, we build machines to approximate such competency. We first mimic human knowledge of the physical world by building an efficient, physics-based simulation engine. Then, we present an analysis-by-synthesis approach to infer properties of the falling object. We further accelerate the process by learning a mapping from a sound wave to object properties, and using the predicted values to initialize the inference. This mapping can be viewed as an approximation of human commonsense learned from past experience. Our model performs well on both synthetic audio clips and real recordings without requiring any annotated data. We conduct behavior studies to compare human responses with ours on estimating object shape, material, and falling height from sound. Our model achieves near-human performance.

************************************

On the Consistency of Quick Shift

Heinrich Jiang

Quick Shift is a popular mode-seeking and clustering algorithm. We present finite sample statistical consistency guarantees for Quick Shift on mode and cluster recovery under mild distributional assumptions. We then apply our results to construct a consistent modal regression algorithm.

************************************

Wasserstein Learning of Deep Generative Point Process Models

Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Le Song, Hongyuan Zha

Point processes are becoming very popular in modeling asynchronous sequential data due to their sound mathematical foundation and strength in modeling a variety of real-world phenomena. Currently, they are often characterized via intensity function which limits model's expressiveness due to unrealistic assumptions on its parametric form used in practice. Furthermore, they are learned via maximum likelihood approach which is prone to failure in multi-modal distributions of sequences. In this paper, we propose an intensity-free approach for point processes modeling that transforms nuisance processes to a target one. Furthermore, we train the model using a likelihood-free leveraging Wasserstein distance between point processes. Experiments on various synthetic and real-world data substantiate the superiority of the proposed point process model over conventional ones.

************************************

Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, Julien Stainer

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.

************************************

Protein Interface Prediction using Graph Convolutional Networks

Alex Fout, Jonathon Byrd, Basir Shariat, Asa Ben-Hur

We consider the prediction of interfaces between proteins, a challenging problem with important applications in drug discovery and design, and examine the performance of existing and newly proposed spatial graph convolution operators for th is task. By performing convolution over a local neighborhood of a node of interest, we are able to stack multiple layers of convolution and learn effective latent representations that integrate information across the graph that represent the three dimensional structure of a protein of interest. An architecture that combines the learned features across pairs of proteins is then used to classify pairs of amino acid residues as part of an interface or not. In our experiments, several graph convolution operators yielded accuracy that is better than the state-of-the-art SVM method in this task.

************************************

Convergence rates of a partition based Bayesian multivariate density estimation method

Linxi Liu, Dangna Li, Wing Hung Wong

We study a class of non-parametric density estimators under Bayesian settings. The estimators are obtained by adaptively partitioning the sample space. Under a suitable prior, we analyze the concentration rate of the posterior distribution, and demonstrate that the rate does not directly depend on the dimension of the problem in several special cases. Another advantage of this class of Bayesian density estimators is that it can adapt to the unknown smoothness of the true density function, thus achieving the optimal convergence rate without artificial conditions on the density. We also validate the theoretical results on a variety of simulated data sets.

************************************

Avoiding Discrimination through Causal Reasoning

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, Bernhard Schölkopf

Recent work on fairness in machine learning has focused on various statistical discrimination criteria and how they trade off. Most of these criteria are observational: They depend only on the joint distribution of predictor, protected attribute, features, and outcome. While convenient to work with, observational criteria have severe inherent limitations that prevent them from resolving matters of fairness conclusively. Going beyond observational criteria, we frame the problem of discrimination based on protected attributes in the language of causal reasoning. This viewpoint shifts attention from "What is the right fairness criterion?" to "What do we want to assume about our model of the causal data generating process?" Through the lens of causality, we make several contributions. First, we crisply articulate why and when observational criteria fail, thus formalizing what was before a matter of opinion. Second, our approach exposes previously ignored subtleties and why they are fundamental to the problem. Finally, we put forward natural causal non-discrimination criteria and develop algorithms that satisfy them.

************************************

Alternating Estimation for Structured High-Dimensional Multi-Response Models

Sheng Chen, Arindam Banerjee

We consider the problem of learning high-dimensional multi-response linear models with structured parameters. By exploiting the noise correlations among different responses, we propose an alternating estimation (AltEst) procedure to estimate the model parameters based on the generalized Dantzig selector (GDS). Under suitable sample size and resampling assumptions, we show that the error of the estimates generated by AltEst, with high probability, converges linearly to certain minimum achievable level, which can be tersely expressed by a few geometric measures, such as Gaussian width of sets related to the parameter structure. To the best of our knowledge, this is the first non-asymptotic statistical guarantee for such AltEst-type algorithm applied to estimation with general structures.

************************************

Multimodal Learning and Reasoning for Visual Question Answering

Ilija Ilievski, Jiashi Feng

Reasoning about entities and their relationships from multimodal data is a key goal of Artificial General Intelligence. The visual question answering (VQA) problem is an excellent way to test such reasoning capabilities of an AI model and its multimodal representation learning. However, the current VQA models are oversimplified deep neural networks, comprised of a long short-term memory (LSTM) unit for question comprehension and a convolutional neural network (CNN) for learning single image representation. We argue that the single visual representation contains a limited and general information about the image contents and thus limits the model reasoning capabilities. In this work we introduce a modular neural network model that learns a multimodal and multifaceted representation of the image and the question. The proposed model learns to use the multimodal representation to reason about the image entities and achieves a new state-of-the-art performance on both VQA benchmark datasets, VQA v1.0 and v2.0, by a wide margin.

************************************

Generative Local Metric Learning for Kernel Regression
Yung-Kyun Noh, Masashi Sugiyama, Kee-Eung Kim, Frank Park, Daniel D. Lee
This paper shows how metric learning can be used with Nadaraya-Watson (NW) kernel regression.  Compared with standard approaches, such as bandwidth selection, we show how metric learning can significantly reduce the mean square error (MSE) in kernel regression, particularly for high-dimensional data.  We propose a method for efficiently learning a good metric function based upon analyzing the performance of the NW estimator for Gaussian-distributed data.  A key feature of our approach is that the NW estimator with a learned metric uses information from both the global and local structure of the training data.  Theoretical and empirical results confirm that the learned metric can considerably reduce the bias and MSE for kernel regression even when the data are not confined to Gaussian.

************************************

Overcoming Catastrophic Forgetting by Incremental Moment Matching
Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, Byoung-Tak Zhang
Catastrophic forgetting is a problem of neural networks that loses the information of the first task after training the second task. Here, we propose a method, i.e. incremental moment matching (IMM), to resolve this problem. IMM incrementally matches the moment of the posterior distribution of the neural network which is trained on the first and the second task, respectively. To make the search space of posterior parameter smooth, the IMM procedure is complemented by various transfer learning techniques including weight transfer, L2-norm of the old and the new parameter, and a variant of dropout with the old parameter. We analyze our approach on a variety of datasets including the MNIST, CIFAR-10, Caltech-UCSD-Birds, and Lifelog datasets. The experimental results show that IMM achieves state-of-the-art performance by balancing the information between an old and a new network.

************************************

Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent
Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, Ji Liu
Most distributed machine learning systems nowadays, including TensorFlow and CNTK, are built in a centralized fashion. One bottleneck of centralized algorithms lies on high communication cost on the central node. Motivated by this, we ask, can decentralized algorithms be faster than its centralized counterpart?  Although decentralized PSGD (D-PSGD) algorithms have been studied by the control community, existing analysis and theory do not show any advantage over centralized PSGD (C-PSGD) algorithms, simply assuming the application scenario where only the decentralized network is available. In this paper, we study a D-PSGD algorithm and provide the first theoretical analysis that indicates a regime in which decentralized algorithms might outperform centralized algorithms for distributed stochastic gradient descent. This is because D-PSGD has comparable total computational complexities to C-PSGD but requires much less communication cost on the busiest node. We further conduct an empirical study to validate our theoretical analysis across multiple frameworks (CNTK and Torch), different network configurations, and computation platforms up to 112 GPUs. On network configurations with low bandwidth or high latency, D-PSGD can be up to one order of magnitude faster than its well-optimized centralized counterparts.

************************************

Gradient Descent Can Take Exponential Time to Escape Saddle Points

Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Aarti Singh, Barnabas Poczos

Although gradient descent (GD) almost always escapes saddle points asymptotically [Lee et al., 2016], this paper shows that even with fairly natural random initialization schemes and non-pathological functions, GD can be significantly slowed down by saddle points, taking exponential time to escape. On the other hand, gradient descent with perturbations [Ge et al., 2015, Jin et al., 2017] is not slowed down by saddle points—it can find an approximate local minimizer in polynomial time. This result implies that GD is inherently slower than perturbed GD, and justifies the importance of adding perturbations for efficient non-convex optimization. While our focus is theoretical, we also present experiments that illustrate our theoretical findings.

************************************

Dual Path Networks

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, Jiashi Feng

In this work, we present a simple, highly efficient and modularized Dual Path Network (DPN) for image classification which presents a new topology of connection paths internally. By revealing the equivalence of the state-of-the-art Residual Network (ResNet) and Densely Convolutional Network (DenseNet) within the HORNN framework, we find that ResNet enables feature re-usage while DenseNet enables new features exploration which are both important for learning good representations. To enjoy the benefits from both path topologies, our proposed Dual Path Network shares common features while maintaining the flexibility to explore new features through dual path architectures. Extensive experiments on three benchmark datasets, ImagNet-1k, Places365 and PASCAL VOC, clearly demonstrate superior performance of the proposed DPN over state-of-the-arts. In particular, on the ImagNet-1k dataset, a shallow DPN surpasses the best ResNeXt-101(64x4d) with 26% smaller model size, 25% less computational cost and 8% lower memory consumption, and a deeper DPN (DPN-131) further pushes the state-of-the-art single model performance with about 2 times faster training speed. Experiments on the Places365 large-scale scene dataset, PASCAL VOC detection dataset, and PASCAL VOC segmentation dataset also demonstrate its consistently better performance than DenseNet, ResNet and the latest ResNeXt model over various applications.

************************************

Model-based Bayesian inference of neural activity and connectivity from all-optical interrogation of a neural circuit

Laurence Aitchison, Lloyd Russell, Adam M. Packer, Jinyao Yan, Philippe Castonguay, Michael Hausser, Srinivas C. Turaga

Population activity measurement by calcium imaging can be combined with cellular resolution optogenetic activity perturbations to enable the mapping of neural connectivity in vivo. This requires accurate inference of perturbed and unperturbed neural activity from calcium imaging measurements, which are noisy and indirect, and can also be contaminated by photostimulation artifacts. We have developed a new fully Bayesian approach to jointly inferring spiking activity and neural connectivity from in vivo all-optical perturbation experiments. In contrast to standard approaches that perform spike inference and analysis in two separate maximum-likelihood phases, our joint model is able to propagate uncertainty in spike inference to the inference of connectivity and vice versa. We use the framework of variational autoencoders to model spiking activity using discrete latent variables, low-dimensional latent common input, and sparse spike-and-slab generalized linear coupling between neurons. Additionally, we model two properties of the optogenetic perturbation: off-target photostimulation and photostimulation transients. Using this model, we were able to fit models on 30 minutes of data in just 10 minutes. We performed an all-optical circuit mapping experiment in primary visual cortex of the awake mouse, and use our approach to predict neural connectivity between excitatory neurons in layer 2/3. Predicted connectivity is sparse and consistent with known correlations with stimulus tuning, spontaneous correlation and distance.

************************************

Universal consistency and minimax rates for online Mondrian Forests

Jaouad Mourtada, Stéphane Gaïffas, Erwan Scornet

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.

************************************

Gradient Episodic Memory for Continual Learning

David Lopez-Paz, Marc'Aurelio Ranzato

One major obstacle towards AI is the poor ability of models to solve new problems quicker, and without forgetting previously acquired knowledge. To better understand this issue, we study the problem of continual learning, where the model observes, once and one by one, examples concerning a sequence of tasks. First, we propose a set of metrics to evaluate models learning over a continuum of data. These metrics characterize models not only by their test accuracy, but also in terms of their ability to transfer knowledge across tasks. Second, we propose a model for continual learning, called Gradient Episodic Memory (GEM) that alleviates forgetting, while allowing beneficial transfer of knowledge to previous tasks. Our experiments on variants of the MNIST and CIFAR-100 datasets demonstrate the strong performance of GEM when compared to the state-of-the-art.

************************************

Variational Inference for Gaussian Process Models with Linear Complexity

Ching-An Cheng, Byron Boots

Large-scale Gaussian process inference has long faced practical challenges due to time and space complexity that is superlinear in dataset size. While sparse variational Gaussian process models are capable of learning from large-scale data, standard strategies for sparsifying the model can prevent the approximation of complex functions. In this work, we propose a novel variational Gaussian process model that decouples the representation of mean and covariance functions in reproducing kernel Hilbert space. We show that this new parametrization generalizes previous models. Furthermore, it yields a variational inference problem that can be solved by stochastic gradient ascent with time and space complexity that is only linear in the number of mean function parameters, regardless of the choice of kernels, likelihoods, and inducing points. This strategy makes the adoption of large-scale expressive Gaussian process models possible. We run several experiments on regression tasks and show that this decoupled approach greatly outperforms previous sparse variational Gaussian process inference procedures.

************************************

The Reversible Residual Network: Backpropagation Without Storing Activations

Aidan N. Gomez, Mengye Ren, Raquel Urtasun, Roger B. Grosse

Residual Networks (ResNets) have demonstrated significant improvement over traditional Convolutional Neural Networks (CNNs) on image classification, increasing in performance as networks grow both deeper and wider.  However, memory consumption becomes a bottleneck as one needs to store all the intermediate activations for calculating gradients using backpropagation. In this work, we present the Reversible Residual Network (RevNet), a variant of ResNets where each layer's activations can be reconstructed exactly from the next layer's. Therefore, the activations for most layers need not be stored in memory during backprop. We demonstrate the effectiveness of RevNets on CIFAR and ImageNet, establishing nearly identical performance to equally-sized ResNets, with activation storage requirements independent of depth.

************************************

Language Modeling with Recurrent Highway Hypernetworks

Joseph Suarez

We present extensive experimental and theoretical support for the efficacy of recurrent highway networks (RHNs) and recurrent hypernetworks complimentary to the original works. Where the original RHN work primarily provides theoretical treatment of the subject, we demonstrate experimentally that RHNs benefit from far better gradient flow than LSTMs in addition to their improved task accuracy. The original hypernetworks work presents detailed experimental results but leaves several theoretical issues unresolved--we consider these in depth and frame severa l feasible solutions that we believe will yield further gains in the future. We demonstrate that these approaches are complementary: by combining RHNs and hypernetworks, we make a significant improvement over current state-of-the-art character-level language modeling performance on Penn Treebank while relying on much simpler regularization. Finally, we argue for RHNs as a drop-in replacement for LSTMs (analogous to LSTMs for vanilla RNNs) and for hypernetworks as a de-facto augmentation (analogous to attention) for recurrent architectures.

************************************

Parametric Simplex Method for Sparse Learning

Haotian Pang, Han Liu, Robert J. Vanderbei, Tuo Zhao

High dimensional sparse learning has imposed a great computational challenge to large scale data analysis. In this paper, we investiage a broad class of sparse learning approaches formulated as linear programs parametrized by a {\em regularization factor}, and solve them by the parametric simplex method (PSM). PSM offers significant advantages over other competing methods: (1) PSM naturally obtains the complete solution path for all values of the regularization parameter; (2) PSM provides a high precision dual certificate stopping criterion; (3) PSM yields sparse solutions through very few iterations, and the solution sparsity significantly reduces the computational cost per iteration. Particularly, we demonstrate the superiority of PSM over various sparse learning approaches, including Dantzig selector for sparse linear regression, sparse support vector machine for sparse linear classification, and sparse differential network estimation. We then provide sufficient conditions under which PSM always outputs sparse solutions such that its computational performance can be significantly boosted. Thorough numerical experiments are provided to demonstrate the outstanding performance of the PSM method.

************************************

Filtering Variational Objectives

Chris J. Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, Yee Teh

When used as a surrogate objective for maximum likelihood estimation in latent variable models, the evidence lower bound (ELBO) produces state-of-the-art results. Inspired by this, we consider the extension of the ELBO to a family of lower bounds defined by a particle filter's estimator of the marginal likelihood, the filtering variational objectives (FIVOs). FIVOs take the same arguments as the ELBO, but can exploit a model's sequential structure to form tighter bounds. We present results that relate the tightness of FIVO's bound to the variance of the particle filter's estimator by considering the generic case of bounds defined as log-transformed likelihood estimators. Experimentally, we show that training with FIVO results in substantial improvements over training the same model architecture with the ELBO on sequential data.

************************************

Cold-Start Reinforcement Learning with Softmax Policy Gradient

Nan Ding, Radu Soricut

Policy-gradient approaches to reinforcement learning have two common and undesirable overhead procedures, namely warm-start training and sample variance reduction. In this paper, we describe a reinforcement learning method based on a softmax value function that requires neither of these procedures. Our method combines the advantages of policy-gradient methods with the efficiency and simplicity of maximum-likelihood approaches. We apply this new cold-start reinforcement learning method in training sequence generation models for structured output prediction problems. Empirical evidence validates this method on automatic summarization and image captioning tasks.

************************************

Bridging the Gap Between Value and Policy Based Reinforcement Learning

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, Dale Schuurmans

We establish a new connection between value and policy based reinforcement learning (RL) based on a relationship between softmax temporal value consistency and policy optimality under entropy regularization. Specifically, we show that softmax consistent action values correspond to optimal entropy regularized policy pro babilities along any action sequence, regardless of provenance. From this observation, we develop a new RL algorithm, Path Consistency Learning (PCL), that minimizes a notion of soft consistency error along multi-step action sequences extracted from both on- and off-policy traces. We examine the behavior of PCL in different scenarios and show that PCL can be interpreted as generalizing both actor-critic and Q-learning algorithms. We subsequently deepen the relationship by showing how a single model can be used to represent both a policy and the corresponding softmax state values, eliminating the need for a separate critic. The experimental evaluation demonstrates that PCL significantly outperforms strong actor-critic and Q-learning baselines across several benchmarks.

************************************

Asynchronous Coordinate Descent under More Realistic Assumptions

Tao Sun, Robert Hannah, Wotao Yin

Asynchronous-parallel algorithms have the potential to vastly speed up algorithms by eliminating costly synchronization. However, our understanding of these algorithms is limited because the current convergence theory of asynchronous block coordinate descent algorithms is based on somewhat unrealistic assumptions. In particular, the age of the shared optimization variables being used to update blocks is assumed to be independent of the block being updated. Additionally, it is assumed that the updates are applied to randomly chosen blocks.  In this paper, we argue that these assumptions either fail to hold or will imply less efficient implementations. We then prove the convergence of asynchronous-parallel block coordinate descent under more realistic assumptions, in particular, always without the independence assumption. The analysis permits both the deterministic (essentially) cyclic and random rules for block choices. Because a bound on the asynchronous delays may or may not be available, we establish convergence for both bounded delays and unbounded delays. The analysis also covers nonconvex, weakly convex, and strongly convex functions. The convergence theory involves a Lyapunov function that directly incorporates both objective progress and delays. A continuous-time ODE is provided to motivate the construction at a high level.

************************************

EEG-GRAPH: A Factor-Graph-Based Model for Capturing Spatial, Temporal, and Observational Relationships in Electroencephalograms

Yogatheesan Varatharajah, Min Jin Chong, Krishnakant Saboo, Brent Berry, Benjamin Brinkmann, Gregory Worrell, Ravishankar Iyer

This paper presents a probabilistic-graphical model that can be used to infer characteristics of instantaneous brain activity by jointly analyzing spatial and temporal dependencies observed in electroencephalograms (EEG). Specifically, we describe a factor-graph-based model with customized factor-functions defined based on domain knowledge, to infer pathologic brain activity with the goal of identifying seizure-generating brain regions in epilepsy patients. We utilize an inference technique based on the graph-cut algorithm to exactly solve graph inference in polynomial time. We validate the model by using clinically collected intracranial EEG data from 29 epilepsy patients to show that the model correctly identifies seizure-generating brain regions. Our results indicate that our model outperforms two conventional approaches used for seizure-onset localization (5-7% better AUC: 0.72, 0.67, 0.65) and that the proposed inference technique provides 3-10% gain in AUC (0.72, 0.62, 0.69) compared to sampling-based alternatives.

************************************

Natural Value Approximators: Learning when to Trust Past Estimates

Zhongwen Xu, Joseph Modayil, Hado P. van Hasselt, Andre Barreto, David Silver, Tom Schaul

Neural networks have a smooth initial inductive bias, such that small changes in input do not lead to large changes in output. However, in reinforcement learning domains with sparse rewards, value functions have non-smooth structure with a characteristic asymmetric discontinuity whenever rewards arrive. We propose a mechanism that learns an interpolation between a direct value estimate and a projected value estimate computed from the encountered reward and the previous estimate. This reduces the need to learn about discontinuities, and thus improves the value function approximation. Furthermore, as the interpolation is learned and s

tate-dependent, our method can deal with heterogeneous observability. We demonstrate that this one change leads to significant improvements on multiple Atari games, when applied to the state-of-the-art A3C algorithm.
************************************

Active Exploration for Learning Symbolic Representations
Garrett Andersen, George Konidaris
We introduce an online active exploration algorithm for data-efficiently learning an abstract symbolic model of an environment. Our algorithm is divided into two parts: the first part quickly generates an intermediate Bayesian symbolic model from the data that the agent has collected so far, which the agent can then use along with the second part to guide its future exploration towards regions of the state space that the model is uncertain about. We show that our algorithm outperforms random and greedy exploration policies on two different computer game domains. The first domain is an Asteroids-inspired game with complex dynamics but basic logical structure. The second is the Treasure Game, with simpler dynamics but more complex logical structure.
************************************

Balancing information exposure in social networks
Kiran Garimella, Aristides Gionis, Nikos Parotsidis, Nikolaj Tatti
Social media has brought a revolution on how people are consuming news. Beyond the undoubtedly large number of advantages brought by social-media platforms, a point of criticism has been the creation of echo chambers and filter bubbles, caused by social homophily and algorithmic personalization.  In this paper we address the problem of balancing the information exposure} in a social network. We assume that two opposing campaigns (or viewpoints) are present in the network, and that network nodes have different preferences towards these campaigns. Our goal is to find two sets of nodes to employ in the respective campaigns, so that the overall information exposure for the two campaigns is balanced. We formally define the problem, characterize its hardness, develop approximation algorithms, and present experimental evaluation results.  Our model is inspired by the literature on influence maximization, but we offer significant novelties. First, balance of information exposure is modeled by a symmetric difference function, which is neither monotone nor submodular, and thus, not amenable to existing approaches. Second, while previous papers consider a setting with selfish agents and provide bounds on best response strategies (i.e., move of the last player), we consider a setting with a centralized agent and provide bounds for a global objective function.
************************************

Nonlinear Acceleration of Stochastic Algorithms
Damien Scieur, Francis Bach, Alexandre d'Aspremont
Extrapolation methods use the last few iterates of an optimization algorithm to produce a better estimate of the optimum. They were shown to achieve optimal convergence rates in a deterministic setting using simple gradient iterates. Here, we study extrapolation methods in a stochastic setting, where the iterates are produced by either a simple or an accelerated stochastic gradient algorithm. We first derive convergence bounds for arbitrary, potentially biased  perturbations, then produce asymptotic bounds using the ratio between the variance of the noise and the accuracy of the current point. Finally, we apply this acceleration technique to stochastic algorithms such as SGD, SAGA, SVRG and Katyusha in different settings, and show significant performance gains.
************************************

Multi-way Interacting Regression via Factorization Machines
Mikhail Yurochkin, XuanLong Nguyen, nikolaos Vasiloglou
We propose a Bayesian regression method that accounts for multi-way interactions of arbitrary orders among the predictor variables. Our model makes use of a factorization mechanism for representing the regression coefficients of interactions among the predictors, while the interaction selection is guided by a prior distribution on random hypergraphs, a construction which generalizes the Finite Feature Model. We present a posterior inference algorithm based on Gibbs sampling, and establish posterior consistency of our regression model. Our method is evalu

ated with extensive experiments on simulated data and demonstrated to be able to identify meaningful interactions in applications in genetics and retail demand forecasting.

************************************

The Expxorcist: Nonparametric Graphical Models Via Conditional Exponential Densities

Arun Suggala, Mladen Kolar, Pradeep K. Ravikumar

Non-parametric multivariate density estimation faces strong statistical and computational bottlenecks, and the more practical approaches impose near-parametric assumptions on the form of the density functions. In this paper, we leverage recent developments to propose a class of non-parametric models which have very attractive computational and statistical properties. Our approach relies on the simple function space assumption that the conditional distribution of each variable conditioned on the other variables has a non-parametric exponential family form.

************************************

Generating steganographic images via adversarial training

Jamie Hayes, George Danezis

Adversarial training has proved to be competitive against supervised learning methods on computer vision tasks. However, studies have mainly been confined to generative tasks such as image synthesis. In this paper, we apply adversarial training techniques to the discriminative task of learning a steganographic algorithm. Steganography is a collection of techniques for concealing the existence of information by embedding it within a non-secret medium, such as cover texts or images. We show that adversarial training can produce robust steganographic techniques: our unsupervised training scheme produces a steganographic algorithm that competes with state-of-the-art steganographic techniques. We also show that supervised training of our adversarial model produces a robust steganalyzer, which performs the discriminative task of deciding if an image contains secret information. We define a game between three parties, Alice, Bob and Eve, in order to simultaneously train both a steganographic algorithm and a steganalyzer. Alice and Bob attempt to communicate a secret message contained within an image, while Eve eavesdrops on their conversation and attempts to determine if secret information is embedded within the image. We represent Alice, Bob and Eve by neural networks, and validate our scheme on two independent image datasets, showing our novel method of studying steganographic problems is surprisingly competitive against established steganographic techniques.

************************************

NeuralFDR: Learning Discovery Thresholds from Hypothesis Features

Fei Xia, Martin J. Zhang, James Y. Zou, David Tse

As datasets grow richer, an important challenge is to leverage the full features in the data to maximize the number of useful discoveries while controlling for false positives. We address this problem in the context of multiple hypotheses testing, where for each hypothesis, we observe a p-value along with a set of features specific to that hypothesis. For example, in genetic association studies, each hypothesis tests the correlation between a variant and the trait. We have a rich set of features for each variant (e.g. its location, conservation, epigenetics etc.) which could inform how likely the variant is to have a true association. However popular testing approaches, such as Benjamini-Hochberg's procedure (BH) and independent hypothesis weighting (IHW), either ignore these features or assume that the features are categorical. We propose a new algorithm, NeuralFDR, which automatically learns a discovery threshold as a function of all the hypothesis features. We parametrize the discovery threshold as a neural network, which enables flexible handling of multi-dimensional discrete and continuous features as well as efficient end-to-end optimization. We prove that NeuralFDR has strong false discovery rate (FDR) guarantees, and show that it makes substantially more discoveries in synthetic and real datasets. Moreover, we demonstrate that the learned discovery threshold is directly interpretable.

************************************

A Scale Free Algorithm for Stochastic Bandits with Bounded Kurtosis

Tor Lattimore

Existing strategies for finite-armed stochastic bandits mostly depend on a parameter of scale that must be known in advance. Sometimes this is in the form of a bound on the payoffs, or the knowledge of a variance or subgaussian parameter. The notable exceptions are the analysis of Gaussian bandits with unknown mean and variance by Cowan and Katehakis [2015a] and of uniform distributions with unknown support [Cowan and Katehakis, 2015b]. The results derived in these specialised cases are generalised here to the non-parametric setup, where the learner knows only a bound on the kurtosis of the noise, which is a scale free measure of the extremity of outliers.

************************************

Value Prediction Network

Junhyuk Oh, Satinder Singh, Honglak Lee

This paper proposes a novel deep reinforcement learning (RL) architecture, called Value Prediction Network (VPN), which integrates model-free and model-based RL methods into a single neural network. In contrast to typical model-based RL methods, VPN learns a dynamics model whose abstract states are trained to make option-conditional predictions of future values (discounted sum of rewards) rather than of future observations. Our experimental results show that VPN has several advantages over both model-free and model-based baselines in a stochastic environment where careful planning is required but building an accurate observation-prediction model is difficult. Furthermore, VPN outperforms Deep Q-Network (DQN) on several Atari games even with short-lookahead planning, demonstrating its potential as a new way of learning a good state representation.

************************************

Detrended Partial Cross Correlation for Brain Connectivity Analysis

Jaime Ide, Fábio Cappabianco, Fabio Faria, Chiang-shan R. Li

Brain connectivity analysis is a critical component of ongoing human connectome projects to decipher the healthy and diseased brain. Recent work has highlighted the power-law (multi-time scale) properties of brain signals; however, there remains a lack of methods to specifically quantify short- vs. long- time range brain connections. In this paper, using detrended partial cross-correlation analysis (DPCCA), we propose a novel functional connectivity measure to delineate brain interactions at multiple time scales, while controlling for covariates. We use a rich simulated fMRI dataset to validate the proposed method, and apply it to a real fMRI dataset in a cocaine dependence prediction task. We show that, compared to extant methods, the DPCCA-based approach not only distinguishes short and long memory functional connectivity but also improves feature extraction and enhances classification accuracy. Together, this paper contributes broadly to new computational methodologies in understanding neural information processing.

************************************