Sign in to GitHub · GitHub
****
Stochastic Optimization with Importance Sampling for Regularized Loss Minimizati
on
Peilin Zhao, Tong Zhang
Uniform sampling of training data has been commonly used in traditional stochast
ic optimization algorithms such as Proximal Stochastic Mirror Descent (prox-SMD)
 and Proximal Stochastic Dual Coordinate Ascent (prox-SDCA). Although uniform sa
mpling can guarantee that the sampled stochastic quantity is an unbiased estimat
e of the corresponding true quantity, the resulting estimator may have a rather
high variance, which negatively affects the convergence of the underlying optimi
zation procedure. In this paper we study stochastic optimization, including prox
-SMD and prox-SDCA, with importance sampling, which improves the convergence rat
e by reducing the stochastic variance. We theoretically analyze the algorithms a
nd empirically validate their effectiveness.
****
Approval Voting and Incentives in Crowdsourcing
Nihar Shah, Dengyong Zhou, Yuval Peres
The growing need for labeled training data has made crowdsourcing an important p
art of machine learning. The quality of crowdsourced labels is, however, adverse
ly affected by three factors: (1) the workers are not experts; (2) the incentive
s of the workers are not aligned with those of the requesters; and (3) the inter
face does not allow workers to convey their knowledge accurately, by forcing the
m to make a single choice among a set of options. In this paper, we address thes
e issues by introducing approval voting to utilize the expertise of workers who
have partial knowledge of the true answer, and coupling it with a ("strictly pro
per") incentive-compatible compensation mechanism. We show rigorous theoretical
guarantees of optimality of our mechanism together with a simple axiomatic chara
cterization. We also conduct preliminary empirical studies on Amazon Mechanical
Turk which validate our approach.
****
A low variance consistent test of relative dependency
Wacha Bounliphone, Arthur Gretton, Arthur Tenenhaus, Matthew Blaschko
We describe a novel non-parametric statistical hypothesis test of relative depen
dence between a source variable and two candidate target variables. Such a test
enables us to determine whether one source variable is significantly more depend
ent on a first target variable or a second. Dependence is measured via the Hilbe
rt-Schmidt Independence Criterion (HSIC), resulting in a pair of empirical depen
dence measures (source-target 1, source-target 2). We test whether the first dep
endence measure is significantly larger than the second. Modeling the covariance
 between these HSIC statistics leads to a provably more powerful test than the c
onstruction of independent HSIC statistics by sub-sampling. The resulting test i
s consistent and unbiased, and (being based on U-statistics) has favorable conve
rgence properties. The test can be computed in quadratic time, matching the comp
utational complexity of standard empirical HSIC estimators. The effectiveness of
 the test is demonstrated on several real-world problems: we identify language g
roups from a multilingual corpus, and we prove that tumor location is more depen
dent on gene expression than chromosomal imbalances. Source code is available fo
r download at https://github.com/wbounliphone/reldep/.
****
An Aligned Subtree Kernel for Weighted Graphs
Lu Bai, Luca Rossi, Zhihong Zhang, Edwin Hancock
In this paper, we develop a new entropic matching kernel for weighted graphs by
aligning depth-based representations. We demonstrate that this kernel can be see
n as an \textbfaligned subtree kernel that incorporates explicit subtree corresp
ondences, and thus addresses the drawback of neglecting the relative locations b
etween substructures that arises in the R-convolution kernels. Experiments on st
andard datasets demonstrate that our kernel can easily outperform state-of-the-a
rt graph kernels in terms of classification accuracy.
****

Spectral Clustering via the Power Method - Provably
Christos Boutsidis, Prabhanjan Kambadur, Alex Gittens
Spectral clustering is one of the most important algorithms in data mining and m
achine intelligence; however, its computational complexity limits its applicatio
n to truly large scale data analysis. The computational bottleneck in spectral c
lustering is computing a few of the top eigenvectors of the (normalized) Laplaci
an matrix corresponding to the graph representing the data to be clustered. One
way to speed up the computation of these eigenvectors is to use the “power metho
d” from the numerical linear algebra literature. Although the power method has b
een empirically used to speed up spectral clustering, the theory behind this app
roach, to the best of our knowledge, remains unexplored. This paper provides the
 first such rigorous theoretical justification, arguing that a small number of p
ower iterations suffices to obtain near-optimal partitionings using the approxim
ate eigenvectors. Specifically, we prove that solving the k-means clustering pro
blem on the approximate eigenvectors obtained via the power method gives an addi
tive-error approximation to solving the k-means problem on the optimal eigenvect
ors.
****

Information Geometry and Minimum Description Length Networks
Ke Sun, Jun Wang, Alexandros Kalousis, Stephan Marchand-Maillet
We study parametric unsupervised mixture learning. We measure the loss of intrin
sic information from the observations to complex mixture models, and then to sim
ple mixture models. We present a geometric picture, where all these representati
ons are regarded as free points in the space of probability distributions. Based
 on minimum description length, we derive a simple geometric principle to learn
all these models together. We present a new learning machine with theories, algo
rithms, and simulations.
****

Efficient Training of LDA on a GPU by Mean-for-Mode Estimation
Jean-Baptiste Tristan, Joseph Tassarotti, Guy Steele
We introduce Mean-for-Mode estimation, a variant of an uncollapsed Gibbs sampler
 that we use to train LDA on a GPU. The algorithm combines benefits of both unco
llapsed and collapsed Gibbs samplers. Like a collapsed Gibbs sampler — and unlik
e an uncollapsed Gibbs sampler — it has good statistical performance, and can us
e sampling complexity reduction techniques such as sparsity. Meanwhile, like an
uncollapsed Gibbs sampler — and unlike a collapsed Gibbs sampler — it is embarra
ssingly parallel, and can use approximate counters.
****

Adaptive Stochastic Alternating Direction Method of Multipliers
Peilin Zhao, Jinwei Yang, Tong Zhang, Ping Li
The Alternating Direction Method of Multipliers (ADMM) has been studied for year
s. Traditional ADMM algorithms need to compute, at each iteration, an (empirical
) expected loss function on all training examples, resulting in a computational
complexity proportional to the number of training examples. To reduce the comple
xity, stochastic ADMM algorithms were proposed to replace the expected loss func
tion with a random loss function associated with one uniformly drawn example plu
s a Bregman divergence term. The Bregman divergence, however, is derived from a
simple 2nd-order proximal function, i.e., the half squared norm, which could be
a suboptimal choice. In this paper, we present a new family of stochastic ADMM a
lgorithms with optimal 2nd-order proximal functions, which produce a new family
of adaptive stochastic ADMM methods. We theoretically prove that the regret boun
ds are as good as the bounds which could be achieved by the best proximal functi
on that can be chosen in hindsight. Encouraging empirical results on a variety o
f real-world datasets confirm the effectiveness and efficiency of the proposed a
lgorithms.
****

A Lower Bound for the Optimization of Finite Sums
Alekh Agarwal, Leon Bottou
This paper presents a lower bound for optimizing a finite sum of n functions, wh
ere each function is L-smooth and the sum is μ-strongly convex. We show that no

algorithm can reach an error εin minimizing all functions from this class in few
er than Ω(n + \sqrtn(κ-1)\log(1/ε)) iterations, where κ=L/μis a surrogate condit
ion number. We then compare this lower bound to upper bounds for recently develo
ped methods specializing to this setting. When the functions involved in this su
m are not arbitrary, but based on i.i.d. random data, then we further contrast t
hese complexity results with those for optimal first-order methods to directly o
ptimize the sum. The conclusion we draw is that a lot of caution is necessary fo
r an accurate comparison, and identify machine learning scenarios where the new
methods help computationally.
****

Learning Word Representations with Hierarchical Sparse Coding
Dani Yogatama, Manaal Faruqui, Chris Dyer, Noah Smith
We propose a new method for learning word representations using hierarchical reg
ularization in sparse coding inspired by the linguistic study of word meanings.
We show an efficient learning algorithm based on stochastic proximal methods tha
t is significantly faster than previous approaches, making it possible to perfor
m hierarchical sparse coding on a corpus of billions of word tokens. Experiments
 on various benchmark tasks—word similarity ranking, syntactic and semantic anal
ogies, sentence completion, and sentiment analysis—demonstrate that the method o
utperforms or is competitive with state-of-the-art methods.
****

Learning Transferable Features with Deep Adaptation Networks
Mingsheng Long, Yue Cao, Jianmin Wang, Michael Jordan
Recent studies reveal that a deep neural network can learn transferable features
 which generalize well to novel tasks for domain adaptation. However, as deep fe
atures eventually transition from general to specific along the network, the fea
ture transferability drops significantly in higher layers with increasing domain
 discrepancy. Hence, it is important to formally reduce the dataset bias and enh
ance the transferability in task-specific layers. In this paper, we propose a ne
w Deep Adaptation Network (DAN) architecture, which generalizes deep convolution
al neural network to the domain adaptation scenario. In DAN, hidden representati
ons of all task-specific layers are embedded in a reproducing kernel Hilbert spa
ce where the mean embeddings of different domain distributions can be explicitly
 matched. The domain discrepancy is further reduced using an optimal multi-kerne
l selection method for mean embedding matching. DAN can learn transferable featu
res with statistical guarantees, and can scale linearly by unbiased estimate of
kernel embedding. Extensive empirical evidence shows that the proposed architect
ure yields state-of-the-art image classification error rates on standard domain
adaptation benchmarks.
****

Robust partially observable Markov decision process
Takayuki Osogami
We seek to find the robust policy that maximizes the expected cumulative reward
for the worst case when a partially observable Markov decision process (POMDP) h
as uncertain parameters whose values are only known to be in a given region. We
prove that the robust value function, which represents the expected cumulative r
eward that can be obtained with the robust policy, is convex with respect to the
 belief state. Based on the convexity, we design a value-iteration algorithm for
 finding the robust policy. We prove that our value iteration converges for an i
nfinite horizon. We also design point-based value iteration for fining the robus
t policy more efficiency possibly with approximation. Numerical experiments show
 that our point-based value iteration can adequately find robust policies.
****

On the Relationship between Sum-Product Networks and Bayesian Networks
Han Zhao, Mazen Melibari, Pascal Poupart
In this paper, we establish some theoretical connections between Sum-Product Net
works (SPNs) and Bayesian Networks (BNs). We prove that every SPN can be convert
ed into a BN in linear time and space in terms of the network size. The key insi
ght is to use Algebraic Decision Diagrams (ADDs) to compactly represent the loca
l conditional probability distributions at each node in the resulting BN by expl

oiting context-specific independence (CSI). The generated BN has a simple direct ed bipartite graphical structure. We show that by applying the Variable Eliminat ion algorithm (VE) to the generated BN with ADD representations, we can recover the original SPN where the SPN can be viewed as a history record or caching of t he VE inference process. To help state the proof clearly, we introduce the notio n of \em normal SPN and present a theoretical analysis of the consistency and de composability properties. We conclude the paper with some discussion of the impl ications of the proof and establish a connection between the depth of an SPN and a lower bound of the tree-width of its corresponding BN.
****

Learning from Corrupted Binary Labels via Class-Probability Estimation
Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, Bob Williamson
Many supervised learning problems involve learning from samples whose labels are corrupted in some way. For example, each sample may have some constant probabil ity of being incorrectly labelled (learning with label noise), or one may have a pool of unlabelled samples in lieu of negative samples (learning from positive and unlabelled data). This paper uses class-probability estimation to study thes e and other corruption processes belonging to the mutually contaminated distribu tions framework (Scott et al., 2013), with three conclusions. First, one can opt imise balanced error and AUC without knowledge of the corruption process paramet ers. Second, given estimates of the corruption parameters, one can minimise a ra nge of classification risks. Third, one can estimate the corruption parameters u sing only corrupted data. Experiments confirm the efficacy of class-probability estimation in learning from corrupted labels.
****

An Explicit Sampling Dependent Spectral Error Bound for Column Subset Selection
Tianbao Yang, Lijun Zhang, Rong Jin, Shenghuo Zhu
In this paper, we consider the problem of column subset selection. We present a novel analysis of the spectral norm reconstruction for a simple randomized algor ithm and establish a new bound that depends explicitly on the sampling probabili ties. The sampling dependent error bound (i) allows us to better understand the tradeoff in the reconstruction error due to sampling probabilities, (ii) exhibit s more insights than existing error bounds that exploit specific probability dis tributions, and (iii) implies better sampling distributions. In particular, we s how that a sampling distribution with probabilities proportional to the square r oot of the statistical leverage scores is better than uniform sampling, and is b etter than leverage-based sampling when the statistical leverage scores are very nonuniform. And by solving a constrained optimization problem related to the er ror bound with an efficient bisection search we are able to achieve better perfo rmance than using either the leverage-based distribution or that proportional to the square root of the statistical leverage scores. Numerical simulations demon strate the benefits of the new sampling distributions for low-rank matrix approx imation and least square approximation compared to state-of-the art algorithms.
****

A Stochastic PCA and SVD Algorithm with an Exponential Convergence Rate
Ohad Shamir
We describe and analyze a simple algorithm for principal component analysis and singular value decomposition, VR-PCA, which uses computationally cheap stochasti c iterations, yet converges exponentially fast to the optimal solution. In contr ast, existing algorithms suffer either from slow convergence, or computationally intensive iterations whose runtime scales with the data size. The algorithm bui lds on a recent variance-reduced stochastic gradient technique, which was previo usly analyzed for strongly convex optimization, whereas here we apply it to an i nherently non-convex problem, using a very different analysis.
****

Attribute Efficient Linear Regression with Distribution-Dependent Sampling
Doron Kukliansky, Ohad Shamir
We consider a budgeted learning setting, where the learner can only choose and o bserve a small subset of the attributes of each training example. We develop eff icient algorithms for Ridge and Lasso linear regression, which utilize the geome

try of the data by a novel distribution-dependent sampling scheme, and have exce ss risk bounds which are better a factor of up to O(d/k) over the state-of-the-a rt, where d is the dimension and k+1 is the number of observed attributes per ex ample. Moreover, under reasonable assumptions, our algorithms are the first in o ur setting which can provably use *less* attributes than full-information algori thms, which is the main concern in budgeted learning. We complement our theoreti cal analysis with experiments which support our claims.
****

Learning Local Invariant Mahalanobis Distances
Ethan Fetaya, Shimon Ullman
For many tasks and data types, there are natural transformations to which the da ta should be invariant or insensitive. For instance, in visual recognition, natu ral images should be insensitive to rotation and translation. This requirement a nd its implications have been important in many machine learning applications, a nd tolerance for image transformations was primarily achieved by using robust fe ature vectors. In this paper we propose a novel and computationally efficient wa y to learn a local Mahalanobis metric per datum, and show how we can learn a loc al invariant metric to any transformation in order to improve performance.
****

Finding Linear Structure in Large Datasets with Scalable Canonical Correlation A nalysis
Zhuang Ma, Yichao Lu, Dean Foster
Canonical Correlation Analysis (CCA) is a widely used spectral technique for fin ding correlation structures in multi-view datasets. In this paper, we tackle the problem of large scale CCA, where classical algorithms, usually requiring compu ting the product of two huge matrices and huge matrix decomposition, are computa tionally and storage expensive. We recast CCA from a novel perspective and propo se a scalable and memory efficient \textitAugmented Approximate Gradient (AppGra d) scheme for finding top k dimensional canonical subspace which only involves l arge matrix multiplying a thin matrix of width k and small matrix decomposition of dimension k\times k. Further, \textitAppGrad achieves optimal storage complex ity O(k(p_1+p_2)), compared with classical algorithms which usually require O(p_ 1^2+p_2^2) space to store two dense whitening matrices. The proposed scheme natu rally generalizes to stochastic optimization regime, especially efficient for hu ge datasets where batch algorithms are prohibitive. The online property of stoch astic \textitAppGrad is also well suited to the streaming scenario, where data c omes sequentially. To the best of our knowledge, it is the first stochastic algo rithm for CCA. Experiments on four real data sets are provided to show the effec tiveness of the proposed methods.
****

Abstraction Selection in Model-based Reinforcement Learning
Nan Jiang, Alex Kulesza, Satinder Singh
State abstractions are often used to reduce the complexity of model-based reinfo rcement learning when only limited quantities of data are available. However, ch oosing the appropriate level of abstraction is an important problem in practice. Existing approaches have theoretical guarantees only under strong assumptions o n the domain or asymptotically large amounts of data, but in this paper we propo se a simple algorithm based on statistical hypothesis testing that comes with a finite-sample guarantee under assumptions on candidate abstractions. Our algorit hm trades off the low approximation error of finer abstractions against the low estimation error of coarser abstractions, resulting in a loss bound that depends only on the quality of the best available abstraction and is polynomial in plan ning horizon.
****

Surrogate Functions for Maximizing Precision at the Top
Purushottam Kar, Harikrishna Narasimhan, Prateek Jain
The problem of maximizing precision at the top of a ranked list, often dubbed Pr ecision@k (prec@k), finds relevance in myriad learning applications such as rank ing, multi-label classification, and learning with severe label imbalance. Howev er, despite its popularity, there exist significant gaps in our understanding of

this problem and its associated performance measure. The most notable of these is the lack of a convex upper bounding surrogate for prec@k. We also lack scalable perceptron and stochastic gradient descent algorithms for optimizing this performance measure. In this paper we make key contributions in these directions. At the heart of our results is a family of truly upper bounding surrogates for prec@k. These surrogates are motivated in a principled manner and enjoy attractive properties such as consistency to prec@k under various natural margin/noise conditions. These surrogates are then used to design a class of novel perceptron algorithms for optimizing prec@k with provable mistake bounds. We also devise scalable stochastic gradient descent style methods for this problem with provable convergence bounds. Our proofs rely on novel uniform convergence bounds which require an in-depth analysis of the structural properties of prec@k and its surrogates. We conclude with experimental results comparing our algorithms with state-of-the-art cutting plane and stochastic gradient algorithms for maximizing prec@k.

****

Optimizing Non-decomposable Performance Measures: A Tale of Two Classes

Harikrishna Narasimhan, Purushottam Kar, Prateek Jain

Modern classification problems frequently present mild to severe label imbalance as well as specific requirements on classification characteristics, and require optimizing performance measures that are non-decomposable over the dataset, such as F-measure. Such measures have spurred much interest and pose specific challenges to learning algorithms since their non-additive nature precludes a direct application of well-studied large scale optimization methods such as stochastic gradient descent. In this paper we reveal that for two large families of performance measures that can be expressed as functions of true positive/negative rates, it is indeed possible to implement point stochastic updates. The families we consider are concave and pseudo-linear functions of TPR, TNR which cover several popularly used performance measures such as F-measure, G-mean and H-mean. Our core contribution is an adaptive linearization scheme for these families, using which we develop optimization techniques that enable truly point-based stochastic updates. For concave performance measures we propose SPADE, a stochastic primal-dual solver; for pseudo-linear measures we propose STAMP, a stochastic alternate maximization procedure. Both methods have crisp convergence guarantees, demonstrate significant speedups over existing methods - often by an order of magnitude or more, and give similar or more accurate predictions on test data.

****

Coresets for Nonparametric Estimation - the Case of DP-Means

Olivier Bachem, Mario Lucic, Andreas Krause

Scalable training of Bayesian nonparametric models is a notoriously difficult challenge. We explore the use of coresets - a data summarization technique originating from computational geometry - for this task. Coresets are weighted subsets of the data such that models trained on these coresets are provably competitive with models trained on the full dataset. Coresets sublinear in the dataset size allow for fast approximate inference with provable guarantees. Existing constructions, however, are limited to parametric problems. Using novel techniques in coreset construction we show the existence of coresets for DP-Means - a prototypical nonparametric clustering problem - and provide a practical construction algorithm. We empirically demonstrate that our algorithm allows us to efficiently trade off computation time and approximation error and thus scale DP-Means to large datasets. For instance, with coresets we can obtain a computational speedup of 45x at an approximation error of only 2.4% compared to solving on the full dataset. In contrast, for the same subsample size, the “naive” approach of uniformly subsampling the data incurs an approximation error of 22.5%.

****

A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits

Pratik Gajane, Tanguy Urvoy, Fabrice Clérot

We study the K-armed dueling bandit problem which is a variation of the classical Multi-Armed Bandit (MAB) problem in which the learner receives only relative feedback about the selected pairs of arms. We propose a new algorithm called Rela

tive Exponential-weight algorithm for Exploration and Exploitation (REX3) to han dle the adversarial utility-based formulation of this problem. This algorithm is a non-trivial extension of the Exponential-weight algorithm for Exploration and Exploitation (EXP3) algorithm. We prove a finite time expected regret upper bou nd of order O(sqrt(K ln(K)T)) for this algorithm and a general lower bound of or der omega(sqrt(KT)). At the end, we provide experimental results using real data from information retrieval applications.

****

Functional Subspace Clustering with Application to Time Series

Mohammad Taha Bahadori, David Kale, Yingying Fan, Yan Liu

Functional data, where samples are random functions, are increasingly common and important in a variety of applications, such as health care and traffic analysi s. They are naturally high dimensional and lie along complex manifolds. These pr operties warrant use of the subspace assumption, but most state-of-the-art subsp ace learning algorithms are limited to linear or other simple settings. To addre ss these challenges, we propose a new framework called Functional Subspace Clust ering (FSC). FSC assumes that functional samples lie in deformed linear subspace s and formulates the subspace learning problem as a sparse regression over opera tors. The resulting problem can be efficiently solved via greedy variable select ion, given access to a fast deformation oracle. We provide theoretical guarantee s for FSC and show how it can be applied to time series with warped alignments. Experimental results on both synthetic data and real clinical time series show t hat FSC outperforms both standard time series clustering and state-of-the-art su bspace clustering.

****

Accelerated Online Low Rank Tensor Learning for Multivariate Spatiotemporal Stre ams

Rose Yu, Dehua Cheng, Yan Liu

Low-rank tensor learning has many applications in machine learning. A series of batch learning algorithms have achieved great successes. However, in many emergi ng applications, such as climate data analysis, we are confronted with large-sca le tensor streams, which poses significant challenges to existing solution in te rms of computational costs and limited response time. In this paper, we propose an online accelerated low-rank tensor learning algorithm (ALTO) to solve the pro blem. At each iteration, we project the current tensor to the subspace of low-ra nk tensors in order to perform efficient tensor decomposition, then recover the decomposition of the new tensor. By randomly glancing at additional subspaces, w e successfully avoid local optima at negligible extra computational cost. We eva luate our method on two tasks in streaming multivariate spatio-temporal analysis : online forecasting and multi-model ensemble, which shows that our method achie ves comparable predictive accuracy with significant boost in run time.

****

Atomic Spatial Processes

Sean Jewell, Neil Spencer, Alexandre Bouchard-Côté

The emergence of compact GPS systems and the establishment of open data initiati ves has resulted in widespread availability of spatial data for many urban centr es. These data can be leveraged to develop data-driven intelligent resource allo cation systems for urban issues such as policing, sanitation, and transportation . We employ techniques from Bayesian non-parametric statistics to develop a proc ess which captures a common characteristic of urban spatial datasets. Specifical ly, our new spatial process framework models events which occur repeatedly at di screte spatial points, the number and locations of which are unknown a priori. W e develop a representation of our spatial process which facilitates posterior si mulation, resulting in an interpretable and computationally tractable model. The framework’s superiority over both empirical grid-based models and Dirichlet pro cess mixture models is demonstrated by fitting, interpreting, and comparing mode ls of graffiti prevalence for both downtown Vancouver and Manhattan.

****

Classification with Low Rank and Missing Data

Elad Hazan, Roi Livni, Yishay Mansour

We consider classification and regression tasks where we have missing data and a
ssume that the (clean) data resides in a low rank subspace. Finding a hidden sub
space is known to be computationally hard. Nevertheless, using a non-proper form
ulation we give an efficient agnostic algorithm that classifies as good as the b
est linear classifier coupled with the best low-dimensional subspace in which th
e data resides. A direct implication is that our algorithm can linearly (and non
-linearly through kernels) classify provably as well as the best classifier that
 has access to the full data.
****

Dynamic Sensing: Better Classification under Acquisition Constraints
Oran Richman, Shie Mannor
In many machine learning applications the quality of the data is limited by reso
urce constraints (may it be power, bandwidth, memory, ...). In such cases, the c
onstraints are on the average resources allocated, therefore there is some contr
ol over each sample’s quality. In most cases this option remains unused and the
data’s quality is uniform over the samples. In this paper we propose to actively
 allocate resources to each sample such that resources are used optimally overal
l. We propose a method to compute the optimal resource allocation. We further de
rive generalization bounds for the case where the problem’s model is unknown. We
 demonstrate the potential benefit of this approach on both simulated and real-l
ife problems.
****

A Modified Orthant-Wise Limited Memory Quasi-Newton Method with Convergence Anal
ysis
Pinghua Gong, Jieping Ye
The Orthant-Wise Limited memory Quasi-Newton (OWL-QN) method has been demonstrat
ed to be very effective in solving the \ell_1-regularized sparse learning proble
m. OWL-QN extends the L-BFGS from solving unconstrained smooth optimization prob
lems to \ell_1-regularized (non-smooth) sparse learning problems. At each iterat
ion, OWL-QN does not involve any \ell_1-regularized quadratic optimization subpr
oblem and only requires matrix-vector multiplications without an explicit use of
 the (inverse) Hessian matrix, which enables OWL-QN to tackle large-scale proble
ms efficiently. Although many empirical studies have demonstrated that OWL-QN wo
rks quite well in practice, several recent papers point out that the existing co
nvergence proof of OWL-QN is flawed and a rigorous convergence analysis for OWL-
QN still remains to be established. In this paper, we propose a modified Orthant
-Wise Limited memory Quasi-Newton (mOWL-QN) algorithm by slightly modifying the
OWL-QN algorithm. As the main technical contribution of this paper, we establish
 a rigorous convergence proof for the mOWL-QN algorithm. To the best of our know
ledge, our work fills the theoretical gap by providing the first rigorous conver
gence proof for the OWL-QN-type algorithm on solving \ell_1-regularized sparse l
earning problems. We also provide empirical studies to show that mOWL-QN works w
ell and is as efficient as OWL-QN.
****

Telling cause from effect in deterministic linear dynamical systems
Naji Shajarisales, Dominik Janzing, Bernhard Schoelkopf, Michel Besserve
Telling a cause from its effect using observed time series data is a major chall
enge in natural and social sciences. Assuming the effect is generated by the cau
se through a linear system, we propose a new approach based on the hypothesis th
at nature chooses the “cause” and the “mechanism generating the effect from the
cause” independently of each other. Specifically we postulate that the power spe
ctrum of the “cause” time series is uncorrelated with the square of the frequenc
y response of the linear filter (system) generating the effect. While most causa
l discovery methods for time series mainly rely on the noise, our method relies
on asymmetries of the power spectral density properties that exist even in deter
ministic systems. We describe mathematical assumptions in a deterministic model
under which the causal direction is identifiable. In particular, we show a scena
rio where the method works but Granger causality fails. Experiments show encoura
ging results on synthetic as well as real-world data. Overall, this suggests tha
t the postulate of Independence of Cause and Mechanism is a promising principle

for causal inference on observed time series.
****

High Dimensional Bayesian Optimisation and Bandits via Additive Models
Kirthevasan Kandasamy, Jeff Schneider, Barnabas Poczos
Bayesian Optimisation (BO) is a technique used in optimising a D-dimensional fun ction which is typically expensive to evaluate. While there have been many succe sses for BO in low dimensions, scaling it to high dimensions has been notoriousl y difficult. Existing literature on the topic are under very restrictive setting s. In this paper, we identify two key challenges in this endeavour. We tackle th ese challenges by assuming an additive structure for the function. This setting is substantially more expressive and contains a richer class of functions than p revious work. We prove that, for additive functions the regret has only linear d ependence on D even though the function depends on all D dimensions. We also dem onstrate several other statistical and computational benefits in our framework. Via synthetic examples, a scientific simulation and a face detection problem we demonstrate that our method outperforms naive BO on additive functions and on se veral examples where the function is not additive.
****

Theory of Dual-sparse Regularized Randomized Reduction
Tianbao Yang, Lijun Zhang, Rong Jin, Shenghuo Zhu
In this paper, we study randomized reduction methods, which reduce high-dimensio nal features into low-dimensional space by randomized methods (e.g., random proj ection, random hashing), for large-scale high-dimensional classification. Previo us theoretical results on randomized reduction methods hinge on strong assumptio ns about the data, e.g., low rank of the data matrix or a large separable margin of classification, which hinder their in broad domains. To address these limita tions, we propose dual-sparse regularized randomized reduction methods that intr oduce a sparse regularizer into the reduced dual problem. Under a mild condition that the original dual solution is a (nearly) sparse vector, we show that the r esulting dual solution is close to the original dual solution and concentrates o n its support set. In numerical experiments, we present an empirical study to su pport the analysis and we also present a novel application of the dual-sparse ra ndomized reduction methods to reducing the communication cost of distributed lea rning from large-scale high-dimensional data.
****

Generalization error bounds for learning to rank: Does the length of document li sts matter?
Ambuj Tewari, Sougata Chaudhuri
We consider the generalization ability of algorithms for learning to rank at a q uery level, a problem also called subset ranking. Existing generalization error bounds necessarily degrade as the size of the document list associated with a qu ery increases. We show that such a degradation is not intrinsic to the problem. For several loss functions, including the cross-entropy loss used in the well kn own ListNet method, there is no degradation in generalization ability as documen t lists become longer. We also provide novel generalization error bounds under \ ell_1 regularization and faster convergence rates if the loss function is smooth .
****

PeakSeg: constrained optimal segmentation and supervised penalty learning for pe ak detection in count data
Toby Hocking, Guillem Rigaill, Guillaume Bourque
Peak detection is a central problem in genomic data analysis, and current algori thms for this task are unsupervised and mostly effective for a single data type and pattern (e.g. H3K4me3 data with a sharp peak pattern). We propose PeakSeg, a new constrained maximum likelihood segmentation model for peak detection with a n efficient inference algorithm: constrained dynamic programming. We investigate unsupervised and supervised learning of penalties for the critical model select ion problem. We show that the supervised method has state-of-the-art peak detect ion across all data sets in a benchmark that includes both sharp H3K4me3 and bro ad H3K36me3 patterns.

****
Mind the duality gap: safer rules for the Lasso
Olivier Fercoq, Alexandre Gramfort, Joseph Salmon
Screening rules allow to early discard irrelevant variables from the optimizatio n in Lasso problems, or its derivatives, making solvers faster. In this paper, w e propose new versions of the so-called \textitsafe rules for the Lasso. Based o n duality gap considerations, our new rules create safe test regions whose diame ters converge to zero, provided that one relies on a converging solver. This pro perty helps screening out more variables, for a wider range of regularization pa rameter values. In addition to faster convergence, we prove that we correctly id entify the active sets (supports) of the solutions in finite time. While our pro posed strategy can cope with any solver, its performance is demonstrated using a coordinate descent algorithm particularly adapted to machine learning use cases . Significant computing time reductions are obtained with respect to previous sa fe rules.
****
A General Analysis of the Convergence of ADMM
Robert Nishihara, Laurent Lessard, Ben Recht, Andrew Packard, Michael Jordan
We provide a new proof of the linear convergence of the alternating direction me thod of multipliers (ADMM) when one of the objective terms is strongly convex. O ur proof is based on a framework for analyzing optimization algorithms introduce d in Lessard et al. (2014), reducing algorithm convergence to verifying the stab ility of a dynamical system. This approach generalizes a number of existing resu lts and obviates any assumptions about specific choices of algorithm parameters. On a numerical example, we demonstrate that minimizing the derived bound on the convergence rate provides a practical approach to selecting algorithm parameter s for particular ADMM instances. We complement our upper bound by constructing a nearly-matching lower bound on the worst-case rate of convergence.
****
Stochastic Primal-Dual Coordinate Method for Regularized Empirical Risk Minimiza tion
Yuchen Zhang, Xiao Lin
We consider a generic convex optimization problem associated with regularized em pirical risk minimization of linear predictors. The problem structure allows us to reformulate it as a convex-concave saddle point problem. We propose a stochas tic primal-dual coordinate method, which alternates between maximizing over one (or more) randomly chosen dual variable and minimizing over the primal variable. We also develop an extension to non-smooth and non-strongly convex loss functio ns, and an extension with better convergence rate on unnormalized data. Both the oretically and empirically, we show that the SPDC method has comparable or bette r performance than several state-of-the-art optimization methods.
****
DiSCO: Distributed Optimization for Self-Concordant Empirical Loss
Yuchen Zhang, Xiao Lin
We propose a new distributed algorithm for empirical risk minimization in machin e learning. The algorithm is based on an inexact damped Newton method, where the inexact Newton steps are computed by a distributed preconditioned conjugate gra dient method. We analyze its iteration complexity and communication efficiency f or minimizing self-concordant empirical loss functions, and discuss the results for distributed ridge regression, logistic regression and binary classification with a smoothed hinge loss. In a standard setting for supervised learning, where the n data points are i.i.d. sampled and when the regularization parameter scal es as 1/\sqrtn, we show that the proposed algorithm is communication efficient: the required round of communication does not increase with the sample size n, an d only grows slowly with the number of machines.
****
Spectral MLE: Top-K Rank Aggregation from Pairwise Comparisons
Yuxin Chen, Changho Suh
This paper explores the preference-based top-K rank aggregation problem. Suppose that a collection of items is repeatedly compared in pairs, and one wishes to r

ecover a consistent ordering that emphasizes the top-K ranked items, based on pa rtially revealed preferences. We focus on the Bradley-Terry-Luce (BTL) model tha t postulates a set of latent preference scores underlying all items, where the o dds of paired comparisons depend only on the relative scores of the items involv ed. We characterize the minimax limits on identifiability of top-K ranked items, in the presence of random and non-adaptive sampling. Our results highlight a se paration measure that quantifies the gap of preference scores between the K-th a nd (K+1)-th ranked items. The minimum sample complexity required for reliable to p-K ranking scales inversely with the separation measure irrespective of other p reference distribution metrics. To approach this minimax limit, we propose a nea rly linear-time ranking scheme, called Spectral MLE, that returns the indices of the top-K items in accordance to a careful score estimate. In a nutshell, Spect ral MLE starts with an initial score estimate with minimal squared loss (obtaine d via a spectral method), and then successively refines each component with the assistance of coordinate-wise MLEs. Encouragingly, Spectral MLE allows perfect t op-K item identification under minimal sample complexity. The practical applicab ility of Spectral MLE is further corroborated by numerical experiments.

****

Paired-Dual Learning for Fast Training of Latent Variable Hinge-Loss MRFs

Stephen Bach, Bert Huang, Jordan Boyd-Graber, Lise Getoor

Latent variables allow probabilistic graphical models to capture nuance and stru cture in important domains such as network science, natural language processing, and computer vision. Naive approaches to learning such complex models can be pr ohibitively expensive—because they require repeated inferences to update beliefs about latent variables—so lifting this restriction for useful classes of models is an important problem. Hinge-loss Markov random fields (HL-MRFs) are graphica l models that allow highly scalable inference and learning in structured domains , in part by representing structured problems with continuous variables. However , this representation leads to challenges when learning with latent variables. W e introduce paired-dual learning, a framework that greatly speeds up training by using tractable entropy surrogates and avoiding repeated inferences. Paired-dua l learning optimizes an objective with a pair of dual inference problems. This a llows fast, joint optimization of parameters and dual variables. We evaluate on social-group detection, trust prediction in social networks, and image reconstru ction, finding that paired-dual learning trains models as accurate as those trai ned by traditional methods in much less time, often before traditional methods m ake even a single parameter update.

****

Structural Maxent Models

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, Umar Syed

We present a new class of density estimation models, Structural Maxent models, w ith feature functions selected from possibly very complex families. The design o f our models is motivated by data-dependent convergence bounds and benefits from new data-dependent learning bounds expressed in terms of the Rademacher complex ities of the sub-families composing the family of features considered. We prove a duality theorem, which we use to derive our Structural Maxent algorithm. We gi ve a full description of our algorithm, including the details of its derivation and report the results of several experiments demonstrating that its performance compares favorably to that of existing regularized Maxent. We further similarly define conditional Structural Maxent models for multi-class classification prob lems. These are conditional probability models making use of possibly complex fe ature families. We also prove a duality theorem for these models which shows the connection between these models and existing binary and multi-class deep boosti ng algorithms.

****

A Provable Generalized Tensor Spectral Method for Uniform Hypergraph Partitionin g

Debarghya Ghoshdastidar, Ambedkar Dukkipati

Matrix spectral methods play an important role in statistics and machine learnin g, and most often the word ‘matrix’ is dropped as, by default, one assumes that

similarities or affinities are measured between two points, thereby resulting in similarity matrices. However, recent challenges in computer vision and text mining have necessitated the use of multi-way affinities in the learning methods, and this has led to a considerable interest in hypergraph partitioning methods in machine learning community. A plethora of “higher-order” algorithms have been proposed in the past decade, but their theoretical guarantees are not well-studied. In this paper, we develop a unified approach for partitioning uniform hypergraphs by means of a tensor trace optimization problem involving the affinity tensor, and a number of existing higher-order methods turn out to be special cases of the proposed formulation. We further propose an algorithm to solve the proposed trace optimization problem, and prove that it is consistent under a planted hypergraph model. We also provide experimental results to validate our theoretical findings.

****

The Benefits of Learning with Strongly Convex Approximate Inference

Ben London, Bert Huang, Lise Getoor

We explore the benefits of strongly convex free energies in variational inference, providing both theoretical motivation and a new meta-algorithm. Using the duality between strong convexity and stability, we prove a high-probability bound on the error of learned marginals that is inversely proportional to the modulus of convexity of the free energy, thereby motivating free energies whose moduli are constant with respect to the size of the graph. We identify sufficient conditions for $\Omega(1)$-strong convexity in two popular variational techniques: tree-reweighted and counting number entropies. Our insights for the latter suggest a novel counting number optimization framework, which guarantees strong convexity for any given modulus. Our experiments demonstrate that learning with a strongly convex free energy, using our optimization framework to guarantee a given modulus, results in substantially more accurate marginal probabilities, thereby validating our theoretical claims and the effectiveness of our framework.

****

Pushing the Limits of Affine Rank Minimization by Adapting Probabilistic PCA

Bo Xin, David Wipf

Many applications require recovering a matrix of minimal rank within an affine constraint set, with matrix completion a notable special case. Because the problem is NP-hard in general, it is common to replace the matrix rank with the nuclear norm, which acts as a convenient convex surrogate. While elegant theoretical conditions elucidate when this replacement is likely to be successful, they are highly restrictive and convex algorithms fail when the ambient rank is too high or when the constraint set is poorly structured. Non-convex alternatives fare somewhat better when carefully tuned; however, convergence to locally optimal solutions remains a continuing source of failure. Against this backdrop we derive a deceptively simple and parameter-free probabilistic PCA-like algorithm that is capable, over a wide battery of empirical tests, of successful recovery even at the theoretical limit where the number of measurements equals the degrees of freedom in the unknown low-rank matrix. Somewhat surprisingly, this is possible even when the affine constraint set is highly ill-conditioned. While proving general recovery guarantees remains evasive for non-convex algorithms, Bayesian-inspired or otherwise, we nonetheless show conditions whereby the underlying cost function has a unique stationary point located at the global optimum; no existing cost function we are aware of satisfies this property. The algorithm has also been successfully deployed on a computer vision application involving image rectification and a standard collaborative filtering benchmark.

****

Budget Allocation Problem with Multiple Advertisers: A Game Theoretic View

Takanori Maehara, Akihiro Yabe, Ken-ichi Kawarabayashi

In marketing planning, advertisers seek to maximize the number of customers by allocating given budgets to each media channel effectively. The budget allocation problem with a bipartite influence model captures this scenario; however, the model is problematic because it assumes there is only one advertiser in the market. In reality, there are many advertisers which are in conflict of advertisement

; thus we must extend the model for such a case. By extending the budget allocat ion problem with a bipartite influence model, we propose a game-theoretic model problem that considers many advertisers. By simulating our model, we can analyze the behavior of a media channel market, e.g., we can estimate which media chann els are allocated by an advertiser, and which customers are influenced by an adv ertiser. Our model has many attractive features. First, our model is a potential game; therefore, it has a pure Nash equilibrium. Second, any Nash equilibrium o f our game has 2-optimal social utility, i.e., the price of anarchy is 2. Finall y, the proposed model can be simulated very efficiently; thus it can be used to analyze large markets.
****

Tracking Approximate Solutions of Parameterized Optimization Problems over Multi -Dimensional (Hyper-)Parameter Domains
Katharina Blechschmidt, Joachim Giesen, Soeren Laue
Many machine learning methods are given as parameterized optimization problems. Important examples of such parameters are regularization- and kernel hyperparame ters. These parameters have to be tuned carefully since the choice of their valu es can have a significant impact on the statistical performance of the learning methods. In most cases the parameter space does not carry much structure and par ameter tuning essentially boils down to exploring the whole parameter space. The case when there is only one parameter received quite some attention over the ye ars. First, algorithms for tracking an optimal solution for several machine lear ning optimization problems over regularization- and hyperparameter intervals had been developed, but since these algorithms can suffer from numerical problems m ore robust and efficient approximate path tracking algorithms have been devised and analyzed recently. By now approximate path tracking algorithms are known for regularization-and kernel hyperparameter paths with optimal path complexities t hat depend only on the prescribed approximation error. Here we extend the work o n approximate path tracking algorithms with approximation guarantees to multi-di mensional parameter domains. We show a lower bound on the complexity of approxim ately exploring a multi-dimensional parameter domain that is the product of the corresponding path complexities. We also show a matching upper bound that can be turned into a theoretically and practically efficient algorithm. Experimental r esults for kernelized support vector machines and the elastic net confirm the th eoretical complexity analysis.
****

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Cov ariate Shift
Sergey Ioffe, Christian Szegedy
Training Deep Neural Networks is complicated by the fact that the distribution o f each layer’s inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates a nd careful parameter initialization, and makes it notoriously hard to train mode ls with saturating nonlinearities. We refer to this phenomenon as internal covar iate shift, and address the problem by normalizing layer inputs. Our method draw s its strength from making normalization a part of the model architecture and pe rforming the normalization for each training mini-batch. Batch Normalization all ows us to use much higher learning rates and be less careful about initializatio n, and in some cases eliminates the need for Dropout. Applied to a stateof-the-a rt image classification model, Batch Normalization achieves the same accuracy wi th 14 times fewer training steps, and beats the original model by a significant margin. Using an ensemble of batch-normalized networks, we improve upon the best published result on ImageNet classification: reaching 4.82% top-5 test error, e xceeding the accuracy of human raters.
****

Distributed Estimation of Generalized Matrix Rank: Efficient Algorithms and Lowe r Bounds
Yuchen Zhang, Martin Wainwright, Michael Jordan
We study the following generalized matrix rank estimation problem: given an n-by -n matrix and a constant c > 0, estimate the number of eigenvalues that are grea

ter than c. In the distributed setting, the matrix of interest is the sum of m m atrices held by separate machines. We show that any deterministic algorithm solv ing this problem must communicate Ω(n^2) bits, which is order-equivalent to tran smitting the whole matrix. In contrast, we propose a randomized algorithm that c ommunicates only O(n) bits. The upper bound is matched by an Ω(n) lower bound on the randomized communication complexity. We demonstrate the practical effective ness of the proposed algorithm with some numerical experiments.

****

Landmarking Manifolds with Gaussian Processes

Dawen Liang, John Paisley

We present an algorithm for finding landmarks along a manifold. These landmarks provide a small set of locations spaced out along the manifold such that they ca pture the low-dimensional non-linear structure of the data embedded in the high-dimensional space. The approach does not select points directly from the dataset , but instead we optimize each landmark by moving along the continuous manifold space (as approximated by the data) according to the gradient of an objective fu nction. We borrow ideas from active learning with Gaussian processes to define t he objective, which has the property that a new landmark is "repelled" by those currently selected, allowing for exploration of the manifold. We derive a stocha stic algorithm for learning with large datasets and show results on several data sets, including the Million Song Dataset and articles from the New York Times.

****

Markov Mixed Membership Models

Aonan Zhang, John Paisley

We present a Markov mixed membership model (Markov M3) for grouped data that lea rns a fully connected graph structure among mixing components. A key feature of Markov M3 is that it interprets the mixed membership assignment as a Markov rand om walk over this graph of nodes. This is in contrast to tree-structured models in which the assignment is done according to a tree structure on the mixing comp onents. The Markov structure results in a simple parametric model that can learn a complex dependency structure between nodes, while still maintaining full conj ugacy for closed-form stochastic variational inference. Empirical results demons trate that Markov M3 performs well compared with tree structured topic models, a nd can learn meaningful dependency structure between topics.

****

A Unified Framework for Outlier-Robust PCA-like Algorithms

Wenzhuo Yang, Huan Xu

We propose a unified framework for making a wide range of PCA-like algorithms – including the standard PCA, sparse PCA and non-negative sparse PCA, etc. – robus t when facing a constant fraction of arbitrarily corrupted outliers. Our theoret ic analysis establishes solid performance guarantees of the proposed framework: its estimation error is upper bounded by a term depending on the intrinsic param eters of the data model, the selected PCA-like algorithm and the fraction of out liers. Comprehensive experiments on synthetic and real-world datasets demonstrat e that the outlier-robust PCA-like algorithms derived from our framework have ou tstanding performance.

****

Streaming Sparse Principal Component Analysis

Wenzhuo Yang, Huan Xu

This paper considers estimating the leading k principal components with at most s non-zero attributes from p-dimensional samples collected sequentially in memor y limited environments. We develop and analyze two memory and computational effi cient algorithms called streaming sparse PCA and streaming sparse ECA for analyz ing data generated according to the spike model and the elliptical model respect ively. In particular, the proposed algorithms have memory complexity O(pk), comp utational complexity O(pk mink,slogp) and sample complexity Θ(s \log p). We prov ide their finite sample performance guarantees, which implies statistical consis tency in the high dimensional regime. Numerical experiments on synthetic and rea l-world datasets demonstrate good empirical performance of the proposed algorith ms.

****
A Divide and Conquer Framework for Distributed Graph Clustering
Wenzhuo Yang, Huan Xu
Graph clustering is about identifying clusters of closely connected nodes, and is a fundamental technique of data analysis with many applications including community detection, VLSI network partitioning, collaborative filtering, and many others. In order to improve the scalability of existing graph clustering algorithms, we propose a novel divide and conquer framework for graph clustering, and establish theoretical guarantees of exact recovery of the clusters. One additional advantage of the proposed framework is that it can identify small clusters – the size of the smallest cluster can be of size o(\sqrtn), in contrast to Ω(\sqrtn) required by standard methods. Extensive experiments on synthetic and real-world datasets demonstrate the efficiency and effectiveness of our framework.
****
How Can Deep Rectifier Networks Achieve Linear Separability and Preserve Distances?
Senjian An, Farid Boussaid, Mohammed Bennamoun
This paper investigates how hidden layers of deep rectifier networks are capable of transforming two or more pattern sets to be linearly separable while preserving the distances with a guaranteed degree, and proves the universal classification power of such distance preserving rectifier networks. Through the nearly isometric nonlinear transformation in the hidden layers, the margin of the linear separating plane in the output layer and the margin of the nonlinear separating boundary in the original data space can be closely related so that the maximum margin classification in the input data space can be achieved approximately via the maximum margin linear classifiers in the output layer. The generalization performance of such distance preserving deep rectifier neural networks can be well justified by the distance-preserving properties of their hidden layers and the maximum margin property of the linear classifiers in the output layer.
****
Improved Regret Bounds for Undiscounted Continuous Reinforcement Learning
K. Lakshmanan, Ronald Ortner, Daniil Ryabko
We consider the problem of undiscounted reinforcement learning in continuous state space. Regret bounds in this setting usually hold under various assumptions on the structure of the reward and transition function. Under the assumption that the rewards and transition probabilities are Lipschitz, for 1-dimensional state space a regret bound of O(T^3/4) after any T steps has been given by Ortner and Ryabko (2012). Here we improve upon this result by using non-parametric kernel density estimation for estimating the transition probability distributions, and obtain regret bounds that depend on the smoothness of the transition probability distributions. In particular, under the assumption that the transition probability functions are smoothly differentiable, the regret bound is shown to be O(T^2/3) asymptotically for reinforcement learning in 1-dimensional state space. Finally, we also derive improved regret bounds for higher dimensional state space.
****
The Fundamental Incompatibility of Scalable Hamiltonian Monte Carlo and Naive Data Subsampling
Michael Betancourt
Leveraging the coherent exploration of Hamiltonian flow, Hamiltonian Monte Carlo produces computationally efficient Monte Carlo estimators, even with respect to complex and high-dimensional target distributions. When confronted with data-intensive applications, however, the algorithm may be too expensive to implement, leaving us to consider the utility of approximations such as data subsampling. In this paper I demonstrate how data subsampling fundamentally compromises the scalability of Hamiltonian Monte Carlo.
****
Faster Rates for the Frank-Wolfe Method over Strongly-Convex Sets
Dan Garber, Elad Hazan
The Frank-Wolfe method (a.k.a. conditional gradient algorithm) for smooth optimization has regained much interest in recent years in the context of large scale

optimization and machine learning. A key advantage of the method is that it avoi ds projections - the computational bottleneck in many applications - replacing i t by a linear optimization step. Despite this advantage, the known convergence r ates of the FW method fall behind standard first order methods for most settings of interest. It is an active line of research to derive faster linear optimizat ion-based algorithms for various settings of convex optimization. In this paper we consider the special case of optimization over strongly convex sets, for whic h we prove that the vanila FW method converges at a rate of \frac1t^2. This give s a quadratic improvement in convergence rate compared to the general case, in w hich convergence is of the order \frac1t, and known to be tight. We show that va rious balls induced by \ell_p norms, Schatten norms and group norms are strongly convex on one hand and on the other hand, linear optimization over these sets i s straightforward and admits a closed-form solution. We further show how several previous fast-rate results for the FW method follow easily from our analysis.
****

Ordered Stick-Breaking Prior for Sequential MCMC Inference of Bayesian Nonparame tric Models

Mrinal Das, Trapit Bansal, Chiranjib Bhattacharyya

This paper introduces ordered stick-breaking process (OSBP), where the atoms in a stick-breaking process (SBP) appear in order. The choice of weights on the ato ms of OSBP ensure that; (1) probability of adding new atoms exponentially decrea se, and (2) OSBP, though non-exchangeable, admit predictive probability function s (PPFs). In a Bayesian nonparametric (BNP) setting, OSBP serves as a natural pr ior over sequential mini-batches, facilitating exchange of relevant statistical information by sharing the atoms of OSBP. One of the major contributions of this paper is SUMO, an MCMC algorithm, for solving the inference problem arising fro m applying OSBP to BNP models. SUMO uses the PPFs of OSBP to obtain a Gibbs-samp ling based truncation-free algorithm which applies generally to BNP models. For large scale inference problems existing algorithms such as particle filtering (P F) are not practical and variational procedures such as TSVI (Wang & Blei, 2012) are the only alternative. For Dirichlet process mixture model (DPMM), SUMO outp erforms TSVI on perplexity by 33% on 3 datasets with million data points, which are beyond the scope of PF, using only 3GB RAM.
****

Online Learning of Eigenvectors

Dan Garber, Elad Hazan, Tengyu Ma

Computing the leading eigenvector of a symmetric real matrix is a fundamental pr imitive of numerical linear algebra with numerous applications. We consider a na tural online extension of the leading eigenvector problem: a sequence of matrice s is presented and the goal is to predict for each matrix a unit vector, with th e overall goal of competing with the leading eigenvector of the cumulative matri x. Existing regret-minimization algorithms for this problem either require to co mpute an \textiteigen decompostion every iteration, or suffer from a large depen dency of the regret bound on the dimension. In both cases the algorithms are not practical for large scale applications. In this paper we present new algorithms that avoid both issues. On one hand they do not require any expensive matrix de compositions and on the other, they guarantee regret rates with a mild dependenc e on the dimension at most. In contrast to previous algorithms, our algorithms a lso admit implementations that enable to leverage sparsity in the data to furthe r reduce computation. We extend our results to also handle non-symmetric matrice s.
****

A Unifying Framework of Anytime Sparse Gaussian Process Regression Models with S tochastic Variational Inference for Big Data

Trong Nghia Hoang, Quang Minh Hoang, Bryan Kian Hsiang Low

This paper presents a novel unifying framework of anytime sparse Gaussian proces s regression (SGPR) models that can produce good predictive performance fast and improve their predictive performance over time. Our proposed unifying framework reverses the variational inference procedure to theoretically construct a non-t rivial, concave functional that is maximized at the predictive distribution of a

ny SGPR model of our choice. As a result, a stochastic natural gradient ascent m
ethod can be derived that involves iteratively following the stochastic natural
gradient of the functional to improve its estimate of the predictive distributio
n of the chosen SGPR model and is guaranteed to achieve asymptotic convergence t
o it. Interestingly, we show that if the predictive distribution of the chosen S
GPR model satisfies certain decomposability conditions, then the stochastic natu
ral gradient is an unbiased estimator of the exact natural gradient and can be c
omputed in constant time (i.e., independent of data size) at each iteration. We
empirically evaluate the trade-off between the predictive performance vs. time e
fficiency of the anytime SGPR models on two real-world million-sized datasets.
****

Yinyang K-Means: A Drop-In Replacement of the Classic K-Means with Consistent Sp
eedup
Yufei Ding, Yue Zhao, Xipeng Shen, Madanlal Musuvathi, Todd Mytkowicz
This paper presents Yinyang K-means, a new algorithm for K-means clustering. By
clustering the centers in the initial stage, and leveraging efficiently maintain
ed lower and upper bounds between a point and centers, it more effectively avoid
s unnecessary distance calculations than prior algorithms. It significantly outp
erforms classic K-means and prior alternative K-means algorithms consistently ac
ross all experimented data sets, cluster numbers, and machine configurations. Th
e consistent, superior performance—plus its simplicity, user-control of overhead
s, and guarantee in producing the same clustering results as the standard K-mean
s does—makes Yinyang K-means a drop-in replacement of the classic K-means with a
n order of magnitude higher performance.
****

Ordinal Mixed Membership Models
Seppo Virtanen, Mark Girolami
We present a novel class of mixed membership models for joint distributions of g
roups of observations that co-occur with ordinal response variables for each gro
up for learning statistical associations between the ordinal response variables
and the observation groups. The class of proposed models addresses a requirement
 for predictive and diagnostic methods in a wide range of practical contemporary
 applications. In this work, by way of illustration, we apply the models to a co
llection of consumer-generated reviews of mobile software applications, where ea
ch review contains unstructured text data accompanied with an ordinal rating, an
d demonstrate that the models infer useful and meaningful recurring patterns of
consumer feedback. We also compare the developed models to relevant existing wor
ks, which rely on improper statistical assumptions for ordinal variables, showin
g significant improvements both in predictive ability and knowledge extraction.
****

Online Tracking by Learning Discriminative Saliency Map with Convolutional Neura
l Network
Seunghoon Hong, Tackgeun You, Suha Kwak, Bohyung Han
We propose an online visual tracking algorithm by learning discriminative salien
cy map using Convolutional Neural Network (CNN). Given a CNN pre-trained on a la
rge-scale image repository in offline, our algorithm takes outputs from hidden l
ayers of the network as feature descriptors since they show excellent representa
tion performance in various general visual recognition problems. The features ar
e used to learn discriminative target appearance models using an online Support
Vector Machine (SVM). In addition, we construct target-specific saliency map by
back-projecting CNN features with guidance of the SVM, and obtain the final trac
king result in each frame based on the appearance model generatively constructed
 with the saliency map. Since the saliency map reveals spatial configuration of
target effectively, it improves target localization accuracy and enables us to a
chieve pixel-level target segmentation. We verify the effectiveness of our track
ing algorithm through extensive experiment on a challenging benchmark, where our
 method illustrates outstanding performance compared to the state-of-the-art tra
cking algorithms.
****

Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods

Seth Flaxman, Andrew Wilson, Daniel Neill, Hannes Nickisch, Alex Smola
Gaussian processes (GPs) are a flexible class of methods with state of the art p
erformance on spatial statistics applications. However, GPs require O(n^3) compu
tations and O(n^2) storage, and popular GP kernels are typically limited to smoo
thing and interpolation. To address these difficulties, Kronecker methods have b
een used to exploit structure in the GP covariance matrix for scalability, while
 allowing for expressive kernel learning (Wilson et al., 2014). However, fast Kr
onecker methods have been confined to Gaussian likelihoods. We propose new scala
ble Kronecker methods for Gaussian processes with non-Gaussian likelihoods, usin
g a Laplace approximation which involves linear conjugate gradients for inferenc
e, and a lower bound on the GP marginal likelihood for kernel learning. Our appr
oach has near linear scaling, requiring O(D n^(D+1)/D) operations and O(D n^2/D)
 storage, for n training data-points on a dense D > 1 dimensional grid. Moreover
, we introduce a log Gaussian Cox process, with highly expressive kernels, for m
odelling spatiotemporal count processes, and apply it to a point pattern (n = 23
3,088) of a decade of crime events in Chicago. Using our model, we discover spat
ially varying multiscale seasonal trends and produce highly accurate long-range
local area forecasts.
****
Statistical and Algorithmic Perspectives on Randomized Sketching for Ordinary Le
ast-Squares
Garvesh Raskutti, Michael Mahoney
We consider statistical and algorithmic aspects of solving large-scale least-squ
ares (LS) problems using randomized sketching algorithms. Prior results show tha
t, from an \emphalgorithmic perspective, when using sketching matrices construct
ed from random projections and leverage-score sampling, if the number of samples
 r much smaller than the original sample size n, then the worst-case (WC) error
is the same as solving the original problem, up to a very small relative error.
From a \emphstatistical perspective, one typically considers the mean-squared er
ror performance of randomized sketching algorithms, when data are generated acco
rding to a statistical linear model. In this paper, we provide a rigorous compar
ison of both perspectives leading to insights on how they differ. To do this, we
 first develop a framework for assessing, in a unified manner, algorithmic and s
tatistical aspects of randomized sketching methods. We then consider the statist
ical prediction efficiency (PE) and the statistical residual efficiency (RE) of
the sketched LS estimator; and we use our framework to provide upper bounds for
several types of random projection and random sampling algorithms. Among other r
esults, we show that the RE can be upper bounded when r is much smaller than n,
while the PE typically requires the number of samples r to be substantially larg
er. Lower bounds developed in subsequent work show that our upper bounds on PE c
an not be improved.
****
On TD(0) with function approximation: Concentration bounds and a centered varian
t with exponential convergence
Nathaniel Korda, Prashanth La
We provide non-asymptotic bounds for the well-known temporal difference learning
 algorithm TD(0) with linear function approximators. These include high-probabil
ity bounds as well as bounds in expectation. Our analysis suggests that a step-s
ize inversely proportional to the number of iterations cannot guarantee optimal
rate of convergence unless we assume (partial) knowledge of the stationary distr
ibution for the Markov chain underlying the policy considered. We also provide b
ounds for the iterate averaged TD(0) variant, which gets rid of the step-size de
pendency while exhibiting the optimal rate of convergence. Furthermore, we propo
se a variant of TD(0) with linear approximators that incorporates a centering se
quence, and establish that it exhibits an exponential rate of convergence in exp
ectation. We demonstrate the usefulness of our bounds on two synthetic experimen
tal settings.
****
Learning Parametric-Output HMMs with Two Aliased States
Roi Weiss, Boaz Nadler

In various applications involving hidden Markov models (HMMs), some of the hidden states are aliased, having identical output distributions. The minimality, identifiability and learnability of such aliased HMMs have been long standing problems, with only partial solutions provided thus far. In this paper we focus on parametric-output HMMs, whose output distributions come from a parametric family, and that have exactly two aliased states. For this class, we present a complete characterization of their minimality and identifiability. Furthermore, for a large family of parametric output distributions, we derive computationally efficient and statistically consistent algorithms to detect the presence of aliasing and learn the aliased HMM transition and emission parameters. We illustrate our theoretical analysis by several simulations.

****

Latent Gaussian Processes for Distribution Estimation of Multivariate Categorical Data

Yarin Gal, Yutian Chen, Zoubin Ghahramani

Multivariate categorical data occur in many applications of machine learning. One of the main difficulties with these vectors of categorical variables is sparsity. The number of possible observations grows exponentially with vector length, but dataset diversity might be poor in comparison. Recent models have gained significant improvement in supervised tasks with this data. These models embed observations in a continuous space to capture similarities between them. Building on these ideas we propose a Bayesian model for the unsupervised task of distribution estimation of multivariate categorical data. We model vectors of categorical variables as generated from a non-linear transformation of a continuous latent space. Non-linearity captures multi-modality in the distribution. The continuous representation addresses sparsity. Our model ties together many existing models, linking the linear categorical latent Gaussian model, the Gaussian process latent variable model, and Gaussian process classification. We derive inference for our model based on recent developments in sampling based variational inference. We show empirically that the model outperforms its linear and discrete counterparts in imputation tasks of sparse data.

****

Improving the Gaussian Process Sparse Spectrum Approximation by Representing Uncertainty in Frequency Inputs

Yarin Gal, Richard Turner

Standard sparse pseudo-input approximations to the Gaussian process (GP) cannot handle complex functions well. Sparse spectrum alternatives attempt to answer this but are known to over-fit. We suggest the use of variational inference for the sparse spectrum approximation to avoid both issues. We model the covariance function with a finite Fourier series approximation and treat it as a random variable. The random covariance function has a posterior, on which a variational distribution is placed. The variational distribution transforms the random covariance function to fit the data. We study the properties of our approximate inference, compare it to alternative ones, and extend it to the distributed and stochastic domains. Our approximation captures complex functions better than standard approaches and avoids over-fitting.

****

Ranking from Stochastic Pairwise Preferences: Recovering Condorcet Winners and Tournament Solution Sets at the Top

Arun Rajkumar, Suprovat Ghoshal, Lek-Heng Lim, Shivani Agarwal

We consider the problem of ranking n items from stochastically sampled pairwise preferences. It was shown recently that when the underlying pairwise preferences are acyclic, several algorithms including the Rank Centrality algorithm, the Matrix Borda algorithm, and the SVM-RankAggregation algorithm succeed in recovering a ranking that minimizes a global pairwise disagreement error (Rajkumar and Agarwal, 2014). In this paper, we consider settings where pairwise preferences can contain cycles. In such settings, one may still like to be able to recover ‘good’ items at the top of the ranking. For example, if a Condorcet winner exists that beats every other item, it is natural to ask that this be ranked at the top. More generally, several tournament solution concepts such as the top cycle, Cope

land set, Markov set and others have been proposed in the social choice literatu re for choosing a set of winners in the presence of cycles. We show that existin g algorithms can fail to perform well in terms of ranking Condorcet winners and various natural tournament solution sets at the top. We then give alternative ra nking algorithms that provably rank Condorcet winners, top cycles, and other tou rnament solution sets of interest at the top. In all cases, we give finite sampl e complexity bounds for our algorithms to recover such winners. As a by-product of our analysis, we also obtain an improved sample complexity bound for the Rank Centrality algorithm to recover an optimal ranking under a Bradley-Terry-Luce (BTL) condition, which answers an open question of Rajkumar and Agarwal (2014).
****

Stochastic Dual Coordinate Ascent with Adaptive Probabilities
Dominik Csiba, Zheng Qu, Peter Richtarik
This paper introduces AdaSDCA: an adaptive variant of stochastic dual coordinate ascent (SDCA) for solving the regularized empirical risk minimization problems. Our modification consists in allowing the method adaptively change the probabil ity distribution over the dual variables throughout the iterative process. AdaSD CA achieves provably better complexity bound than SDCA with the best fixed proba bility distribution, known as importance sampling. However, it is of a theoretic al character as it is expensive to implement. We also propose AdaSDCA+: a practi cal variant which in our experiments outperforms existing non-adaptive methods.
****

Vector-Space Markov Random Fields via Exponential Families
Wesley Tansey, Oscar Hernan Madrid Padilla, Arun Sai Suggala, Pradeep Ravikumar
We present Vector-Space Markov Random Fields (VS-MRFs), a novel class of undirec ted graphical models where each variable can belong to an arbitrary vector space . VS-MRFs generalize a recent line of work on scalar-valued, uni-parameter expon ential family and mixed graphical models, thereby greatly broadening the class o f exponential families available (e.g., allowing multinomial and Dirichlet distr ibutions). Specifically, VS-MRFs are the joint graphical model distributions whe re the node-conditional distributions belong to generic exponential families wit h general vector space domains. We also present a sparsistent M-estimator for le arning our class of MRFs that recovers the correct set of edges with high probab ility. We validate our approach via a set of synthetic data experiments as well as a real-world case study of over four million foods from the popular diet trac king app MyFitnessPal. Our results demonstrate that our algorithm performs well empirically and that VS-MRFs are capable of capturing and highlighting interesti ng structure in complex, real-world data. All code for our algorithm is open sou rce and publicly available.
****

JUMP-Means: Small-Variance Asymptotics for Markov Jump Processes
Jonathan Huggins, Karthik Narasimhan, Ardavan Saeedi, Vikash Mansinghka
Markov jump processes (MJPs) are used to model a wide range of phenomenon from d isease progression to RNA path folding. However, existing methods suffer from a number of shortcomings: degenerate trajectories in the case of ML estimation of parametric models and poor inferential performance in the case of nonparametric models. We take a small-variance asymptotics (SVA) approach to overcome these li mitations. We derive the small-variance asymptotics for parametric and nonparame tric MJPs for both directly observed and hidden state models. In the parametric case we obtain a novel objective function which leads to non-degenerate trajecto ries. To derive the nonparametric version we introduce the gamma-gamma process, a novel extension to the gamma-exponential process. We propose algorithms for ea ch of these formulations, which we call \emphJUMP-means. Our experiments demonst rate that JUMP-means is competitive with or outperforms widely used MJP inferenc e approaches in terms of both speed and reconstruction accuracy.
****

Low Rank Approximation using Error Correcting Coding Matrices
Shashanka Ubaru, Arya Mazumdar, Yousef Saad
Low-rank matrix approximation is an integral component of tools such as principa l component analysis (PCA), as well as is an important instrument used in applic

ations like web search models, text mining and computer vision, e.g., face recog nition. Recently, randomized algorithms were proposed to effectively construct l ow rank approximations of large matrices. In this paper, we show how matrices fr om error correcting codes can be used to find such low rank approximations. The benefits of using these code matrices are the following: (i) They are easy to ge nerate and they reduce randomness significantly. (ii) Code matrices have low coh erence and have a better chance of preserving the geometry of an entire subspace of vectors; (iii) Unlike Fourier transforms or Hadamard matrices, which require sampling O(k\log k) columns for a rank-k approximation, the log factor is not n ecessary in the case of code matrices. (iv) Under certain conditions, the approx imation errors can be better and the singular values obtained can be more accura te, than those obtained using Gaussian random matrices and other structured rand om matrices.

****

Off-policy Model-based Learning under Unknown Factored Dynamics

Assaf Hallak, Francois Schnitzler, Timothy Mann, Shie Mannor

Off-policy learning in dynamic decision problems is essential for providing stro ng evidence that a new policy is better than the one in use. But how can we prov e superiority without testing the new policy? To answer this question, we introd uce the G-SCOPE algorithm that evaluates a new policy based on data generated by the existing policy. Our algorithm is both computationally and sample efficient because it greedily learns to exploit factored structure in the dynamics of the environment. We present a finite sample analysis of our approach and show throu gh experiments that the algorithm scales well on high-dimensional problems with few samples.

****

Log-Euclidean Metric Learning on Symmetric Positive Definite Manifold with Appli cation to Image Set Classification

Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xianqiu Li, Xilin Chen

The manifold of Symmetric Positive Definite (SPD) matrices has been successfully used for data representation in image set classification. By endowing the SPD m anifold with Log-Euclidean Metric, existing methods typically work on vector-for ms of SPD matrix logarithms. This however not only inevitably distorts the geome trical structure of the space of SPD matrix logarithms but also brings low effic iency especially when the dimensionality of SPD matrix is high. To overcome this limitation, we propose a novel metric learning approach to work directly on log arithms of SPD matrices. Specifically, our method aims to learn a tangent map th at can directly transform the matrix logarithms from the original tangent space to a new tangent space of more discriminability. Under the tangent map framework , the novel metric learning can then be formulated as an optimization problem of seeking a Mahalanobis-like matrix, which can take the advantage of traditional metric learning techniques. Extensive evaluations on several image set classific ation tasks demonstrate the effectiveness of our proposed metric learning method .

****

Asymmetric Transfer Learning with Deep Gaussian Processes

Melih Kandemir

We introduce a novel Gaussian process based Bayesian model for asymmetric transf er learning. We adopt a two-layer feed-forward deep Gaussian process as the task learner of source and target domains. The first layer projects the data onto a separate non-linear manifold for each task. We perform knowledge transfer by pro jecting the target data also onto the source domain and linearly combining its r epresentations on the source and target domain manifolds. Our approach achieves the state-of-the-art in a benchmark real-world image categorization task, and im proves on it in cross-tissue tumor detection from histopathology tissue slide im ages.

****

Towards a Lower Sample Complexity for Robust One-bit Compressed Sensing

Rongda Zhu, Quanquan Gu

In this paper, we propose a novel algorithm based on nonconvex sparsity-inducing

penalty for one-bit compressed sensing. We prove that our algorithm has a sampl
e complexity of O(s/ε^2) for strong signals, and O(s\log d/ε^2) for weak signals
, where s is the number of nonzero entries in the signal vector, d is the signal
 dimension and εis the recovery error. For general signals, the sample complexit
y of our algorithm lies between O(s/ε^2) and O(s\log d/ε^2). This is a remarkabl
e improvement over the existing best sample complexity O(s\log d/ε^2). Furthermo
re, we show that our algorithm achieves exact support recovery with high probabi
lity for strong signals. Our theory is verified by extensive numerical experimen
ts, which clearly illustrate the superiority of our algorithm for both approxima
te signal and support recovery in the noisy setting.
****

BilBOWA: Fast Bilingual Distributed Representations without Word Alignments
Stephan Gouws, Yoshua Bengio, Greg Corrado
We introduce BilBOWA (Bilingual Bag-of-Words without Alignments), a simple and c
omputationally-efficient model for learning bilingual distributed representation
s of words which can scale to large monolingual datasets and does not require wo
rd-aligned parallel training data. Instead it trains directly on monolingual dat
a and extracts a bilingual signal from a smaller set of raw-text sentence-aligne
d data. This is achieved using a novel sampled bag-of-words cross-lingual object
ive, which is used to regularize two noise-contrastive language models for effic
ient cross-lingual feature learning. We show that bilingual embeddings learned u
sing the proposed model outperforms state-of-the-art methods on a cross-lingual
document classification task as well as a lexical translation task on the WMT11
data.
****

Multi-view Sparse Co-clustering via Proximal Alternating Linearized Minimization
Jiangwen Sun, Jin Lu, Tingyang Xu, Jinbo Bi
When multiple views of data are available for a set of subjects, co-clustering a
ims to identify subject clusters that agree across the different views. We explo
re the problem of co-clustering when the underlying clusters exist in different
subspaces of each view. We propose a proximal alternating linearized minimizatio
n algorithm that simultaneously decomposes multiple data matrices into sparse ro
w and columns vectors. This approach is able to group subjects consistently acro
ss the views and simultaneously identify the subset of features in each view tha
t are associated with the clusters. The proposed algorithm can globally converge
 to a critical point of the problem. A simulation study validates that the propo
sed algorithm can identify the hypothesized clusters and their associated featur
es. Comparison with several latest multi-view co-clustering methods on benchmark
 datasets demonstrates the superior performance of the proposed approach.
****

Cascading Bandits: Learning to Rank in the Cascade Model
Branislav Kveton, Csaba Szepesvari, Zheng Wen, Azin Ashkan
A search engine usually outputs a list of K web pages. The user examines this li
st, from the first web page to the last, and chooses the first attractive page.
This model of user behavior is known as the cascade model. In this paper, we pro
pose cascading bandits, a learning variant of the cascade model where the object
ive is to identify K most attractive items. We formulate our problem as a stocha
stic combinatorial partial monitoring problem. We propose two algorithms for sol
ving it, CascadeUCB1 and CascadeKL-UCB. We also prove gap-dependent upper bounds
 on the regret of these algorithms and derive a lower bound on the regret in cas
cading bandits. The lower bound matches the upper bound of CascadeKL-UCB up to a
 logarithmic factor. We experiment with our algorithms on several problems. The
algorithms perform surprisingly well even when our modeling assumptions are viol
ated.
****

Latent Topic Networks: A Versatile Probabilistic Programming Framework for Topic
 Models
James Foulds, Shachi Kumar, Lise Getoor
Topic models have become increasingly prominent text-analytic machine learning t
ools for research in the social sciences and the humanities. In particular, cust

om topic models can be developed to answer specific research questions. The desi gn of these models requires a non-trivial amount of effort and expertise, motiva ting general-purpose topic modeling frameworks. In this paper we introduce laten t topic networks, a flexible class of richly structured topic models designed to facilitate applied research. Custom models can straightforwardly be developed i n our framework with an intuitive first-order logical probabilistic programming language. Latent topic networks admit scalable training via a parallelizable EM algorithm which leverages ADMM in the M-step. We demonstrate the broad applicabi lity of the models with case studies on modeling influence in citation networks, and U.S. Presidential State of the Union addresses.

****

Random Coordinate Descent Methods for Minimizing Decomposable Submodular Functio ns

Alina Ene, Huy Nguyen

Submodular function minimization is a fundamental optimization problem that aris es in several applications in machine learning and computer vision. The problem is known to be solvable in polynomial time, but general purpose algorithms have high running times and are unsuitable for large-scale problems. Recent work have used convex optimization techniques to obtain very practical algorithms for min imizing functions that are sums of “simple” functions. In this paper, we use ran dom coordinate descent methods to obtain algorithms with faster \emphlinear conv ergence rates and cheaper iteration costs. Compared to alternating projection me thods, our algorithms do not rely on full-dimensional vector operations and they converge in significantly fewer iterations.

****

Alpha-Beta Divergences Discover Micro and Macro Structures in Data

Karthik Narayan, Ali Punjani, Pieter Abbeel

Although recent work in non-linear dimensionality reduction investigates multipl e choices of divergence measure during optimization \citeyang2013icml,bunte2012n euro, little work discusses the direct effects that divergence measures have on visualization. We study this relationship, theoretically and through an empirica l analysis over 10 datasets. Our works shows how the αand βparameters of the gen eralized alpha-beta divergence can be chosen to discover hidden macro-structures (categories, e.g. birds) or micro-structures (fine-grained classes, e.g. toucan s). Our method, which generalizes t-SNE \citetsne, allows us to discover such st ructure without extensive grid searches over (α, β) due to our theoretical analy sis: such structure is apparent with particular choices of (α, β) that generaliz e across datasets. We also discuss efficient parallel CPU and GPU schemes which are non-trivial due to the tree-structures employed in optimization and the larg e datasets that do not fully fit into GPU memory. Our method runs 20x faster tha n the fastest published code \citefmm. We conclude with detailed case studies on the following very large datasets: ILSVRC 2012, a standard computer vision data set with 1.2M images; SUSY, a particle physics dataset with 5M instances; and HI GGS, another particle physics dataset with 11M instances. This represents the la rgest published visualization attained by SNE methods. We have open-sourced our visualization code: \texttthttp://rll.berkeley.edu/absne/.

****

Fictitious Self-Play in Extensive-Form Games

Johannes Heinrich, Marc Lanctot, David Silver

Fictitious play is a popular game-theoretic model of learning in games. However, it has received little attention in practical applications to large problems. T his paper introduces two variants of fictitious play that are implemented in beh avioural strategies of an extensive-form game. The first variant is a full-width process that is realization equivalent to its normal-form counterpart and there fore inherits its convergence guarantees. However, its computational requirement s are linear in time and space rather than exponential. The second variant, Fict itious Self-Play, is a machine learning framework that implements fictitious pla y in a sample-based fashion. Experiments in imperfect-information poker games co mpare our approaches and demonstrate their convergence to approximate Nash equil ibria.

****
Counterfactual Risk Minimization: Learning from Logged Bandit Feedback
Adith Swaminathan, Thorsten Joachims
We develop a learning principle and an efficient algorithm for batch learning fr
om logged bandit feedback. This learning setting is ubiquitous in online systems
 (e.g., ad placement, web search, recommendation), where an algorithm makes a pr
ediction (e.g., ad ranking) for a given input (e.g., query) and observes bandit
feedback (e.g., user clicks on presented ads). We first address the counterfactu
al nature of the learning problem through propensity scoring. Next, we prove gen
eralization error bounds that account for the variance of the propensity-weighte
d empirical risk estimator. These constructive bounds give rise to the Counterfa
ctual Risk Minimization (CRM) principle. We show how CRM can be used to derive a
 new learning method – called Policy Optimizer for Exponential Models (POEM) – f
or learning stochastic linear rules for structured output prediction. We present
 a decomposition of the POEM objective that enables efficient stochastic gradien
t optimization. POEM is evaluated on several multi-label classification problems
 showing substantially improved robustness and generalization performance compar
ed to the state-of-the-art.
****
The Hedge Algorithm on a Continuum
Walid Krichene, Maximilian Balandat, Claire Tomlin, Alexandre Bayen
We consider an online optimization problem on a subset S of R^n (not necessarily
 convex), in which a decision maker chooses, at each iteration t, a probability
distribution x^(t) over S, and seeks to minimize a cumulative expected loss, whe
re each loss is a Lipschitz function revealed at the end of iteration t. Buildin
g on previous work, we propose a generalized Hedge algorithm and show a O(\sqrtt
 \log t) bound on the regret when the losses are uniformly Lipschitz and S is un
iformly fat (a weaker condition than convexity). Finally, we propose a generaliz
ation to the dual averaging method on the set of Lebesgue-continuous distributio
ns over S.
****
A Linear Dynamical System Model for Text
David Belanger, Sham Kakade
Low dimensional representations of words allow accurate NLP models to be trained
 on limited annotated data. While most representations ignore words’ local conte
xt, a natural way to induce context-dependent representations is to perform infe
rence in a probabilistic latent-variable sequence model. Given the recent succes
s of continuous vector space word representations, we provide such an inference
procedure for continuous states, where words’ representations are given by the p
osterior mean of a linear dynamical system. Here, efficient inference can be per
formed using Kalman filtering. Our learning algorithm is extremely scalable, ope
rating on simple co-occurrence counts for both parameter initialization using th
e method of moments and subsequent iterations of EM. In our experiments, we empl
oy our inferred word embeddings as features in standard tagging tasks, obtaining
 significant accuracy improvements. Finally, the Kalman filter updates can be se
en as a linear recurrent neural network. We demonstrate that using the parameter
s of our model to initialize a non-linear recurrent neural network language mode
l reduces its training time by a day and yields lower perplexity.
****
Unsupervised Learning of Video Representations using LSTMs
Nitish Srivastava, Elman Mansimov, Ruslan Salakhudinov
We use Long Short Term Memory (LSTM) networks to learn representations of video
sequences. Our model uses an encoder LSTM to map an input sequence into a fixed
length representation. This representation is decoded using single or multiple d
ecoder LSTMs to perform different tasks, such as reconstructing the input sequen
ce, or predicting the future sequence. We experiment with two kinds of input seq
uences – patches of image pixels and high-level representations (“percepts") of
video frames extracted using a pretrained convolutional net. We explore differen
t design choices such as whether the decoder LSTMs should condition on the gener
ated output. We analyze the outputs of the model qualitatively to see how well t

he model can extrapolate the learned video representation into the future and in to the past. We further evaluate the representations by finetuning them for a su pervised learning problem – human action recognition on the UCF-101 and HMDB-51 datasets. We show that the representations help improve classification accuracy, especially when there are only few training examples. Even models pretrained on unrelated datasets (300 hours of YouTube videos) can help action recognition pe rformance.

****

Message Passing for Collective Graphical Models

Tao Sun, Dan Sheldon, Akshat Kumar

Collective graphical models (CGMs) are a formalism for inference and learning ab out a population of independent and identically distributed individuals when onl y noisy aggregate data are available. We highlight a close connection between ap proximate MAP inference in CGMs and marginal inference in standard graphical mod els. The connection leads us to derive a novel Belief Propagation (BP) style alg orithm for collective graphical models. Mathematically, the algorithm is a stric t generalization of BP—it can be viewed as an extension to minimize the Bethe fr ee energy plus additional energy terms that are non-linear functions of the marg inals. For CGMs, the algorithm is much more efficient than previous approaches t o inference. We demonstrate its performance on two synthetic experiments concern ing bird migration and collective human mobility.

****

DP-space: Bayesian Nonparametric Subspace Clustering with Small-variance Asympto tics

Yining Wang, Jun Zhu

Subspace clustering separates data points approximately lying on union of affine subspaces into several clusters. This paper presents a novel nonparametric Baye sian subspace clustering model that infers both the number of subspaces and the dimension of each subspace from the observed data. Though the posterior inferenc e is hard, our model leads to a very efficient deterministic algorithm, DP-space , which retains the nonparametric ability under a small-variance asymptotic anal ysis. DP-space monotonically minimizes an intuitive objective with an explicit t radeoff between data fitness and model complexity. Experimental results demonstr ate that DP-space outperforms various competitors in terms of clustering accurac y and at the same time it is highly efficient.

****

HawkesTopic: A Joint Model for Network Inference and Topic Modeling from Text-Ba sed Cascades

Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor, Yan Liu

Understanding the diffusion of information in social network and social media re quires modeling the text diffusion process. In this work, we develop the HawkesT opic model (HTM) for analyzing text-based cascades, such as "retweeting a post" or "publishing a follow-up blog post". HTM combines Hawkes processes and topic m odeling to simultaneously reason about the information diffusion pathways and th e topics characterizing the observed textual information. We show how to jointly infer them with a mean-field variational inference algorithm and validate our a pproach on both synthetic and real-world data sets, including a news media datas et for modeling information diffusion, and an ArXiv publication dataset for mode ling scientific influence. The results show that HTM is significantly more accur ate than several baselines for both tasks.

****

MADE: Masked Autoencoder for Distribution Estimation

Mathieu Germain, Karol Gregor, Iain Murray, Hugo Larochelle

There has been a lot of recent interest in designing neural network models to es timate a distribution from a set of examples. We introduce a simple modification for autoencoder neural networks that yields powerful generative models. Our met hod masks the autoencoder’s parameters to respect autoregressive constraints: ea ch input is reconstructed only from previous inputs in a given ordering. Constra ined this way, the autoencoder outputs can be interpreted as a set of conditiona l probabilities, and their product, the full joint probability. We can also trai

n a single network that can decompose the joint probability in multiple differen t orderings. Our simple framework can be applied to multiple architectures, incl uding deep ones. Vectorized implementations, such as on GPUs, are simple and fas t. Experiments demonstrate that this approach is competitive with state-of-the-a rt tractable distribution estimators. At test time, the method is significantly faster and scales better than other autoregressive estimators.
****

An Online Learning Algorithm for Bilinear Models
Yuanbin Wu, Shiliang Sun
We investigate the bilinear model, which is a matrix form linear model with the rank 1 constraint. A new online learning algorithm is proposed to train the mode l parameters. Our algorithm runs in the manner of online mirror descent, and gra dients are computed by the power iteration. To analyze it, we give a new second order approximation of the squared spectral norm, which helps us to get a regret bound. Experiments on two sequential labelling tasks give positive results.
****

Adaptive Belief Propagation
Georgios Papachristoudis, John Fisher
Graphical models are widely used in inference problems. In practice, one may con struct a single large-scale model to explain a phenomenon of interest, which may be utilized in a variety of settings. The latent variables of interest, which c an differ in each setting, may only represent a small subset of all variables. T he marginals of variables of interest may change after the addition of measureme nts at different time points. In such adaptive settings, naive algorithms, such as standard belief propagation (BP), may utilize many unnecessary computations b y propagating messages over the entire graph. Here, we formulate an efficient in ference procedure, termed adaptive BP (AdaBP), suitable for adaptive inference s ettings. We show that it gives exact results for trees in discrete and Gaussian Markov Random Fields (MRFs), and provide an extension to Gaussian loopy graphs. We also provide extensions on finding the most likely sequence of the entire lat ent graph. Lastly, we compare the proposed method to standard BP and to that of (Sumer et al., 2011), which tackles the same problem. We show in synthetic and r eal experiments that it outperforms standard BP by orders of magnitude and explo re the settings that it is advantageous over (Sumer et al., 2011).
****

Large-scale log-determinant computation through stochastic Chebyshev expansions
Insu Han, Dmitry Malioutov, Jinwoo Shin
Logarithms of determinants of large positive definite matrices appear ubiquitous ly in machine learning applications including Gaussian graphical and Gaussian pr ocess models, partition functions of discrete graphical models, minimum-volume e llipsoids and metric and kernel learning. Log-determinant computation involves t he Cholesky decomposition at the cost cubic in the number of variables (i.e., th e matrix dimension), which makes it prohibitive for large-scale applications. We propose a linear-time randomized algorithm to approximate log-determinants for very large-scale positive definite and general non-singular matrices using a sto chastic trace approximation, called the Hutchinson method, coupled with Chebyshe v polynomial expansions that both rely on efficient matrix-vector multiplication s. We establish rigorous additive and multiplicative approximation error bounds depending on the condition number of the input matrix. In our experiments, the p roposed algorithm can provide very high accuracy solutions at orders of magnitud e faster time than the Cholesky decomposition and Shur completion, and enables u s to compute log-determinants of matrices involving tens of millions of variable s.
****

Differentially Private Bayesian Optimization
Matt Kusner, Jacob Gardner, Roman Garnett, Kilian Weinberger
Bayesian optimization is a powerful tool for fine-tuning the hyper-parameters of a wide variety of machine learning models. The success of machine learning has led practitioners in diverse real-world settings to learn classifiers for practi cal problems. As machine learning becomes commonplace, Bayesian optimization bec

omes an attractive method for practitioners to automate the process of classifie r hyper-parameter tuning. A key observation is that the data used for tuning mod els in these settings is often sensitive. Certain data such as genetic predispos ition, personal email statistics, and car accident history, if not properly priv ate, may be at risk of being inferred from Bayesian optimization outputs. To add ress this, we introduce methods for releasing the best hyper-parameters and clas sifier accuracy privately. Leveraging the strong theoretical guarantees of diffe rential privacy and known Bayesian optimization convergence bounds, we prove tha t under a GP assumption these private quantities are often near-optimal. Finally , even if this assumption is not satisfied, we can use different smoothness guar antees to protect privacy.
****

A Nearly-Linear Time Framework for Graph-Structured Sparsity
Chinmay Hegde, Piotr Indyk, Ludwig Schmidt
We introduce a framework for sparsity structures defined via graphs. Our approac h is flexible and generalizes several previously studied sparsity models. Moreov er, we provide efficient projection algorithms for our sparsity model that run i n nearly-linear time. In the context of sparse recovery, we show that our framew ork achieves an information-theoretically optimal sample complexity for a wide r ange of parameters. We complement our theoretical analysis with experiments demo nstrating that our algorithms improve on prior work also in practice.
****

Support Matrix Machines
Luo Luo, Yubo Xie, Zhihua Zhang, Wu-Jun Li
In many classification problems such as electroencephalogram (EEG) classificatio n and image classification, the input features are naturally represented as matr ices rather than vectors or scalars. In general, the structure information of th e original feature matrix is useful and informative for data analysis tasks such as classification. One typical structure information is the correlation between columns or rows in the feature matrix. To leverage this kind of structure infor mation, we propose a new classification method that we call support matrix machi ne (SMM). Specifically, SMM is defined as a hinge loss plus a so-called spectral elastic net penalty which is a spectral extension of the conventional elastic n et over a matrix. The spectral elastic net enjoys a property of grouping effect, i.e., strongly correlated columns or rows tend to be selected altogether or not . Since the optimization problem for SMM is convex, this encourages us to devise an alternating direction method of multipliers algorithm for solving the proble m. Experimental results on EEG and face image classification data show that our model is more robust and efficient than the state-of-the-art methods.
****

Rademacher Observations, Private Data, and Boosting
Richard Nock, Giorgio Patrini, Arik Friedman
The minimization of the logistic loss is a popular approach to batch supervised learning. Our paper starts from the surprising observation that, when fitting li near classifiers, the minimization of the logistic loss is \textitequivalent to the minimization of an exponential \textitrado-loss computed (i) over transforme d data that we call Rademacher observations (rados), and (ii) over the \textitsa me classifier as the one of the logistic loss. Thus, a classifier learnt from ra dos can be \textitdirectly used to classify \textitobservations. We provide a le arning algorithm over rados with boosting-compliant convergence rates on the \te xtitlogistic loss (computed over examples). Experiments on domains with up to mi llions of examples, backed up by theoretical arguments, display that learning ov er a small set of random rados can challenge the state of the art that learns ov er the \textitcomplete set of examples. We show that rados comply with various p rivacy requirements that make them good candidates for machine learning in a pri vacy framework. We give several algebraic, geometric and computational hardness results on reconstructing examples from rados. We also show how it is possible t o craft, and efficiently learn from, rados in a differential privacy framework. Tests reveal that learning from differentially private rados brings non-trivial privacy vs accuracy tradeoffs.

****
From Word Embeddings To Document Distances
Matt Kusner, Yu Sun, Nicholas Kolkin, Kilian Weinberger
We present the Word Mover’s Distance (WMD), a novel distance function between te xt documents. Our work is based on recent results in word embeddings that learn semantically meaningful representations for words from local co-occurrences in s entences. The WMD distance measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document. We show that this d istance metric can be cast as an instance of the Earth Mover’s Distance, a well studied transportation problem for which several highly efficient solvers have b een developed. Our metric has no hyperparameters and is straight-forward to impl ement. Further, we demonstrate on eight real world document classification data sets, in comparison with seven state-of-the-art baselines, that the WMD metric l eads to unprecedented low k-nearest neighbor document classification error rates .
****
Bayesian and Empirical Bayesian Forests
Taddy Matthew, Chun-Sheng Chen, Jun Yu, Mitch Wyle
We derive ensembles of decision trees through a nonparametric Bayesian model, al lowing us to view such ensembles as samples from a posterior distribution. This insight motivates a class of Bayesian Forest (BF) algorithms that provide small gains in performance and large gains in interpretability. Based on the BF framew ork, we are able to show that high-level tree hierarchy is stable in large sampl es. This motivates an empirical Bayesian Forest (EBF) algorithm for building app roximate BFs on massive distributed datasets and we show that EBFs outperform su b-sampling based alternatives by a large margin.
****
Inferring Graphs from Cascades: A Sparse Recovery Framework
Jean Pouget-Abadie, Thibaut Horel
In the Graph Inference problem, one seeks to recover the edges of an unknown gra ph from the observations of cascades propagating over this graph. In this paper, we approach this problem from the sparse recovery perspective. We introduce a g eneral model of cascades, including the voter model and the independent cascade model, for which we provide the first algorithm which recovers the graph’s edges with high probability and O(s log m) measurements where s is the maximum degree of the graph and m is the number of nodes. Furthermore, we show that our algori thm also recovers the edge weights (the parameters of the diffusion process) and is robust in the context of approximate sparsity. Finally we prove an almost ma tching lower bound of $\Omega(s \log m/s)$ and validate our approach empirically on syn thetic graphs.
****
Distributed Box-Constrained Quadratic Optimization for Dual Linear SVM
Ching-Pei Lee, Dan Roth
Training machine learning models sometimes needs to be done on large amounts of data that exceed the capacity of a single machine, motivating recent works on de veloping algorithms that train in a distributed fashion. This paper proposes an efficient box-constrained quadratic optimization algorithm for distributedly tra ining linear support vector machines (SVMs) with large data. Our key technical c ontribution is an analytical solution to the problem of computing the optimal st ep size at each iteration, using an efficient method that requires only O(1) com munication cost to ensure fast convergence. With this optimal step size, our app roach is superior to other methods by possessing global linear convergence, or, equivalently, $O(\log(1/\varepsilon))$ iteration complexity for an epsilon-accurate solution , for distributedly solving the non-strongly-convex linear SVM dual problem. Exp eriments also show that our method is significantly faster than state-of- the-ar t distributed linear SVM algorithms including DSVM-AVE, DisDCA and TRON.
****
Safe Exploration for Optimization with Gaussian Processes
Yanan Sui, Alkis Gotovos, Joel Burdick, Andreas Krause

We consider sequential decision problems under uncertainty, where we seek to opt
imize an unknown function from noisy samples. This requires balancing exploratio
n (learning about the objective) and exploitation (localizing the maximum), a pr
oblem well-studied in the multi-armed bandit literature. In many applications, h
owever, we require that the sampled function values exceed some prespecified "sa
fety" threshold, a requirement that existing algorithms fail to meet. Examples i
nclude medical applications where patient comfort must be guaranteed, recommende
r systems aiming to avoid user dissatisfaction, and robotic control, where one s
eeks to avoid controls causing physical harm to the platform. We tackle this nov
el, yet rich, set of problems under the assumption that the unknown function sat
isfies regularity conditions expressed via a Gaussian process prior. We develop
an efficient algorithm called SafeOpt, and theoretically guarantee its convergen
ce to a natural notion of optimum reachable under safety constraints. We evaluat
e SafeOpt on synthetic data, as well as two real applications: movie recommendat
ion, and therapeutic spinal cord stimulation.
****

The Ladder: A Reliable Leaderboard for Machine Learning Competitions
Avrim Blum, Moritz Hardt
The organizer of a machine learning competition faces the problem of maintaining
 an accurate leaderboard that faithfully represents the quality of the best subm
ission of each competing team. What makes this estimation problem particularly c
hallenging is its sequential and adaptive nature. As participants are allowed to
 repeatedly evaluate their submissions on the leaderboard, they may begin to ove
rfit to the holdout data that supports the leaderboard. Few theoretical results
give actionable advice on how to design a reliable leaderboard. Existing approac
hes therefore often resort to poorly understood heuristics such as limiting the
bit precision of answers and the rate of re-submission. In this work, we introdu
ce a notion of leaderboard accuracy tailored to the format of a competition. We
introduce a natural algorithm called the Ladder and demonstrate that it simultan
eously supports strong theoretical guarantees in a fully adaptive model of estim
ation, withstands practical adversarial attacks, and achieves high utility on re
al submission files from a Kaggle competition. Notably, we are able to sidestep
a powerful recent hardness result for adaptive risk estimation that rules out al
gorithms such as ours under a seemingly very similar notion of accuracy. On a pr
actical note, we provide a completely parameter-free variant of our algorithm th
at can be deployed in a real competition with no tuning required whatsoever.
****

Enabling scalable stochastic gradient-based inference for Gaussian processes by
employing the Unbiased LInear System SolvEr (ULISSE)
Maurizio Filippone, Raphael Engler
In applications of Gaussian processes where quantification of uncertainty is of
primary interest, it is necessary to accurately characterize the posterior distr
ibution over covariance parameters. This paper proposes an adaptation of the Sto
chastic Gradient Langevin Dynamics algorithm to draw samples from the posterior
distribution over covariance parameters with negligible bias and without the nee
d to compute the marginal likelihood. In Gaussian process regression, this has t
he enormous advantage that stochastic gradients can be computed by solving linea
r systems only. A novel unbiased linear systems solver based on parallelizable c
ovariance matrix-vector products is developed to accelerate the unbiased estimat
ion of gradients. The results demonstrate the possibility to enable scalable and
 exact (in a Monte Carlo sense) quantification of uncertainty in Gaussian proces
ses without imposing any special structure on the covariance or reducing the num
ber of input vectors.
****

Finding Galaxies in the Shadows of Quasars with Gaussian Processes
Roman Garnett, Shirley Ho, Jeff Schneider
We develop an automated technique for detecting damped Lyman-αabsorbers (DLAs) a
long spectroscopic sightlines to quasi-stellar objects (QSOs or quasars). The de
tection of DLAs in large-scale spectroscopic surveys such as SDSS–III is critica
l to address outstanding cosmological questions, such as the nature of galaxy fo

rmation. We use nearly 50000 QSO spectra to learn a tailored Gaussian process mo del for quasar emission spectra, which we apply to the DLA detection problem via Bayesian model selection. We demonstrate our method’s effectiveness with a larg e-scale validation experiment on over 100000 spectra, with excellent performance .

****

Following the Perturbed Leader for Online Structured Learning

Alon Cohen, Tamir Hazan

We investigate a new Follow the Perturbed Leader (FTPL) algorithm for online str uctured prediction problems. We show a regret bound which is comparable to the s tate of the art of FTPL algorithms and is comparable with the best possible regr et in some cases. To better understand FTPL algorithms for online structured lea rning, we present a lower bound on the regret for a large and natural class of F TPL algorithms that use logconcave perturbations. We complete our investigation with an online shortest path experiment and empirically show that our algorithm is both statistically and computationally efficient.

****

Reified Context Models

Jacob Steinhardt, Percy Liang

A classic tension exists between exact inference in a simple model and approxima te inference in a complex model. The latter offers expressivity and thus accurac y, but the former provides coverage of the space, an important property for conf idence estimation and learning with indirect supervision. In this work, we intro duce a new approach, reified context models, to reconcile this tension. Specific ally, we let the choice of factors in a graphical model (the contexts) be random variables inside the model itself. In this sense, the contexts are reified and can be chosen in a data-dependent way. Empirically, we show that our approach ob tains expressivity and coverage on three sequence modeling tasks.

****

Large-Scale Markov Decision Problems with KL Control Cost and its Application to Crowdsourcing

Yasin Abbasi-Yadkori, Peter Bartlett, Xi Chen, Alan Malek

We study average and total cost Markov decision problems with large state spaces . Since the computational and statistical costs of finding the optimal policy sc ale with the size of the state space, we focus on searching for near-optimality in a low-dimensional family of policies. In particular, we show that for problem s with a Kullback-Leibler divergence cost function, we can reduce policy optimiz ation to a convex optimization and solve it approximately using a stochastic sub gradient algorithm. We show that the performance of the resulting policy is clos e to the best in the low-dimensional family. We demonstrate the efficacy of our approach by controlling the important crowdsourcing application of budget alloca tion in crowd labeling.

****

Learning Fast-Mixing Models for Structured Prediction

Jacob Steinhardt, Percy Liang

Markov Chain Monte Carlo (MCMC) algorithms are often used for approximate infere nce inside learning, but their slow mixing can be difficult to diagnose and the resulting approximate gradients can seriously degrade learning. To alleviate the se issues, we define a new model family using strong Doeblin Markov chains, whos e mixing times can be precisely controlled by a parameter. We also develop an al gorithm to learn such models, which involves maximizing the data likelihood unde r the induced stationary distribution of these chains. We show empirical improve ments on two challenging inference tasks.

****

A Probabilistic Model for Dirty Multi-task Feature Selection

Daniel Hernandez-Lobato, Jose Miguel Hernandez-Lobato, Zoubin Ghahramani

Multi-task feature selection methods often make the hypothesis that learning tas ks share relevant and irrelevant features. However, this hypothesis may be too r estrictive in practice. For example, there may be a few tasks with specific rele vant and irrelevant features (outlier tasks). Similarly, a few of the features m

ay be relevant for only some of the tasks (outlier features). To account for thi s, we propose a model for multi-task feature selection based on a robust prior d istribution that introduces a set of binary latent variables to identify outlier tasks and outlier features. Expectation propagation can be used for efficient a pproximate inference under the proposed prior. Several experiments show that a m odel based on the new robust prior provides better predictive performance than o ther benchmark methods.
****
On Deep Multi-View Representation Learning
Weiran Wang, Raman Arora, Karen Livescu, Jeff Bilmes
We consider learning representations (features) in the setting in which we have access to multiple unlabeled views of the data for representation learning while only one view is available at test time. Previous work on this problem has prop osed several techniques based on deep neural networks, typically involving eithe r autoencoder-like networks with a reconstruction objective or paired feedforwar d networks with a correlation-based objective. We analyze several techniques bas ed on prior work, as well as new variants, and compare them experimentally on vi sual, speech, and language domains. To our knowledge this is the first head-to-h ead comparison of a variety of such techniques on multiple tasks. We find an adv antage for correlation-based representation learning, while the best results on most tasks are obtained with our new variant, deep canonically correlated autoen coders (DCCAE).
****
Learning Program Embeddings to Propagate Feedback on Student Code
Chris Piech, Jonathan Huang, Andy Nguyen, Mike Phulsuksombati, Mehran Sahami, Le onidas Guibas
Providing feedback, both assessing final work and giving hints to stuck students , is difficult for open-ended assignments in massive online classes which can ra nge from thousands to millions of students. We introduce a neural network method to encode programs as a linear mapping from an embedded precondition space to a n embedded postcondition space and propose an algorithm for feedback at scale us ing these linear maps as features. We apply our algorithm to assessments from th e Code.org Hour of Code and Stanford University’s CS1 course, where we propagate human comments on student assignments to orders of magnitude more submissions.
****
Safe Subspace Screening for Nuclear Norm Regularized Least Squares Problems
Qiang Zhou, Qi Zhao
Nuclear norm regularization has been shown very promising for pursing a low rank matrix solution in various machine learning problems. Many efforts have been de voted to develop efficient algorithms for solving the optimization problem in nu clear norm regularization. Solving it for large-scale matrix variables, however, is still a challenging task since the complexity grows fast with the size of ma trix variable. In this work, we propose a novel method called safe subspace scre ening (SSS), to improve the efficiency of the solver for nuclear norm regularize d least squares problems. Motivated by the fact that the low rank solution can b e represented by a few subspaces, the proposed method accurately discards a pred ominant percentage of inactive subspaces prior to solving the problem to reduce problem size. Consequently, a much smaller problem is required to solve, making it more efficient than optimizing the original problem. The proposed SSS is safe , in that its solution is identical to the solution from the solver. In addition , the proposed SSS can be used together with any existing nuclear norm solver si nce it is independent of the solver. Extensive results on several synthetic and real data sets show that the proposed SSS is very effective in inactive subspace screening.
****
Efficient Learning in Large-Scale Combinatorial Semi-Bandits
Zheng Wen, Branislav Kveton, Azin Ashkan
A stochastic combinatorial semi-bandit is an online learning problem where at ea ch step a learning agent chooses a subset of ground items subject to combinatori al constraints, and then observes stochastic weights of these items and receives

their sum as a payoff. In this paper, we consider efficient learning in large-scale combinatorial semi-bandits with linear generalization, and as a solution, propose two learning algorithms called Combinatorial Linear Thompson Sampling (CombLinTS) and Combinatorial Linear UCB (CombLinUCB). Both algorithms are computationally efficient as long as the offline version of the combinatorial problem can be solved efficiently. We establish that CombLinTS and CombLinUCB are also provably statistically efficient under reasonable assumptions, by developing regret bounds that are independent of the problem scale (number of items) and sublinear in time. We also evaluate CombLinTS on a variety of problems with thousands of items. Our experiment results demonstrate that CombLinTS is scalable, robust to the choice of algorithm parameters, and significantly outperforms the best of our baselines.

****

Swept Approximate Message Passing for Sparse Estimation

Andre Manoel, Florent Krzakala, Eric Tramel, Lenka Zdeborovà

Approximate Message Passing (AMP) has been shown to be a superior method for inference problems, such as the recovery of signals from sets of noisy, lower-dimensionality measurements, both in terms of reconstruction accuracy and in computational efficiency. However, AMP suffers from serious convergence issues in contexts that do not exactly match its assumptions. We propose a new approach to stabilizing AMP in these contexts by applying AMP updates to individual coefficients rather than in parallel. Our results show that this change to the AMP iteration can provide theoretically expected, but hitherto unobtainable, performance for problems on which the standard AMP iteration diverges. Additionally, we find that the computational costs of this swept coefficient update scheme is not unduly burdensome, allowing it to be applied efficiently to signals of large dimensionality.

****

Simple regret for infinitely many armed bandits

Alexandra Carpentier, Michal Valko

We consider a stochastic bandit problem with infinitely many arms. In this setting, the learner has no chance of trying all the arms even once and has to dedicate its limited number of samples only to a certain number of arms. All previous algorithms for this setting were designed for minimizing the cumulative regret of the learner. In this paper, we propose an algorithm aiming at minimizing the simple regret. As in the cumulative regret setting of infinitely many armed bandits, the rate of the simple regret will depend on a parameter βcharacterizing the distribution of the near-optimal arms. We prove that depending on β, our algorithm is minimax optimal either up to a multiplicative constant or up to a \log(n) factor. We also provide extensions to several important cases: when $\beta$is unknown, in a natural setting where the near-optimal arms have a small variance, and in the case of unknown time horizon.

****

Exponential Integration for Hamiltonian Monte Carlo

Wei-Lun Chao, Justin Solomon, Dominik Michels, Fei Sha

We investigate numerical integration of ordinary differential equations (ODEs) for Hamiltonian Monte Carlo (HMC). High-quality integration is crucial for designing efficient and effective proposals for HMC. While the standard method is leapfrog (Stormer-Verlet) integration, we propose the use of an exponential integrator, which is robust to stiff ODEs with highly-oscillatory components. This oscillation is difficult to reproduce using leapfrog integration, even with carefully selected integration parameters and preconditioning. Concretely, we use a Gaussian distribution approximation to segregate stiff components of the ODE. We integrate this term analytically for stability and account for deviation from the approximation using variation of constants. We consider various ways to derive Gaussian approximations and conduct extensive empirical studies applying the proposed “exponential HMC” to several benchmarked learning problems. We compare to state-of-the-art methods for improving leapfrog HMC and demonstrate the advantages of our method in generating many effective samples with high acceptance rates in short running times.

****
Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays
Junpei Komiyama, Junya Honda, Hiroshi Nakagawa
We discuss a multiple-play multi-armed bandit (MAB) problem in which several arms are selected at each round. Recently, Thompson sampling (TS), a randomized algorithm with a Bayesian spirit, has attracted much attention for its empirically excellent performance, and it is revealed to have an optimal regret bound in the standard single-play MAB problem. In this paper, we propose the multiple-play Thompson sampling (MP-TS) algorithm, an extension of TS to the multiple-play MAB problem, and discuss its regret analysis. We prove that MP-TS has the optimal regret upper bound that matches the regret lower bound provided by Anantharam et al.\,(1987). Therefore, MP-TS is the first computationally efficient algorithm with optimal regret. A set of computer simulations was also conducted, which compared MP-TS with state-of-the-art algorithms. We also propose a modification of MP-TS, which is shown to have better empirical performance.
****
Faster cover trees
Mike Izbicki, Christian Shelton
The cover tree data structure speeds up exact nearest neighbor queries over arbitrary metric spaces. This paper makes cover trees even faster. In particular, we provide (1) a simpler definition of the cover tree that reduces the number of nodes from O(n) to exactly n, (2) an additional invariant that makes queries faster in practice, (3) algorithms for constructing and querying the tree in parallel on multiprocessor systems, and (4) a more cache efficient memory layout. On standard benchmark datasets, we reduce the number of distance computations by 10–50%. On a large-scale bioinformatics dataset, we reduce the number of distance computations by 71%. On a large-scale image dataset, our parallel algorithm with 16 cores reduces tree construction time from 3.5 hours to 12 minutes.
****
Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization
Tyler Johnson, Carlos Guestrin
By reducing optimization to a sequence of small subproblems, working set methods achieve fast convergence times for many challenging problems. Despite excellent performance, theoretical understanding of working sets is limited, and implementations often resort to heuristics to determine subproblem size, makeup, and stopping criteria. We propose Blitz, a fast working set algorithm accompanied by useful guarantees. Making no assumptions on data, our theory relates subproblem size to progress toward convergence. This result motivates methods for optimizing algorithmic parameters and discarding irrelevant variables as iterations progress. Applied to L1-regularized learning, Blitz convincingly outperforms existing solvers in sequential, limited-memory, and distributed settings. Blitz is not specific to L1-regularized learning, making the algorithm relevant to many applications involving sparsity or constraints.
****
Unsupervised Domain Adaptation by Backpropagation
Yaroslav Ganin, Victor Lempitsky
Top-performing deep architectures are trained on massive amounts of labeled data. In the absence of labeled data for a certain task, domain adaptation often provides an attractive option given that labeled data of similar nature but from a different domain (e.g. synthetic images) are available. Here, we propose a new approach to domain adaptation in deep architectures that can be trained on large amount of labeled data from the source domain and large amount of unlabeled data from the target domain (no labeled target-domain data is necessary). As the training progresses, the approach promotes the emergence of "deep" features that are (i) discriminative for the main learning task on the source domain and (ii) invariant with respect to the shift between the domains. We show that this adaptation behaviour can be achieved in almost any feed-forward model by augmenting it with few standard layers and a simple new gradient reversal layer. The resulting augmented architecture can be trained using standard backpropagation. Overall,

the approach can be implemented with little effort using any of the deep-learnin g packages. The method performs very well in a series of image classification ex periments, achieving adaptation effect in the presence of big domain shifts and outperforming previous state-of-the-art on Office datasets.
****

Non-Linear Cross-Domain Collaborative Filtering via Hyper-Structure Transfer
Yan-Fu Liu, Cheng-Yu Hsu, Shan-Hung Wu
The Cross Domain Collaborative Filtering (CDCF) exploits the rating matrices fro m multiple domains to make better recommendations. Existing CDCF methods adopt t he sub-structure sharing technique that can only transfer linearly correlated kn owledge between domains. In this paper, we propose the notion of Hyper-Structure Transfer (HST) that requires the rating matrices to be explained by the project ions of some more complex structure, called the hyper-structure, shared by all d omains, and thus allows the non-linearly correlated knowledge between domains to be identified and transferred. Extensive experiments are conducted and the resu lts demonstrate the effectiveness of our HST models empirically.
****

Manifold-valued Dirichlet Processes
Hyunwoo Kim, Jia Xu, Baba Vemuri, Vikas Singh
Statistical models for manifold-valued data permit capturing the intrinsic natur e of the curved spaces in which the data lie and have been a topic of research f or several decades. Typically, these formulations use geodesic curves and distan ces defined locally for most cases - this makes it hard to design parametric mod els globally on smooth manifolds. Thus, most (manifold specific) parametric mode ls available today assume that the data lie in a small neighborhood on the manif old. To address this 'locality' problem, we propose a novel nonparametric model which unifies multivariate general linear models (MGLMs) using multiple tangent spaces. Our framework generalizes existing work on (both Euclidean and non-Eucli dean) general linear models providing a recipe to globally extend the locally-de fined parametric models (using a mixture of local models). By grouping observati ons into sub-populations at multiple tangent spaces, our method provides insight s into the hidden structure (geodesic relationships) in the data. This yields a framework to group observations and discover geodesic relationships between cova riates X and manifold-valued responses Y, which we call Dirichlet process mixtur es of multivariate general linear models (DP-MGLM) on Riemannian manifolds. Fina lly, we present proof of concept experiments to validate our model.
****

Multi-Task Learning for Subspace Segmentation
Yu Wang, David Wipf, Qing Ling, Wei Chen, Ian Wassell
Subspace segmentation is the process of clustering a set of data points that are assumed to lie on the union of multiple linear or affine subspaces, and is incr easingly being recognized as a fundamental tool for data analysis in high dimens ional settings. Arguably one of the most successful approaches is based on the o bservation that the sparsest representation of a given point with respect to a d ictionary formed by the others involves nonzero coefficients associated with poi nts originating in the same subspace. Such sparse representations are computed i ndependently for each data point via \ell_1-norm minimization and then combined into an affinity matrix for use by a final spectral clustering step. The downsid e of this procedure is two-fold. First, unlike canonical compressive sensing sce narios with ideally-randomized dictionaries, the data-dependent dictionaries her e are unavoidably highly structured, disrupting many of the favorable properties of the \ell_1 norm. Secondly, by treating each data point independently, we ign ore useful relationships between points that can be leveraged for jointly comput ing such sparse representations. Consequently, we motivate a multi-task learning -based framework for learning coupled sparse representations leading to a segmen tation pipeline that is both robust against correlation structure and tailored t o generate an optimal affinity matrix. Theoretical analysis and empirical tests are provided to support these claims.
****

Markov Chain Monte Carlo and Variational Inference: Bridging the Gap

Tim Salimans, Diederik Kingma, Max Welling
Recent advances in stochastic gradient variational inference have made it possib
le to perform variational Bayesian inference with posterior approximations conta
ining auxiliary random variables. This enables us to explore a new synthesis of
variational inference and Monte Carlo methods where we incorporate one or more s
teps of MCMC into our variational approximation. By doing so we obtain a rich cl
ass of inference algorithms bridging the gap between variational methods and MCM
C, and offering the best of both worlds: fast posterior approximation through th
e maximization of an explicit objective, with the option of trading off addition
al computation for additional accuracy. We describe the theoretical foundations
that make this possible and show some promising first results.
****

Scalable Model Selection for Large-Scale Factorial Relational Models
Chunchen Liu, Lu Feng, Ryohei Fujimaki, Yusuke Muraoka
With a growing need to understand large-scale networks, factorial relational mod
els, such as binary matrix factorization models (BMFs), have become important in
 many applications. Although BMFs have a natural capability to uncover overlappi
ng group structures behind network data, existing inference techniques have issu
es of either high computational cost or lack of model selection capability, and
this limits their applicability. For scalable model selection of BMFs, this pape
r proposes stochastic factorized asymptotic Bayesian (sFAB) inference that combi
nes concepts in two recently-developed techniques: stochastic variational infere
nce (SVI) and FAB inference. sFAB is a highly-efficient algorithm, having both s
calability and an inherent model selection capability in a single inference fram
ework. Empirical results show the superiority of sFAB/BMF in both accuracy and s
calability over state-of-the-art inference methods for overlapping relational mo
dels.
****

The Power of Randomization: Distributed Submodular Maximization on Massive Datas
ets
Rafael Barbosa, Alina Ene, Huy Nguyen, Justin Ward
A wide variety of problems in machine learning, including exemplar clustering, d
ocument summarization, and sensor placement, can be cast as constrained submodul
ar maximization problems. Unfortunately, the resulting submodular optimization p
roblems are often too large to be solved on a single machine. We consider a dist
ributed, greedy algorithm that combines previous approaches with randomization.
The result is an algorithm that is embarrassingly parallel and achieves provable
, constant factor, worst-case approximation guarantees. In our experiments, we d
emonstrate its efficiency in large problems with different kinds of constraints
with objective values always close to what is achievable in the centralized sett
ing.
****

Dealing with small data: On the generalization of context trees
Ralf Eggeling, Mikko Koivisto, Ivo Grosse
Context trees (CT) are a widely used tool in machine learning for representing c
ontext-specific independences in conditional probability distributions. Parsimon
ious context trees (PCTs) are a recently proposed generalization of CTs that can
 enable statistically more efficient learning due to a higher structural flexibi
lity, which is particularly useful for small-data settings. However, this comes
at the cost of a computationally expensive structure learning algorithm, which i
s feasible only for domains with small alphabets and tree depths. In this work,
we investigate to which degree CTs can be generalized to increase statistical ef
ficiency while still keeping the learning computationally feasible. Approaching
this goal from two different angles, we (i) propose algorithmic improvements to
the PCT learning algorithm, and (ii) study further generalizations of CTs, which
 are inspired by PCTs, but trade structural flexibility for computational effici
ency. By empirical studies both on simulated and real-world data, we demonstrate
 that the synergy of combining of both orthogonal approaches yields a substantia
l improvement in obtaining statistically efficient and computationally feasible
generalizations of CTs.

****
Non-Gaussian Discriminative Factor Models via the Max-Margin Rank-Likelihood
Xin Yuan, Ricardo Henao, Ephraim Tsalik, Raymond Langley, Lawrence Carin
We consider the problem of discriminative factor analysis for data that are in general non-Gaussian. A Bayesian model based on the ranks of the data is proposed. We first introduce a max-margin version of the rank-likelihood. A discriminative factor model is then developed, integrating the new max-margin rank-likelihood and (linear) Bayesian support vector machines, which are also built on the max-margin principle. The discriminative factor model is further extended to the nonlinear case through mixtures of local linear classifiers, via Dirichlet processes. Fully local conjugacy of the model yields efficient inference with both Markov Chain Monte Carlo and variational Bayes approaches. Extensive experiments on benchmark and real data demonstrate superior performance of the proposed model and its potential for applications in computational biology.
****
A Bayesian nonparametric procedure for comparing algorithms
Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon
A fundamental task in machine learning is to compare the performance of multiple algorithms. This is typically performed by frequentist tests (usually the Friedman test followed by a series of multiple pairwise comparisons). This implies dealing with null hypothesis significance tests and p-values, although the shortcomings of such methods are well known. First, we propose a nonparametric Bayesian version of the Friedman test using a Dirichlet process (DP) based prior. Our derivations show that, from a Bayesian perspective, the Friedman test is an inference for a multivariate mean based on an ellipsoid inclusion test. Second, we derive a joint procedure for the analysis of the multiple comparisons which accounts for their dependencies and which is based on the posterior probability computed through the DP. The proposed approach allows verifying the null hypothesis, not only rejecting it. Third, we apply our test to perform algorithms racing, i.e., the problem of identifying the best algorithm among a large set of candidates. We show by simulation that our approach is competitive both in terms of accuracy and speed in identifying the best algorithm.
****
Convergence rate of Bayesian tensor estimator and its minimax optimality
Taiji Suzuki
We investigate the statistical convergence rate of a Bayesian low-rank tensor estimator, and derive the minimax optimal rate for learning a low-rank tensor. Our problem setting is the regression problem where the regression coefficient forms a tensor structure. This problem setting occurs in many practical applications, such as collaborative filtering, multi-task learning, and spatio-temporal data analysis. The convergence rate of the Bayes tensor estimator is analyzed in terms of both in-sample and out-of-sample predictive accuracies. It is shown that a fast learning rate is achieved without any strong convexity of the observation. Moreover, we show that the method has adaptivity to the unknown rank of the true tensor, that is, the near optimal rate depending on the true rank is achieved even if it is not known a priori. Finally, we show the minimax optimal learning rate for the tensor estimation problem, and thus show that the derived bound of the Bayes estimator is tight and actually near minimax optimal.
****
On Identifying Good Options under Combinatorially Structured Feedback in Finite Noisy Environments
Yifan Wu, Andras Gyorgy, Csaba Szepesvari
We consider the problem of identifying a good option out of finite set of options under combinatorially structured, noisy feedback about the quality of the options in a sequential process: In each round, a subset of the options, from an available set of subsets, can be selected to receive noisy information about the quality of the options in the chosen subset. The goal is to identify the highest quality option, or a group of options of the highest quality, with a small error probability, while using the smallest number of measurements. The problem generalizes best-arm identification problems. By extending previous work, we design ne

w algorithms that are shown to be able to exploit the combinatorial structure of the problem in a nontrivial fashion, while being unimprovable in special cases. The algorithms call a set multi-covering oracle, hence their performance and efficiency is strongly tied to whether the associated set multi-covering problem can be efficiently solved.

****

Nested Sequential Monte Carlo Methods

Christian Naesseth, Fredrik Lindsten, Thomas Schon

We propose nested sequential Monte Carlo (NSMC), a methodology to sample from sequences of probability distributions, even where the random variables are high-dimensional. NSMC generalises the SMC framework by requiring only approximate, properly weighted, samples from the SMC proposal distribution, while still resulting in a correct SMC algorithm. Furthermore, NSMC can in itself be used to produce such properly weighted samples. Consequently, one NSMC sampler can be used to construct an efficient high-dimensional proposal distribution for another NSMC sampler, and this nesting of the algorithm can be done to an arbitrary degree. This allows us to consider complex and high-dimensional models using SMC. We show results that motivate the efficacy of our approach on several filtering problems with dimensions in the order of 100 to 1000.

****

Sparse Variational Inference for Generalized GP Models

Rishit Sheth, Yuyang Wang, Roni Khardon

Gaussian processes (GP) provide an attractive machine learning model due to their non-parametric form, their flexibility to capture many types of observation data, and their generic inference procedures. Sparse GP inference algorithms address the cubic complexity of GPs by focusing on a small set of pseudo-samples. To date, such approaches have focused on the simple case of Gaussian observation likelihoods. This paper develops a variational sparse solution for GPs under general likelihoods by providing a new characterization of the gradients required for inference in terms of individual observation likelihood terms. In addition, we propose a simple new approach for optimizing the sparse variational approximation using a fixed point computation. We demonstrate experimentally that the fixed point operator acts as a contraction in many cases and therefore leads to fast convergence. An experimental evaluation for count regression, classification, and ordinal regression illustrates the generality and advantages of the new approach.

****

Universal Value Function Approximators

Tom Schaul, Daniel Horgan, Karol Gregor, David Silver

Value functions are a core component of reinforcement learning. The main idea is to to construct a single function approximator V(s; theta) that estimates the long-term reward from any state s, using parameters θ. In this paper we introduce universal value function approximators (UVFAs) V(s,g;theta) that generalise not just over states s but also over goals g. We develop an efficient technique for supervised learning of UVFAs, by factoring observed values into separate embedding vectors for state and goal, and then learning a mapping from s and g to these factored embedding vectors. We show how this technique may be incorporated into a reinforcement learning algorithm that updates the UVFA solely from observed rewards. Finally, we demonstrate that a UVFA can successfully generalise to previously unseen goals.

****

Approximate Dynamic Programming for Two-Player Zero-Sum Markov Games

Julien Perolat, Bruno Scherrer, Bilal Piot, Olivier Pietquin

This paper provides an analysis of error propagation in Approximate Dynamic Programming applied to zero-sum two-player Stochastic Games. We provide a novel and unified error propagation analysis in L_p-norm of three well-known algorithms adapted to Stochastic Games (namely Approximate Value Iteration, Approximate Policy Iteration and Approximate Generalized Policy Iteration). We show that we can achieve a stationary policy which is \frac2γ(1 - γ)^2 ε+ \frac1(1 - γ)^2ε’-optimal, where εis the value function approximation error and ε’ is the approximate gr

eedy operator error. In addition, we provide a practical algorithm (AGPI-Q) to s
olve infinite horizon γ-discounted two-player zero-sum stochastic games in a bat
ch setting. It is an extension of the Fitted-Q algorithm (which solves Markov De
cisions Processes in a batch setting) and can be non-parametric. Finally, we dem
onstrate experimentally the performance of AGPI-Q on a simultaneous two-player g
ame, namely Alesia.
****

On Greedy Maximization of Entropy
Dravyansh Sharma, Ashish Kapoor, Amit Deshpande
Submodular function maximization is one of the key problems that arise in many m
achine learning tasks. Greedy selection algorithms are the proven choice to solv
e such problems, where prior theoretical work guarantees (1 - 1/e) approximation
 ratio. However, it has been empirically observed that greedy selection provides
 almost optimal solutions in practice. The main goal of this paper is to explore
 and answer why the greedy selection does significantly better than the theoreti
cal guarantee of (1 - 1/e). Applications include, but are not limited to, sensor
 selection tasks which use both entropy and mutual information as a maximization
 criteria. We give a theoretical justification for the nearly optimal approximat
ion ratio via detailed analysis of the curvature of these objective functions fo
r Gaussian RBF kernels.
****

Metadata Dependent Mondrian Processes
Yi Wang, Bin Li, Yang Wang, Fang Chen
Stochastic partition processes in a product space play an important role in mode
ling relational data. Recent studies on the Mondrian process have introduced mor
e flexibility into the block structure in relational models. A side-effect of su
ch high flexibility is that, in data sparsity scenarios, the model is prone to o
verfit. In reality, relational entities are always associated with meta informat
ion, such as user profiles in a social network. In this paper, we propose a meta
data dependent Mondrian process (MDMP) to incorporate meta information into the
stochastic partition process in the product space and the entity allocation proc
ess on the resulting block structure. MDMP can not only encourage homogeneous re
lational interactions within blocks but also discourage meta-label diversity wit
hin blocks. Regularized by meta information, MDMP becomes more robust in data sp
arsity scenarios and easier to converge in posterior inference. We apply MDMP to
 link prediction and rating prediction and demonstrate that MDMP is more effecti
ve than the baseline models in prediction accuracy with a more parsimonious mode
l structure.
****

Complex Event Detection using Semantic Saliency and Nearly-Isotonic SVM
Xiaojun Chang, Yi Yang, Eric Xing, Yaoliang Yu
We aim to detect complex events in long Internet videos that may last for hours.
 A major challenge in this setting is that only a few shots in a long video are
relevant to the event of interest while others are irrelevant or even misleading
. Instead of indifferently pooling the shots, we first define a novel notion of
semantic saliency that assesses the relevance of each shot with the event of int
erest. We then prioritize the shots according to their saliency scores since sho
ts that are semantically more salient are expected to contribute more to the fin
al event detector. Next, we propose a new isotonic regularizer that is able to e
xploit the semantic ordering information. The resulting nearly-isotonic SVM clas
sifier exhibits higher discriminative power. Computationally, we develop an effi
cient implementation using the proximal gradient algorithm, and we prove new, cl
osed-form proximal steps. We conduct extensive experiments on three real-world v
ideo datasets and confirm the effectiveness of the proposed approach.
****

Rebuilding Factorized Information Criterion: Asymptotically Accurate Marginal Li
kelihood
Kohei Hayashi, Shin-ichi Maeda, Ryohei Fujimaki
Factorized information criterion (FIC) is a recently developed approximation tec
hnique for the marginal log-likelihood, which provides an automatic model select

ion framework for a few latent variable models (LVMs) with tractable inference a lgorithms. This paper reconsiders FIC and fills theoretical gaps of previous FIC studies. First, we reveal the core idea of FIC that allows generalization for a broader class of LVMs, including continuous LVMs, in contrast to previous FICs, which are applicable only to binary LVMs. Second, we investigate the model sele ction mechanism of the generalized FIC. Our analysis provides a formal justifica tion of FIC as a model selection criterion for LVMs and also a systematic proced ure for pruning redundant latent variables that have been removed heuristically in previous studies. Third, we provide an interpretation of FIC as a variational free energy and uncover previously-unknown their relationship. A demonstrative study on Bayesian principal component analysis is provided and numerical experim ents support our theoretical results.

****

Double Nyström Method: An Efficient and Accurate Nyström Scheme for Large-Scale Data Sets

Woosang Lim, Minhwan Kim, Haesun Park, Kyomin Jung

The Nyström method has been one of the most effective techniques for kernel-base d approach that scales well to large data sets. Since its introduction, there ha s been a large body of work that improves the approximation accuracy while maint aining computational efficiency. In this paper, we present a novel Nyström metho d that improves both accuracy and efficiency based on a new theoretical analysis . We first provide a generalized sampling scheme, CAPS, that minimizes a novel e rror bound based on the subspace distance. We then present our double Nyström me thod that reduces the size of the decomposition in two stages. We show that our method is highly efficient and accurate compared to other state-of-the-art Nystr öm methods by evaluating them on a number of real data sets.

****

The Composition Theorem for Differential Privacy

Peter Kairouz, Sewoong Oh, Pramod Viswanath

Interactive querying of a database degrades the privacy level. In this paper we answer the fundamental question of characterizing the level of privacy degradati on as a function of the number of adaptive interactions and the differential pri vacy levels maintained by the individual queries. Our solution is complete: the privacy degradation guarantee is true for every privacy mechanism, and further, we demonstrate a sequence of privacy mechanisms that do degrade in the character ized manner. The key innovation is the introduction of an operational interpreta tion (involving hypothesis testing) to differential privacy and the use of the c orresponding data processing inequalities. Our result improves over the state of the art and has immediate applications to several problems studied in the liter ature.

****

Convex Formulation for Learning from Positive and Unlabeled Data

Marthinus Du Plessis, Gang Niu, Masashi Sugiyama

We discuss binary classification from only from positive and unlabeled data (PU classification), which is conceivable in various real-world machine learning pro blems. Since unlabeled data consists of both positive and negative data, simply separating positive and unlabeled data yields a biased solution. Recently, it wa s shown that the bias can be canceled by using a particular non-convex loss such as the ramp loss. However, classifier training with a non-convex loss is not st raightforward in practice. In this paper, we discuss a convex formulation for PU classification that can still cancel the bias. The key idea is to use different loss functions for positive and unlabeled samples. However, in this setup, the hinge loss is not permissible. As an alternative, we propose the double hinge lo ss. Theoretically, we prove that the estimators converge to the optimal solution s at the optimal parametric rate. Experimentally, we demonstrate that PU classif ication with the double hinge loss performs as accurate as the non-convex method , with a much lower computational cost.

****

Threshold Influence Model for Allocating Advertising Budgets

Atsushi Miyauchi, Yuni Iwamasa, Takuro Fukunaga, Naonori Kakimura

We propose a new influence model for allocating budgets to advertising channels. Our model captures customer’s sensitivity to advertisements as a threshold behavior; a customer is expected to be influenced if the influence he receives exceeds his threshold. Over the threshold model, we discuss two optimization problems. The first one is the budget-constrained influence maximization. We propose two greedy algorithms based on different strategies, and analyze the performance when the influence is submodular. We then introduce a new characteristic to measure the cost-effectiveness of a marketing campaign, that is, the proportion of the resulting influence to the cost spent. We design an almost linear-time approximation algorithm to maximize the cost-effectiveness. Furthermore, we design a better-approximation algorithm based on linear programming for a special case. We conduct thorough experiments to confirm that our algorithms outperform baseline algorithms.
****

Strongly Adaptive Online Learning
Amit Daniely, Alon Gonen, Shai Shalev-Shwartz
Strongly adaptive algorithms are algorithms whose performance on every time interval is close to optimal. We present a reduction that can transform standard low-regret algorithms to strongly adaptive. As a consequence, we derive simple, yet efficient, strongly adaptive algorithms for a handful of problems.
****

CUR Algorithm for Partially Observed Matrices
Miao Xu, Rong Jin, Zhi-Hua Zhou
CUR matrix decomposition computes the low rank approximation of a given matrix by using the actual rows and columns of the matrix. It has been a very useful tool for handling large matrices. One limitation with the existing algorithms for CUR matrix decomposition is that they cannot deal with entries in a \it partially observed matrix, while incomplete matrices are found in many real world applications. In this work, we alleviate this limitation by developing a CUR decomposition algorithm for partially observed matrices. In particular, the proposed algorithm computes the low rank approximation of the target matrix based on (i) the randomly sampled rows and columns, and (ii) a subset of observed entries that are randomly sampled from the matrix. Our analysis shows the relative error bound, measured by spectral norm, for the proposed algorithm when the target matrix is of full rank. We also show that only O(n r\ln r) observed entries are needed by the proposed algorithm to perfectly recover a rank r matrix of size n\times n, which improves the sample complexity of the existing algorithms for matrix completion. Empirical studies on both synthetic and real-world datasets verify our theoretical claims and demonstrate the effectiveness of the proposed algorithm.
****

A Deterministic Analysis of Noisy Sparse Subspace Clustering for Dimensionality-reduced Data
Yining Wang, Yu-Xiang Wang, Aarti Singh
Subspace clustering groups data into several lowrank subspaces. In this paper, we propose a theoretical framework to analyze a popular optimization-based algorithm, Sparse Subspace Clustering (SSC), when the data dimension is compressed via some random projection algorithms. We show SSC provably succeeds if the random projection is a subspace embedding, which includes random Gaussian projection, uniform row sampling, FJLT, sketching, etc. Our analysis applies to the most general deterministic setting and is able to handle both adversarial and stochastic noise. It also results in the first algorithm for privacy-preserved subspace clustering.
****

MRA-based Statistical Learning from Incomplete Rankings
Eric Sibony, Stéphan Clemençon, Jérémie Jakubowicz
Statistical analysis of rank data describing preferences over small and variable subsets of a potentially large ensemble of items 1, ..., n is a very challenging problem. It is motivated by a wide variety of modern applications, such as recommender systems or search engines. However, very few inference methods have been documented in the literature to learn a ranking model from such incomplete ran

k data. The goal of this paper is twofold: it develops a rigorous mathematical f
ramework for the problem of learning a ranking model from incomplete rankings an
d introduces a novel general statistical method to address it. Based on an origi
nal concept of multi-resolution analysis (MRA) of incomplete rankings, it finely
 adapts to any observation setting, leading to a statistical accuracy and an alg
orithmic complexity that depend directly on the complexity of the observed data.
 Beyond theoretical guarantees, we also provide experimental results that show i
ts statistical performance.
****

Risk and Regret of Hierarchical Bayesian Learners
Jonathan Huggins, Josh Tenenbaum
Common statistical practice has shown that the full power of Bayesian methods is
 not realized until hierarchical priors are used, as these allow for greater “ro
bustness” and the ability to “share statistical strength.” Yet it is an ongoing
challenge to provide a learning-theoretically sound formalism of such notions th
at: offers practical guidance concerning when and how best to utilize hierarchic
al models; provides insights into what makes for a good hierarchical prior; and,
 when the form of the prior has been chosen, can guide the choice of hyperparame
ter settings. We present a set of analytical tools for understanding hierarchica
l priors in both the online and batch learning settings. We provide regret bound
s under log-loss, which show how certain hierarchical models compare, in retrosp
ect, to the best single model in the model class. We also show how to convert a
Bayesian log-loss regret bound into a Bayesian risk bound for any bounded loss,
a result which may be of independent interest. Risk and regret bounds for Studen
t’s t and hierarchical Gaussian priors allow us to formalize the concepts of “ro
bustness” and “sharing statistical strength.” Priors for feature selection are i
nvestigated as well. Our results suggest that the learning-theoretic benefits of
 using hierarchical priors can often come at little cost on practical problems.
****

Towards a Learning Theory of Cause-Effect Inference
David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, Iliya Tolstikhin
We pose causal inference as the problem of learning to classify probability dist
ributions. In particular, we assume access to a collection {(S_i,l_i)}_i=1^n, wh
ere each S_i is a sample drawn from the probability distribution of X_i \times Y
_i, and l_i is a binary label indicating whether “X_i \to Y_i” or “X_i ←Y_i”. Gi
ven these data, we build a causal inference rule in two steps. First, we featuri
ze each S_i using the kernel mean embedding associated with some characteristic
kernel. Second, we train a binary classifier on such embeddings to distinguish b
etween causal directions. We present generalization bounds showing the statistic
al consistency and learning rates of the proposed approach, and provide a simple
 implementation that achieves state-of-the-art cause-effect inference. Furthermo
re, we extend our ideas to infer causal relationships between more than two vari
ables.
****

DRAW: A Recurrent Neural Network For Image Generation
Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, Daan Wierstra
This paper introduces the Deep Recurrent Attentive Writer (DRAW) architecture fo
r image generation with neural networks. DRAW networks combine a novel spatial a
ttention mechanism that mimics the foveation of the human eye, with a sequential
 variational auto-encoding framework that allows for the iterative construction
of complex images. The system substantially improves on the state of the art for
 generative models on MNIST, and, when trained on the Street View House Numbers
dataset, it is able to generate images that are indistinguishable from real data
 with the naked eye.
****

Multiview Triplet Embedding: Learning Attributes in Multiple Maps
Ehsan Amid, Antti Ukkonen
For humans, it is usually easier to make statements about the similarity of obje
cts in relative, rather than absolute terms. Moreover, subjective comparisons of
 objects can be based on a number of different and independent attributes. For e

xample, objects can be compared based on their shape, color, etc. In this paper, we consider the problem of uncovering these hidden attributes given a set of relative distance judgments in the form of triplets. The attribute that was used to generate a particular triplet in this set is unknown. Such data occurs, e.g., in crowdsourcing applications where the triplets are collected from a large group of workers. We propose the Multiview Triplet Embedding (MVTE) algorithm that produces a number of low-dimensional maps, each corresponding to one of the hidden attributes. The method can be used to assess how many different attributes were used to create the triplets, as well as to assess the difficulty of a distance comparison task, and find objects that have multiple interpretations in relation to the other objects.
****

Distributed Gaussian Processes
Marc Deisenroth, Jun Wei Ng
To scale Gaussian processes (GPs) to large data sets we introduce the robust Bayesian Committee Machine (rBCM), a practical and scalable product-of-experts model for large-scale distributed GP regression. Unlike state-of-the-art sparse GP approximations, the rBCM is conceptually simple and does not rely on inducing or variational parameters. The key idea is to recursively distribute computations to independent computational units and, subsequently, recombine them to form an overall result. Efficient closed-form inference allows for straightforward parallelisation and distributed computations with a small memory footprint. The rBCM is independent of the computational graph and can be used on heterogeneous computing infrastructures, ranging from laptops to clusters. With sufficient computing resources our distributed GP model can handle arbitrarily large data sets.
****

Guaranteed Tensor Decomposition: A Moment Approach
Gongguo Tang, Parikshit Shah
We develop a theoretical and computational framework to perform guaranteed tensor decomposition, which also has the potential to accomplish other tensor tasks such as tensor completion and denoising. We formulate tensor decomposition as a problem of measure estimation from moments. By constructing a dual polynomial, we demonstrate that measure optimization returns the correct CP decomposition under an incoherence condition on the rank-one factors. To address the computational challenge, we present a hierarchy of semidefinite programs based on sums-of-squares relaxations of the measure optimization problem. By showing that the constructed dual polynomial is a sum-of-squares modulo the sphere, we prove that the smallest SDP in the relaxation hierarchy is exact and the decomposition can be extracted from the solution under the same incoherence condition. One implication is that the tensor nuclear norm can be computed exactly using the smallest SDP as long as the rank-one factors of the tensor are incoherent. Numerical experiments are conducted to test the performance of the moment approach.
****

\ell_1,p-Norm Regularization: Error Bounds and Convergence Rate Analysis of First-Order Methods
Zirui Zhou, Qi Zhang, Anthony Man-Cho So
Recently, \ell_1,p-regularization has been widely used to induce structured sparsity in the solutions to various optimization problems. Motivated by the desire to analyze the convergence rate of first-order methods, we show that for a large class of \ell_1,p-regularized problems, an error bound condition is satisfied when $p\in[1,2]$ or $p=\infty$but fails to hold for any $p\in(2,\infty)$. Based on this result, we show that many first-order methods enjoy an asymptotic linear rate of convergence when applied to \ell_1,p-regularized linear or logistic regression with $p\in[1,2]$ or $p=\infty$. By contrast, numerical experiments suggest that for the same class of problems with $p\in(2,\infty)$, the aforementioned methods may not converge linearly.
****

Consistent estimation of dynamic and multi-layer block models
Qiuyi Han, Kevin Xu, Edoardo Airoldi
Significant progress has been made recently on theoretical analysis of estimators for the stochastic block model (SBM). In this paper, we consider the multi-gra

ph SBM, which serves as a foundation for many application settings including dyn amic and multi-layer networks. We explore the asymptotic properties of two estim ators for the multi-graph SBM, namely spectral clustering and the maximum-likeli hood estimate (MLE), as the number of layers of the multi-graph increases. We de rive sufficient conditions for consistency of both estimators and propose a vari ational approximation to the MLE that is computationally feasible for large netw orks. We verify the sufficient conditions via simulation and demonstrate that th ey are practical. In addition, we apply the model to two real data sets: a dynam ic social network and a multi-layer social network with several types of relatio ns.

****

On the Rate of Convergence and Error Bounds for LSTD(λ)

Manel Tagorti, Bruno Scherrer

We consider LSTD(λ), the least-squares temporal-difference algorithm with eligib ility traces algorithm proposed by Boyan (2002). It computes a linear approximat ion of the value function of a fixed policy in a large Markov Decision Process. Under a β-mixing assumption, we derive, for any value of λ∈(0,1), a high-probabi lity bound on the rate of convergence of this algorithm to its limit. We deduce a high-probability bound on the error of this algorithm, that extends (and sligh tly improves) that derived by Lazaric et al. (2012) in the specific case where λ =0. In the context of temporal-difference algorithms with value function approxi mation, this analysis is to our knowledge the first to provide insight on the ch oice of the eligibility-trace parameter λwith respect to the approximation quali ty of the space and the number of samples.

****

Variational Inference with Normalizing Flows

Danilo Rezende, Shakir Mohamed

The choice of the approximate posterior distribution is one of the core problems in variational inference. Most applications of variational inference employ sim ple families of posterior approximations in order to allow for efficient inferen ce, focusing on mean-field or other simple structured approximations. This restr iction has a significant impact on the quality of inferences made using variatio nal methods. We introduce a new approach for specifying flexible, arbitrarily co mplex and scalable approximate posterior distributions. Our approximations are d istributions constructed through a normalizing flow, whereby a simple initial de nsity is transformed into a more complex one by applying a sequence of invertibl e transformations until a desired level of complexity is attained. We use this v iew of normalizing flows to develop categories of finite and infinitesimal flows and provide a unified view of approaches for constructing rich posterior approx imations. We demonstrate that the theoretical advantages of having posteriors th at better match the true posterior, combined with the scalability of amortized v ariational approaches, provides a clear improvement in performance and applicabi lity of variational inference.

****

Controversy in mechanistic modelling with Gaussian processes

Benn Macdonald, Catherine Higham, Dirk Husmeier

Parameter inference in mechanistic models based on non-affine differential equat ions is computationally onerous, and various faster alternatives based on gradie nt matching have been proposed. A particularly promising approach is based on no nparametric Bayesian modelling with Gaussian processes, which exploits the fact that a Gaussian process is closed under differentiation. However, two alternativ e paradigms have been proposed. The first paradigm, proposed at NIPS 2008 and AI STATS 2013, is based on a product of experts approach and a marginalization over the derivatives of the state variables. The second paradigm, proposed at ICML 2 014, is based on a probabilistic generative model and a marginalization over the state variables. The claim has been made that this leads to better inference re sults. In the present article, we offer a new interpretation of the second parad igm, which highlights the underlying assumptions, approximations and limitations . In particular, we show that the second paradigm suffers from an intrinsic iden tifiability problem, which the first paradigm is not affected by.

****
Convex Learning of Multiple Tasks and their Structure
Carlo Ciliberto, Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco
Reducing the amount of human supervision is a key problem in machine learning an d a natural approach is that of exploiting the relations (structure) among diffe rent tasks. This is the idea at the core of multi-task learning. In this context a fundamental question is how to incorporate the tasks structure in the learnin g problem. We tackle this question by studying a general computational framework that allows to encode a-priori knowledge of the tasks structure in the form of a convex penalty; in this setting a variety of previously proposed methods can b e recovered as special cases, including linear and non-linear approaches. Within this framework, we show that tasks and their structure can be efficiently learn ed considering a convex optimization problem that can be approached by means of block coordinate methods such as alternating minimization and for which we prove convergence to the global minimum.
****
K-hyperplane Hinge-Minimax Classifier
Margarita Osadchy, Tamir Hazan, Daniel Keren
We explore a novel approach to upper bound the misclassification error for probl ems with data comprising a small number of positive samples and a large number o f negative samples. We assign the hinge-loss to upper bound the misclassificatio n error of the positive examples and use the minimax risk to upper bound the mis classification error with respect to the worst case distribution that generates the negative examples. This approach is computationally appealing since the majo rity of training examples (belonging to the negative class) are represented by t he statistics of their distribution, in contrast to kernel SVM which produces a very large number of support vectors in such settings. We derive empirical risk bounds for linear and non-linear classification and show that they are dimension ally independent and decay as 1/\sqrtm for m samples. We propose an efficient al gorithm for training an intersection of finite number of hyperplane and demonstr ate its effectiveness on real data, including letter and scene recognition.
****
Non-Stationary Approximate Modified Policy Iteration
Boris Lesner, Bruno Scherrer
We consider the infinite-horizon γ-discounted optimal control problem formalized by Markov Decision Processes. Running any instance of Modified Policy Iteration —a family of algorithms that can interpolate between Value and Policy Iteration— with an error εat each iteration is known to lead to stationary policies that ar e at least \frac2γε(1-γ)^2-optimal. Variations of Value and Policy Iteration, th at build \ell-periodic non-stationary policies, have recently been shown to disp lay a better \frac2γε(1-γ)(1-γ^\ell)-optimality guarantee. Our first contributio n is to describe a new algorithmic scheme, Non-Stationary Modified Policy Iterat ion, a family of algorithms parameterized by two integers m \ge 0 and \ell \ge 1 that generalizes all the above mentionned algorithms. While m allows to interpo late between Value-Iteration-style and Policy-Iteration-style updates, \ell spec ifies the period of the non-stationary policy that is output. We show that this new family of algorithms also enjoys the improved \frac2γε(1-γ)(1-γ^\ell)-optima lity guarantee. Perhaps more importantly, we show, by exhibiting an original pro blem instance, that this guarantee is tight for all m and \ell; this tightness w as to our knowledge only proved two specific cases, Value Iteration (m=0,\ell=1) and Policy Iteration (m=∞,\ell=1).
****
Entropy evaluation based on confidence intervals of frequency estimates : Applic ation to the learning of decision trees
Mathieu Serrurier, Henri Prade
Entropy gain is widely used for learning decision trees. However, as we go deepe r downward the tree, the examples become rarer and the faithfulness of entropy d ecreases. Thus, misleading choices and over-fitting may occur and the tree has t o be adjusted by using an early-stop criterion or post pruning algorithms. Howev er, these methods still depends on the choices previously made, which may be uns

atisfactory. We propose a new cumulative entropy function based on confidence in tervals on frequency estimates that together considers the entropy of the probab ility distribution and the uncertainty around the estimation of its parameters. This function takes advantage of the ability of a possibility distribution to up per bound a family of probabilities previously estimated from a limited set of e xamples and of the link between possibilistic specificity order and entropy. The proposed measure has several advantages over the classical one. It performs sig nificant choices of split and provides a statistically relevant stopping criteri on that allows the learning of trees whose size is well-suited w.r.t. the availa ble data. On the top of that, it also provides a reasonable estimator of the per formances of a decision tree. Finally, we show that it can be used for designing a simple and efficient online learning algorithm.
****

Geometric Conditions for Subspace-Sparse Recovery
Chong You, Rene Vidal
Given a dictionary \Pi and a signal ξ= \Pi \mathbf x generated by a few \textitl inearly independent columns of \Pi, classical sparse recovery theory deals with the problem of uniquely recovering the sparse representation \mathbf x of ξ. In this work, we consider the more general case where ξlies in a low-dimensional su bspace spanned by a few columns of \Pi, which are possibly \textitlinearly depen dent. In this case, \mathbf x may not unique, and the goal is to recover any sub set of the columns of \Pi that spans the subspace containing ξ. We call such a r epresentation \mathbf x \textitsubspace-sparse. We study conditions under which existing pursuit methods recover a subspace-sparse representation. Such conditio ns reveal important geometric insights and have implications for the theory of c lassical sparse recovery as well as subspace clustering.
****

An Empirical Study of Stochastic Variational Inference Algorithms for the Beta B ernoulli Process
Amar Shah, David Knowles, Zoubin Ghahramani
Stochastic variational inference (SVI) is emerging as the most promising candida te for scaling inference in Bayesian probabilistic models to large datasets. How ever, the performance of these methods has been assessed primarily in the contex t of Bayesian topic models, particularly latent Dirichlet allocation (LDA). Deri ving several new algorithms, and using synthetic, image and genomic datasets, we investigate whether the understanding gleaned from LDA applies in the setting o f sparse latent factor models, specifically beta process factor analysis (BPFA). We demonstrate that the big picture is consistent: using Gibbs sampling within SVI to maintain certain posterior dependencies is extremely effective. However, we also show that different posterior dependencies are important in BPFA relativ e to LDA.
****

Long Short-Term Memory Over Recursive Structures
Xiaodan Zhu, Parinaz Sobihani, Hongyu Guo
The chain-structured long short-term memory (LSTM) has showed to be effective in a wide range of problems such as speech recognition and machine translation. In this paper, we propose to extend it to tree structures, in which a memory cell can reflect the history memories of multiple child cells or multiple descendant cells in a recursive process. We call the model S-LSTM, which provides a princip led way of considering long-distance interaction over hierarchies, e.g., languag e or image parse structures. We leverage the models for semantic composition to understand the meaning of text, a fundamental problem in natural language unders tanding, and show that it outperforms a state-of-the-art recursive model by repl acing its composition layers with the S-LSTM memory blocks. We also show that ut ilizing the given structures is helpful in achieving a performance better than t hat without considering the structures.
****

Weight Uncertainty in Neural Network
Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, Daan Wierstra
We introduce a new, efficient, principled and backpropagation-compatible algorit

hm for learning a probability distribution on the weights of a neural network, c
alled Bayes by Backprop. It regularises the weights by minimising a compression
cost, known as the variational free energy or the expected lower bound on the ma
rginal likelihood. We show that this principled kind of regularisation yields co
mparable performance to dropout on MNIST classification. We then demonstrate how
 the learnt uncertainty in the weights can be used to improve generalisation in
non-linear regression problems, and how this weight uncertainty can be used to d
rive the exploration-exploitation trade-off in reinforcement learning.
****

Learning Submodular Losses with the Lovasz Hinge
Jiaqian Yu, Matthew Blaschko
Learning with non-modular losses is an important problem when sets of prediction
s are made simultaneously. The main tools for constructing convex surrogate loss
 functions for set prediction are margin rescaling and slack rescaling. In this
work, we show that these strategies lead to tight convex surrogates iff the unde
rlying loss function is increasing in the number of incorrect predictions. Howev
er, gradient or cutting-plane computation for these functions is NP-hard for non
-supermodular loss functions. We propose instead a novel convex surrogate loss f
unction for submodular losses, the Lovasz hinge, which leads to O(p log p) compl
exity with O(p) oracle accesses to the loss function to compute a gradient or cu
tting-plane. As a result, we have developed the first tractable convex surrogate
s in the literature for submodular losses. We demonstrate the utility of this no
vel convex surrogate through a real world image labeling task.
****

Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Se
lection
Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, Hoyt Koepke
There has been significant recent work on the theory and application of randomiz
ed coordinate descent algorithms, beginning with the work of Nesterov [SIAM J.
Optim., 22(2), 2012], who showed that a random-coordinate selection rule achieve
s the same convergence rate as the Gauss-Southwell selection rule. This result s
uggests that we should never use the Gauss-Southwell rule, as it is typically mu
ch more expensive than random selection. However, the empirical behaviours of th
ese algorithms contradict this theoretical result: in applications where the com
putational costs of the selection rules are comparable, the Gauss-Southwell sele
ction rule tends to perform substantially better than random coordinate selectio
n. We give a simple analysis of the Gauss-Southwell rule showing that—except in
extreme cases—it’s convergence rate is faster than choosing random coordinates.
Further, in this work we (i) show that exact coordinate optimization improves th
e convergence rate for certain sparse problems, (ii) propose a Gauss-Southwell-L
ipschitz rule that gives an even faster convergence rate given knowledge of the
Lipschitz constants of the partial derivatives, (iii) analyze the effect of appr
oximate Gauss-Southwell rules, and (iv) analyze proximal-gradient variants of th
e Gauss-Southwell rule.
****

Hashing for Distributed Data
Cong Leng, Jiaxiang Wu, Jian Cheng, Xi Zhang, Hanqing Lu
Recently, hashing based approximate nearest neighbors search has attracted much
attention. Extensive centralized hashing algorithms have been proposed and achie
ved promising performance. However, due to the large scale of many applications,
 the data is often stored or even collected in a distributed manner. Learning ha
sh functions by aggregating all the data into a fusion center is infeasible beca
use of the prohibitively expensive communication and computation overhead. In th
is paper, we develop a novel hashing model to learn hash functions in a distribu
ted setting. We cast a centralized hashing model as a set of subproblems with co
nsensus constraints. We find these subproblems can be analytically solved in par
allel on the distributed compute nodes. Since no training data is transmitted ac
ross the nodes in the learning process, the communication cost of our model is i
ndependent to the data size. Extensive experiments on several large scale datase
ts containing up to 100 million samples demonstrate the efficacy of our method.

****
Large-scale Distributed Dependent Nonparametric Trees
Zhiting Hu, Ho Qirong, Avinava Dubey, Eric Xing
Practical applications of Bayesian nonparametric (BNP) models have been limited, due to their high computational complexity and poor scaling on large data. In this paper, we consider dependent nonparametric trees (DNTs), a powerful infinite model that captures time-evolving hierarchies, and develop a large-scale distributed training system. Our major contributions include: (1) an effective memoized variational inference for DNTs, with a novel birth-merge strategy for exploring the unbounded tree space; (2) a model-parallel scheme for concurrent tree growing/pruning and efficient model alignment, through conflict-free model partitioning and lightweight synchronization; (3) a data-parallel scheme for variational parameter updates that allows distributed processing of massive data. Using 64 cores in 36 hours, our system learns a 10K-node DNT topic model on 8M documents that captures both high-frequency and long-tail topics. Our data and model scales are orders-of-magnitude larger than recent results on the hierarchical Dirichlet process, and the near-linear scalability indicates great potential for even bigger problem sizes.
****
Qualitative Multi-Armed Bandits: A Quantile-Based Approach
Balazs Szorenyi, Robert Busa-Fekete, Paul Weng, Eyke Hüllermeier
We formalize and study the multi-armed bandit (MAB) problem in a generalized stochastic setting, in which rewards are not assumed to be numerical. Instead, rewards are measured on a qualitative scale that allows for comparison but invalidates arithmetic operations such as averaging. Correspondingly, instead of characterizing an arm in terms of the mean of the underlying distribution, we opt for using a quantile of that distribution as a representative value. We address the problem of quantile-based online learning both for the case of a finite (pure exploration) and infinite time horizon (cumulative regret minimization). For both cases, we propose suitable algorithms and analyze their properties. These properties are also illustrated by means of first experimental studies.
****
Deep Edge-Aware Filters
Li Xu, Jimmy Ren, Qiong Yan, Renjie Liao, Jiaya Jia
There are many edge-aware filters varying in their construction forms and filtering properties. It seems impossible to uniformly represent and accelerate them in a single framework. We made the attempt to learn a big and important family of edge-aware operators from data. Our method is based on a deep convolutional neural network with a gradient domain training procedure, which gives rise to a powerful tool to approximate various filters without knowing the original models and implementation details. The only difference among these operators in our system becomes merely the learned parameters. Our system enables fast approximation for complex edge-aware filters and achieves up to 200x acceleration, regardless of their originally very different implementation. Fast speed can also be achieved when creating new effects using spatially varying filter or filter combination, bearing out the effectiveness of our deep edge-aware filters.
****
A Convex Optimization Framework for Bi-Clustering
Shiau Hong Lim, Yudong Chen, Huan Xu
We present a framework for biclustering and clustering where the observations are general labels. Our approach is based on the maximum likelihood estimator and its convex relaxation, and generalizes recent works in graph clustering to the biclustering setting. In addition to standard biclustering setting where one seeks to discover clustering structure simultaneously in two domain sets, we show that the same algorithm can be as effective when clustering structure only occurs in one domain. This allows for an alternative approach to clustering that is more natural in some scenarios. We present theoretical results that provide sufficient conditions for the recovery of the true underlying clusters under a generalized stochastic block model. These are further validated by our empirical results on both synthetic and real data.

****
Is Feature Selection Secure against Training Data Poisoning?
Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, Fabio Roli
Learning in adversarial settings is becoming an important task for application domains where attackers may inject malicious data into the training set to subvert normal operation of data-driven technologies. Feature selection has been widely used in machine learning for security applications to improve generalization and computational efficiency, although it is not clear whether its use may be beneficial or even counterproductive when training data are poisoned by intelligent attackers. In this work, we shed light on this issue by providing a framework to investigate the robustness of popular feature selection methods, including LASSO, ridge regression and the elastic net. Our results on malware detection show that feature selection methods can be significantly compromised under attack (we can reduce LASSO to almost random choices of feature sets by careful insertion of less than 5% poisoned training samples), highlighting the need for specific countermeasures.
****
Predictive Entropy Search for Bayesian Optimization with Unknown Constraints
Jose Miguel Hernandez-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, Zoubin Ghahramani
Unknown constraints arise in many types of expensive black-box optimization problems. Several methods have been proposed recently for performing Bayesian optimization with constraints, based on the expected improvement (EI) heuristic. However, EI can lead to pathologies when used with constraints. For example, in the case of decoupled constraints—i.e., when one can independently evaluate the objective or the constraints—EI can encounter a pathology that prevents exploration. Additionally, computing EI requires a current best solution, which may not exist if none of the data collected so far satisfy the constraints. By contrast, information-based approaches do not suffer from these failure modes. In this paper, we present a new information-based method called Predictive Entropy Search with Constraints (PESC). We analyze the performance of PESC and show that it compares favorably to EI-based approaches on synthetic and benchmark problems, as well as several real-world examples. We demonstrate that PESC is an effective algorithm that provides a promising direction towards a unified solution for constrained Bayesian optimization.
****
A Theoretical Analysis of Metric Hypothesis Transfer Learning
Michaël Perrot, Amaury Habrard
We consider the problem of transferring some a priori knowledge in the context of supervised metric learning approaches. While this setting has been successfully applied in some empirical contexts, no theoretical evidence exists to justify this approach. In this paper, we provide a theoretical justification based on the notion of algorithmic stability adapted to the regularized metric learning setting. We propose an on-average-replace-two-stability model allowing us to prove fast generalization rates when an auxiliary source metric is used to bias the regularizer. Moreover, we prove a consistency result from which we show the interest of considering biased weighted regularized formulations and we provide a solution to estimate the associated weight. We also present some experiments illustrating the interest of the approach in standard metric learning tasks and in a transfer learning problem where few labelled data are available.
****
Generative Moment Matching Networks
Yujia Li, Kevin Swersky, Rich Zemel
We consider the problem of learning deep generative models from data. We formulate a method that generates an independent sample via a single feedforward pass through a multilayer preceptron, as in the recently proposed generative adversarial networks (Goodfellow et al., 2014). Training a generative adversarial network, however, requires careful optimization of a difficult minimax program. Instead, we utilize a technique from statistical hypothesis testing known as maximum me

an discrepancy (MMD), which leads to a simple objective that can be interpreted as matching all orders of statistics between a dataset and samples from the model, and can be trained by backpropagation. We further boost the performance of this approach by combining our generative network with an auto-encoder network, using MMD to learn to generate codes that can then be decoded to produce samples. We show that the combination of these techniques yields excellent generative models compared to baseline approaches as measured on MNIST and the Toronto Face Database.

****

Stay on path: PCA along graph paths

Megasthenis Asteris, Anastasios Kyrillidis, Alex Dimakis, Han-Gyol Yi, Bharath Chandrasekaran

We introduce a variant of (sparse) PCA in which the set of feasible support sets is determined by a graph. In particular, we consider the following setting: given a directed acyclic graph G on p vertices corresponding to variables, the non-zero entries of the extracted principal component must coincide with vertices lying along a path in G. From a statistical perspective, information on the underlying network may potentially reduce the number of observations required to recover the population principal component. We consider the canonical estimator which optimally exploits the prior knowledge by solving a non-convex quadratic maximization on the empirical covariance. We introduce a simple network and analyze the estimator under the spiked covariance model for sparse PCA. We show that side information potentially improves the statistical complexity. We propose two algorithms to approximate the solution of the constrained quadratic maximization, and recover a component with the desired properties. We empirically evaluate our schemes on synthetic and real datasets.

****

Deep Learning with Limited Numerical Precision

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, Pritish Narayanan

Training of large-scale deep neural networks is often constrained by the available computational resources. We study the effect of limited precision data representation and computation on neural network training. Within the context of low-precision fixed-point computations, we observe the rounding scheme to play a crucial role in determining the network’s behavior during training. Our results show that deep networks can be trained using only 16-bit wide fixed-point number representation when using stochastic rounding, and incur little to no degradation in the classification accuracy. We also demonstrate an energy-efficient hardware accelerator that implements low-precision fixed-point arithmetic with stochastic rounding

****

Safe Screening for Multi-Task Feature Learning with Multiple Data Matrices

Jie Wang, Jieping Ye

Multi-task feature learning (MTFL) is a powerful technique in boosting the predictive performance by learning multiple related classification/regression/clustering tasks simultaneously. However, solving the MTFL problem remains challenging when the feature dimension is extremely large. In this paper, we propose a novel screening rule—that is based on the dual projection onto convex sets (DPC)—to quickly identify the inactive features—that have zero coefficients in the solution vectors across all tasks. One of the appealing features of DPC is that: it is safe in the sense that the detected inactive features are guaranteed to have zero coefficients in the solution vectors across all tasks. Thus, by removing the inactive features from the training phase, we may have substantial savings in the computational cost and memory usage without sacrificing accuracy. To the best of our knowledge, it is the first screening rule that is applicable to sparse models with multiple data matrices. A key challenge in deriving DPC is to solve a nonconvex problem. We show that we can solve for the global optimum efficiently via a properly chosen parametrization of the constraint set. Moreover, DPC has very low computational cost and can be integrated with any existing solvers. We have evaluated the proposed DPC rule on both synthetic and real data sets. The experiments indicate that DPC is very effective in identifying the inactive feature

s—especially for high dimensional data—which leads to a speedup up to several orders of magnitude.
****

Harmonic Exponential Families on Manifolds
Taco Cohen, Max Welling
In a range of fields including the geosciences, molecular biology, robotics and computer vision, one encounters problems that involve random variables on manifolds. Currently, there is a lack of flexible probabilistic models on manifolds that are fast and easy to train. We define an extremely flexible class of exponential family distributions on manifolds such as the torus, sphere, and rotation groups, and show that for these distributions the gradient of the log-likelihood can be computed efficiently using a non-commutative generalization of the Fast Fourier Transform (FFT). We discuss applications to Bayesian camera motion estimation (where harmonic exponential families serve as conjugate priors), and modelling of the spatial distribution of earthquakes on the surface of the earth. Our experimental results show that harmonic densities yield a significantly higher likelihood than the best competing method, while being orders of magnitude faster to train.
****

Training Deep Convolutional Neural Networks to Play Go
Christopher Clark, Amos Storkey
Mastering the game of Go has remained a long-standing challenge to the field of AI. Modern computer Go programs rely on processing millions of possible future positions to play well, but intuitively a stronger and more ’humanlike’ way to play the game would be to rely on pattern recognition rather than brute force computation. Following this sentiment, we train deep convolutional neural networks to play Go by training them to predict the moves made by expert Go players. To solve this problem we introduce a number of novel techniques, including a method of tying weights in the network to ’hard code’ symmetries that are expected to exist in the target function, and demonstrate in an ablation study they considerably improve performance. Our final networks are able to achieve move prediction accuracies of 41.1% and 44.4% on two different Go datasets, surpassing previous state of the art on this task by significant margins. Additionally, while previous move prediction systems have not yielded strong Go playing programs, we show that the networks trained in this work acquired high levels of skill. Our convolutional neural networks can consistently defeat the well known Go program GNU Go and win some games against state of the art Go playing program Fuego while using a fraction of the play time.
****

Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP)
Andrew Wilson, Hannes Nickisch
We introduce a new structured kernel interpolation (SKI) framework, which generalises and unifies inducing point methods for scalable Gaussian processes (GPs). SKI methods produce kernel approximations for fast computations through kernel interpolation. The SKI framework clarifies how the quality of an inducing point approach depends on the number of inducing (aka interpolation) points, interpolation strategy, and GP covariance kernel. SKI also provides a mechanism to create new scalable kernel methods, through choosing different kernel interpolation strategies. Using SKI, with local cubic kernel interpolation, we introduce KISS-GP, which is 1) more scalable than inducing point alternatives, 2) naturally enables Kronecker and Toeplitz algebra for substantial additional gains in scalability, without requiring any grid data, and 3) can be used for fast and expressive kernel learning. KISS-GP costs O(n) time and storage for GP inference. We evaluate KISS-GP for kernel matrix approximation, kernel learning, and natural sound modelling.
****

Learning Deep Structured Models
Liang-Chieh Chen, Alexander Schwing, Alan Yuille, Raquel Urtasun
Many problems in real-world applications involve predicting several random variables that are statistically related. Markov random fields (MRFs) are a great mat

hematical tool to encode such dependencies. The goal of this paper is to combine MRFs with deep learning to estimate complex representations while taking into account the dependencies between the output random variables. Towards this goal, we propose a training algorithm that is able to learn structured models jointly with deep features that form the MRF potentials. Our approach is efficient as it blends learning and inference and makes use of GPU acceleration. We demonstrate the effectiveness of our algorithm in the tasks of predicting words from noisy images, as well as tagging of Flickr photographs. We show that joint learning of the deep features and the MRF parameters results in significant performance gains.

****

Community Detection Using Time-Dependent Personalized PageRank
Haim Avron, Lior Horesh

Local graph diffusions have proven to be valuable tools for solving various graph clustering problems. As such, there has been much interest recently in efficient local algorithms for computing them. We present an efficient local algorithm for approximating a graph diffusion that generalizes both the celebrated personalized PageRank and its recent competitor/companion - the heat kernel. Our algorithm is based on writing the diffusion vector as the solution of an initial value problem, and then using a waveform relaxation approach to approximate the solution. Our experimental results suggest that it produces rankings that are distinct and competitive with the ones produced by high quality implementations of personalized PageRank and localized heat kernel, and that our algorithm is a useful addition to the toolset of localized graph diffusions.

****

Scalable Variational Inference in Log-supermodular Models
Josip Djolonga, Andreas Krause

We consider the problem of approximate Bayesian inference in log-supermodular models. These models encompass regular pairwise MRFs with binary variables, but allow to capture high order interactions, which are intractable for existing approximate inference techniques such as belief propagation, mean field and variants. We show that a recently proposed variational approach to inference in log-supermodular models – L-Field – reduces to the widely studied minimum norm problem for submodular minimization. This insight allows to leverage powerful existing tools, and allows solving the variational problem orders of magnitude more efficiently than previously possible. We then provide another natural interpretation of L-Field, demonstrating that it exactly minimizes a specific type of Renyi divergence measure. This insight sheds light on the nature of the variational approximations produced by L-Field. Furthermore, we show how to perform parallel inference as message passing in a suitable factor graph at a linear convergence rate, without having to sum up over all the configurations of the factor. Finally, we apply our approach to a challenging image segmentation task. Our experiments confirm scalability of our approach, high quality of the marginals and the benefit of incorporating higher order potentials.

****

Variational Inference for Gaussian Process Modulated Poisson Processes
Chris Lloyd, Tom Gunter, Michael Osborne, Stephen Roberts

We present the first fully variational Bayesian inference scheme for continuous Gaussian-process-modulated Poisson processes. Such point processes are used in a variety of domains, including neuroscience, geo-statistics and astronomy, but their use is hindered by the computational cost of existing inference schemes. Our scheme: requires no discretisation of the domain; scales linearly in the number of observed events; and is many orders of magnitude faster than previous sampling based approaches. The resulting algorithm is shown to outperform standard methods on synthetic examples, coal mining disaster data and in the prediction of Malaria incidences in Kenya.

****

Scalable Deep Poisson Factor Analysis for Topic Modeling
Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, Lawrence Carin

A new framework for topic modeling is developed, based on deep graphical models,

where interactions between topics are inferred through deep latent binary hierarchies. The proposed multi-layer model employs a deep sigmoid belief network or restricted Boltzmann machine, the bottom binary layer of which selects topics for use in a Poisson factor analysis model. Under this setting, topics live on the bottom layer of the model, while the deep specification serves as a flexible prior for revealing topic structure. Scalable inference algorithms are derived by applying Bayesian conditional density filtering algorithm, in addition to extending recently proposed work on stochastic gradient thermostats. Experimental results on several corpora show that the proposed approach readily handles very large collections of text documents, infers structured topic representations, and obtains superior test perplexities when compared with related models.
****

Hidden Markov Anomaly Detection
Nico Goernitz, Mikio Braun, Marius Kloft
We introduce a new anomaly detection methodology for data with latent dependency structure. As a particular instantiation, we derive a hidden Markov anomaly detector that extends the regular one-class support vector machine. We optimize the approach, which is non-convex, via a DC (difference of convex functions) algorithm, and show that the parameter v can be conveniently used to control the number of outliers in the model. The empirical evaluation on artificial and real data from the domains of computational biology and computational sustainability shows that the approach can achieve significantly higher anomaly detection performance than the regular one-class SVM.
****

Robust Estimation of Transition Matrices in High Dimensional Heavy-tailed Vector Autoregressive Processes
Huitong Qiu, Sheng Xu, Fang Han, Han Liu, Brian Caffo
Gaussian vector autoregressive (VAR) processes have been extensively studied in the literature. However, Gaussian assumptions are stringent for heavy-tailed time series that frequently arises in finance and economics. In this paper, we develop a unified framework for modeling and estimating heavy-tailed VAR processes. In particular, we generalize the Gaussian VAR model by an elliptical VAR model that naturally accommodates heavy-tailed time series. Under this model, we develop a quantile-based robust estimator for the transition matrix of the VAR process. We show that the proposed estimator achieves parametric rates of convergence in high dimensions. This is the first work in analyzing heavy-tailed high dimensional VAR processes. As an application of the proposed framework, we investigate Granger causality in the elliptical VAR process, and show that the robust transition matrix estimator induces sign-consistent estimators of Granger causality. The empirical performance of the proposed methodology is demonstrated by both synthetic and real data. We show that the proposed estimator is robust to heavy tails, and exhibit superior performance in stock price prediction.
****

Convex Calibrated Surrogates for Hierarchical Classification
Harish Ramaswamy, Ambuj Tewari, Shivani Agarwal
Hierarchical classification problems are multiclass supervised learning problems with a pre-defined hierarchy over the set of class labels. In this work, we study the consistency of hierarchical classification algorithms with respect to a natural loss, namely the tree distance metric on the hierarchy tree of class labels, via the usage of calibrated surrogates. We first show that the Bayes optimal classifier for this loss classifies an instance according to the deepest node in the hierarchy such that the total conditional probability of the subtree rooted at the node is greater than \frac12. We exploit this insight to develop new consistent algorithm for hierarchical classification, that makes use of an algorithm known to be consistent for the “multiclass classification with reject option (MCRO)” problem as a sub-routine. Our experiments on a number of benchmark datasets show that the resulting algorithm, which we term OvA-Cascade, gives improved performance over other state-of-the-art hierarchical classification algorithms.
****

Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks

Jose Miguel Hernandez-Lobato, Ryan Adams
Large multilayer neural networks trained with backpropagation have recently achieved state-of-the-art results in a wide range of problems. However, using backprop for neural net learning still has some disadvantages, e.g., having to tune a large number of hyperparameters to the data, lack of calibrated probabilistic predictions, and a tendency to overfit the training data. In principle, the Bayesian approach to learning neural networks does not have these problems. However, existing Bayesian techniques lack scalability to large dataset and network sizes. In this work we present a novel scalable method for learning Bayesian neural networks, called probabilistic backpropagation (PBP). Similar to classical backpropagation, PBP works by computing a forward propagation of probabilities through the network and then doing a backward computation of gradients. A series of experiments on ten real-world datasets show that PBP is significantly faster than other techniques, while offering competitive predictive abilities. Our experiments also show that PBP provides accurate estimates of the posterior variance on the network weights.
****

Active Nearest Neighbors in Changing Environments
Christopher Berlind, Ruth Urner
While classic machine learning paradigms assume training and test data are generated from the same process, domain adaptation addresses the more realistic setting in which the learner has large quantities of labeled data from some source task but limited or no labeled data from the target task it is attempting to learn. In this work, we give the first formal analysis showing that using active learning for domain adaptation yields a way to address the statistical challenges inherent in this setting. We propose a novel nonparametric algorithm, ANDA, that combines an active nearest neighbor querying strategy with nearest neighbor prediction. We provide analyses of its querying behavior and of finite sample convergence rates of the resulting classifier under covariate shift. Our experiments show that ANDA successfully corrects for dataset bias in multi-class image categorization.
****

Bipartite Edge Prediction via Transductive Learning over Product Graphs
Hanxiao Liu, Yiming Yang
This paper addresses the problem of predicting the missing edges of a bipartite graph where each side of the vertices has its own intrinsic structure. We propose a new optimization framework to map the two sides of the intrinsic structures onto the manifold structure of the edges via a graph product, and to reduce the original problem to vertex label propagation over the product graph. This framework enjoys flexible choices in the formulation of graph products, and supports a rich family of graph transduction schemes with scalable inference. Experiments on benchmark datasets for collaborative filtering, citation network analysis and prerequisite prediction of online courses show advantageous performance of the proposed approach over other state-of-the-art methods.
****

Trust Region Policy Optimization
John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, Philipp Moritz
In this article, we describe a method for optimizing control policies, with guaranteed monotonic improvement. By making several approximations to the theoretically-justified scheme, we develop a practical algorithm, called Trust Region Policy Optimization (TRPO). This algorithm is effective for optimizing large nonlinear policies such as neural networks. Our experiments demonstrate its robust performance on a wide variety of tasks: learning simulated robotic swimming, hopping, and walking gaits; and playing Atari games using images of the screen as input. Despite its approximations that deviate from the theory, TRPO tends to give monotonic improvement, with little tuning of hyperparameters.
****

Discovering Temporal Causal Relations from Subsampled Data
Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, Philipp Geiger
Granger causal analysis has been an important tool for causal analysis for time

series in various fields, including neuroscience and economics, and recently it has been extended to include instantaneous effects between the time series to explain the contemporaneous dependence in the residuals. In this paper, we assume that the time series at the true causal frequency follow the vector autoregressive model. We show that when the data resolution becomes lower due to subsampling, neither the original Granger causal analysis nor the extended one is able to discover the underlying causal relations. We then aim to answer the following question: can we estimate the temporal causal relations at the right causal frequency from the subsampled data? Traditionally this suffers from the identifiability problems: under the Gaussianity assumption of the data, the solutions are generally not unique. We prove that, however, if the noise terms are non-Gaussian, the underlying model for the high frequency data is identifiable from subsampled data under mild conditions. We then propose an Expectation-Maximization (EM) approach and a variational inference approach to recover temporal causal relations from such subsampled data. Experimental results on both simulated and real data are reported to illustrate the performance of the proposed approaches.
****

Preference Completion: Large-scale Collaborative Ranking from Pairwise Comparisons

Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, Inderjit Dhillon

In this paper we consider the collaborative ranking setting: a pool of users each provides a set of pairwise preferences over a small subset of the set of d possible items; from these we need to predict each user’s preferences for items s/he has not yet seen. We do so via fitting a rank r score matrix to the pairwise data, and provide two main contributions: (a) We show that an algorithm based on convex optimization provides good generalization guarantees once each user provides as few as O(r \log^2 d) pairwise comparisons — essentially matching the sample complexity required in the related matrix completion setting (which uses actual numerical as opposed to pairwise information), and also matching a lower bound we establish here. (b) We develop a large-scale non-convex implementation, which we call AltSVM, which trains a factored form of the matrix via alternating minimization (which we show reduces to alternating SVM problems), and scales and parallelizes very well to large problem settings. It also outperforms common baselines on many moderately large popular collaborative filtering datasets in both NDCG and other measures of ranking performance.
****

Causal Inference by Identification of Vector Autoregressive Processes with Hidden Components

Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, Dominik Janzing

A widely applied approach to causal inference from a time series X, often referred to as “(linear) Granger causal analysis”, is to simply regress present on past and interpret the regression matrix \hatB causally. However, if there is an unmeasured time series Z that influences X, then this approach can lead to wrong causal conclusions, i.e., distinct from those one would draw if one had additional information such as Z. In this paper we take a different approach: We assume that X together with some hidden Z forms a first order vector autoregressive (VAR) process with transition matrix A, and argue why it is more valid to interpret A causally instead of \hatB. Then we examine under which conditions the most important parts of A are identifiable or almost identifiable from only X. Essentially, sufficient conditions are (1) non-Gaussian, independent noise or (2) no influence from X to Z. We present two estimation algorithms that are tailored towards conditions (1) and (2), respectively, and evaluate them on synthetic and real-world data. We discuss how to check the model using X.
****

On Symmetric and Asymmetric LSHs for Inner Product Search

Behnam Neyshabur, Nathan Srebro

We consider the problem of designing locality sensitive hashes (LSH) for inner product similarity, and of the power of asymmetric hashes in this context. Shrivastava and Li (2014a) argue that there is no symmetric LSH for the problem and propose an asymmetric LSH based on different mappings for query and database point

s. However, we show there does exist a simple symmetric LSH that enjoys stronger guarantees and better empirical performance than the asymmetric LSH they suggest. We also show a variant of the settings where asymmetry is in-fact needed, but there a different asymmetric LSH is required.
****

The Kendall and Mallows Kernels for Permutations
Yunlong Jiao, Jean-Philippe Vert
We show that the widely used Kendall tau correlation coefficient is a positive definite kernel for permutations. It offers a computationally attractive alternative to more complex kernels on the symmetric group to learn from rankings, or to learn to rank. We show how to extend it to partial rankings or rankings with uncertainty, and demonstrate promising results on high-dimensional classification problems in biomedical applications.
****

Bayesian Multiple Target Localization
Purnima Rajan, Weidong Han, Raphael Sznitman, Peter Frazier, Bruno Jedynak
We consider the problem of quickly localizing multiple targets by asking questions of the form “How many targets are within this set" while obtaining noisy answers. This setting is a generalization to multiple targets of the game of 20 questions in which only a single target is queried. We assume that the targets are points on the real line, or in a two dimensional plane for the experiments, drawn independently from a known distribution. We evaluate the performance of a policy using the expected entropy of the posterior distribution after a fixed number of questions with noisy answers. We derive a lower bound for the value of this problem and study a specific policy, named the dyadic policy. We show that this policy achieves a value which is no more than twice this lower bound when answers are noise-free, and show a more general constant factor approximation guarantee for the noisy setting. We present an empirical evaluation of this policy on simulated data for the problem of detecting multiple instances of the same object in an image. Finally, we present experiments on localizing multiple faces simultaneously on real images.
****

Submodularity in Data Subset Selection and Active Learning
Kai Wei, Rishabh Iyer, Jeff Bilmes
We study the problem of selecting a subset of big data to train a classifier while incurring minimal performance loss. We show the connection of submodularity to the data likelihood functions for Naive Bayes (NB) and Nearest Neighbor (NN) classifiers, and formulate the data subset selection problems for these classifiers as constrained submodular maximization. Furthermore, we apply this framework to active learning and propose a novel scheme filtering active submodular selection (FASS), where we combine the uncertainty sampling method with a submodular data subset selection framework. We extensively evaluate the proposed framework on text categorization and handwritten digit recognition tasks with four different classifiers, including Deep Neural Network (DNN) based classifiers. Empirical results indicate that the proposed framework yields significant improvement over the state-of-the-art algorithms on all classifiers.
****

Variational Generative Stochastic Networks with Collaborative Shaping
Philip Bachman, Doina Precup
We develop an approach to training generative models based on unrolling a variational auto-encoder into a Markov chain, and shaping the chain’s trajectories using a technique inspired by recent work in Approximate Bayesian computation. We show that the global minimizer of the resulting objective is achieved when the generative model reproduces the target distribution. To allow finer control over the behavior of the models, we add a regularization term inspired by techniques used for regularizing certain types of policy search in reinforcement learning. We present empirical results on the MNIST and TFD datasets which show that our approach offers state-of-the-art performance, both quantitatively and from a qualitative point of view.
****

Adding vs. Averaging in Distributed Primal-Dual Optimization
Chenxin Ma, Virginia Smith, Martin Jaggi, Michael Jordan, Peter Richtarik, Marti n Takac
Distributed optimization methods for large-scale machine learning suffer from a communication bottleneck. It is difficult to reduce this bottleneck while still efficiently and accurately aggregating partial work from different machines. In this paper, we present a novel generalization of the recent communication-effici ent primal-dual framework (COCOA) for distributed optimization. Our framework, C OCOA+, allows for additive combination of local updates to the global parameters at each iteration, whereas previous schemes only allow conservative averaging. We give stronger (primal-dual) convergence rate guarantees for both COCOA as wel l as our new variants, and generalize the theory for both methods to cover non-s mooth convex loss functions. We provide an extensive experimental comparison tha t shows the markedly improved performance of COCOA+ on several real-world distri buted datasets, especially when scaling up the number of machines.
****
Feature-Budgeted Random Forest
Feng Nan, Joseph Wang, Venkatesh Saligrama
We seek decision rules for \it prediction-time cost reduction, where complete da ta is available for training, but during prediction-time, each feature can only be acquired for an additional cost. We propose a novel random forest algorithm t o minimize prediction error for a user-specified \it average feature acquisition budget. While random forests yield strong generalization performance, they do n ot explicitly account for feature costs and furthermore require low correlation among trees, which amplifies costs. Our random forest grows trees with low acqui sition cost and high strength based on greedy minimax cost-weighted-impurity spl its. Theoretically, we establish near-optimal acquisition cost guarantees for ou r algorithm. Empirically, on a number of benchmark datasets we demonstrate compe titive accuracy-cost curves against state-of-the-art prediction-time algorithms.
****
Entropic Graph-based Posterior Regularization
Maxwell Libbrecht, Michael Hoffman, Jeff Bilmes, William Noble
Graph smoothness objectives have achieved great success in semi-supervised learn ing but have not yet been applied extensively to unsupervised generative models. We define a new class of entropic graph-based posterior regularizers that augme nt a probabilistic model by encouraging pairs of nearby variables in a regulariz ation graph to have similar posterior distributions. We present a three-way alte rnating optimization algorithm with closed-form updates for performing inference on this joint model and learning its parameters. This method admits updates lin ear in the degree of the regularization graph, exhibits monotone convergence and is easily parallelizable. We are motivated by applications in computational bio logy in which temporal models such as hidden Markov models are used to learn a h uman-interpretable representation of genomic data. On a synthetic problem, we sh ow that our method outperforms existing methods for graph-based regularization a nd a comparable strategy for incorporating long-range interactions using existin g methods for approximate inference. Using genome-scale functional genomics data , we integrate genome 3D interaction data into existing models for genome annota tion and demonstrate significant improvements in predicting genomic activity.
****
Unsupervised Riemannian Metric Learning for Histograms Using Aitchison Transform ations
Tam Le, Marco Cuturi
Many applications in machine learning handle bags of features or histograms rath er than simple vectors. In that context, defining a proper geometry to compare h istograms can be crucial for many machine learning algorithms. While one might b e tempted to use a default metric such as the Euclidean metric, empirical eviden ce shows this may not be the best choice when dealing with observations that lie in the probability simplex. Additionally, it might be desirable to choose a met ric adaptively based on data. We consider in this paper the problem of learning a Riemannian metric on the simplex given unlabeled histogram data. We follow the

approach of Lebanon(2006), who proposed to estimate such a metric within a parametric family by maximizing the inverse volume of a given data set of points under that metric. The metrics we consider on the multinomial simplex are pull-back metrics of the Fisher information parameterized by operations within the simplex known as Aitchison(1982) transformations. We propose an algorithmic approach to maximize inverse volumes using sampling and contrastive divergences. We provide experimental evidence that the metric obtained under our proposal outperforms alternative approaches.

****

Low-Rank Matrix Recovery from Row-and-Column Affine Measurements

Or Zuk, Avishai Wagner

We propose and study a row-and-column affine measurement scheme for low-rank matrix recovery. Each measurement is a linear combination of elements in one row or one column of a matrix X. This setting arises naturally in applications from different domains. However, current algorithms developed for standard matrix recovery problems do not perform well in our case, hence the need for developing new algorithms and theory for our problem. We propose a simple algorithm for the problem based on Singular Value Decomposition (SVD) and least-squares (LS), which we term alg. We prove that (a simplified version of) our algorithm can recover X exactly with the minimum possible number of measurements in the noiseless case. In the general noisy case, we prove performance guarantees on the reconstruction accuracy under the Frobenius norm. In simulations, our row-and-column design and alg algorithm show improved speed, and comparable and in some cases better accuracy compared to standard measurements designs and algorithms. Our theoretical and experimental results suggest that the proposed row-and-column affine measurements scheme, together with our recovery algorithm, may provide a powerful framework for affine matrix reconstruction.

****

Algorithms for the Hard Pre-Image Problem of String Kernels and the General Problem of String Prediction

Sébastien Giguère, Amélie Rolland, Francois Laviolette, Mario Marchand

We address the pre-image problem encountered in structured output prediction and the one of finding a string maximizing the prediction function of various kernel-based classifiers and regressors. We demonstrate that these problems reduce to a common combinatorial problem valid for many string kernels. For this problem, we propose an upper bound on the prediction function which has low computational complexity and which can be used in a branch and bound search algorithm to obtain optimal solutions. We also show that for many string kernels, the complexity of the problem increases significantly when the kernel is normalized. On the optical word recognition task, the exact solution of the pre-image problem is shown to significantly improve the prediction accuracy in comparison with an approximation found by the best known heuristic. On the task of finding a string maximizing the prediction function of kernel-based classifiers and regressors, we highlight that existing methods can be biased toward long strings that contain many repeated symbols. We demonstrate that this bias is removed when using normalized kernels. Finally, we present results for the discovery of lead compounds in drug discovery. The source code can be found at https://github.com/a-ro/preimage

****

A Multitask Point Process Predictive Model

Wenzhao Lian, Ricardo Henao, Vinayak Rao, Joseph Lucas, Lawrence Carin

Point process data are commonly observed in fields like healthcare and social science. Designing predictive models for such event streams is an under-explored problem, due to often scarce training data. In this work we propose a multitask point process model, leveraging information from all tasks via a hierarchical Gaussian process (GP). Nonparametric learning functions implemented by a GP, which map from past events to future rates, allow analysis of flexible arrival patterns. To facilitate efficient inference, we propose a sparse construction for this hierarchical model, and derive a variational Bayes method for learning and inference. Experimental results are shown on both synthetic data and an application on real electronic health records.

****
A Hybrid Approach for Probabilistic Inference using Random Projections
Michael Zhu, Stefano Ermon
We introduce a new meta-algorithm for probabilistic inference in graphical models based on random projections. The key idea is to use approximate inference algorithms for an (exponentially) large number of samples, obtained by randomly projecting the original statistical model using universal hash functions. In the case where the approximate inference algorithm is a variational approximation, this approach can be viewed as interpolating between sampling-based and variational techniques. The number of samples used controls the trade-off between the accuracy of the approximate inference algorithm and the variance of the estimator. We show empirically that by using random projections, we can improve the accuracy of common approximate inference algorithms.
****
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio
Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. We validate the use of attention with state-of-the-art performance on three benchmark datasets: Flickr8k, Flickr30k and MS COCO.
****
Learning to Search Better than Your Teacher
Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, John Langford
Methods for learning to search for structured prediction typically imitate a reference policy, with existing theoretical guarantees demonstrating low regret compared to that reference. This is unsatisfactory in many applications where the reference policy is suboptimal and the goal of learning is to improve upon it. Can learning to search work even when the reference is poor? We provide a new learning to search algorithm, LOLS, which does well relative to the reference policy, but additionally guarantees low regret compared to deviations from the learned policy: a local-optimality guarantee. Consequently, LOLS can improve upon the reference policy, unlike previous algorithms. This enables us to develop structured contextual bandits, a partial information structured prediction setting with many potential applications.
****
Gated Feedback Recurrent Neural Networks
Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, Yoshua Bengio
In this work, we propose a novel recurrent neural network (RNN) architecture. The proposed RNN, gated-feedback RNN (GF-RNN), extends the existing approach of stacking multiple recurrent layers by allowing and controlling signals flowing from upper recurrent layers to lower layers using a global gating unit for each pair of layers. The recurrent signals exchanged between layers are gated adaptively based on the previous hidden states and the current input. We evaluated the proposed GF-RNN with different types of recurrent units, such as tanh, long short-term memory and gated recurrent units, on the tasks of character-level language modeling and Python program evaluation. Our empirical evaluation of different RNN units, revealed that in both tasks, the GF-RNN outperforms the conventional approaches to build deep stacked RNNs. We suggest that the improvement arises because the GF-RNN can adaptively assign different layers to different timescales and layer-to-layer interactions (including the top-down ones which are not usually present in a stacked RNN) by learning to gate these interactions.
****
Context-based Unsupervised Data Fusion for Decision Making
Erfan Soltanmohammadi, Mort Naraghi-Pour, Mihaela Schaar

Big Data received from sources such as social media, in-stream monitoring systems, networks, and markets is often mined for discovering patterns, detecting anomalies, and making decisions or predictions. In distributed learning and real-time processing of Big Data, ensemble-based systems in which a fusion center (FC) is used to combine the local decisions of several classifiers, have shown to be superior to single expert systems. However, optimal design of the FC requires knowledge of the accuracy of the individual classifiers which, in many cases, is not available. Moreover, in many applications supervised training of the FC is not feasible since the true labels of the data set are not available. In this paper, we propose an unsupervised joint estimation-detection scheme to estimate the accuracies of the local classifiers as functions of data context and to fuse the local decisions of the classifiers. Numerical results show the dramatic improvement of the proposed method as compared with the state of the art approaches.
****

Phrase-based Image Captioning
Remi Lebret, Pedro Pinheiro, Ronan Collobert
Generating a novel textual description of an image is an interesting problem that connects computer vision and natural language processing. In this paper, we present a simple model that is able to generate descriptive sentences given a sample image. This model has a strong focus on the syntax of the descriptions. We train a purely linear model to embed an image representation (generated from a previously trained Convolutional Neural Network) into a multimodal space that is common to the images and the phrases that are used to described them. The system is then able to infer phrases from a given image sample. Based on the sentence description statistics, we propose a simple language model that can produce relevant descriptions for a given test image using the phrases inferred. Our approach, which is considerably simpler than state-of-the-art models, achieves comparable results in two popular datasets for the task: Flickr30k and the recently proposed Microsoft COCO.
****

Celeste: Variational inference for a generative model of astronomical images
Jeffrey Regier, Andrew Miller, Jon McAuliffe, Ryan Adams, Matt Hoffman, Dustin Lang, David Schlegel, Mr Prabhat
We present a new, fully generative model of optical telescope image sets, along with a variational procedure for inference. Each pixel intensity is treated as a Poisson random variable, with a rate parameter dependent on latent properties of stars and galaxies. Key latent properties are themselves random, with scientific prior distributions constructed from large ancillary data sets. We check our approach on synthetic images. We also run it on images from a major sky survey, where it exceeds the performance of the current state-of-the-art method for locating celestial bodies and measuring their colors.
****

Distributional Rank Aggregation, and an Axiomatic Analysis
Adarsh Prasad, Harsh Pareek, Pradeep Ravikumar
The rank aggregation problem has been studied with varying desiderata in varied communities such as Theoretical Computer Science, Statistics, Information Retrieval and Social Welfare Theory. We introduce a variant of this problem we call distributional rank aggregation, where the ranking data is only available via the induced distribution over the set of all permutations. We provide a novel translation of the usual social welfare theory axioms to this setting. As we show this allows for a more quantitative characterization of these axioms: which then are not only less prone to misinterpretation, but also allow simpler proofs for some key impossibility theorems. Most importantly, these quantitative characterizations lead to natural and novel relaxations of these axioms, which as we show, allow us to get around celebrated impossibility results in social choice theory. We are able to completely characterize the class of positional scoring rules with respect to our axioms and show that Borda Count is optimal in a certain sense.
****

Gradient-based Hyperparameter Optimization through Reversible Learning
Dougal Maclaurin, David Duvenaud, Ryan Adams

Tuning hyperparameters of learning algorithms is hard because gradients are usua
lly unavailable. We compute exact gradients of cross-validation performance with
 respect to all hyperparameters by chaining derivatives backwards through the en
tire training procedure. These gradients allow us to optimize thousands of hyper
parameters, including step-size and momentum schedules, weight initialization di
stributions, richly parameterized regularization schemes, and neural network arc
hitectures. We compute hyperparameter gradients by exactly reversing the dynamic
s of stochastic gradient descent with momentum.
****

Bimodal Modelling of Source Code and Natural Language
Miltos Allamanis, Daniel Tarlow, Andrew Gordon, Yi Wei
We consider the problem of building probabilistic models that jointly model shor
t natural language utterances and source code snippets. The aim is to bring toge
ther recent work on statistical modelling of source code and work on bimodal mod
els of images and natural language. The resulting models are useful for a variet
y of tasks that involve natural language and source code. We demonstrate their p
erformance on two retrieval tasks: retrieving source code snippets given a natur
al language query, and retrieving natural language descriptions given a source c
ode query (i.e., source code captioning). The experiments show there to be promi
se in this direction, and that modelling the structure of source code is helpful
 towards the retrieval tasks.
****

Cheap Bandits
Manjesh Hanawal, Venkatesh Saligrama, Michal Valko, Remi Munos
We consider stochastic sequential learning problems where the learner can observ
e the average reward of several actions. Such a setting is interesting in many a
pplications involving monitoring and surveillance, where the set of the actions
to observe represent some (geographical) area. The importance of this setting is
 that in these applications, it is actually cheaper to observe average reward of
 a group of actions rather than the reward of a single action. We show that when
 the reward is smooth over a given graph representing the neighboring actions, w
e can maximize the cumulative reward of learning while minimizing the sensing co
st. In this paper we propose CheapUCB, an algorithm that matches the regret guar
antees of the known algorithms for this setting and at the same time guarantees
a linear cost again over them. As a by-product of our analysis, we establish a $\Omega
(\sqrt(dT))$ lower bound on the cumulative regret of spectral bandits for a class
 of graphs with effective dimension d.
****

Subsampling Methods for Persistent Homology
Frederic Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rina
ldo, Larry Wasserman
Persistent homology is a multiscale method for analyzing the shape of sets and f
unctions from point cloud data arising from an unknown distribution supported on
 those sets. When the size of the sample is large, direct computation of the per
sistent homology is prohibitive due to the combinatorial nature of the existing
algorithms. We propose to compute the persistent homology of several subsamples
of the data and then combine the resulting estimates. We study the risk of two e
stimators and we prove that the subsampling approach carries stable topological
information while achieving a great reduction in computational complexity.
****

An embarrassingly simple approach to zero-shot learning
Bernardino Romera-Paredes, Philip Torr
Zero-shot learning consists in learning how to recognize new concepts by just ha
ving a description of them. Many sophisticated approaches have been proposed to
address the challenges this problem comprises. In this paper we describe a zero-
shot learning approach that can be implemented in just one line of code, yet it
is able to outperform state of the art approaches on standard datasets. The appr
oach is based on a more general framework which models the relationships between
 features, attributes, and classes as a two linear layers network, where the wei
ghts of the top layer are not learned but are given by the environment. We furth

er provide a learning bound on the generalization error of this kind of approach es, by casting them as domain adaptation methods. In experiments carried out on three standard real datasets, we found that our approach is able to perform sign ificantly better than the state of art on all of them, obtaining a ratio of impr ovement up to 17%.
****

Binary Embedding: Fundamental Limits and Fast Algorithm
Xinyang Yi, Constantine Caramanis, Eric Price
Binary embedding is a nonlinear dimension reduction methodology where high dimen sional data are embedded into the Hamming cube while preserving the structure of the original space. Specifically, for an arbitrary N distinct points in \mathbb S^p-1, our goal is to encode each point using m-dimensional binary strings such that we can reconstruct their geodesic distance up to δuniform distortion. Exist ing binary embedding algorithms either lack theoretical guarantees or suffer fro m running time O(mp). We make three contributions: (1) we establish a lower boun d that shows any binary embedding oblivious to the set of points requires m =Ω(\ frac1δ^2\logN) bits and a similar lower bound for non-oblivious embeddings into Hamming distance; (2) we propose a novel fast binary embedding algorithm with pr ovably optimal bit complexity m = O(\frac1 δ^2\logN) and near linear running tim e O(p \log p) whenever \log N ■δ\sqrtp, with a slightly worse running time for l arger \log N; (3) we also provide an analytic result about embedding a general s et of points K ⊂\mathbbS^p-1 with even infinite size. Our theoretical findings a re supported through experiments on both synthetic and real data sets.
****

Scalable Bayesian Optimization Using Deep Neural Networks
Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, Ryan Adams
Bayesian optimization is an effective methodology for the global optimization of functions with expensive evaluations. It relies on querying a distribution over functions defined by a relatively cheap surrogate model. An accurate model for this distribution over functions is critical to the effectiveness of the approac h, and is typically fit using Gaussian processes (GPs). However, since GPs scale cubically with the number of observations, it has been challenging to handle ob jectives whose optimization requires many evaluations, and as such, massively pa rallelizing the optimization. In this work, we explore the use of neural network s as an alternative to GPs to model distributions over functions. We show that p erforming adaptive basis function regression with a neural network as the parame tric form performs competitively with state-of-the-art GP-based approaches, but scales linearly with the number of data rather than cubically. This allows us to achieve a previously intractable degree of parallelism, which we apply to large scale hyperparameter optimization, rapidly finding competitive models on benchm ark object recognition tasks using convolutional networks, and image caption gen eration using neural language models.
****

How Hard is Inference for Structured Prediction?
Amir Globerson, Tim Roughgarden, David Sontag, Cafer Yildirim
Structured prediction tasks in machine learning involve the simultaneous predict ion of multiple labels. This is often done by maximizing a score function on the space of labels, which decomposes as a sum of pairwise elements, each depending on two specific labels. The goal of this paper is to develop a theoretical expl anation of the empirical effectiveness of heuristic inference algorithms for sol ving such structured prediction problems. We study the minimum-achievable expect ed Hamming error in such problems, highlighting the case of 2D grid graphs, whic h are common in machine vision applications. Our main theorems provide tight upp er and lower bounds on this error, as well as a polynomial-time algorithm that a chieves the bound.
****

Online Time Series Prediction with Missing Data
Oren Anava, Elad Hazan, Assaf Zeevi
We consider the problem of time series prediction in the presence of missing dat

a. We cast the problem as an online learning problem in which the goal of the le
arner is to minimize prediction error. We then devise an efficient algorithm for
 the problem, which is based on autoregressive model, and does not assume any st
ructure on the missing data nor on the mechanism that generates the time series.
 We show that our algorithm’s performance asymptotically approaches the performa
nce of the best AR predictor in hindsight, and corroborate the theoretic results
 with an empirical study on synthetic and real-world data.
****

Proteins, Particles, and Pseudo-Max-Marginals: A Submodular Approach
Jason Pacheco, Erik Sudderth
Variants of max-product (MP) belief propagation effectively find modes of many c
omplex graphical models, but are limited to discrete distributions. Diverse part
icle max-product (D-PMP) robustly approximates max-product updates in continuous
 MRFs using stochastically sampled particles, but previous work was specialized
to tree-structured models. Motivated by the challenging problem of protein side
chain prediction, we extend D-PMP in several key ways to create a generic MAP in
ference algorithm for loopy models. We define a modified diverse particle select
ion objective that is provably submodular, leading to an efficient greedy algori
thm with rigorous optimality guarantees, and corresponding max-marginal error bo
unds. We further incorporate tree-reweighted variants of the MP algorithm to all
ow provable verification of global MAP recovery in many models. Our general-purp
ose Matlab library is applicable to a wide range of pairwise graphical models, a
nd we validate our approach using optical flow benchmarks. We further demonstrat
e superior side chain prediction accuracy compared to baseline algorithms from t
he state-of-the-art Rosetta package.
****

A Fast Variational Approach for Learning Markov Random Field Language Models
Yacine Jernite, Alexander Rush, David Sontag
Language modelling is a fundamental building block of natural language processin
g. However, in practice the size of the vocabulary limits the distributions appl
icable for this task: specifically, one has to either resort to local optimizati
on methods, such as those used in neural language models, or work with heavily c
onstrained distributions. In this work, we take a step towards overcoming these
difficulties. We present a method for global-likelihood optimization of a Markov
 random field language model exploiting long-range contexts in time independent
of the corpus size. We take a variational approach to optimizing the likelihood
and exploit underlying symmetries to greatly simplify learning. We demonstrate t
he efficiency of this method both for language modelling and for part-of-speech
tagging.
****

Removing systematic errors for exoplanet search via latent causes
Bernhard Schölkopf, David Hogg, Dun Wang, Dan Foreman-Mackey, Dominik Janzing, C
arl-Johann Simon-Gabriel, Jonas Peters
We describe a method for removing the effect of confounders in order to reconstr
uct a latent quantity of interest. The method, referred to as half-sibling regre
ssion, is inspired by recent work in causal inference using additive noise model
s. We provide a theoretical justification and illustrate the potential of the me
thod in a challenging astronomy application.
****

Scalable Nonparametric Bayesian Inference on Point Processes with Gaussian Proce
sses
Yves-Laurent Kom Samo, Stephen Roberts
In this paper we propose an efficient, scalable non-parametric Gaussian process
model for inference on Poisson point processes. Our model does not resort to gri
dding the domain or to introducing latent thinning points. Unlike competing mode
ls that scale as O(n^3) over n data points, our model has a complexity O(nk^2) w
here k << n. We propose a MCMC sampler and show that the model obtained is faste
r, more accurate and generates less correlated samples than competing approaches
 on both synthetic and real-life data. Finally, we show that our model easily ha
ndles data sizes not considered thus far by alternate approaches.

****
Correlation Clustering in Data Streams
KookJin Ahn, Graham Cormode, Sudipto Guha, Andrew McGregor, Anthony Wirth
In this paper, we address the problem of \emphcorrelation clustering in the dyna mic data stream model. The stream consists of updates to the edge weights of a g raph on n nodes and the goal is to find a node-partition such that the end-point s of negative-weight edges are typically in different clusters whereas the end-p oints of positive-weight edges are typically in the same cluster. We present pol ynomial-time, O(n·\textpolylog n)-space approximation algorithms for natural pro blems that arise. We first develop data structures based on linear sketches that allow the “quality” of a given node-partition to be measured. We then combine t hese data structures with convex programming and sampling techniques to solve th e relevant approximation problem. However the standard LP and SDP formulations a re not obviously solvable in O(n·\textpolylog n)-space. Our work presents space- efficient algorithms for the convex programming required, as well as approaches to reduce the adaptivity of the sampling. Note that the improved space and runni ng-time bounds achieved from streaming algorithms are also useful for offline se ttings such as MapReduce models.
****
Learning Scale-Free Networks by Dynamic Node Specific Degree Prior
Qingming Tang, Siqi Sun, Jinbo Xu
Learning network structure underlying data is an important problem in machine le arning. This paper presents a novel degree prior to study the inference of scale -free networks, which are widely used to model social and biological networks. I n particular, this paper formulates scale-free network inference using Gaussian Graphical model (GGM) regularized by a node degree prior. Our degree prior not o nly promotes a desirable global degree distribution, but also exploits the estim ated degree of an individual node and the relative strength of all the edges of a single node. To fulfill this, this paper proposes a ranking-based method to dy namically estimate the degree of a node, which makes the resultant optimization problem challenging to solve. To deal with this, this paper presents a novel ADM M (alternating direction method of multipliers) procedure. Our experimental resu lts on both synthetic and real data show that our prior not only yields a scale- free network, but also produces many more correctly predicted edges than existin g scale-free inducing prior, hub-inducing prior and the l_1 norm.
****
Deep Unsupervised Learning using Nonequilibrium Thermodynamics
Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli
A central problem in machine learning involves modeling complex data-sets using highly flexible families of probability distributions in which learning, samplin g, inference, and evaluation are still analytically or computationally tractable . Here, we develop an approach that simultaneously achieves both flexibility and tractability. The essential idea, inspired by non-equilibrium statistical physi cs, is to systematically and slowly destroy structure in a data distribution thr ough an iterative forward diffusion process. We then learn a reverse diffusion p rocess that restores structure in data, yielding a highly flexible and tractable generative model of the data. This approach allows us to rapidly learn, sample from, and evaluate probabilities in deep generative models with thousands of lay ers or time steps, as well as to compute conditional and posterior probabilities under the learned model. We additionally release an open source reference imple mentation of the algorithm.
****
Modeling Order in Neural Word Embeddings at Scale
Andrew Trask, David Gilmore, Matthew Russell
Natural Language Processing (NLP) systems commonly leverage bag-of-words co-occu rrence techniques to capture semantic and syntactic word relationships. The resu lting word-level distributed representations often ignore morphological informat ion, though character-level embeddings have proven valuable to NLP tasks. We pro pose a new neural language model incorporating both word order and character ord er in its embedding. The model produces several vector spaces with meaningful su

bstructure, as evidenced by its performance of 85.8% on a recent word-analogy ta
sk, exceeding best published syntactic word-analogy scores by a 58% error margin
. Furthermore, the model includes several parallel training methods, most notabl
y allowing a skip-gram network with 160 billion parameters to be trained overnig
ht on 3 multi-core CPUs, 14x larger than the previous largest neural network.
****

Distributed Inference for Dirichlet Process Mixture Models
Hong Ge, Yutian Chen, Moquan Wan, Zoubin Ghahramani
Bayesian nonparametric mixture models based on the Dirichlet process (DP) have b
een widely used for solving problems like clustering, density estimation and top
ic modelling. These models make weak assumptions about the underlying process th
at generated the observed data. Thus, when more data are collected, the complexi
ty of these models can change accordingly. These theoretical properties often le
ad to superior predictive performance when compared to traditional finite mixtur
e models. However, despite the increasing amount of data available, the applicat
ion of Bayesian nonparametric mixture models is so far limited to relatively sma
ll data sets. In this paper, we propose an efficient distributed inference algor
ithm for the DP and the HDP mixture model. The proposed method is based on a var
iant of the slice sampler for DPs. Since this sampler does not involve a pre-det
ermined truncation, the stationary distribution of the sampling algorithm is unb
iased. We provide both local thread-level and distributed machine-level parallel
 implementations and study the performance of this sampler through an extensive
set of experiments on image and text data. When compared to existing inference a
lgorithms, the proposed method exhibits state-of-the-art accuracy and strong sca
lability with up to 512 cores.
****

Compressing Neural Networks with the Hashing Trick
Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, Yixin Chen
As deep nets are increasingly used in applications suited for mobile devices, a
fundamental dilemma becomes apparent: the trend in deep learning is to grow mode
ls to absorb ever-increasing data set sizes; however mobile devices are designed
 with very little memory and cannot store such large models. We present a novel
network architecture, HashedNets, that exploits inherent redundancy in neural ne
tworks to achieve drastic reductions in model sizes. HashedNets uses a low-cost
hash function to randomly group connection weights into hash buckets, and all co
nnections within the same hash bucket share a single parameter value. These para
meters are tuned to adjust to the HashedNets weight sharing architecture with st
andard backprop during training. Our hashing procedure introduces no additional
memory overhead, and we demonstrate on several benchmark data sets that HashedNe
ts shrink the storage requirements of neural networks substantially while mostly
 preserving generalization performance.
****

Intersecting Faces: Non-negative Matrix Factorization With New Guarantees
Rong Ge, James Zou
Non-negative matrix factorization (NMF) is a natural model of admixture and is w
idely used in science and engineering. A plethora of algorithms have been develo
ped to tackle NMF, but due to the non-convex nature of the problem, there is lit
tle guarantee on how well these methods work. Recently a surge of research have
focused on a very restricted class of NMFs, called separable NMF, where provably
 correct algorithms have been developed. In this paper, we propose the notion of
 subset-separable NMF, which substantially generalizes the property of separabil
ity. We show that subset-separability is a natural necessary condition for the f
actorization to be unique or to have minimum volume. We developed the Face-Inter
sect algorithm which provably and efficiently solves subset-separable NMF under
natural conditions, and we prove that our algorithm is robust to small noise. We
 explored the performance of Face-Intersect on simulations and discuss settings
where it empirically outperformed the state-of-art methods. Our work is a step t
owards finding provably correct algorithms that solve large classes of NMF probl
ems.
****

Scaling up Natural Gradient by Sparsely Factorizing the Inverse Fisher Matrix
Roger Grosse, Ruslan Salakhudinov
Second-order optimization methods, such as natural gradient, are difficult to ap ply to high-dimensional problems, because they require approximately solving lar ge linear systems. We present FActorized Natural Gradient (FANG), an approximati on to natural gradient descent where the Fisher matrix is approximated with a Ga ussian graphical model whose precision matrix can be computed efficiently. We an alyze the Fisher matrix for a small RBM and derive an extremely sparse graphical model which is a good match to the covariance of the sufficient statistics. Our experiments indicate that FANG allows RBMs to be trained more efficiently compa red with stochastic gradient descent. Additionally, our analysis yields insight into the surprisingly good performance of the “centering trick” for training RBM s.
****

A Deeper Look at Planning as Learning from Replay
Harm Vanseijen, Rich Sutton
In reinforcement learning, the notions of experience replay, and of planning as learning from replayed experience, have long been used to find good policies wit h minimal training data. Replay can be seen either as model-based reinforcement learning, where the store of past experiences serves as the model, or as a way t o avoid a conventional model of the environment altogether. In this paper, we lo ok more deeply at how replay blurs the line between model-based and model-free m ethods. First, we show for the first time an exact equivalence between the seque nce of value functions found by a model-based policy-evaluation method and by a model-free method with replay. Second, we present a general replay method that c an mimic a spectrum of methods ranging from the explicitly model-free (TD(0)) to the explicitly model-based (linear Dyna). Finally, we use insights gained from these relationships to design a new model-based reinforcement learning algorithm for linear function approximation. This method, which we call forgetful LSTD(la mbda), improves upon regular LSTD(lambda) because it extends more naturally to o nline control, and improves upon linear Dyna because it is a multi-step method, enabling it to perform well even in non-Markov problems or, equivalently, in pro blems with significant function approximation.
****

Optimal and Adaptive Algorithms for Online Boosting
Alina Beygelzimer, Satyen Kale, Haipeng Luo
We study online boosting, the task of converting any weak online learner into a strong online learner. Based on a novel and natural definition of weak online le arnability, we develop two online boosting algorithms. The first algorithm is an online version of boost-by-majority. By proving a matching lower bound, we show that this algorithm is essentially optimal in terms of the number of weak learn ers and the sample complexity needed to achieve a specified accuracy. The second algorithm is adaptive and parameter-free, albeit not optimal.
****

Global Convergence of Stochastic Gradient Descent for Some Non-convex Matrix Pro blems
Christopher De Sa, Christopher Re, Kunle Olukotun
Stochastic gradient descent (SGD) on a low-rank factorization is commonly employ ed to speed up matrix problems including matrix completion, subspace tracking, a nd SDP relaxation. In this paper, we exhibit a step size scheme for SGD on a low -rank least-squares problem, and we prove that, under broad sampling conditions, our method converges globally from a random starting point within O(ε^-1 n \log n) steps with constant probability for constant-rank problems. Our modification of SGD relates it to stochastic power iteration. We also show some experiments to illustrate the runtime and convergence of the algorithm.
****

An Empirical Exploration of Recurrent Network Architectures
Rafal Jozefowicz, Wojciech Zaremba, Ilya Sutskever
The Recurrent Neural Network (RNN) is an extremely powerful sequence model that is often difficult to train. The Long Short-Term Memory (LSTM) is a specific RNN

architecture whose design makes it much easier to train. While wildly successful in practice, the LSTM’s architecture appears to be ad-hoc so it is not clear if it is optimal, and the significance of its individual components is unclear. In this work, we aim to determine whether the LSTM architecture is optimal or whether much better architectures exist. We conducted a thorough architecture search where we evaluated over ten thousand different RNN architectures, and identified an architecture that outperforms both the LSTM and the recently-introduced Gated Recurrent Unit (GRU) on some but not all tasks. We found that adding a bias of 1 to the LSTM’s forget gate closes the gap between the LSTM and the GRU.
****

Complete Dictionary Recovery Using Nonconvex Optimization
Ju Sun, Qing Qu, John Wright
We consider the problem of recovering a complete (i.e., square and invertible) dictionary mb A_0, from mb Y = mb A_0 mb X_0 with mb Y ∈\mathbb R^n \times p. This recovery setting is central to the theoretical understanding of dictionary learning. We give the first efficient algorithm that provably recovers mb A_0 when mb X_0 has O(n) nonzeros per column, under suitable probability model for mb X_0. Prior results provide recovery guarantees when mb X_0 has only O(\sqrtn) nonzeros per column. Our algorithm is based on nonconvex optimization with a spherical constraint, and hence is naturally phrased in the language of manifold optimization. Our proofs give a geometric characterization of the high-dimensional objective landscape, which shows that with high probability there are no spurious local minima. Experiments with synthetic data corroborate our theory. Full version of this paper is available online: \urlhttp://arxiv.org/abs/1504.06785.
****

Safe Policy Search for Lifelong Reinforcement Learning with Sublinear Regret
Haitham Bou Ammar, Rasul Tutunov, Eric Eaton
Lifelong reinforcement learning provides a promising framework for developing versatile agents that can accumulate knowledge over a lifetime of experience and rapidly learn new tasks by building upon prior knowledge. However, current lifelong learning methods exhibit non-vanishing regret as the amount of experience increases, and include limitations that can lead to suboptimal or unsafe control policies. To address these issues, we develop a lifelong policy gradient learner that operates in an adversarial setting to learn multiple tasks online while enforcing safety constraints on the learned policies. We demonstrate, for the first time, sublinear regret for lifelong policy search, and validate our algorithm on several benchmark dynamical systems and an application to quadrotor control.
****

PASSCoDe: Parallel ASynchronous Stochastic dual Co-ordinate Descent
Cho-Jui Hsieh, Hsiang-Fu Yu, Inderjit Dhillon
Stochastic Dual Coordinate Descent (DCD) is one of the most efficient ways to solve the family of L2-regularized empirical risk minimization problems, including linear SVM, logistic regression, and many others. The vanilla implementation of DCD is quite slow; however, by maintaining primal variables while updating dual variables, the time complexity of DCD can be significantly reduced. Such a strategy forms the core algorithm in the widely-used LIBLINEAR package. In this paper, we parallelize the DCD algorithms in LIBLINEAR. In recent research, several synchronized parallel DCD algorithms have been proposed, however, they fail to achieve good speedup in the shared memory multi-core setting. In this paper, we propose a family of parallel asynchronous stochastic dual coordinate descent algorithms (PASSCoDe). Each thread repeatedly selects a random dual variable and conducts coordinate updates using the primal variables that are stored in the shared memory. We analyze the convergence properties of DCD when different locking/atomic mechanisms are applied. For implementation with atomic operations, we show linear convergence under mild conditions. For implementation without any atomic operations or locking, we present a novel error analysis for PASSCoDe under the multi-core environment, showing that the converged solution is the exact solution for a primal problem with a perturbed regularizer. Experimental results show that our methods are much faster than previous parallel coordinate descent solvers.

****
High Confidence Policy Improvement
Philip Thomas, Georgios Theocharous, Mohammad Ghavamzadeh
We present a batch reinforcement learning (RL) algorithm that provides probabilistic guarantees about the quality of each policy that it proposes, and which has no hyper-parameter that requires expert tuning. Specifically, the user may select any performance lower-bound and confidence level and our algorithm will ensure that the probability that it returns a policy with performance below the lower bound is at most the specified confidence level. We then propose an incremental algorithm that executes our policy improvement algorithm repeatedly to generate multiple policy improvements. We show the viability of our approach with a simple 4 x 4 gridworld and the standard mountain car problem, as well as with a digital marketing application that uses real world data.
****
Fixed-point algorithms for learning determinantal point processes
Zelda Mariet, Suvrit Sra
Determinantal point processes (DPPs) offer an elegant tool for encoding probabilities over subsets of a ground set. Discrete DPPs are parametrized by a positive semidefinite matrix (called the DPP kernel), and estimating this kernel is key to learning DPPs from observed data. We consider the task of learning the DPP kernel, and develop for it a surprisingly simple yet effective new algorithm. Our algorithm offers the following benefits over previous approaches: (a) it is much simpler; (b) it yields equally good and sometimes even better local maxima; and (c) it runs an order of magnitude faster on large problems. We present experimental results on both real and simulated data to illustrate the numerical performance of our technique.
****
Consistent Multiclass Algorithms for Complex Performance Measures
Harikrishna Narasimhan, Harish Ramaswamy, Aadirupa Saha, Shivani Agarwal
This paper presents new consistent algorithms for multiclass learning with complex performance measures, defined by arbitrary functions of the confusion matrix. This setting includes as a special case all loss-based performance measures, which are simply linear functions of the confusion matrix, but also includes more complex performance measures such as the multiclass G-mean and micro F_1 measures. We give a general framework for designing consistent algorithms for such performance measures by viewing the learning problem as an optimization problem over the set of feasible confusion matrices, and give two specific instantiations based on the Frank-Wolfe method for concave performance measures and on the bisection method for ratio-of-linear performance measures. The resulting algorithms are provably consistent and outperform a multiclass version of the state-of-the-art SVMperf method in experiments; for large multiclass problems, the algorithms are also orders of magnitude faster than SVMperf.
****
Optimizing Neural Networks with Kronecker-factored Approximate Curvature
James Martens, Roger Grosse
We propose an efficient method for approximating natural gradient descent in neural networks which we call Kronecker-factored Approximate Curvature (K-FAC). K-FAC is based on an efficiently invertible approximation of a neural network’s Fisher information matrix which is neither diagonal nor low-rank, and in some cases is completely non-sparse. It is derived by approximating various large blocks of the Fisher (corresponding to entire layers) as being the Kronecker product of two much smaller matrices. While only several times more expensive to compute than the plain stochastic gradient, the updates produced by K-FAC make much more progress optimizing the objective, which results in an algorithm that can be much faster than stochastic gradient descent with momentum in practice. And unlike some previously proposed approximate natural-gradient/Newton methods which use high-quality non-diagonal curvature matrices (such as Hessian-free optimization), K-FAC works very well in highly stochastic optimization regimes. This is because the cost of storing and inverting K-FAC’s approximation to the curvature matrix does not depend on the amount of data used to estimate it, which is a feature t

ypically associated only with diagonal or low-rank approximations to the curvatu
re matrix.
****

A Convex Exemplar-based Approach to MAD-Bayes Dirichlet Process Mixture Models
En-Hsu Yen, Xin Lin, Kai Zhong, Pradeep Ravikumar, Inderjit Dhillon
MAD-Bayes (MAP-based Asymptotic Derivations) has been recently proposed as a gen
eral technique to derive scalable algorithm for Bayesian Nonparametric models. H
owever, the combinatorial nature of objective functions derived from MAD-Bayes r
esults in hard optimization problem, for which current practice employs heuristi
c algorithms analogous to k-means to find local minimum. In this paper, we consi
der the exemplar-based version of MAD-Bayes formulation for DP and Hierarchical
DP (HDP) mixture model. We show that an exemplar-based MAD-Bayes formulation can
 be relaxed to a convex structural-regularized program that, under cluster-separ
ation conditions, shares the same optimal solution to its combinatorial counterp
art. An algorithm based on Alternating Direction Method of Multiplier (ADMM) is
then proposed to solve such program. In our experiments on several benchmark dat
a sets, the proposed method finds optimal solution of the combinatorial problem
and significantly improves existing methods in terms of the exemplar-based objec
tive.
****

Multi-instance multi-label learning in the presence of novel class instances
Anh Pham, Raviv Raich, Xiaoli Fern, Jesús Pérez Arriaga
Multi-instance multi-label learning (MIML) is a framework for learning in the pr
esence of label ambiguity. In MIML, experts provide labels for groups of instanc
es (bags), instead of directly providing a label for every instance. When labeli
ng efforts are focused on a set of target classes, instances outside this set wi
ll not be appropriately modeled. For example, ornithologists label bird audio re
cordings with a list of species present. Other additional sound instances, e.g.,
 a rain drop or a moving vehicle sound, are not labeled. The challenge is due to
 the fact that for a given bag, the presence or absence of novel instances is la
tent. In this paper, this problem is addressed using a discriminative probabilis
tic model that accounts for novel instances. We propose an exact and efficient i
mplementation of the maximum likelihood approach to determine the model paramete
rs and consequently learn an instance-level classifier for all classes including
 the novel class. Experiments on both synthetic and real datasets illustrate the
 effectiveness of the proposed approach.
****

Entropy-Based Concentration Inequalities for Dependent Variables
Liva Ralaivola, Massih-Reza Amini
We provide new concentration inequalities for functions of dependent variables.
The work extends that of Janson (2004), which proposes concentration inequalitie
s using a combination of the Laplace transform and the idea of fractional graph
coloring, as well as many works that derive concentration inequalities using the
 entropy method (see, e.g., (Boucheron et al., 2003)). We give inequalities for
fractionally sub-additive and fractionally self-bounding functions. In the way,
we prove a new Talagrand concentration inequality for fractionally sub-additive
functions of dependent variables. The results allow us to envision the derivatio
n of generalization bounds for various applications where dependent variables na
turally appear, such as in bipartite ranking.
****

PU Learning for Matrix Completion
Cho-Jui Hsieh, Nagarajan Natarajan, Inderjit Dhillon
In this paper, we consider the matrix completion problem when the observations a
re one-bit measurements of some underlying matrix M , and in particular the obse
rved samples consist only of ones and no zeros. This problem is motivated by mod
ern applications such as recommender systems and social networks where only “lik
es” or “friendships” are observed. The problem is an instance of PU (positive-un
labeled) learning, i.e. learning from only positive and unlabeled examples that
has been studied in the context of binary classification. Under the assumption t
hat M has bounded nuclear norm, we provide recovery guarantees for two different

observation models: 1) M parameterizes a distribution that generates a binary matrix, 2) M is thresholded to obtain a binary matrix. For the first case, we propose a “shifted matrix completion” method that recovers M using only a subset of indices corresponding to ones; for the second case, we propose a “biased matrix completion” method that recovers the (thresholded) binary matrix. Both methods yield strong error bounds — if M ∈R^n \times n, the error is bounded as O(1-ρ) , where 1-ρdenotes the fraction of ones observed. This implies a sample complexity of O(n log n) ones to achieve a small error, when M is dense and n is large. We extend our analysis to the inductive matrix completion problem, where rows and columns of M have associated features. We develop efficient and scalable optimization procedures for both the proposed methods and demonstrate their effectiveness for link prediction (on real-world networks consisting of over 2 million nodes and 90 million links) and semi-supervised clustering tasks.

****

An Asynchronous Distributed Proximal Gradient Method for Composite Convex Optimization

Necdet Aybat, Zi Wang, Garud Iyengar

We propose a distributed first-order augmented Lagrangian (DFAL) algorithm to minimize the sum of composite convex functions, where each term in the sum is a private cost function belonging to a node, and only nodes connected by an edge can directly communicate with each other. This optimization model abstracts a number of applications in distributed sensing and machine learning. We show that any limit point of DFAL iterates is optimal; and for any eps > 0, an eps-optimal and eps-feasible solution can be computed within O(log(1/eps)) DFAL iterations, which require O(\psi_\textmax^1.5/d_\textmin ·1/ε) proximal gradient computations and communications per node in total, where \psi_\textmax denotes the largest eigenvalue of the graph Laplacian, and d_\textmin is the minimum degree of the graph. We also propose an asynchronous version of DFAL by incorporating randomized block coordinate descent methods; and demonstrate the efficiency of DFAL on large scale sparse-group LASSO problems.

****

Sparse Subspace Clustering with Missing Entries

Congyuan Yang, Daniel Robinson, Rene Vidal

We consider the problem of clustering incomplete data drawn from a union of subspaces. Classical subspace clustering methods are not applicable to this problem because the data are incomplete, while classical low-rank matrix completion methods may not be applicable because data in multiple subspaces may not be low rank. This paper proposes and evaluates two new approaches for subspace clustering and completion. The first one generalizes the sparse subspace clustering algorithm so that it can obtain a sparse representation of the data using only the observed entries. The second one estimates a suitable kernel matrix by assuming a random model for the missing entries and obtains the sparse representation from this kernel. Experiments on synthetic and real data show the advantages and disadvantages of the proposed methods, which all outperform the natural approach (low-rank matrix completion followed by sparse subspace clustering) when the data matrix is high-rank or the percentage of missing entries is large.

****

Moderated and Drifting Linear Dynamical Systems

Jinyan Guan, Kyle Simek, Ernesto Brau, Clayton Morrison, Emily Butler, Kobus Barnard

We consider linear dynamical systems, particularly coupled linear oscillators, where the parameters represent meaningful values in a domain theory and thus learning what affects them contributes to explanation. Rather than allow perturbations of latent states, we assume that temporal variation beyond noise is explained by parameter drift, and variation across coupled systems is a function of moderating variables. This change of focus reduces opportunities for efficient inference, and we propose sampling procedures to learn and fit the models. We test our approach on a real dataset of physiological measures of heterosexual couples engaged in a conversation about a potentially emotional topic, with body mass index (BMI) being considered as a moderator. We evaluate several models on their abi

lity to predict future conversation dynamics (the last 20% of the data for each test couple), with shared parameters being learned using held out data. As proof of concept, we validate the hypothesis that BMI affects the conversation dynamic in the experimentally chosen topic.

****

Boosted Categorical Restricted Boltzmann Machine for Computational Prediction of Splice Junctions

Taehoon Lee, Sungroh Yoon

Splicing refers to the elimination of non-coding regions in transcribed pre-messenger ribonucleic acid (RNA). Discovering splice sites is an important machine learning task that helps us not only to identify the basic units of genetic heredity but also to understand how different proteins are produced. Existing methods for splicing prediction have produced promising results, but often show limited robustness and accuracy. In this paper, we propose a deep belief network-based methodology for computational splice junction prediction. Our proposal includes a novel method for training restricted Boltzmann machines for class-imbalanced prediction. The proposed method addresses the limitations of conventional contrastive divergence and provides regularization for datasets that have categorical features. We tested our approach using public human genome datasets and obtained significantly improved accuracy and reduced runtime compared to state-of-the-art alternatives. The proposed approach was less sensitive to the length of input sequences and more robust for handling false splicing signals. Furthermore, we could discover non-canonical splicing patterns that were otherwise difficult to recognize using conventional methods. Given the efficiency and robustness of our methodology, we anticipate that it can be extended to the discovery of primary structural patterns of other subtle genomic elements.

****

Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo

Yu-Xiang Wang, Stephen Fienberg, Alex Smola

We consider the problem of Bayesian learning on sensitive datasets and present two simple but somewhat surprising results that connect Bayesian learning to “differential privacy”, a cryptographic approach to protect individual-level privacy while permitting database-level utility. Specifically, we show that under standard assumptions, getting one sample from a posterior distribution is differentially private “for free”; and this sample as a statistical estimator is often consistent, near optimal, and computationally tractable. Similarly but separately, we show that a recent line of work that use stochastic gradient for Hybrid Monte Carlo (HMC) sampling also preserve differentially privacy with minor or no modifications of the algorithmic procedure at all, these observations lead to an “anytime” algorithm for Bayesian learning under privacy constraint. We demonstrate that it performs much better than the state-of-the-art differential private methods on synthetic and real datasets.

****

A trust-region method for stochastic variational inference with applications to streaming data

Lucas Theis, Matt Hoffman

Stochastic variational inference allows for fast posterior inference in complex Bayesian models. However, the algorithm is prone to local optima which can make the quality of the posterior approximation sensitive to the choice of hyperparameters and initialization. We address this problem by replacing the natural gradient step of stochastic varitional inference with a trust-region update. We show that this leads to generally better results and reduced sensitivity to hyperparameters. We also describe a new strategy for variational inference on streaming data and show that here our trust-region method is crucial for getting good performance.

****

Inference in a Partially Observed Queuing Model with Applications in Ecology

Kevin Winner, Garrett Bernstein, Dan Sheldon

We consider the problem of inference in a probabilistic model for transient populations where we wish to learn about arrivals, departures, and population size o

ver all time, but the only available data are periodic counts of the population size at specific observation times. The underlying model arises in queueing theory (as an M/G/inf queue) and also in ecological models for short-lived animals such as insects. Our work applies to both systems. Previous work in the ecology literature focused on maximum likelihood estimation and made a simplifying independence assumption that prevents inference over unobserved random variables such as arrivals and departures. The contribution of this paper is to formulate a latent variable model and develop a novel Gibbs sampler based on Markov bases to perform inference using the correct, but intractable, likelihood function. We empirically validate the convergence behavior of our sampler and demonstrate the ability of our model to make much finer-grained inferences than the previous approach.

****

Deterministic Independent Component Analysis

Ruitong Huang, Andras Gyorgy, Csaba Szepesvári

We study independent component analysis with noisy observations. We present, for the first time in the literature, consistent, polynomial-time algorithms to recover non-Gaussian source signals and the mixing matrix with a reconstruction error that vanishes at a 1/\sqrtT rate using T observations and scales only polynomially with the natural parameters of the problem. Our algorithms and analysis also extend to deterministic source signals whose empirical distributions are approximately independent.

****

On the Optimality of Multi-Label Classification under Subset Zero-One Loss for Distributions Satisfying the Composition Property

Maxime Gasse, Alexandre Aussem, Haytham Elghazel

The benefit of exploiting label dependence in multi-label classification is known to be closely dependent on the type of loss to be minimized. In this paper, we show that the subsets of labels that appear as irreducible factors in the factorization of the conditional distribution of the label set given the input features play a pivotal role for multi-label classification in the context of subset Zero-One loss minimization, as they divide the learning task into simpler independent multi-class problems. We establish theoretical results to characterize and identify these irreducible label factors for any given probability distribution satisfying the Composition property. The analysis lays the foundation for generic multi-label classification and optimal feature subset selection procedures under this subclass of distributions. Our conclusions are supported by carefully designed experiments on synthetic and benchmark data.

****

Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization

Roy Frostig, Rong Ge, Sham Kakade, Aaron Sidford

We develop a family of accelerated stochastic algorithms that optimize sums of convex functions. Our algorithms improve upon the fastest running time for empirical risk minimization (ERM), and in particular linear least-squares regression, across a wide range of problem settings. To achieve this, we establish a framework, based on the classical proximal point algorithm, useful for accelerating recent fast stochastic algorithms in a black-box fashion. Empirically, we demonstrate that the resulting algorithms exhibit notions of stability that are advantageous in practice. Both in theory and in practice, the provided algorithms reap the computational benefits of adding a large strongly convex regularization term, without incurring a corresponding bias to the original ERM problem.

****

A New Generalized Error Path Algorithm for Model Selection

Bin Gu, Charles Ling

Model selection with cross validation (CV) is very popular in machine learning. However, CV with grid and other common search strategies cannot guarantee to find the model with minimum CV error, which is often the ultimate goal of model selection. Recently, various solution path algorithms have been proposed for several important learning algorithms including support vector classification, Lasso,

and so on. However, they still do not guarantee to find the model with minimum CV error.In this paper, we first show that the solution paths produced by various algorithms have the property of piecewise linearity. Then, we prove that a large class of error (or loss) functions are piecewise constant, linear, or quadratic w.r.t. the regularization parameter, based on the solution path. Finally, we propose a new generalized error path algorithm (GEP), and prove that it will find the model with minimum CV error for the entire range of the regularization parameter. The experimental results on a variety of datasets not only confirm our theoretical findings, but also show that the best model with our GEP has better generalization error on the test data, compared to the grid search, manual search, and random search.
****