Ideas for projects 276-2017:

1. Come up with your own idea. Must be relevant to the class.
2. Join frontier research by presenting and adding to current research in Dechter’s group examples areas are
	1. Abstraction sampling
	2. Height vs width pseudo-trees: what are the tradeoffs?
	3. Approximate counting (Sciex)
	4. Heuristic for mini-bucket partitionings.

**[R173]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r173.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r173.pdf) |[Color PDF](http://www.ics.uci.edu/~dechter/publications/r173-color.pdf)
**Emma Rollon and Rina Dechter.** "New Mini-Bucket Partitioning Heuristics for Bounding the Probability of Evidence" *forthcoming in AAAI 2010*

|  |
| --- |
| **[R235]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r235.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r235.pdf) | [Slides](http://www.ics.uci.edu/~dechter/publications/r235-slides.pdf) **Radu Marinescu, Junkyu Lee, Alexander Ihler, and Rina Dechter.** "Anytime Best+Depth-First Search for Bounding Marginal MAP" *in Proceedings of AAAI 2017.*  **[R233]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r233.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r233.pdf) | [Slides](http://www.ics.uci.edu/~dechter/publications/r233-slides.pptx) **Qi Lou, Rina Dechter, and Alexander Ihler.** "Anytime Anyspace AND/OR Search for Bounding the Partition Function" *in Proceedings of AAAI 2017.*  |
| **2016**  |
| **[R232]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r232.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r232.pdf) | [Poster](http://www.ics.uci.edu/~dechter/publications/r232-poster.pdf) **William Lam, Kalev Kask, Rina Dechter, and Javier Larrosa** "On the Impact of Subproblem Orderings on Anytime AND/OR Best-First Search for Lower Bounds" *in Proceedings of the European Conference on Artificial Intelligence 2016 (ECAI 2016).* Longer version: **[R232a]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r232a.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r232a.pdf) **[R231]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r231.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r231.pdf) | [Slides](http://www.ics.uci.edu/~dechter/publications/r231-slides.pdf) **Rodrigo de Salvo Braz, Ciaran O'Reilly, Vibhav Gogate, and Rina Dechter.** "Probabilistic Inference Modulo Theories" *in Proceedings of the International Joint Conference on Artificial Intelligence 2016 (IJCAI 2016).* **[R230]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r230.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r230.pdf) | [Slides](http://www.ics.uci.edu/~dechter/publications/r230-slides.pdf) **Javier Larrosa, Emma Rollon, and Rina Dechter.** "Limited Discrepancy AND/OR Search and its Application to Optimization Tasks in Graphical Models" *in Proceedings of the International Joint Conference on Artificial Intelligence 2016 (IJCAI 2016).*  **[R228]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r228.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r228.pdf) **Natalia Flerova, Radu Marinescu, and Rina Dechter.** "Weighted heuristic anytime search: new schemes for optimization over graphical models" *in Annals of Mathematics and Artificial Intelligence.* **[R227]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r227.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r227.pdf) | [Slides](http://www.ics.uci.edu/~dechter/publications/r227-slides.pdf) **Junkyu Lee, Radu Marinescu, and Rina Dechter.** "Applying Search Based Probabilistic Inference Algorithms to Probabilistic Conformant Planning: Preliminary Results" *in proceedings of International Symposium on Artificial Intelligence and Mathematics (ISAIM 2016).* **[R226]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r226.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r226.pdf) | [Slides](http://www.ics.uci.edu/~dechter/publications/r226-slides.pdf) **Rina Dechter, Kalev Kask, William Lam, and Javier Larrosa.** "Look-ahead with Mini-Bucket Heuristics for MPE" *in proceedings of AAAI 2016.* **[R225]** [Abstract](http://www.ics.uci.edu/~dechter/publications/r225.html) | [PDF](http://www.ics.uci.edu/~dechter/publications/r225.pdf) | [Slides](http://www.ics.uci.edu/~dechter/publications/r225-slides.pdf) **Junkyu Lee, Radu Marinescu, Rina Dechter and Alexander Ihler.** "From Exact to Anytime Solutions for Marginal MAP" *in proceedings of AAAI 2016.*  |  |
|  |
|  |

1. Recent papers from a subset of researchers (search on your own):
	1. Vibhav Gogate
	2. Rodrigo Del-vaz
	3. David Poole
	4. Adnan Darwiche
	5. Guy van der…
2. Based on papers from UAI 2016: http://www.auai.org/uai2016/accepted.php

**UAI 2016 - Accepted Papers**

|  |  |
| --- | --- |
| **ID: 1**  | **Characterizing Tightness of LP Relaxations by Forbidding Signed Minors** +We consider binary pairwise graphical models and provide an exact characterization (necessary and sufficient conditions observing signs of potentials) of tightness for the LP relaxation on the triplet-consistent polytope of the MAP inference problem, by forbidding an odd-K5 (complete graph on 5 variables with all edges odd) as a signed minor in the signed suspension graph. This captures signs of both singleton and edge potentials in a compact and efficiently testable condition, and improves significantly on earlier results. We provide other results on tightness of LP relaxations by forbidding minors, draw connections and suggest paths for future research.*Adrian Weller, University of Cambridge* |
|  |  |
| **ID: 5**  | **Correlated Tag Learning in Topic Model** +It is natural to expect that the documents in a corpus will be correlated, and these correlations are reflected by not only the words but also the observed tags in each document. Most previous works model these type of corpus without considering the correlations among the tags. In this work, we develop a Correlated Tag Learning (CTL) model for semi-structured documents based on the topic model to enable the construction of the correlation graph among tags via a logistic normal participation process. For inference of the CTL model, we devise a variational inference algorithm to approximate the posterior. In experiments, we visualize the tag correlation graph generated by CTL on the DBLP corpus and for the tasks of document retrieval and classification, the correlation graph among tags is helpful to improve the generalization performance compared with the state-of-the-art baselines.*Shuangyin Li, HKUST; Rong Pan, Sun Yat-sen University; Yu Zhang, ; Qiang Yang, Hong Kong University of Science and Technology* |
|  |  |
| **ID: 10**  | **Lighter-Communication Distributed Machine Learning via Sufficient Factor Broadcasting** +Matrix-parametrized models (MPMs) are widely used in machine learning (ML) applications. In large-scale ML problems, the parameter matrix of a MPM can grow at an unexpected rate, resulting in high communication and parameter synchronization costs. To address this issue, we offer two contributions: first, we develop a computation model for a large family of MPMs, which share the following property: the parameter update computed on each data sample is a rank-1 matrix, \ie the outer product of two ``sufficient factors (SFs). Second, we implement a decentralized, peer-to-peer system, Sufficient Factor Broadcasting (SFB), which broadcasts the SFs among worker machines, and reconstructs the update matrices locally at each worker. SFB takes advantage of small rank-1 matrix updates and efficient partial broadcasting strategies to dramatically improve communication efficiency. We propose a graph optimization based partial broadcasting scheme, which minimizes the delay of information dissemination under the constraint that each machine only communicates with a subset rather than all of machines. Furthermore, we provide theoretical analysis to show that SFB guarantees convergence of algorithms without requiring a centralized synchronization mechanism. Experiments corroborate SFB's efficiency on four MPMs.*Pengtao Xie, Carnegie Mellon University; Jin Kyu Kim, Carnegie Mellon University; Yi Zhou, Syracuse University; Qirong Ho, Institute for Infocomm Research, ASTAR; Abhimanu Kumar, Groupon Inc.; Yaoliang Yu, ; Eric Xing, Carnegie Mellon University* |
|  |  |
| **ID: 11**  | **Unsupervised Discovery of El Nino Using Causal Feature Learning on Microlevel Climate Data** +We show that the climate phenomena of El Nino and La Nina arise naturally as states of macro-variables when the recent causal feature learning framework of Chalupka et al. (2015a,b) is applied to micro-level measures of zonal wind (ZW) and sea surface temperatures (SST) taken over the equatorial band of the Pacific Ocean. The method identifies these unusual climate states on the basis of the relation between ZW and SST patterns without any input about past occurrences of El Nino or La Nina. The simpler alternatives of (i) clustering the SST fields while disregarding their relationship with ZW patterns, or (ii) clustering the joint ZW-SST patterns, do not discover El Nino. We discuss the degree to which our method supports a causal interpretation and using a low-dimensional toy example we explain its success over other clustering approaches. Finally, we propose a new robust and scalable alternative to the algorithm of Chalupka et al. (2015b), which circumvents the need for high-dimensional density learning.*Krzysztof Chalupka, Caltech; Tobias Bischoff, Caltech; Frederick Eberhardt, ; Pietro Perona, Caltech* |
|  |  |
| **ID: 13**  | **Online Bayesian Multiple Kernel Bipartite Ranking** +Bipartite ranking aims to maximize the area under the ROC curve (AUC) of a decision function. To tackle this problem when the data appears sequentially, existing online AUC maximization methods focus on seeking a point estimate of the decision function in a linear or predefined single kernel space, and cannot learn effective kernels automatically from the streaming data. In this paper, we first develop a Bayesian multiple kernel bipartite ranking model, which circumvents the kernel selection problem by estimating a posterior distribution over the model weights. To make our model applicable to streaming data, we then present a kernelized online Bayesian passive-aggressive learning framework by maintaining a variational approximation to the posterior based on data augmentation. Furthermore, to efficiently deal with large-scale data, we design a fixed budget strategy which can effectively control online model complexity. Extensive experimental studies confirm the superiority of our Bayesian multi-kernel approach.*Changde Du, ; Changying Du, Institute of Software, CAS; Ali Luo, ; Guoping Long, ; Qing He, ; Yucheng Li,* |
|  |  |
| **ID: 14**  | **Alternative Markov and Causal Properties for Acyclic Directed Mixed Graphs** +We extend Andersson-Madigan-Perlman chain graphs by (i) relaxing the semidirected acyclity constraint so that only directed cycles are forbidden, and (ii) allowing up to two edges between any pair of nodes. We introduce global, and ordered local and pairwise Markov properties for the new models. We show the equivalence of these properties for strictly positive probability distributions. We also show that when the random variables are continuous, the new models can be interpreted as systems of structural equations with correlated errors. This enables us to adapt Pearl's do-calculus to them. Finally, we describe an exact algorithm for learning the new models from observational and interventional data via answer set programming.*Jose Pena,* |
|  |  |
| **ID: 15**  | **Overdispersed Black-Box Variational Inference** +We introduce overdispersed black-box variational inference, a method to reduce the variance of the Monte Carlo estimator of the gradient in black-box variational inference. Instead of taking samples from the variational distribution, we use importance sampling to take samples from an overdispersed distribution in the same exponential family as the variational approximation. Our approach is general since it can be readily applied to any exponential family distribution, which is the typical choice for the variational approximation. We run experiments on two non-conjugate probabilistic models to show that our method effectively reduces the variance, and the overhead introduced by the computation of the proposal parameters and the importance weights is negligible. We find that our overdispersed importance sampling scheme provides lower variance than black-box variational inference, even when the latter uses twice the number of samples. This results in faster convergence of the black-box inference procedure.*Francisco Ruiz, Columbia University; Michalis Titsias, Athens University of Economics and Business; david Blei, Columbia University* |
|  |  |
| **ID: 16**  | **Optimal Denoising Matrix in Dantzig Selector** +Dantzig Selector (DS), whose empirical and theoretical performances are both comparable to LASSO, is widely used in compressed sensing and sparse learning for feature selection and sparse signal recovery. Since DS can be reduced to a linear programming problem, many matured linear programming solvers can be applied for scaling up. The DS formulation can be expressed as a basis pursuit denoising problem, wherein the data matrix (or measurement matrix) is employed as the denoising matrix to eliminate the observation noise. However, we notice that the data matrix may not be the optimal denoising matrix for sparse signal recovery from the denoising perspective, as shown by a simple counter-example. To compute the optimal denoising matrix, we first formulate the computation of the optimal denoising matrix as a minimax optimization, which turns out to be an NP-hard problem. To make the problem computationally tractable, we propose a novel algorithm, termed as Optimal Denoising Dantzig Selector (ODDS), to approximately compute the optimal denoising matrix. Empirical experiments validate the proposed method. Finally, a novel sparse reinforcement learning algorithm is formulated by extending the proposed ODDS algorithm to temporal difference learning, and empirical experimental results demonstrate to outperform the conventional ``vanilla'' DS-TD algorithm.*Bo Liu, Auburn University; Luwan Zhang, Department of Statistics, University of Wisconsin-Madison; Ji Liu, University of Rochester* |
|  |  |
| **ID: 20**  | **Thompson Sampling is Asymptotically Optimal in General Environments** +We discuss a variant of Thompson sampling for nonparametric reinforcement learning in a countable classes of general stochastic environments. These environments can be non-Markov, non-ergodic, and partially observable. We show that Thompson sampling learns the environment class in the sense that (1) asymptotically its value converges to the optimal value in mean and (2) given a recoverability assumption regret is sublinear.*Jan Leike, Australian National University; Tor Lattimore, University of Alberta; Laurent Orseau, Google DeepMind; Marcus Hutter,* |
|  |  |
| **ID: 31**  | **Bounded Rational Decision-Making in Feedforward Neural Networks** +Bounded rational decision-makers transform sensory input into motor output under limited computational resources. Mathematically, such decision-makers can be modeled as information-theoretic channels with limited transmission rate. Here, we apply this formalism for the first time to multilayer feedforward neural networks. We derive synaptic weight update rules for two scenarios, where either each neuron is considered as a bounded rational decision-maker or the network as a whole. In the update rules, bounded rationality translates into information-theoretically motivated types of regularization in weight space. In experiments on the MNIST benchmark classification task for handwritten digits, we show that such information-theoretic regularization successfully prevents overfitting across different architectures and attains results that are competitive with other recent techniques like dropout, dropconnect and Bayes by backprop, for both ordinary and convolutional neural networks.*Felix Leibfried, Max Planck Society; Daniel Braun, Max Planck Society* |
|  |  |
| **ID: 34**  | **An Efficient Multi-Class Selective Sampling Algorithm on Graphs** +A graph based multi-class classification problem is usually converted into a collection of binary classification tasks using the one-vs-all strategy, and then solved by applying some binary classification algorithms. Unlike the one-vs-all strategy, we propose a unified framework that operates directly on the multi-class problem without reducing it to a collection of binary-classification tasks. Moreover, this framework makes active learning possible for multi-class problems, while the one-vs-all strategy can not. Specifically, we employ a novel randomized query approach to prioritize the informative instances. This query technique based on the criterion of ``margin and ``uncertainty can achieve a comparable mistake bound with its fully-supervised counterpart. To take full advantage of correctly predicted labels discarded in traditional conservative algorithms, we propose an aggressive selective sampling that can update the model even if no error occurs. Thanks to the aggressive updating strategy, the aggressive algorithm achieves a lower mistake bound than its conservative competitor in expectation. Encouraging experimental results on real-world graph databases show that the proposed technique by querying an extremely small ratio of labels can achieve a better classification accuracy.*Peng Yang, Institute for Infocomm Researc; Peilin Zhao, ; Zhen Hai, Institute for Infocomm Research; Wei Liu, ; Xiao-Li Li, ; Steven C.H. Hoi,* |
|  |  |
| **ID: 45**  | **On the Theory and Practice of Privacy-Preserving Bayesian Data Analysis** +Bayesian inference has great promise for the privacy-preserving analysis of sensitive data, as posterior sampling automatically preserves differential privacy, an algorithmic notion of data privacy, under certain conditions (Wang et al., 2015). While Wang et al. (2015)'s one posterior sample (OPS) approach elegantly provides privacy for free, it is data inefficient in the sense of asymptotic relative efficiency (ARE). We show that a simple alternative based on the Laplace mechanism, the workhorse technique of differential privacy, is as asymptotically efficient as non-private posterior inference, under general assumptions. The Laplace mechanism has additional practical advantages including efficient use of the privacy budget for MCMC. We demonstrate the practicality of our approach on a time-series analysis of sensitive military records from the Afghanistan and Iraq wars disclosed by the Wikileaks organization.*James Foulds, UC San Diego; Joseph Geumlek, UC San Diego; Max Welling, University of Amsterdam; Kamalika Chaudhuri,* |
|  |  |
| **ID: 46**  | **A Characterization of Markov Equivalence Classes for Relational Causal Model with Path Semantics** +Relational Causal Models (RCM) generalize Causal Bayesian Networks so as to extend causal discovery to relational domains. We provide a novel characterization of the Markov equivalence of RCMs. We show that the resulting characterization is quite elegant in the case of path semantics but not in the case of bridge burning semantics. We introduce a novel representation of unshielded triples that allows efficient determination of the Markov equivalence of one RCM with another. We provide a sound and complete algorithm for recovering the structure of an RCM from conditional independence queries. Our analysis also suggests ways to improve the orientation recall of algorithms for learning the structure of RCM under the bridge burning semantics.*Sanghack Lee, Penn State University; Vasant Honavar, Penn State University* |
|  |  |
| **ID: 61**  | **Political Dimensionality Estimation Using a Probabilistic Graphical Model** +This paper attempts to move beyond the left-right characterization of political ideologies. We propose a trait based probabilistic model for estimating the manifold of political opinion. We demonstrate the efficacy of our model on two novel and large scale datasets of public opinion. Our experiments show that although the political spectrum is richer than a simple left-right structure, peoples' opinions on seemingly unrelated political issues are very correlated, so fewer than 10 dimensions are enough to represent peoples' entire political opinion.*Yoad Lewenberg, The Hebrew University of Jerus; Yoram Bachrach, Microsoft Research ; Lucas Bordeaux, ; Pushmeet Kohli,* |
|  |  |
| **ID: 64**  | **Hierarchical learning of grids of microtopics** +The counting grid is a grid of microtopics, sparse word/feature distributions. The generative model associated with the grid does not use these microtopics individually. Rather, it groups them in overlapping rectangular windows and uses these grouped microtopics as either mixture or admixture components. This paper builds upon the basic counting grid model and it shows that hierarchical reasoning helps avoid bad local minima, produces better classification accuracy and, most interestingly, allows for extraction of large numbers of coherent microtopics even from small datasets. We evaluate this in terms of consistency, diversity and clarity of the indexed content, as well as in a user study on word intrusion tasks. We demonstrate that these models work well as a technique for embedding raw images and discuss interesting parallels between hierarchical CG models and other deep architectures.*Alessandro Perina, Microsoft; Nebojsa Jojic,* |
|  |  |
| **ID: 66**  | **Pruning Rules for Learning Parsimonious Context Trees** +We give a novel algorithm for finding a parsimonious context tree (PCT) that best fits a given data set. PCTs extend traditional context trees by allowing context-specific grouping of the states of a context variable, also enabling skipping the variable. However, they gain statistical efficiency at the cost of computational efficiency, as the search space of PCTs is of tremendous size. We propose pruning rules based on efficiently computable score upper bounds with the aim of reducing this search space significantly. While our concrete bounds exploit properties of the BIC score, the ideas apply also to other scoring functions. Empirical results show that our algorithm is typically an order-of-magnitude faster than a recently proposed memory-intensive algorithm [Eggeling et al. ICML 2015], or alternatively, about equally fast but using dramatically less memory.*Ralf Eggeling, University of Helsinki; Mikko Koivisto,* |
|  |  |
| **ID: 67**  | **Improving Imprecise Compressive Sensing Models** +Random sampling in compressive sensing (CS) enables the compression of large amounts of input signals in an efficient manner, which is useful for many applications. CS reconstructs the compressed signals exactly with overwhelming probability when incoming data can be sparsely represented with a few components. However, the theory of CS framework including random sampling has been focused on exact recovery of signal; impreciseness in signal recovery has been neglected. This can be problematic when there is uncertainty in the number of sparse components such as signal sparsity in dynamic systems that can change over time. We present a new theoretical framework that handles uncertainty in signal recovery from the perspective of recovery success and quality. We show that the signal recovery success in our model is more accurate than the success probability analysis in the CS framework. Our model is then extended to the case where the success or failure of signal recovery can be relaxed. We represent the number of components included in signal recovery with a right-tailed distribution and focus on recovery quality. Experimental results confirm the accuracy of our model in dynamic systems.*Dongeun Lee, UNIST; Rafael de Lima, ; Jaesik Choi,* |
|  |  |
| **ID: 68**  | **Safely Interruptible Agents** +Reinforcement learning agents in interaction with a complex environment like the real world are unlikely to behave optimally all the time. If such an agent is operating in real-time under human supervision, now and then it may be necessary for a human operator to press the big red button to prevent the agent from continuing a harmful sequence of actions---harmful either for the agent or for the environment---and lead the agent into a safer situation. However, if the learning agent expects to receive rewards from this sequence, it may learn in the long run to avoid such interruptions, for example by disabling the red button---which is an undesirable outcome. This paper explores a way to make sure a learning agent will /not/ learn to prevent (or seek!) being interrupted by the environment or a human operator. We provide a formal definition of safe interruptibility and exploit the off-policy learning property to prove that either some agents are already safely interruptible, like Q-learning, or can easily be made so, like Sarsa. We show that even ideal, uncomputable reinforcement learning agents for (deterministic) general computable environments can be made safely interruptible.*Laurent Orseau, Google DeepMind; Stuart Armstrong, Future of Humanity Institute, Oxford, UK* |
|  |  |
| **ID: 71**  | **Merging Strategies for Sum-Product Networks: From Trees to Graphs** +Learning the structure of sum-product networks (SPNs) – arithmetic circuits over latent and observed variables – has been the subject of much recent research. These networks admit linear time exact inference, and thus help alleviate one of the chief disadvantage of probabilistic graphical models: the high computational complexity and low accuracy of probabilistic inference. Although, algorithms for learning them from data have come quite far and often outperform algorithms that learn probabilistic graphical models, a key issue with existing approaches is that they induce tree SPNs, a small sub-class of SPNs. Thus, the power of SPNs has not yet been fully realized. In this paper, we address this limitation by developing post-processing approaches that induce graph SPNs from tree SPNs by merging similar sub-structures in the SPN. The key benefits of graph SPNs over tree SPNs include smaller computational complexity which facilitates faster online inference, and better generalization accuracy because of reduced variance, at the cost of slight increase in the learning time. We demonstrate experimentally that our merging technique significantly improves the accuracy of SPNs, achieving state-of-the-art performance on several benchmark datasets.*Tahrima Rahman, University of Texas at Dallas; Vibhav Gogate, The University of Texas at Dallas* |
|  |  |
| **ID: 73**  | **Bayesian Hyperparameter Optimization for Ensemble Learning** +In this paper, we bridge the gap between hyperparameter optimization and ensemble learning by performing Bayesian optimization of an ensemble with regards to its hyperparameters. Our method consists in building a fixed-size ensemble, optimizing the configuration of one classifier of the ensemble at each iteration of the hyperparameter optimization algorithm, taking into consideration the interaction with the other models when evaluating potential performances. We also consider the case where the ensemble is to be reconstructed at the end of the hyperparameter optimization phase, through a greedy selection over the pool of models generated during the optimization. We study the performance of our proposed method on three different hyperparameter spaces, showing that our approach is better than both the best single model and a greedy ensemble construction over the models produced by a standard Bayesian optimization.*Julien-Charles Levesque, Universite Laval; Christian Gagne, Universite Laval; Robert Sabourin, Ecole de Technologie Superieure* |
|  |  |
| **ID: 74**  | **Interpretable Policies for Dynamic Product Recommendations** +In many applications, it may be better to compute a good interpretable policy instead of a complex optimal one. For example, a recommendation engine might perform better when accounting for user profiles, but in the absence of such loyalty data, assumptions would have to be made that increase the complexity of the recommendation policy. A simple greedy recommendation could be implemented based on aggregated user data, but another simple policy can improve on this by accounting for the fact that users come from different segments of a population. In this paper, we study the problem of computing an optimal policy that is interpretable. In particular, we consider a policy to be interpretable if the decisions (e.g., recommendations) depend only on a small number of simple state attributes (e.g., the currently viewed product). This novel model is a general Markov decision problem with action constraints over states. We show that this problem is NP hard and develop a MILP formulation that gives an exact solution when policies are restricted to being deterministic. We demonstrate the effectiveness of the approach on a real-world business case for a European tour operator's recommendation engine.*Marek Petrik, IBM; Ronny Luss, IBM* |
|  |  |
| **ID: 77**  | **Convergence Rates for Greedy Kaczmarz Algorithms, and Randomized Kaczmarz Rules Using the Orthogonality Graph** +The Kaczmarz method is an iterative algorithm for solving systems of linear equalities and inequalities that iteratively projects onto the constraints. Recently, Strohmer and Vershynin [J. Fourier Anal. Appl., 15(2):262-278, 2009] gave the first non-asymptotic convergence rate analysis for this algorithm, spurring numerous extensions and generalizations of the Kaczmarz method. Rather than the randomized selection rule analyzed in that work, in this paper we instead discuss greedy and approximate greedy selection rules. We show that in some applications the computational costs of greedy and random selection are comparable, and that (except in extreme cases) greedy selection rules give faster convergence rates than random selection rules. Further, we give the first multi-step analysis of Kaczmarz methods for a particular greedy rule, and propose a provably-faster randomized selection rule for matrices with many orthogonal rows.*Julie Nutini, University of British Columbia; Behrooz Sepehry, University of British Columbia; Issam Laradji, University of British Columbia; Alim Virani, ; Mark Schmidt, University of British Columbia; Hoyt Koepke, Dato* |
|  |  |
| **ID: 79**  | **Structured Prediction: From Gaussian Perturbations to Linear-Time Principled Algorithms** +Margin-based structured prediction commonly uses a maximum loss over all possible structured outputs \cite{Altun03,Collins04b,Taskar03}. In natural language processing, recent work \cite{Zhang14,Zhang15} has proposed the use of the maximum loss over random structured outputs sampled independently from some proposal distribution. This method is linear-time in the number of random structured outputs and trivially parallelizable. We study this family of loss functions in the PAC-Bayes framework under Gaussian perturbations \cite{McAllester07}. Under some technical conditions and up to statistical accuracy, we show that this family of loss functions produces a tighter upper bound of the Gibbs decoder distortion than commonly used methods. Thus, using the maximum loss over random structured outputs is a principled way of learning the parameter of structured prediction models. Besides explaining the experimental success of \cite{Zhang14,Zhang15}, our theoretical results show that more general techniques are possible.*Jean Honorio, Purdue University; Tommi Jaakkola,* |
|  |  |
| **ID: 83**  | **Efficient Observation Selection in Probabilistic Graphical Models Using Bayesian Lower Bounds** +Real-world data often includes rich relational information, which can be leveraged to help predict unknown variables using a small amount of observed variables via a propagation effect. We consider the problem of selecting the best subset of variables to observe to maximize the overall prediction accuracy. Under the Bayesian framework, the optimal subset should be chosen to minimize the Bayesian optimal error rate, which, unfortunately, is critically challenging to calculate when the variables follow complex and high dimensional probabilistic distributions such as graphical models. In this paper, we propose to use a class of Bayesian lower bounds, including Bayesian Cramer Rao bounds as well as a novel extension of it to discrete graphical models, as surrogate criteria for optimal subset selection, providing a set of computationally efficient algorithms. Extensive experiments are presented to demonstrate our algorithm on both simulated and real-world datasets.*Dilin Wang, Dartmouth College; John Fisher III, MIT; Qiang Liu, Dartmouth College* |
|  |  |
| **ID: 86**  | **Efficient Feature Group Sequencing for Anytime Linear Prediction** +We consider extit{anytime} linear prediction in the common machine learning setting where features are in groups that have costs. We achieve anytime or interruptible predictions by sequencing computation of feature groups and reporting results using the computed features at interruption. We extend Orthogonal Matching Pursuit (OMP) and Forward Regression (FR) to learn the sequencing greedily under this group setting with costs. We theoretically guarantee that our algorithms achieve near-optimal linear predictions at each budget when a feature group is chosen. With a novel analysis of OMP, we improve its theoretical bound to the same strength as that of FR. In addition, we develop a novel algorithm that consumes cost $4B$ to approximate the optimal performance of extit{any} cost $B$, and prove that with cost less than $4B$, such an approximation is impossible. To our knowledge, these are the first anytime bounds at extit{all} budgets. We experiment our algorithms on two real-world data-sets and evaluate them in terms of anytime linear prediction performance against cost-weighted Group Lasso, and alternative greedy algorithms.*Hanzhang Hu, Carnegie Mellon University; Alexander Grubb, ; J. Andrew Bagnell, ; Martial Hebert,* |
|  |  |
| **ID: 87**  | **A Formal Solution to the Grain of Truth Problem** +A Bayesian agent acting in a multi-agent environment learns to predict the other agents' policies if its prior assigns positive probability to them (in other words, its prior contains a grain of truth). Finding a reasonably large class of policies that contains the Bayes-optimal policies with respect to this class is known as the grain of truth problem. Only small classes are known to have a grain of truth and the literature contains several related impossibility results. In this paper we present a formal and general solution to the full grain of truth problem: we construct a class of policies that contains all computable policies as well as Bayes-optimal policies for every lower semicomputable prior over the class. When the environment is unknown, Bayes-optimal agents may fail to act optimally even asymptotically. However, agents based on Thompson sampling converge to play ε-Nash equilibria in arbitrary unknown computable multi-agent environments. While these results are purely theoretical, we show that they can be computationally approximated arbitrarily closely.*Jan Leike, Australian National University; Jessica Taylor, Machine Intelligence Research Institute; Benya Fallenstein, Machine Intelligence Research Institute* |
|  |  |
| **ID: 90**  | **Convex Relaxation Regression: Black-Box Optimization of Smooth Functions by Learning Their Convex Envelopes** +Finding efficient and provable methods to solve non-convex optimization problems is an outstanding challenge in machine learning and optimization theory. A popular approach used to tackle non-convex problems is to use convex relaxation techniques to find a convex surrogate for the problem. Unfortunately, convex relaxations typically must be found on a problem-by-problem basis. Thus, providing a general-purpose strategy to estimate a convex relaxation would have a wide reaching impact. Here, we introduce Convex Relaxation Regression (CoRR), an approach for learning convex relaxations for a class of smooth functions. The main idea behind our approach is to estimate the convex envelope of a function f by evaluating f at a set of T random points and then fitting a convex function to these function evaluations. We prove that with probability greater than 1-\delta, the solution of our algorithm converges to the global optimizer of f with error O((\log(1/\delta) /T)^lpha ) for some lpha > 0. Our approach enables the use of convex optimization tools to solve a class of non-convex optimization problems.*Mohamm Gheshlaghi Azar, Northwestern University; Eva Dyer, Northwestern University; Konrad Kording, Northwestern University* |
|  |  |
| **ID: 91**  | **Model-Free Reinforcement Learning with Skew-Symmetric Bilinear Utilities** +In reinforcement learning, policies are typically evaluated according to the expectation of cumulated rewards. Researchers in decision theory have argued that more sophisticated decision criteria can better model the preferences of a decision maker. In particular, Skew-Symmetric Bilinear (SSB) utility functions generalize von Neumann and Morgenstern's expected utility (EU) theory to encompass rational decision behaviors that EU cannot accommodate. We adopt an SSB utility function to compare policies in the reinforcement learning setting. We provide a model-free SSB reinforcement learning algorithm, SSB Q-learning, and prove its convergence towards a policy that is epsilon-optimal according to SSB. The proposed algorithm is an adaptation of fictitious play combined with techniques from stochastic approximation. We also present some experimental results which evaluate our approach in a variety of settings.*Hugo Gilbert, LIP6-UPMC; Bruno Zanuttini, University of Caen, France; Paul Weng, SYSU-CMU JIE; Paolo Viappiani, Lip6, Paris; Esther Nicart, Cordon Electronics DS2i* |
|  |  |
| **ID: 96**  | **Cascading Bandits for Large-Scale Recommendation Problems** +Most recommender systems recommend a list of items. The user examines the list, from the first item to the last, and often chooses the first attractive item and does not examine the rest. This type of user behavior can be modeled by the cascade model. In this work, we study cascading bandits, an online learning variant of the cascade model where the goal is to recommend K most attractive items from a large set of L candidate items. We propose two algorithms for solving this problem, which are based on the idea of linear generalization. The key idea in our solutions is that we learn a predictor of the attraction probabilities of items from their features, as opposing to learning the attraction probability of each item independently as in the existing work. This results in practical learning algorithms whose regret does not depend on the number of items L. We bound the regret of one algorithm and comprehensively evaluate the other on a range of recommendation problems. The algorithm performs well and outperforms all baselines.*Shi Zong, CMU; Hao Ni, CMU; Kenny Sung, CMU; Rosemary Ke, University of Montreal; Zheng Wen, Adobe Research; Branislav Kveton, Adobe Research* |
|  |  |
| **ID: 99**  | **Training Neural Nets to Aggregate Crowdsourced Responses** +We propose a new method for aggregating crowdsourced responses, based on a deep neural network. Once trained, the aggregator network gets as inputs the responses of multiple participants to a the same set of questions, and outputs its prediction for the correct response to each question. We empirically evaluate our approach on a dataset of responses to a standard IQ questionnaire, and show it outperforms existing state-of-the-art methods.*Alex Gaunt, Microsoft Research; Diana Borsa, UCL; Yoram Bachrach, Microsoft Research* |
|  |  |
| **ID: 102**  | **Quasi-Newton Hamiltonian Monte Carlo** +The Hamiltonian Monte Carlo (HMC) method has become significantly popular in recent years. It is the state-of-the-art MCMC sampler due to its more efficient exploration to the parameter space than the standard random-walk based proposal. The key idea behind HMC is that it makes use of first-order gradient information about the target distribution. In this paper, we propose a novel dynamics. The new dynamics uses second-order geometric information about the desired distribution. The second-order information is estimated by using a quasi-Newton method (say, the BFGS method), so it does not bring heavy computational burden. Moreover, our theoretical analysis guarantees that this dynamics remains the target distribution invariant. As a result, the proposed quasi-Newton Hamiltonian Monte Carlo (QNHMC) algorithm traverses the parameter space more efficiently than the standard HMC and produces a less correlated series of samples. Finally, empirical evaluation on simulated data verifies the effectiveness and efficiency of our approach. We also conduct applications of QNHMC in Bayesian logistic regression and online Bayesian matrix factorization problems.*Tianfan Fu, Shanghai Jiao Tong University; Luo Luo, Shanghai Jiao Tong University; Zhihua Zhang, Shanghai Jiao Tong University* |
|  |  |
| **ID: 105**  | **Utilize Old Coordinates: Faster Doubly Stochastic Gradients for Kernel Methods** +To address the scalability issue of kernel methods, random features are commonly used for kernel approximation (Rahimi and Recht, 2007). They map the input data to a randomized low-dimensional feature space and apply fast linear learning algorithms on it. However, to achieve high precision results, one might still need large number of random features, which is infeasible in large-scale applications. Dai et al. (2014) address this issue by recomputing the random features of small batches in each iteration instead of pre-generating for the whole dataset and keeping them in the memory. The algorithm increases the number of random features linearly with iterations, which can reduce the approximation error to arbitrarily small. A drawback of this approach is that the large number of random features slows down the prediction and gradient evaluation after several iterations. We propose two algorithms to remedy this situation by “utilizing” old random features instead of adding new features in certain iterations. By checking the expected descent amount, the proposed algorithm selects “important” old features to update. The resulting procedure is surprisingly simple without enhancing the complexity of the original algorithm but effective in practice. We conduct empirical studies on both medium and large-scale datasets, such as ImageNet, to demonstrate the power of the proposed algorithms.*Chun-Liang Li, Carnegie Mellon University; Barnabas Poczos, Carnegie Mellon University* |
|  |  |
| **ID: 106**  | **Adversarial Inverse Optimal Control for General Imitation Learning Losses and Embodiment Transfer** +We develop a general framework for inverse optimal control that distinguishes between rationalizing demonstrated behavior and imitating inductively inferred behavior. This enables learning for more general imitative evaluation measures and differences between the capabilities of the demonstrator and those of the learner (i.e., differences in embodiment). Our formulation takes the form of a zero-sum game between a predictor attempting to minimize an imitative loss measure, and an adversary attempting to maximize the loss by approximating the demonstrated examples in limited ways. We establish the consistency and generalization guarantees of this approach and il- lustrate its benefits on real and synthetic imitation learning tasks.*Xiangli Chen, UIC; Mathew Monfort, ; Brian Ziebart, ; Peter Carr,* |
|  |  |
| **ID: 110**  | **Budgeted Semi-supervised Support Vector Machine** +Due to the prevalence of unlabeled data, semi-supervised learning has drawn significant attention and has been found applicable in many real-world applications. In this paper, we present the so-called Budgeted Semi-supervised Support Vector Machine (BS3VM), a method that leverages the excellent generalization capacity of kernel-based method with the adjacent and distributive information carried in a spectral graph for semi-supervised learning purpose. The fact that the optimization problem of BS3VM can be solved directly in the primal form makes it fast and efficient in memory usage. We validate the proposed method on several benchmark datasets to demonstrate its accuracy and efficiency. The experimental results show that BS3VM can scale up efficiently to the large-scale datasets where it yields a comparable classification accuracy while simultaneously achieving a significant computational speed-up comparing with the baselines.*Mi Dinh, HCMc University of Pedagogy, Vietnam; Phuong Duong, HCMc University of Pedagogy, Vietnam; Trung Le, HCMc University of Pedagogy; Tu Nguyen, PRaDA, Deakin university, Australia; Vu Nguyen, PRaDA, Deakin university, Australia; Dinh Phung, PRaDA, Deakin university, Australia* |
|  |  |
| **ID: 116**  | **Online Forest Density Estimation** +Online density estimation is the problem of predicting a sequence of outcomes, revealed one at a time, almost as well as the best expert chosen from a reference class of probabilistic models. The performance of each expert is measured with the log-likelihood loss. The class of experts examined in this paper is the family of discrete, acyclic graphical models, also known as Markov forests. By coupling Bayesian mixtures with symmetric Dirichlet priors for parameter learning, and a variant of ``Follow the Perturbed Leader'' strategy for structure learning, we derive an online forest density estimation algorithm that achieves a low regret, with a per-round time complexity that is quasi-quadratic in the input dimension. Using simple and flexible update rules, this algorithm can be easily adapted to predict with Markov trees or mixtures of Markov forests. Empirical results indicate that our online algorithm is a practical alternative to the state-of-the-art batch algorithms for learning tree-structured graphical models.*Frederic Koriche, CRIL* |
|  |  |
| **ID: 120**  | **Markov Beta Processes for Time Evolving Dictionary Learning** +We develop Markov beta processes (MBP) as a model suitable for data which can be represented by a sparse set of latent features which evolve over time. Most time evolving nonparametric latent feature models in the literature vary feature usage, but maintain a constant set of features over time. We show that being able to model features which themselves evolve over time results in the MBP outperforming other beta process based models. Our construction utilizes Poisson process operations, which leave each transformed beta process marginally beta process distributed. This allows one to analytically marginalize out latent beta processes, exploiting conjugacy when we couple them with Bernoulli processes, leading to a surprisingly elegant Gibbs MCMC scheme considering the expressiveness of the prior. We apply the model to the task of denoising and interpolating noisy image sequences and in predicting time evolving gene expression data, demonstrating superior performance to other beta process based methods.*Amar Shah, University of Cambridge; Zoubin Ghahramani, Cambridge University* |
|  |  |
| **ID: 124**  | **Content-based Modeling of Reciprocal Relationships using Hawkes and Gaussian Processes** +There has been growing interest in inferring implicit social structures using interaction data. This approach is motivated by the fact that entities organize themselves into groups having frequent interactions between each other. Unlike previous approaches that focused on subjectively declared relationships, the idea is to exploit the actual evidence at hand to reach conclusions about group formations, resulting in more objective data-driven inferences. To this end, Blundell et. al. (2012) have employed Hawkes processes, and proposed a Hawkes IRM model to infer social structures from interaction data. A major factor that encourages the use of Hawkes processes is the capability to model reciprocity in the interaction between social entities. However, reciprocation is dynamically conditioned upon two key factors: the significance of each message sent by the sender, and the receptivity to each message received by the receiver. In the model proposed by Blundell et. al. (2012), reciprocity is not affected by either of these factors, since the content of each message is not taken into account. In this paper, we extend the work of Blundell et. al. (2012) by introducing Gaussian processes (GPs) into the Hawkes IRM model: based on the content of each message, GPs are used to model the message significance as well as receptivity. This allows us to more accurately capture the interactions among entities. The application of GPs also allows us to flexibly model the rates of reciprocal activities between two entities, allowing asymmetry in reciprocity to be captured more accurately. This leads to better cluster detection capability. Our model outperforms previous Hawkes and Poisson process-based models at predicting verbal, email, and citation activities.*Xi Tan, Purdue University; Syed Naqvi, Purdue University; Yuan (Alan) Qi, Purdue University; Katherine Heller, Duke University; Vinayak Rao, Purdue University;* |
|  |  |
| **ID: 135**  | **Forward Backward Greedy Algorithms for Multi-Task Learning with Faster Rates** +A large body of algorithms have been proposed for multi-task learning. However, the effectiveness of many multi-task learning algorithms highly depends on the structural regularization, which incurs bias in the resulting estimators and leads to slower convergence rate. In this paper, we aim at developing a multi-task learning algorithm with faster convergence rate. In particular, we propose a general estimator for multi-task learning with row sparsity constraint on the parameter matrix, i.e., the number of nonzero rows in the parameter matrix being small. The proposed estimator is a nonconvex optimization problem. In order to solve it, we develop a forward backward greedy algorithm with provable guarantee. More specifically, we prove that the approximate solution output by the greedy algorithm attains a sharper estimation error bound than many state-of-the-art multi-task learning methods. Moreover, our estimator enjoys model selection consistency under a mild condition. Thorough experiments on both synthetic and real-world data demonstrate the effectiveness of our method and back up our theory.*Lu Tian, University of Virginia; Pan Xu, University of Virginia; Quanquan Gu,* |
|  |  |
| **ID: 138**  | **Learning to Smooth with Bidirectional Predictive State Inference Machines** +In this paper we present Learning to Smooth (L2S), a learning algorithm that is designed for optimizing the performance of predicting latent information using all past and future observations, on time series that can be modeled by chain Conditional Random Fields (chain-CRFs) with latent states. By leveraging Predictive State Representations (PSRs), we model the beliefs on latent states through predictive states—an alternative but equivalent representation that only consists of observable quantities. Predictive states enable the use of well developed supervised learning approaches to learn regressors or classifiers as predictors that can approximate forward and backward message passing, as well as the computation of the marginalization step, in the predictive state space. Instead of parameterizing the graphical models, we parametrize and restrict the class of predictors to encode the underlying graphical models. We show that L2S can achieve theoretical guarantees on smoothing performance in the agnostic setting. Additional empirical verification on dynamical system benchmarks, a sequential image recognition task and a natural language processing domain showcases the efficacy of L2S.*Wen Sun, Carnegie Mellon University; Roberto Capobianco, Sapienza University of Rome; J. Andrew Bagnell, ; Byron Boots, Georgia Institute of Technology; Geoffrey J. Gordon, Carnegie Mellon University* |
|  |  |
| **ID: 143**  | **Adaptive Algorithms and Data-Dependent Guarantees for Bandit Convex Optimization** +We present adaptive algorithms with strong data-dependent regret guarantees for the problem of bandit convex optimization. In the process, we develop a general framework from which all previous main results in this setting can be recovered. The key method is the introduction of adaptive regularization. By appropriately adapting the exploration scheme, we show that one can derive regret guarantees which can be significantly more favorable than those previously known. Moreover, our analysis also modularizes the problematic quantities in achieving the conjectured minimax optimal rates in the most general setting of the problem.*Scott Yang, Courant Institute of Mathemati; Mehryar Mohri, Courant Institute of Mathematical Sciences* |
|  |  |
| **ID: 145**  | **Bayesian Learning of Kernel Embeddings** +Kernel methods are one of the mainstays of machine learning, but the problem of kernel learning remains challenging, with only a few heuristics and very little theory. This is of particular importance in methods based on estimation of kernel mean embeddings of probability measures. For characteristic kernels, which include most commonly used ones, the kernel mean embedding uniquely determines its probability measure, so it can be used to design a powerful statistical testing framework, which includes nonparametric two-sample and independence tests. In practice, however, the performance of these tests can be very sensitive to the choice of kernel and its lengthscale parameters. To address this central issue, we propose a new probabilistic model for kernel mean embeddings, the Bayesian Kernel Embedding model, combining a Gaussian process prior over the Reproducing Kernel Hilbert Space containing the mean embedding with a conjugate likelihood function, thus yielding a closed form posterior over the mean embedding. The posterior mean of our model is closely related to recently proposed shrinkage estimators for kernel mean embeddings, while the posterior uncertainty is a new, interesting feature with various possible applications. Critically for the purposes of effective and principled kernel learning, our model gives a simple, closed form marginal likelihood of the observed data given the kernel hyperparameters. This marginal likelihood can either be optimized to inform the hyperparameter choice or fully Bayesian inference can be used.*Seth Flaxman, Oxford; Dino Sejdinovic, University of Oxford; John Cunningham, Columbia University; Sarah Filippi, Oxford* |
|  |  |
| **ID: 155**  | **A Generative Block-Diagonal Model for Clustering** +Probabilistic mixture models are among the most important clustering methods. These models assume that the feature vectors of the samples can be described by a mixture of several components. Each of these component follows a distribution of a certain form. In recent years, there has been an increasing amount of interest and work in similarity-matrix-based methods. Rather than considering the feature vectors, these methods learn patterns by observing the similarity matrix that describes the pairwise relative similarity between each pair of data points. However, there is limited work in probabilistic mixture model for clustering with input data in the form of a similarity matrix. Observing this, we propose a generative model for clustering that finds the block-diagonal structure of the similarity matrix to ensure that the samples within the same cluster (diagonal block) are similar while the samples from different clusters (off-diagonal block) are less similar. In this model, we assume the elements in the similarity matrix follow one of beta distributions, depending on whether the element belongs to one of the diagonal blocks or to off-diagonal blocks. The assignment of the element to a block is determined by the cluster indicators that follow categorical distributions. Experiments on both synthetic and real data show that the performance of our method is comparable to the state-of-the-art methods.*Junxiang Chen, Northeastern University; Jennifer Dy,* |
|  |  |
| **ID: 157**  | **Elliptical Slice Sampling with Expectation Propagation** +Markov Chain Monte Carlo techniques remain the gold standard for approximate Bayesian inference, but their practical issues -- including onerous runtime and sensitivity to tuning parameters -- often lead researchers to use faster but typically less accurate deterministic approximations. Here we couple the fast but biased deterministic approximation offered by expectation propagation with elliptical slice sampling, a state-of-the-art MCMC method. We extend our hybrid deterministic-MCMC method to include recycled samples and analytical slices, and we rigorously prove the validity of each enhancement. Taken together, we show that these advances provide an order of magnitude gain in efficiency beyond existing state-of-the-art sampling techniques in Bayesian classification and multivariate gaussian quadrature problems.*Francois Fagan, Columbia University; Jalaj Bhandari, Columbia University; John Cunningham, Columbia University* |
|  |  |
| **ID: 160**  | **Scalable Joint Modeling of Longitudinal and Point Process Data for Disease Trajectory Prediction and Improving Management of Chronic Kidney Disease** +A major goal in personalized medicine is the ability to provide individualized predictions about the future trajectory of a disease. Moreover, for many complex chronic diseases, patients simultaneously have additional comorbid conditions. Accurate determination of the risk of developing serious complications associated with a disease or its comorbidities may be more clinically useful than prediction of future disease trajectory in such cases. We propose a novel probabilistic generative model that can provide individualized predictions of future disease progression while jointly modeling the pattern of related recurrent adverse events. We fit our model using a scalable variational inference algorithm and apply our method to a large dataset of longitudinal electronic patient health records. Our model gives superior performance in terms of both prediction of future disease trajectories and of future serious events when compared to non-joint models. Our predictions are currently being utilized by our local accountable care organization during chart reviews of high risk patients.*Joseph Futoma, Duke University; Mark Sendak, Duke University; Blake Cameron, Duke University; Katherine Heller, Duke University* |
|  |  |
| **ID: 161**  | **Probabilistic Size-constrained Microclustering** +Microclustering refers to clustering models that produce small clusters or, equivalently, to models where the size of the clusters grows sublinearly with the number of samples. We formulate probabilistic microclustering models by assigning a prior distribution on the size of the clusters, and in particular consider microclustering models with explicit bounds on the size of the clusters. The combinatorial constraints make full Bayesian inference complicated, but we manage to develop a Gibbs sampling algorithm that can efficiently sample from the joint cluster allocation of all data points. We empirically demonstrate the computational efficiency of the algorithm for problem instances of varying difficulty.*Arto Klami, ; Aditya Jitta, University of Helsinki* |
|  |  |
| **ID: 163**  | **Active Uncertainty Calibration in Bayesian ODE Solvers** +There is resurging interest, in statistics and machine learning, in solvers for ordinary differential equations (ODEs) that return probability measures instead of point estimates. Recently, Conrad et al. introduced a sampling-based class of methods that are 'well-calibrated' in a specific sense. But the computational cost of these methods is significantly above that of classic methods. On the other hand, Schober et al. pointed out a precise connection between classic Runge-Kutta ODE solvers and Gaussian filters, which gives only a rough probabilistic calibration, but at negligible cost overhead. By formulating the solution of ODEs as approximate inference in linear Gaussian SDEs, we investigate a range of probabilistic ODE solvers, that bridge the trade-off between computational cost and probabilistic calibration, and identify the inaccurate gradient measurement as the crucial source of uncertainty. We propose the novel filtering-based method Bayesian Quadrature filtering (BQF) which uses Bayesian quadrature to actively learn the imprecision in the gradient measurement by collecting multiple gradient evaluations.*Hans Kersting, MPI for Intelligent Systems; Philipp Hennig, MPI for Intelligent Systems* |
|  |  |
| **ID: 167**  | **Gradient Methods for Stackelberg Games** +Stackelberg games are two-stage games in which the first player (called the leader) commits to a strategy, after which the other player (the follower) selects a best-response. These types of games have seen numerous practical application in security settings, where the leader (in this case, a defender) must allocate resources to protect various targets. Real world applications include the scheduling of US federal air marshals to international flights, and resource allocation at LAX airport. However, the best known algorithm for solving general Stackelberg games requires solving Integer Programs, and fails to scale beyond a few (significantly smaller than 100) number of leader actions, or follower types. In this paper, we present a new approach for solving large Stackelberg games in the security settings, using techniques from AI. Large-scale control problems are often times solved by restricting the controller to a rich, but parameterized, class of policies; the optimal control can then be computed using Monte Carlo gradient methods. We demonstrate that the same approach can be taken in a strategic setting. We evaluate our strategies empirically, demonsrating they have negligible regret against the leader's true equilibrium strategy, while scaling to large games.*Kareem Amin, University of Michigan; Michael Wellman, ; Satinder Singh,* |
|  |  |
| **ID: 168**  | **Optimal Stochastic Strongly Convex Optimization with a Logarithmic Number of Projections** +We consider stochastic strongly convex optimization with a complex inequality constraint. The inequality constraint may lead to computationally expensive projections in algorithmic iterations, when the stochastic gradient descent method is employed as the solver. To reduce the computational costs pertaining to the projections, we propose the Epoch-Projection Stochastic Gradient Descent method (Epro-SGD), which consists of a sequence of epochs and only performs projections at the end of each epoch. To prevent the intermediate solutions from violating the constraint and ensure that the projection at the end of each epoch does not deviate largely from the objective value, we hinge the constraint into the objective function (with a fixed Lagrangian multiplier) and apply SGD to the augmented objective function at each iteration using a stochastic gradient of the original objective function and a subgradient of the hinged constraint function. We show that the proposed Epro-SGD method enjoys an optimal convergence rate for stochastic strongly convex optimization, i.e., O(1/T), while it only requires to perform log(T) projections for a total number of iterations T. For illustration we apply Epro-SGD for solving a distance metric learning formulation with a positive definite constraint. The empirical studies on real data sets verify the effectiveness of the proposed Epro-SGD method.*Jianhui Chen, Yahoo; Tianbao Yang, University of Iowa; Lijun Zhang, Nanjing University; Qihang Lin, ; Yi Chang, Yahoo!* |
|  |  |
| **ID: 183**  | **Subspace Clustering with a Twist** +Subspace segmentation or clustering can be defined as the process of assigning subspace labels to a set of data points assumed to lie on the union of multiple low-dimensional, linear subspaces. Given that each point can be efficiently expressed using a linear combination of other points from the same subspace, a variety of segmentation algorithms built upon $\ell\_1$, nuclear norm, and other convex penalties have recently shown state-of-the-art robustness on multiple benchmarks. However, what if instead of observing the original data points, we instead only have access to transformed, or `twisted' so to speak, measurements? Here we consider underdetermined affine transformations that may arise in computer vision applications such as bidirectional reflectance distribution function (BRDF) estimation. Unfortunately most existing approaches, convex or otherwise, do not address this highly useful generalization. To fill this void, we proceed by deriving a probabilistic model that simultaneously estimates the latent data points and subspace memberships using simple EM update rules. Moreover, in certain restricted settings this approach is guaranteed to produce the correct clustering. Finally a wide range of corroborating empirical evidence, including a BRDF estimation task, speaks to the practical efficacy of this algorithm.*David Wipf, Microsoft Research; Yue Dong, Microsoft Research; Bo Xin, Peking University* |
|  |  |
| **ID: 185**  | **Bounded Rationality in Wagering Mechanisms** +Wagering mechanisms allow decision makers to inexpensively collect forecasts from groups of experts who reveal their information via bets with one another. Such mechanisms naturally induce a game in which strategic considerations come into play. What happens in the game depends on the reasoning power of the experts. At one extreme, if experts are fully rational, no-trade theorems imply no participation. At the other extreme, if experts ignore strategic considerations, even the least informed will wager as if his beliefs are correct. Economists have analyzed the former case and decision theorists the latter, but both are arguably unrealistic. In this paper, we adopt an intermediate model of bounded rationality in wagering mechanisms based on level-k reasoning. Under this model, overconfidence allows some participation to be sustained, but experts who realize they are at a relative disadvantage do bow out. We derive conditions on the particular wagering mechanism used under which participation is unbiased, and show that unbiasedness always implies truthful reports. We show that if participation is unbiased, then participation rates unavoidably fall as players' rationality increases, vanishing for large k. Finally, we zoom in on one particular information structure to give a complete characterization specifying the conditions under which mechanisms are unbiased and show how to maximize participation rates among all unbiased mechanisms.*David Pennock, Microsoft Research; Vasilis Syrgkanis, Microsoft Research; Jennifer Wortman Vaughan, Microsoft Research* |
|  |  |
| **ID: 186**  | **Bridging Heterogeneous Domains With Parallel Transport For Vision and Multimedia Applications** +Accounting for different feature types across datasets is a relatively under-studied problem in domain adaptation. We address this heterogeneous adaptation setting using principles from parallel transport and hierarchical sparse coding. By learning generative subspaces from each domain, we first perform label-independent cross-domain feature mapping using parallel transport, and obtain a collection of paths (bridges) that could compensate domain shifts. We encode the information contained in these bridges into an expanded prior, and then integrate the prior into a hierarchical sparse coding framework to learn a selective subset of codes representing holistic data properties that are robust to domain change and feature type variations. We then utilize label information on the sparse codes to perform classification, or in the absence of labels perform clustering, and obtain improved results on several previously studied heterogeneous adaptation datasets. We highlight the flexibility of our approach by accounting for multiple heterogeneous domains in training as well as in testing, and by considering the zero-shot domain transfer scenario where there are data categories in testing which are not seen during training. In that process we also empirically show how existing heterogeneous adaptation solutions can benefit from the findings of our study.*Raghuraman Gopalan, AT&T Research* |
|  |  |
| **ID: 193**  | **A General Statistical Framework for Designing Strategy-proof Assignment Mechanisms** +We develop a statistical framework for designing strategy-proof assignment mechanisms from a sample of agent type profiles. Our goal is to find a strategy-proof mechanism that closely approximates a given outcome rule. Our framework is quite flexible, does not require specialized characterization results, and allows a designer to control the space of strategy-proof mechanisms searched over by providing a rule class with appropriate capacity. We are also able to handle both assignment problems with and without money within the same framework. Our approach involves solving a sample-based optimization problem over a chosen space of agent-independent rules, subject to a feasibility constraint on the sample. A transformation is applied to the obtained rule to ensure feasibility on all type profiles. We derive a general sample complexity bound in terms of the capacity of the chosen class and provide applications for our results.*Harikrishna Narasimhan, Harvard University; David Parkes, Harvard University* |
|  |  |
| **ID: 202**  | **Accelerated Stochastic Block Coordinate Gradient Descent for Sparsity Constrained Nonconvex Optimization** +We propose an accelerated stochastic block coordinate descent algorithm for nonconvex optimization under sparsity constraint in the high dimensional regime. At the core of our algorithm is leveraging both stochastic partial gradient and full partial gradient restricted to each coordinate block to accelerate the convergence. We prove that the algorithm converges to the unknown true parameter at a linear rate, up to the statistical error of the underlying model. Experiments on both synthetic and real datasets backup our theory.*Jinghui Chen, University of Virginia; Quanquan Gu,* |
|  |  |
| **ID: 204**  | **Incremental Preference Elicitation for Decision Making Under Risk with the Rank-Dependent Utility Model** +This work concerns decision making under risk with the rank-dependent utility model (RDU), a generalization of expected utility providing enhanced descriptive possibilities. We introduce a new incremental decision procedure, involving monotone regression spline functions to model both components of RDU, namely the probability weighting function and the utility function. First, assuming the utility function is known, we propose an elicitation procedure that incrementally collects preference information in order to progressively specify the probability weighting function until the optimal choice can be identified. Then, we present two elicitation procedures for the construction of a utility function as a monotone spline. Finally, numerical tests are provided to show the practical efficiency of the proposed methods.*Patrice Perny, LIP6; Paolo Viappiani, Lip6, Paris; abdellah Boukhatem, LIP6* |
|  |  |
| **ID: 212**  | **Online learning with Erdos-Renyi side-observation graphs** +We consider adversarial multi-armed bandit problems where the learner is allowed to observe losses of a number of arms beside the arm that it actually chose. We study the case where all non-chosen arms reveal their loss with an unknown probability r\_t, independently of each other and the action of the learner. Moreover, we allow r\_t to change in every round t, which rules out the possibility of estimating r\_t by a well-concentrated sample average. We propose an algorithm which operates under the assumption that r\_t is large enough to warrant at least one side observation with high probability. We show that after T rounds in a bandit problem with N arms, the expected regret of our algorithm is of order O(\sqrt{\sum\_{t=1}^T (1/r\_t) \log N }), given that r\_t\ge \log T / (2N-2) for all t. All our bounds are within logarithmic factors of the best achievable performance of any algorithm that is even allowed to know exact values of r\_t.*Tomáš Kocák, Inria Lille - Nord Europe; gergely Neu, ; Michal Valko, Inria Lille - Nord Europe* |
|  |  |
| **ID: 214**  | **Stability of Causal Inference** +We consider the sensitivity of causal identification to small perturbations in the input. A long line of work culminating in papers by Shpitser and Pearl and Huang and Valtorta led to a complete procedure for the causal identification problem. In our main result in this paper, we show that the identification function computed by these procedures is in some cases extremely unstable numerically. Specifically, the ``condition number'' of causal identification can be of the order of Ω(exp(n^{0.49})) on an identifiable semi-Markovian model with n visible nodes. That is, in order to give an output accurate to d bits, the empirical probabilities of the observable events need to be obtained to accuracy d+Ω(n^{0.49}) bits.*Leonard Schulman, California Institute of Technology; Piyush Srivastava, California Institute of Techno* |
|  |  |
| **ID: 217**  | **Inferring Causal Direction from Relational Data** +Inferring the direction of causal dependence from observational data is a fundamental problem in many scientific fields. Significant progress has been made in inferring causal direction from data that are independent and identically distributed (i.i.d.), but little is understood about this problem in the more general relational setting. This work examines the task of inferring the causal direction of peer dependence in relational data. We show that, in contrast to the i.i.d. setting, the direction of peer dependence can be inferred using simple procedures, regardless of the form of the underlying distribution, and we provide a theoretical characterization on the identifiability of direction. We then examine the conditions under which the presence of confounding can be detected. Finally, we demonstrate the efficacy of the proposed methods with synthetic experiments, and we provide an application on real-world data.*David Arbour, University of Massachusetts Am; Katerina Marazopoulou, University of Massachusetts Amhe; David Jensen,* |
|  |  |
| **ID: 218**  | **Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions** +Several recent works have explored stochastic gradient methods for variational inference that exploit the geometry of the variational-parameter space. However, the theoretical properties of these methods are not well-understood and they typically only apply to conditionally-conjugate models. We present a new stochastic method for variational inference which exploits the geometry of the variational-parameter space, and yields simple closed-form updates even for nonconjugate models. We also give a convergence rate analysis of our method, as well as many other previous methods which exploit the geometry of the space. Our analysis generalizes existing convergence results for stochastic mirror descent on non-convex objectives by using a more general class of divergence functions. Beyond giving a theoretical justification for a variety of recent methods, our experiments show that new algorithms derived in this framework lead to state of the art results on a variety of problems. Further, we expect that the generality of our theoretical analysis could also apply to a variety of other applications.*Reza Babanezhad Harikandeh, UBC; Mark Schmidt, University of British Columbia; Mohammad Emtiyaz Khan, ; Wu Lin, ; Masashi Sugiyama,* |
|  |  |
| **ID: 219**  | **Taming the Noise in Reinforcement Learning via Soft Updates** +Model-free reinforcement learning algorithms, such as Q-learning, perform poorly in the early stages of learning in noisy environments, because much effort is spent unlearning biased estimates of the state-action value function. The bias results from selecting, among several noisy estimates, the apparent optimum, which may actually be suboptimal. We propose G-learning, a new off-policy learning algorithm that regularizes the value estimates by penalizing deterministic policies at the beginning of the learning process. We show that this method reduces the bias of the value-function estimation, leading to faster convergence to the optimal value and the optimal policy. Moreover, G-learning enables the natural incorporation of prior domain knowledge, when available. The stochastic nature of G-learning also makes it avoid some exploration costs, a property usually attributed only to on-policy algorithms. We illustrate these ideas in several examples, where G-learning results in significant improvements of the convergence rate and the cost of the learning process.*Roy Fox, HUJI; Ari Pakman, Columbia University; Naftali Tishby, HUJI* |
|  |  |
| **ID: 223**  | **Conjugate Conformal Prediction for Online Binary Classification** +Binary classification (rain or shine, disease or not, increase or decrease) is a fundamental problem in machine learning. We present an algorithm that can take any standard online binary classification algorithm and provably improve its performance under very weak assumptions. The extent of improvement will depend on the data size, stability of the algorithm, and room for improvement in the algorithms performance. Our experiments on standard machine learning data sets and standard algorithms (k-nearest neighbors and random forests) show the effectiveness of our approach, even beyond what is possible using previous work on conformal predictors which our approach is also based on. Though we focus on binary classification, our theory extends to multiway classification and can be applied to general machine learning under concept drift. Our code and data will be housed on a permanent website for purposes of reproducibility testing and use by the community.*Mustafa Kocak, NYU Tandon SoE; Dennis Shasha, Courant Institute of Mathematical Sciences New York University; Elza Erkip, NYU Tandon School of Engineering* |
|  |  |
| **ID: 225**  | **Large-scale Submodular Greedy Exemplar Selection with Structured Similarity Matrices** +Exemplar clustering attempts to find a subset of data-points that summarizes the entire data-set in the sense of minimizing the sum of distances from each point to its closest exemplar. It has many important applications in machine learning including document and video summarization, data compression, scalability of kernel methods and Gaussian processes, active learning and feature selection. A key challenge in the adoption of exemplar clustering to large-scale applications has been the availability of accurate and scalable algorithms. We propose an approach that combines structured similarity matrix representations with submodular greedy maximization that can dramatically increase the scalability of exemplar clustering and still enjoy good approximation guarantees. Exploiting structured similarity matrices within the context of submodular greedy algorithms is by no means trivial, as naive approaches still require computing all the entries of the matrix. We propose a randomized approach based on sampling sign-patterns of columns of the similarity matrix and establish accuracy guarantees. We demonstrate significant computational speed-ups while still achieving highly accurate solutions, and solve a problem with millions of data-points in about a minute on a single commodity computer.*Dmitry Malioutov, IBM Research; Abhishek Kumar, IBM Research; Ian Yen, University of Texas at Austin* |
|  |  |
| **ID: 226**  | **Finite Sample Complexity of Rare Pattern Anomaly Detection** +Anomaly detection is a fundamental problem for which a wide variety of algorithms have been developed. However, compared to supervised learning, there has been very little work aimed at understanding the sample complexity of anomaly detection. In this paper, we take a step in this direction by introducing a Probably Approximately Correct (PAC) framework for anomaly detection based on the identification of rare patterns. In analogy with the PAC framework for supervised learning, we develop sample complexity results that relate the complexity of the pattern space to the data requirements needed for PAC guarantees. We instantiate the general result for a number of pattern spaces, some of which are implicit in current state-of-the-art anomaly detectors. Finally, we design a new simple anomaly detection algorithm motivated by our analysis and show experimentally on several benchmark problems that it is competitive with a state-of-the-art detector using the same pattern space.*Md Amran Siddiqui, Oregon Sate University; Alan Fern, ; Thomas Dietterich, Oregon State University; Shubhomoy Das, Oregon State University* |
|  |  |
| **ID: 227**  | **Towards a Theoretical Understanding of Negative Transfer in Collective Matrix Factorization** +Collective matrix factorization (CMF) is a popular technique to boost the overall factorization quality of related matrices, under the assumption that these matrices share a same factor. It is widely admitted that CMF suffers from performance degeneration when this assumption fails, an effect called i{negative transfer (n.t.)}. However, the theoretical nature of this effect remains a mysterious to date. This paper aims to advance our understanding on negative transfer in theory. Under the statistical mini-max framework, we derive a lower bound for the CMF estimator and gain two important insights. First, the $n.t.$ effect can be explained as the rise of a bias term in the standard lower bound. In particular, the bias depends only on the structure of factor space, suggesting that $n.t.$ may only be resolved at the phase of model construction but not model selection. Second, the $n.t.$ effect can be explained as the rise of an $n\_{th}$-root function on the learning rate. To be specific, through a finer lower bound we obtain a learning rate of $\Omega(1/|\omega|^{rac{1}{d}})$ for CMF, where $d$ is the dimension of a Grassmannian containing the ranges of factors. This rate is $d\_{th}$-root more slowly than the standard $n.t.$-free rate $\Omega(1/|\omega|)$, suggesting that $n.t.$ may be much more effectively addressed by refining $d$ instead of increasing sample. This discovery is also supported in simulation.*Chao Lan, University of Kansas; Jun Huan, University of Kansas* |
|  |  |
| **ID: 229**  | **Importance Weighted Consensus Monte Carlo for Distributed Bayesian Inference** +The recent explosion in big data has created a significant challenge for efficient and scalable Bayesian inference. In this paper, we consider a divide-and-conquer setting in which case the data is partitioned into different subsets with communication constraints, and a proper combination strategy is used to aggregate the Monte Carlo samples from each of the subsets. We propose a new importance weighted consensus Monte Carlo method for efficient Bayesian inference in this setting. Our method significantly outperforms the previous one-shot combination strategies in terms of accuracy, and is also much more efficient than the previous iterative combination methods that require repeated re-sampling and communication. We provide two practical versions of our method, and illustrate their properties both theoretically and empirically.*Qiang Liu, Dartmouth College* |
|  |  |
| **ID: 235**  | **A Correlated Worker Model for Grouped, Imbalanced and Multitask Data** +We consider the important crowdsourcing problem of estimating worker confusion matrices, or sensitivities and specificities for binary classification tasks. In addition to providing diagnostic insights into worker performance, such estimates enable robust online task routing for classification tasks exhibiting imbalance and asymmetric costs. However, labeled data is often expensive and hence estimates must be made without much of it. This poses a challenge to existing methods. In this paper, we propose a novel model that captures the correlations between entries in confusion matrices. We applied this model in two practical scenarios: (1) an imbalanced classification task in which workers are known to belong to groups and (2) a multitask scenario in which labels for the same workers are available in more than one labeling task. We derive an efficient variational inference approach that scales to large datasets. Experiments on two real world citizen science datasets (biomedical citation screening and galaxy morphological classification) demonstrate consistent improvement over competitive baselines.*An Nguyen, University of Texas at Austin; Byron Wallace, University of Texas at Austin; Matthew Lease, University of Texas at Austin* |
|  |  |
| **ID: 236**  | **The Mondrian Kernel** +We introduce the Mondrian kernel, a fast random feature approximation to the expensive problem of learning with a Laplace kernel. It is suitable for both batch and online learning, and admits a fast kernel width selection procedure as the random features can be re-used efficiently for all kernel widths. The features are constructed by sampling trees from a Mondrian process [Roy and Teh, 2009], and we highlight the connection to Mondrian forests [Lakshminarayanan et al., 2014], where trees are also sampled from a Mondrian process, but fit independently. This link provides a new insight into the relationship between kernel methods and random forests.*Matej Balog, University of Cambridge; Balaji Lakshminarayanan, ; Daniel Roy, University of Toronto; Yee Whye Teh, University of Oxford* |
|  |  |
| **ID: 239**  | **Learning Network of Multivariate Hawkes Processes: A Time Series Approach** +Learning the influence structure of multiple time series data is of great interest to many disciplines. This paper studies the problem of recovering the causal structure in network of multivariate linear Hawkes processes. In such processes, the occurrence of an event in one process affects the probability of occurrence of new events in some other processes. Thus, a natural notion of causality exists between such processes captured by the support of the excitation matrix. We show that the resulting causal influence network is equivalent to the Directed Information graph (DIG) of the processes, which encodes the causal factorization of the joint distribution of the processes. Furthermore, we present an algorithm for learning the support of excitation matrix (or equivalently the DIG). The performance of the algorithm is evaluated on synthesized multivariate Hawkes networks as well as a stock market and MemeTracker real-world dataset.*Jalal Etesami, UIUC; negar kiyavash, uiuc; Kun Zhang, cmu; Kushagra Singhal, uiuc* |
|  |  |
| **ID: 242**  | **Bayesian Estimators As Voting Rules** +We investigate the fairness of Bayesian estimators (BEs) by viewing them as (irresolute) voting rules and evaluating them by satisfaction of desirable social choice axioms. We characterize the class of BEs that satisfy neutrality by the class of BEs with neutral structures. We prove that a BE with a neutral structure is a minimax rule if it further satisfies parameter connectivity. We prove that no BE satisfies strict Condorcet criterion. We also propose three new BEs of natural frameworks and investigate their computational complexity and satisfaction of monotonicity and Condorcet criterion.*Lirong Xia,* |
|  |  |
| **ID: 246**  | **Budget Allocation using Weakly Coupled, Constrained Markov Decision Processes** +We consider the problem of budget (or other resource) allocation in sequential decision problems involving a large number of concurrently running sub-processes, whose only interaction is through their consumption of budget. Our first contribution is the introduction of budgeted MDPs (BMDPs), an MDP model in which policies/values are a function of available budget, (c.f. constrained MDPs which are solved given a fixed budget). BMDPs allow one to explicitly trade off allocated budget and expected value. We show that optimal value functions are concave, non-decreasing in budget, and piecewise-linear in the finite horizon case, and can be computed by dynamic programming (and support ready approximation). Our second contribution is method that exploits BMDP solutions to allocate budget to a large number of independent BMDPs, coupled only by their common budget pool. The problem can be cast as a multiple-choice knapsack problem, which admits an efficient, optimal greedy algorithm. Empirical results in an online advertising domain confirm the efficacy of our methods.*Craig Boutilier, Google; Tyler Lu, Google* |
|  |  |
| **ID: 247**  | **A Kernel Test for Three-Variable Interactions with Random Processes** +We apply a wild bootstrap method to the Lancaster three-variable interaction measure in order to detect factorisation of the joint distribution on three variables forming a stationary random process, for which the existing permutation bootstrap method fails. As in the i.i.d. case, the Lancaster test is found to outperform existing tests in cases for which two independent variables individually have a weak influence on a third, but that when considered jointly the influence is strong. The main contributions of this paper are twofold: first, we prove that the Lancaster statistic satisfies the conditions required to estimate the quantiles of the null distribution using the wild bootstrap; second, the manner in which this is proved is novel, simpler than existing methods, and can further be applied to other statistics.*Paul Rubenstein, Cambridge/MPI Tuebingen; Kacper Chwialkowski, UCL / Gatsby Unit; Arthur Gretton, Gatsby Unit, University College London* |
|  |  |
| **ID: 253**  | **Context-dependent feature analysis with random forests** +For many problems, feature selection is often more complicated than identifying a single subset of input variables that would together explain the output. There may be interactions that depend on contextual information, i.e., variables that reveal to be relevant only in some specific circumstances. In this setting, the contribution of this paper is to extend the random forest variable importances framework in order (i) to identify variables whose relevance is context-dependent and (ii) to characterize as precisely as possible the effect of contextual information on these variables. The usage and the relevance of our framework for highlighting context-dependent variables is illustrated on both artificial and real datasets.*Antonio Sutera, University of Liège; Gilles Louppe, CERN; Vân Anh Huynh-Thu, University of Liège; Louis Wehenkel, University of Liège; Pierre Geurts, University of Liège* |
|  |  |
| **ID: 255**  | **Analysis of Nystrom method with sequential ridge leverage scores** +Large-scale kernel ridge regression (KRR) is limited by the need to store a large kernel matrix $K\_t$. To avoid storing the entire matrix $K\_t$, Nystrom methods subsample a subset of columns of the kernel matrix, and efficiently find an approximate KRR solution on the reconstructed $ ilde K\_t$. The chosen subsampling distribution in turn affects the statistical and computational tradeoffs. For KRR problems, Alaoui et al. (2015) show that a sampling distribution proportional to the ridge leverage scores (RLSs) provides strong reconstruction guarantees. While exact RLSs are as difficult to compute as a KRR solution, we may be able to approximate them well enough. In this paper, we study KRR problems in a sequential setting and introduce the INK-Estimate algorithm, that incrementally computes the RLSs estimates. INK-Estimate maintains a small sketch of $K\_t$, that at each step is used to compute an intermediate estimate of the RLSs. First, our sketch update does not require access to previously seen columns, and therefore a single pass over the kernel matrix is sufficient. Second, the algorithm requires a fixed, small space budget to run dependent only on the effective dimension of the kernel matrix. Finally, our sketch provides strong approximation guarantees on the distance $|K\_t - ilde K\_t|\_2$, and on the statistical risk of the approximate KRR solution at any time, because all our guarantees hold at any intermediate step.*Daniele Calandriello, INRIA Lille - Nord Europe; Alessandro Lazaric, ; Michal Valko, Inria Lille - Nord Europe* |
|  |  |
| **ID: 257**  | **Degrees of Freedom in Deep Neural Networks** +In this paper, we explore degrees of freedom in deep sigmoidal neural networks. We show that the degrees of freedom in these models is related to the expected optimism, which is the expected difference between test error and training error. We provide an efficient Monte-Carlo method to estimate the degrees of freedom for multi-class classification methods. We show degrees of freedom are lower than the parameter count in a simple XOR network. We extend these results to neural nets trained on synthetic and real data, and investigate impact of network's architecture and different regularization choices. The degrees of freedom in deep networks are dramatically smaller than the number of parameters, in some real datasets several orders of magnitude. Further, we observe that for fixed number of parameters, deeper networks have less degrees of freedom exhibiting a regularization-by-depth.*Tianxiang Gao, University of North Carolina a; Vladimir Jojic, University of North Carolina at Chapel Hill* |
|  |  |
| **ID: 262**  | **Scalable Nonparametric Bayesian Multilevel Clustering** +Multilevel clustering problems where the content and contextual information are jointly clustered are ubiquitous in modern data sets. Existing work on this problem are limited to small datasets due to the use of the Gibbs sampler. We address the problem of scaling up multilevel clustering under a Bayesian nonparametric setting, extending the MC2 model proposed in (Nguyen et al., 2014). We ground our approach in mean-field and stochastic variational inference (SVI) theory. However, the interplay between content and context modeling makes naive mean-field approach inefficient. We develop a tree-structured SVI algorithm that avoids the need to repeatedly go through the corpus as in Gibbs sampler. More crucially, our method is immediately amendable to parallelization, facilitating a scalable distributed implementation of our algorithm on the Apache Spark platform. We conducted extensive experiments in a variety of domains including text, images, and real-world user application activities. Direct comparison with the Gibbs-sampler demonstrates that our method is an order-of-magnitude faster without loss of model quality. Our Spark-based implementation gains another order-of-magnitude speed up and can scale to large real-world data sets containing millions of documents and groups.*Viet Huynh, Deakin University; Dinh Phung, PRaDA, Deakin university, Australia; Svetha Venkatesh, Deakin University; Long Nguyen, ; Matthew Hoffman, ; Hung Bui, Adobe Research* |
|  |  |
| **ID: 267**  | **On Hyper-Parameter Estimation In Empirical Bayes: A Revisit of The MacKay Algorithm** +An iterative procedure introduced in MacKay's evidence framework is often used for estimating the hyper-parameter in empirical Bayes. Despite its effectiveness, the procedure has stayed primarily as a heuristic to date. This paper formally investigates the mathematical nature of this procedure and justifies it as a well-principled algorithm framework. This framework, which we call the MacKay algorithm, is shown to be closely related to the EM algorithm under certain Gaussian assumption.*Chune Li, Beihang University; Yongyi Mao, University of Ottawa; Richong Zhang, Beihang University; Jinpeng Huai, Beihang University* |
|  |  |
| **ID: 268**  | **Modeling Transitivity in Complex Networks** +An important source of high clustering coefficient in real-world networks is transitivity. However, existing algorithms which model transitivity suffer from at least one of the following problems: i) they produce graphs of a specific class like bipartite graphs, ii) they do not give an analytical argument for the high clustering coefficient of the model, and iii) their clustering coefficient is still significantly lower than real-world networks. In this paper, we propose a new model for complex networks which is based on adding transitivity to scale-free models. We theoretically analyze the model and provide analytical arguments for its different properties. In particular, we calculate a lower bound on the clustering coefficient of the model which is independent of the network size, as seen in real-world networks. More than theoretical analysis, the main properties of the model are evaluated empirically and it is shown that the model can precisely simulate real-world networks from different domains with and different specifications.*Morteza Haghir Chehreghani, Xerox Research Centre Europe; Mostafa Haghir Chehreghani, KU Leuven* |
|  |  |
| **ID: 269**  | **Sparse Gaussian Processes for Bayesian Optimization** +Bayesian optimization schemes often rely on Gaussian processes (GP). GP models are very flexible, but are known to scale poorly with the number of training points. While several efficient sparse GP models are known, they have limitations when applied in optimization settings. We propose a novel Bayesian optimization framework that uses sparse online Gaussian processes (GPs). We introduce a new updating scheme for the online GP that accounts for our preference during optimization for regions with better performance. We apply this method to optimize the performance of a free-electron laser, and demonstrate empirically that the weighted updating scheme leads to both improved results and greater consistency, which is particularly desirable when the cost of optimization is high.*Mitchell McIntire, Stanford University; Daniel Ratner, SLAC National Accelerator Laboratory; Stefano Ermon,* |
|  |  |
| **ID: 270**  | **Sequential Nonparametric Testing with the Law of the Iterated Logarithm** +We propose a new algorithmic framework for sequential hypothesis testing with i.i.d. data, which includes A/B testing, nonparametric two-sample testing, and independence testing as special cases. It is novel in several ways: (a) it takes linear time and constant space to compute on the fly, (b) it has the same power guarantee as a nonsequential version of the test with the same computational constraints up to a small factor, and (c) it accesses only as many samples as are required – its stopping time adapts to the unknown difficulty of the problem. All our test statistics are constructed to be zero-mean martingales under the null hypothesis, and the rejection threshold is governed by a uniform non-asymptotic law of the iterated logarithm (LIL). For the case of nonparametric two-sample mean testing, we also provide a finite sample power analysis, and the first non-asymptotic stopping time calculations for this class of problems. We verify our predictions for type I and II errors and stopping times using simulations.*Akshay Balsubramani, UC San Diego; Aaditya Ramdas, UC Berkeley* |
|  |  |
| **ID: 284**  | **Stochastic Portfolio Theory: A Machine Learning Approach** +In this paper we propose a novel application of Gaussian processes (GPs) to financial asset allocation. Our approach is deeply rooted in Stochastic Portfolio Theory (SPT), a stochastic analysis framework recently introduced by Robert E. Fernholz that aims at flexibly analysing the performance of certain investment strategies in stock markets relative to benchmark indices. In particular, SPT has exhibited some investment strategies based on company sizes that, under realistic assumptions, outperform benchmark indices with probability 1 over certain time horizons. Galvanised by this result, we consider the inverse problem that consists of learning (from historical data) an optimal investment strategy based on any given set of trading characteristics, and using a user-specified optimality criterion that may go beyond outperforming a benchmark index. Although the inverse problem is of the utmost interest to investment management practitioners, it can hardly be tackled using the SPT framework. We show that our machine learning approach learns investment strategies that considerably outperform existing SPT strategies in the US stock market.*Yves-Laurent Kom Samo, University of Oxford; Alexander Vervuurt, University of Oxford* |
|  |  |
| **ID: 286**  | **Individual Planning in Open and Typed Agent Systems** +Open agent systems are multiagent systems in which one or more agents may leave the system at any time possibly resuming after some interval and in which new agents may also join. Planning in such systems becomes challenging in the absence of inter-agent communication because agents must predict if others have left the system or new agents are now present to decide on possibly choosing a different line of action. In this paper, we prioritize open systems where agents of differing types may leave and possibly reenter but new agents do not join. With the help of a realistic domain -- wildfire suppression -- we motivate the need for individual planning in open environments and present a first approach for robust decision-theoretic planning in such multiagent systems. Evaluations in domain simulations clearly demonstrate the improved performance compared to previous methods that disregard the openness.*Muthukumaran Chandrasekaran, University of Georgia; Adam Eck, University of Nebraska; Prashant Doshi, University of Georgia; Leenkiat Soh, University of Nebraska* |
|  |  |
| **ID: 293**  | **Random Features for Online Sampling with a Reservoir** +We introduce an alternative to reservoir sampling, a classic and popular algorithm for drawing a fixed-size subsample from streaming data in a single pass. Rather than draw a random sample, our approach performs an online optimization which aims to select the subset which provides the best overall approximation to the full data set, as judged using a kernel two-sample test. This produces subsets which minimize the worst-case relative error when computing expectations of functions in a specified function class, using just the samples from the subset. Kernel functions are approximated using random Fourier features, and the subset of samples itself is stored in a random projection tree, allowing for an algorithm which runs in a single pass through the whole data set, with only a logarithmic time complexity in the size of the subset at each iteration. These “super-samples” subsampled from the full data provide a concise summary, as demonstrated empirically on mixture models and the MNIST dataset.*Brooks Paige, University of Oxford; Dino Sejdinovic, University of Oxford; Frank Wood, University of Oxford* |
|  |  |
| **ID: 294**  | **MDPs with Unawareness in Robotics** +We formalize decision-making problems in robotics and automated control using continuous MDPs and actions that take place over continuous time intervals. We then approximate the continuous MDP using finer and finer discretizations. Doing this results in a family of systems, each of which has an extremely large action space, although only a few actions are “interesting”. This can be modeled using MDPUs— MDPs with unawareness. where the action space is much smaller. As we show, MDPUs can be used as a general framework for learning tasks in robotic problems. We prove results on the difficulty of learning a near-optimal policy in an an MDPU for a continuous task. We apply these ideas to the problem of having a humanoid robot learn on its own how to walk.*Nan Rong, Cornell University; Joseph Halpern, Cornell University; Ashutosh Saxena, Cornell University* |
|  |  |
| **ID: 305**  | **On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection** +We study the identification and estimation of functional causal models under selection bias, with a focus on th situation where the selection depends solely on the effect variable, which is known as outcome-dependent selection. We address two questions of identifiability: the identifiability of the causal direction between two variables in the presence of selection bias, and, given the causal direction, the identifiability of the model with outcome-dependent selection. Regarding the first, we show that in the framework of post-nonlinear causal models, once outcome-dependent selection is properly modeled, the causal direction between two variables is generically identifiable; regarding the second, we identify some mild conditions under which an additive noise causal model with outcome-dependent selection is to a large extent identifiable. We also propose two methods for estimating an additive noise model from data that are generated with outcome-dependent selection.*Kun Zhang, CMU; Jiji Zhang, ; Biwei Huang, MPI; Bernhard Schoelkopf, ; Clark Glymour, CMU* |
|  |  |
| **ID: 308**  | **Non-parametric Domain Approximation for Scalable Gibbs Sampling in MLNs** +MLNs utilize relational structures that are ubiquitous in real-world situations to represent large probabilistic graphical models compactly. However, as is now well-known, inference complexity is one of the main bottlenecks in MLNs. Recently, several approaches have been proposed that exploit approximate symmetries in the MLN to reduce inference complexity. These approaches approximate large domains containing many objects with much smaller domains of meta-objects (or cluster-centers), so that inference is considerably faster and more scalable. However, a drawback in most of these approaches is that it is typically very hard to tune the parameters (e.g., number of clusters) such that inference is both efficient and accurate. Here, we propose a novel non-parametric approach that trades-off solution quality with efficiency to automatically learn the optimal domain approximation. Further, we show how to perform Gibbs sampling effectively in a domain-approximated MLN by adapting the sampler according to the approximation. Our results on several benchmarks show that our approach is scalable, accurate and converges faster than existing methods.*Deepak Venugopal, The University of Memphis; Somdeb Sarkhel, ; Kyle Cherry,* |
|  |  |
| **ID: 319**  | **The deterministic information bottleneck** +Compression fundamentally involves a decision about what is relevant and what is not. The information bottleneck (IB) by Tishby, Pereira, and Bialek formalized this notion as an information-theoretic optimization problem and proposed an optimal tradeoff between throwing away as many bits as possible, and selectively keeping those that are most important. Here, we introduce an alternative formulation, the deterministic information bottleneck (DIB), that we argue better captures this notion of compression. As suggested by its name, the solution to the DIB problem is a deterministic encoder, as opposed to the stochastic encoder that is optimal under the IB. We then compare the IB and DIB on synthetic data, showing that the IB and DIB perform similarly in terms of the IB cost function, but that the DIB vastly outperforms the IB in terms of the DIB cost function. Moreover, the DIB offered a 1-2 order of magnitude speedup over the IB in our experiments. Our derivation of the DIB also offers a method for continuously interpolating between the soft clustering of the IB and the hard clustering of the DIB.*DJ Strouse, Princeton University; david Schwab, Northwestern University* |