Report from Dagstuhl Seminar 15361

Mathematical and Computational Foundations of Learning Theory

Edited by

Matthias Hein[1], Gabor Lugosi[2], and Lorenzo Rosasco[3]

1 **Universität des Saarlandes, DE,** hein@cs.uni-saarland.de
2 **UPF – Barcelona, ES,** gabor.lugosi@gmail.com
3 **MIT – Cambridge, US,** lrosasco@mit.edu

Abstract

Machine learning has become a core field in computer science. Over the last decade the statistical machine learning approach has been successfully applied in many areas such as bioinformatics, computer vision, robotics and information retrieval. The main reasons for the success of machine learning are its strong theoretical foundations and its multidisciplinary approach integrating aspects of computer science, applied mathematics, and statistics among others. The goal of the seminar was to bring together again experts from computer science, mathematics and statistics to discuss the state of the art in machine learning and identify and formulate the key challenges in learning which have to be addressed in the future. The main topics of this seminar were:

- Interplay between Optimization and Learning,
- Learning Data Representations.

Seminar August 30 to September 4, 2015 – http://www.dagstuhl.de/15361
1998 ACM Subject Classification G.1.2 Approximation, G.1.6 Optimization, G.3 Probability and Statistics, I.5 Pattern Recognition, I.2.6 Learning
Keywords and phrases learning theory, non-smooth optimization (convex and non-convex), signal processing
Digital Object Identifier 10.4230/DagRep.5.8.54

1 Executive Summary

Matthias Hein
Gabor Lugosi
Lorenzo Rosasco

Machine learning is nowadays a central field in computer science. Over the last decade the statistical learning approach has been successfully applied in many areas such as bioinformatics, computer vision, robotics and information retrieval. We believe that the main reasons for the success of machine learning are its strong theoretical foundations and its multidisciplinary approach integrating aspects of computer science, applied mathematics, and statistics among others.

Two very successful conferences titled "Mathematical Foundations of Learning Theory" in Barcelona 2004 and Paris 2006 have been inspired by this point of view on the foundations of machine learning. In 2011 the Dagstuhl seminar "Mathematical and Computational Foundations of Learning Theory" has been organized in the same spirit, bringing together

Mathematical and Computational Foundations of Learning Theory, *Dagstuhl Reports*, Vol. 5, Issue 8, pp. 54–73
Editors: Matthias Hein, Gabor Lugosi, and Lorenzo Rosasco
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

leading researchers from computer science and mathematics to discuss the state of the art and future challenges in machine learning. The 2011 Dagstuhl seminar has been the first to cover a wide range of facets of modern learning theory and has been unanimously considered a success by the participants. Since 2011 new challenges have emerged largely motivated by the availability of data-sets of unprecedented size and complexity. It is now common in many applied domains of science and technology to have datasets with thousands and even millions data-points, features and attributes/categories. For example ImageNet (http://image-net.org) is a computer vision database for object recognition including one million images of one thousands different objects, and image representations are often of the order of several tens of thousands features. Datasets of analogous complexity are customary in biology and information science (e.g. text classification). The need of analyzing and extracting information from this kind of data has posed a host of new challenges and open questions.

The second Dagstuhl seminar on "Mathematical and Computational Foundations of Learning Theory" covered broadly recent developments in the area of learning. The main focus was on two topics:

- **Interplay between Optimization and Learning**
  While statistical modeling and computational aspects have for a long time been considered separate steps in the design of learning algorithms, dealing effectively with big data requires developing new strategies where statistical and computational complexities are taken simultaneously into account. In other words, the trade-off between optimization error and generalization error has to be exploited. On the other hand it has very recently been noticed that several non-convex NP-hard learning problems (sparse recovery, compressed sensing, dictionary learning, matrix factorization etc.) can be solved efficiently and optimally (in a global sense) under conditions on the data resp. the chosen model or under the use of additional constraints.
- **Learning Data Representations**
  Data representation (e.g. the choice of kernels or features) is widely acknowledged to be the crucial step in solving learning problems. Provided with a suitable data representation, and enough labeled data, supervised algorithms, such as Support Vector Machines or Boosting, can provide good generalization performance. While data representations are often designed ad hoc for specific problems, availability of large/huge amount of unlabeled data have recently motivated the development of data driven techniques, e.g. dictionary learning, to adaptively solve the problem. Indeed, although novel tools for efficient data labeling have been developed (e.g. Amazon Mechanical Turk– http://mturk.com) most available data are unlabeled and reducing the amount of (human) supervision needed to effectively solve a task remains an important open challenge. While up-to-now the theory of supervised learning has become a mature field, an analogous theory of unsupervised and semi-supervised learning of data representation is still in its infancy and progress in the field is often assessed on a purely empirical basis.

The seminar featured a series of talks on both topics with interesting and exciting new results which lead to insights in both areas as well as a lot of discussion and interaction between the participants which for sure will manifest in several follow-up papers. Also it became obvious during the seminar that there are close connections between these two topics. Apart from these two main topics several other aspects of learning theory were discussed, leading to a quite complete picture on the current state-of-the-art in the field.

Acknowledgements. We would like to thank Dagmar Glaser and the staff at Schloss Dagstuhl for their continuous support and great hospitality which was the basis for the success of this seminar.

2 Table of Contents

3 Overview of Talks

3.1 Convex Risks, Calibrated Surrogates, Consistency, and Their Relationship with Nonparametric Estimation

Shivani Agarwal (Indian Institute of Science – Bangalore, IN)

Joint work of Agarwal, Shivani; Ramaswamy, Harish G.
Main reference H. G. Ramaswamy, S. Agarwal, "Convex calibration dimension for multiclass loss matrices," to appear in the Journal of Machine Learning Research; pre-print available as arXiv:1408.2764v2 [cs.LG], 2015.
URL http://arxiv.org/abs/1408.2764v2

In the theoretical analysis of supervised learning, the notions of PAC learning and universally Bayes consistent learning are often treated separately. We argue that classical PAC learning can essentially be viewed as a form of parametric estimation, while universally Bayes consistent learning can be viewed as a form of nonparametric estimation. A popular framework for achieving universal Bayes consistency is to minimize a (convex) calibrated surrogate risk; this is well understood for binary classification and a few selected multiclass problems, but a general understanding has remained elusive. We discuss our recent work on developing a unified framework for designing convex calibrated surrogates for general multiclass learning problems. In particular, we introduce the notion of 'convex calibration dimension' of a general multiclass loss matrix, which is the smallest number of dimensions in which one can define a convex calibrated surrogate, and give a general recipe for designing low-dimensional convex calibrated surrogates for learning problems with low-rank loss matrices. We also discuss connections between calibrated surrogates and property elicitation. In particular, we show how calibrated surrogates in supervised learning can essentially be viewed as strictly proper scoring rules for estimating certain useful properties of the conditional label distribution. These results help to shed light on how to design universally Bayes consistent algorithms for general multiclass problems, while also pointing to many open directions.

References

1 Harish G. Ramaswamy and Shivani Agarwal. *Convex calibration dimension for multiclass loss matrices*. Journal of Machine Learning Research, 2015. To appear.

2 Harish G. Ramaswamy, Shivani Agarwal and Ambuj Tewari. *Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses*. NIPS 2013.

3 Arpit Agarwal and Shivani Agarwal. *On consistent surrogate risk minimization and property elicitation*. COLT 2015.

3.2 Dictionary learning using tensor methods

Animashree Anandkumar (University of California – Irvine, US)

URL http://newport.eecs.uci.edu/anandkumar/#publications

The dictionary learning problem posits that the input data is a combination of unknown dictionary elements. Traditional methods are based on alternating minimization between the dictionary elements and coefficients. We present alternative methods based on tensor decomposition which recover the dictionary elements. These methods can consistently recover the dictionary elements when the coefficients are independent or sufficiently uncorrelated. We also present recent extensions to the convolutional setting, where shift invariance constraints are imposed.

3.3 Optimal online prediction with quadratic loss

Peter L. Bartlett (University of California – Berkeley, US)

Joint work of Bartlett, Peter L.; Koolen, Wouter M.; Malek, Alan; Takimoto, Eiji; Warmuth, Manfred K.
Main reference P. L. Bartlett, W. M. Koolen, A. Malek, E. Takimoto, M. K. Warmuth, "Minimax Fixed-Design Linear Regression," in Proc. of the 28th Conf. on Learning Theory (COLT'15), JMLR Proceedings, Vol. 40, pp. 226–239, 2015.
URL http://jmlr.org/proceedings/papers/v40/Bartlett15.html

We consider a linear regression game in which the covariates are known in advance: at each round, the learner predicts a real value, the adversary reveals a label, and the learner incurs a squared error loss. The aim is to minimize the difference between the cumulative loss and that of the linear predictor that is best in hindsight. For a variety of constraints on the adversary's labels, we obtain an explicit expression for the minimax regret and we show that the minimax optimal strategy is linear, with a parameter choice that is reminiscent of ordinary least squares. This strategy is easy to compute and does not require knowledge of the constraint set.

We also consider the case of adversarial design, and exhibit constraint sets of covariate sequences for which the same strategy is minimax optimal.

3.4 Learning to cluster – a statistical framework for incorporating domain knowledge in clustering.

Shai Ben-David (University of Waterloo, CA)

Joint work of Ben-David, Shai; Ashtiani, Hassan
Main reference H. Ashtiani, S. Ben-David, "Representation Learning for Clustering: A Statistical Framework," in Proc. of the 31st Conf. on Uncertainty in Artificial Intelligence (UAI'15), paper ID 305, 10 pages, 2015; pre-print available as arXiv:1506.05900v1 [stat.ML], 2015.
URL http://auai.org/uai2015/proceedings/papers/305.pdf
URL http://arxiv.org/abs/1506.05900v1

Clustering is an area of huge practical relevance but rather meager theoretical foundations. The multitude of clustering algorithms (and their possible parameter settings) and the diversity of the results they may yield, call for incorporation of domain expertise in the process of selecting a clustering algorithm and setting up its parameters. I outlined recent progress made along this direction. In particular, I described a novel statistical/machine-learning approach to that challenge; a model selection algorithm that is based on interactions with the clustering user. I analyzed the statistical complexity of the proposed approach. I also mentioned some common misconceptions and potential pitfalls, aiming to stimulate discussions and highlight open questions.

References

1 Hassan Ashtiani and Shai Ben-David. *Representation Learning for Clustering: A Statistical Framework*. Proceedings of UAI 2015 and CoRR abs/1506.05900, 2015.

3.5 Is adaptive early stopping possible in statistical inverse problems?

Gilles Blanchard (Universität Potsdam, DE)

We consider a standard (mathematically idealized) setting of statistical inverse problems, taking the form of the "Gaussian sequence model" $Y_i = \lambda_i \mu_i + \varepsilon_i$, $i = 1, \ldots, D$, the random noise variables ε_i are i.i.d. Gaussian with (known) variance σ^2 , the coefficients λ_i are known, and the goal is to recover as well as possible (in the sense of squared risk) the "signal sequence" $(\mu_i)_{1 \le i \le D}$.

Consider the simple family of "keep or kill" estimators depending on a cutoff index k_0, that is, the corresponding estimate sequence $(\hat{\mu}_i^{(k_0)})_{1 \le i \le D}$ is simply equal to $\lambda_i^{-1} Y_i$ for $i < k_0$ and 0 for $k_0 \le i \le D$. The question of *adaptivity* is the following: is it possible to choose $\hat{k}_0$ from the data only, in such a way that the performance obtained is comparable (whithin a multiplicative constant) to the best possible deterministic, a priori choice of k_0 minimising the average squared risk (usually called "oracle", since it depends on the unknown signal)?

There exist a number of well-known methods achieving oracle adaptivity, such as penalization or Lepski's method. However, they have in common that the estimators for *all* values of k_0 have to be computed first and compared to each other in some way. Contrast this to an "early stopping" approach where we would like to compute iteratively the estimators for $k_0 = 1, 2, \ldots$ and have to decide to stop at some point $\hat{k}_0$ without being allowed to compute the other estimators. Is oracle adaptivity possible then? This question is motivated by settings where computing estimators for larger k_0 requires more computational cost; furthermore some form of early stopping is most often used in practice.

After careful mathematical formalization of the problem, our first result is that, if one must base the early stopping decision at index k_0 on the sole information of $Y_i, i \le k_0$, then adaptive early stopping is not possible in general. A more realistic scenario is when we are additionally allowed to use the information of the *residual* $\sum_{i=k_0+1}^{D} Y_i^2$ to decide to stop at k_0 (or not). In that case, partial oracle adaptation is possible, essentially when the oracle stopping time k_0^* is larger in order than $\sqrt{D}$ (remember D is the maximum considered dimension). This adaptive stopping can be achieved by a simple "discrepancy principle" commanding to stop when the residual becomes smaller than $D\sigma^2$, a type of rule which is often used in practice. We establish lower and upper bounds, in particular showing that if the oracle k_0^* is of order strictly smaller than $\sqrt{D}$, oracle adaptation is *not* possible in general.

3.6 Adaptive tail index estimation

Stéphane Boucheron (Paris Diderot University, FR)

Joint work of Stéphane Boucheron, Maud Thomas
Main reference S. Boucheron, M. Thomas, "Tail index estimation, concentration and adaptivity," arXiv:1503.05077v3 [math.ST], 2015.
URL http://arxiv.org/abs/1503.05077v3

Assume data $X_1, \ldots, X_n$ are collected from a univariate distribution F and we want to estimate $\overline{F}(x) = 1 - F(x)$ where $x > \overline{F}(\max(X_1, \ldots, X_n)$ or estimate a quantile of order $1 - 1/t$ for $t > n$. In order the face this challenge with a reasonable of possibility of success, a tail regularity assumption is necessary. In the so-called heavy tail domains, this assumption reads as: for all $x > 0$, $\lim_{t\to\infty} \overline{F}(tx)/\overline{F}(t) = x^{-1/\gamma}$ for some $\gamma > 0$ which is called the tail (or extreme value) index. In words, $\overline{F}$ is assumed to be regularly varying with index $-1/\gamma$. Estimating γ from a sample is called the tail index estimation problem (see [4] for a presentation of Extreme Value Theory). Many tail index estimators (Hill, Pickands, Moments, ...) consist of computing statistics from the k largest order statistics. Practitioners face an estimator selection problem: picking k so as to achieve a good trade-off between variance (large values of k) and bias (small values of k). We present an adaptive version of the Hill estimator based on Lespki's model selection method (which has been used in learning theory in order to achieve adaptivity in classification, see [2, 3]). This simple data-driven index selection method is shown to satisfy an kind of oracle inequality and is checked to achieve the lower risk bound recently derived by [1]. In order to establish the (pseudo)-oracle inequality, we derive non-asymptotic variance bounds and concentration inequalities for Hill estimators. These concentration inequalities are derived from Talagrand's concentration inequality for smooth functions of independent exponentially distributed random variables combined with three tools of Extreme Value Theory: the quantile transform, Karamata's representation of slowly varying functions, and Rényi's characterisation for the order statistics of exponential samples.

References

1 Carpentier, Alexandra and Kim, Arlene K.H. *Adaptive and minimax optimal estimation of the tail coefficient*. arXiv:1309.2585v1, 2013.

2 Tsybakov, Alexandre B. *Optimal aggregation of classifiers in statistical learning*. Annals of Statistics 32, 135–166, 2004.

3 Boucheron, Stéphane and Bousquet, Olivier and Lugosi, Gábor. *Theory of Classification: a Survey of Some Recent Advances*. ESAIM: Probability and Statistics 9, 329–375, 2005.

4 Beirlant, Jan and Goegebeur, Yuri and Segers, Johan and Teugels, Jozef. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

3.7 Multi-scale exploration of convex functions and bandit convex optimization

Sébastien Bubeck (Microsoft Research – Redmond, US)

Joint work of Bubeck, Sébastien; Eldan, Ronen
Main reference S. Bubeck, R. Eldan, "Multi-scale exploration of convex functions and bandit convex optimization," arXiv:1507.06580v1 [math.MG], 2015.
URL http://arxiv.org/abs/1507.06580v1

We construct a new map from a convex function to a distribution on its domain, with the property that this distribution is a multi-scale exploration of the function. We use this map to solve a decade-old open problem in adversarial bandit convex optimization by showing that the minimax regret for this problem is $\bar{O}(poly(n)\sqrt{T})$, where n is the dimension and T the number of rounds. This bound is obtained by studying the dual Bayesian maximin regret via the information ratio analysis of Russo and Van Roy, and then using the multi-scale exploration to solve the Bayesian problem.

3.8 Information theory of algorithms

Joachim M. Buhmann (ETH Zürich, CH)

Main reference J. M. Buhmann, "SIMBAD: Emergence of Pattern Similarity, Similarity-Based Pattern Analysis and Recognition," in M. Pelillo (ed.), "Similarity-Based Pattern Analysis and Recognition – Part I", pp. 45–64, Springer, 2013.
URL http://dx.doi.org/10.1007/978-1-4471-5628-4_3

Algorithms map input spaces to output spaces where inputs are possibly affected by fluctuations. Beside run time and memory consumption, an algorithm might be characterized by its sensitivity to the signal in the input and its robustness to input fluctuations. The achievable precision of an algorithm, i.e., the attainable resolution in output space, is determined by its capability to extract predictive information in the input relative to its output. I will present an information theoretic framework for algorithm analysis where an algorithm is characterized as computational evolution of a (possibly contracting) posterior distribution over the output space. The tradeoff between precision and stability is controlled by an input sensitive generalization capacity (GC). GC measures how much the posteriors on two different problem instances agree despite the noise in the input. Thereby, GC objectively ranks different algorithms for the same data processing task based on the bit rate of their respective capacities. Information theoretic algorithm selection is demonstrated for minimum spanning tree algorithms and for greedy MaxCut algorithms. The method can rank centroid based and spectral clustering methods, e.g. k-means, pairwise clustering, normalized cut, adaptive ratio cut and dominant set clustering.

3.9 Fast algorithms and (other) minimax optimal algorithms for mixed regression

Constantine Caramanis (Univ. of Texas at Austin, US)

Joint work of Caramanis, Constantine; Chen, Yudong; Yi, Xinyang
Main reference Y. Chen, X. Yi, C. Caramanis, "A Convex Formulation for Mixed Regression with Two Components: Minimax Optimal Rates," in Proc. of the 27th Conf. on Learning Theory (COLT'14), JMLR Proceedings, Vol. 35, pp. 560–604, 2014; pre-print available as arXiv:1312.7006v2 [stat.ML], 2015.
URL http://jmlr.org/proceedings/papers/v35/chen14.html
URL http://arxiv.org/abs/1312.7006v2

Mixture models represent the superposition of statistical processes, and are natural in machine learning and statistics. In mixed regression, the relationship between input and output is given by one of possibly several different (noisy) linear functions. Thus the solution encodes a combinatorial selection problem, and hence computing it is difficult in the worst case. Even in the average case, little is known in the realm of efficient algorithms with strong statistical guarantees.

We give general conditions for linear convergence of an EM-like (and hence fast) algorithm for latent-variable problems in high dimensions, and show this implies that for sparse (or low-rank) mixed regression, EM converges linearly, in a neighborhood of the optimal solution, in the high-SNR regime. For the low-SNR regime, we show that a new behavior emerges. Here, we give a convex optimization formulation that provably recovers the true solution, and we provide upper bounds on the recovery errors for both arbitrary noise and stochastic noise settings. We also give matching minimax lower bounds, showing that our algorithm is information-theoretically optimal.

Our results represent what is, as far as we know, the only tractable algorithm guaranteeing successful recovery with tight bounds on recovery errors and sample complexity.

References

1 Yudong Chen, Xinyang Yi, Constantine Caramanis. *A Convex Formulation for Mixed Regression with Two Components: Minimax Optimal Rates*. JMLR W&CP, 35:560–604, 2014

2 Xinyang Yi, Constantine Caramanis. *Regularized EM Algorithms: A Unified Framework and Statistical Guarantees*. To appear at NIPS 2015

3.10 Sparse and spurious: dictionary learning with noise and outliers

Rémi Gribonval (INRIA Rennes – Bretagne Atlantique, FR)

Joint work of Gribonval, Rémi; Jenatton, Rodolphe; Bach, Francis; Kleinsteuber, Martin; Seibert, Matthias;

Main reference R. Gribonval, R. Jenatton, F. R. Bach, "Sparse and Spurious: Dictionary Learning With Noise and Outliers," IEEE Transactions on Information Theory, 61(11):6298–6319, 2015: pre-print available as arXiv:1407.5155v4 [cs.LG], 2015.

URL http://dx.doi.org/10.1109/TIT.2015.2472522

URL http://arxiv.org/abs/1407.5155v4

In this talk I draw a panorama of dictionary learning for low-dimensional modeling. After reviewing the basic empirical principles of dictionary learning and related matrix factorizations such as PCA, K-means and NMF, I discuss techniques to learn dictionaries with controlled computational efficiency, as well as a series of recent theoretical results establishing the statistical significance of learned dictionaries even in the presence of noise and outliers.

References

1 Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, Matthias Seibert. *Sample Complexity of Dictionary Learning and other Matrix Factorizations*. IEEE Transactions on Information Theory, 2015

2 Rémi Gribonval, Rodolphe Jenatton, Francis Bach. *Sparse and spurious: dictionary learning with noise and outliers*. IEEE Transactions on Information Theory, 2015

3.11 Empirical portfolio selections and a problem on aggregation

László Györfi (Budapest University of Technology & Economics, HU)

Main reference L. Györfi, Gy. Ottucsák, A. Urbán, "Empirical log-optimal portfolio selections: a survey," in L. Györfi, G. Ottucsák, H. Walk (eds.), "Machine Learning for Financial Engineering," pp. 79–116, Imperial College Press, 2012.

URL http://www.worldscientific.com/worldscibooks/10.1142/p818

This talk provides a survey of discrete time, multi period, equential investment strategies for financial markets. Under memoryless assumption on the underlying process generating the asset prices the Best Constantly Rebalanced Portfolio is studied, called log-optimal portfolio, which achieves the maximal asymptotic average growth rate. For generalized dynamic portfolio selection, when asset prices are generated by a stationary and ergodic process, growth optimal empirical strategies are shown, where some principles of nonparametric regression estimation and of machine learning aggregation are applied. The empirical performance of the methods is illustrated for NYSE data. An open problem is presented, too, which means that the consistency has been proved if the learning parameter for the aggregation is between 0 and 1, while the empirical results are better if the learning parameter is larger than 1. The problem is to extend the consistency to this case.

3.12 Train faster, generalize better: Stability of stochastic gradient descent

Moritz Hardt (Google Research – Mountain View, US)

We show that any model trained by a stochastic gradient method with few iterations has vanishing generalization error. We prove this by showing the method is algorithmically stable in the sense of Bousquet and Elisseeff. Our analysis only employs elementary tools from convex and continuous optimization. Our results apply to both convex and non-convex optimization under standard Lipschitz and smoothness assumptions.

Applying our results to the convex case, we provide new explanations for why multiple epochs of stochastic gradient descent generalize well in practice. In the nonconvex case, we provide a new interpretation of common practices in neural networks, and provide a formal rationale for stability-promoting mechanisms in training large, deep models. Conceptually, our findings underscore the importance of reducing training time beyond its obvious benefit.

3.13 Robust Regression via Hard Thresholding

Prateek Jain (Microsoft Research India – Bangalore, IN)

Joint work of Jain, Prateek; Bhatia Kush; Kar Purushottam
Main reference K. Bhatia, P. Jain, P. Kar, "Robust Regression via Hard Thresholding," to appear in Proc. of the 29th Annual Conf. on Neural Information Processing Systems (NIPS'15): pre-print available as arXiv:1506.02428v1 [cs.LG], 2105.
URL http://arxiv.org/abs/1506.02428v1

In this talk, we will discuss the problem of Robust Least Squares Regression (RLSR) where several response variables can be adversarially corrupted. More specifically, for a data matrix $X \in R^{p \times n}$ and an underlying model w^*, the response vector is generated as $y = X'w^* + b$ where $b \in R^n$ is the corruption vector supported over at most $C\,n$ coordinates. Existing exact recovery results for RLSR focus solely on L1-penalty based convex formulations and impose relatively strict model assumptions such as requiring the corruptions b to be selected independently of X. In this talk, we will focus on a simple hard-thresholding algorithm that we call TORRENT which, under mild conditions on X, can recover w^* exactly even if b corrupts the response variables in an adversarial manner, i.e. both the support and entries of b are selected adversarially after observing X and w^*. We will also discuss certain extensions of TORRENT that can scale efficiently to large scale problems, such as high dimensional sparse recovery. We will present empirical results that show that TORRENT, and more so its extensions, offer significantly faster recovery than the state-of-the-art L1 solvers. For instance, even on moderate-sized datasets (with $p = 50K$) with around 40% corrupted responses, a variant of our proposed method called TORRENT-HYB is more than 20x faster than the best L1 solver.

See http://arxiv.org/abs/1506.02428 for more details.

3.14 Optimizing decomposable submodular functions

Stefanie Jegelka (MIT – Cambridge, US)

Submodular functions capture a spectrum of discrete problems in machine learning, signal processing and computer vision. In these areas, practical algorithms are a major concern that motivates to exploit structure in addition to submodularity. A simple example of such a structure are functions that decompose as a sum of "simple" submodular functions. For this setting, several algorithms arise from relations between submodularity and convexity. In particular, this talk will focus on a class of algorithms that solve submodular minimization as a best approximation problem. These algorithms are easy to use and to parallelize, and solve both a convex relaxation and the original discrete problem. We observe that the algorithms work well in practice, and analyze their convergence properties.

References

1 S. Jegelka, F. Bach, S. Sra. *Reflection methods for user-friendly submodular optimization.* NIPS 2013

2 R. Nishihara, S. Jegelka, M.I. Jordan. *On the linear convergence rate of decomposable submodular function minimization.* NIPS 2014

3.15 Matrix factorization with binary components – uniqueness in a randomized model

Felix Krahmer (TU München, DE)

Joint work of Hein, Matthias; James, David; Krahmer, Felix

Motivated by an application in computational biology, we consider low-rank matrix factorization with $\{0, 1\}$-constraints on the first of the factors and optionally convex constraints on the second one. Despite apparent intractability, it has been shown by Hein et al. [1] that one can provably recover the underlying factorization, provided there exists a unique solution. We conjecture that by choosing a sparse Bernoulli random model for the binary factor, there will be a unique solution with high probability. Due to limited applicability of Littlewood-Offord inequalities, previous results do not generalize. We present partial progress for limited rank.

References

1 M. Slawski, M. Hein, and P. Lutsik, *Matrix Factorization with Binary Components.* NIPS 2013

3.16 Variational Inference in Probabilistic Submodular Models

Andreas Krause (ETH Zürich, CH)

Joint work of Djolonga, Josip; Krause, Andreas
Main reference J. Djolonga, A. Krause, "From MAP to Marginals: Variational Inference in Bayesian Submodular Models," in Proc. of the 28th Annual Conf. on Advances in Neural Information Processing Systems (NIPS'14), pp. 244–252, 2014.
URL http://papers.nips.cc/paper/5492-from-map-to-marginals-variational-inference-in-bayesian-submodular-models

As a discrete analogue of convexity, submodularity has profound implications for optimization. In recent years, submodular optimization has found many new applications, such as in machine learning and network analysis. These include active learning, dictionary learning, data summarization, influence maximization and network structure inference. In this talk, I will present our recent work on quantifying uncertainty in submodular optimization. In particular, we carry out the first systematic investigation of inference and learning in probabilistic submodular models (PSMs). These are probabilistic models defined through submodular functions – log-sub/supermodular distributions – generalizing regular binary Markov Random Fields and Determinantal Point Processes. They express natural notions such as attractiveness and repulsion and allow to capture long-range, high-order dependencies among the variables. I will present our recently discovered variational approach towards inference in general PSMs based on sub- and supergradients. We obtain both lower and upper bounds on the log- partition function, which enables computing probability intervals for marginals, conditionals and marginal likelihoods. We also obtain fully factorized approximate posteriors, at essentially the same computational cost as ordinary submodular optimization. Our framework results in convex problems for optimizing over differentials of submodular functions, which we show how to optimally solve. Our approximation is exact at the mode (for log-supermodular distributions), and we provide bounds on the approximation quality of the log-partition function with respect to the curvature of the function. We further establish natural relations between our variational approach and the classical mean-field method from statistical physics. Exploiting additive structure in the objective leads to highly scalable, parallelizable message passing algorithms. We empirically demonstrate the accuracy of our inference scheme on several PSMs arising in computer vision and network analysis.

3.17 Learning Representations from Incomplete Data

Robert D. Nowak (University of Wisconsin – Madison, US)

Joint work of Nowak, Robert D; Pimentel, Daniel; Boston, Nigel
Main reference D. L. Pimentel-Alarcón, N. Boston, R. D. Nowak, "A Characterization of Deterministic Sampling Patterns for Low-Rank Matrix Completion," arXiv:1503.02596v2 [stat.ML] , 2015.
URL http://arxiv.org/abs/1503.02596v2

Low-rank matrix completion (LRMC) problems arise in a wide variety of applications. Previous theory mainly provides conditions for completion under missing-at-random samplings. This talk presents deterministic conditions for completion. An incomplete $d \times N$ matrix is finitely rank-r completable if there are at most finitely many rank-r matrices that agree with all its observed entries. Finite completability is the tipping point in LRMC, as a few additional samples of a finitely completable matrix guarantee its unique completability.

The main contribution the talk is a characterization of finitely completable observation sets. We use this characterization to derive sufficient deterministic sampling conditions for unique completability. We also show that under uniform random sampling schemes, these conditions are satisfied with high probability if $O(\max\{r, \log d\})$ entries per column are observed. Extensions of these results to subspace clustering with missing data are also given.

Further details can be found in the following papers: arXiv:1503.02596, arXiv:1410.0633

3.18 Tight convex relaxations for sparse matrix factorization

Guillaume Obozinski (ENPC – Marne-la-Vallée, FR)

Joint work of Richard, Emile; Vert, Jean-Philippe

In this talk, I will consider statistical learning problems in which the parameter is a matrix which is the sum of a small number of sparse rank one (non-orthogonal) factors, and which can be viewed as generalizations of the sparse PCA problem with multiple factors. Based on an assumption that the sparsity of the factors is fixed and known, I will present a matrix norm which provides an tight although NP-hard convex relaxation of the learning problem. I will discuss the sample complexity of learning the matrix in the rank one case and show that considering a computationally more expensive convex relaxation leads to an improvement of the sample complexity by an order of magnitude as compared with the usual convex regularization considered, like combinations of the L1-norm and the trace norm. I will then describe an algorithm, relying on a rank-one sparse PCA oracle to solve the convex problems considered and illustrate that, in practice, when state-of-the-art heuristic algorithms for rank one sparse PCA are used as surrogates for the oracle, our algorithm outperforms other existing methods.

3.19 Active Regression

Sivan Sabato (Ben Gurion University – Beer Sheva, IL)

Joint work of Rémi Munos
Main reference S. Sabato, R. Munos, "Active Regression by Stratification," in Proc. of the 28th Annual Conf. on Advances in Neural Information Processing Systems (NIPS'14), pp. 269–477, 2014.
URL http://papers.nips.cc/paper/5468-active-regression-by-stratification

We propose a new active learning algorithm for parametric linear regression with random design. We provide finite sample convergence guarantees for general distributions in the misspecified model. This is the first active learner for this setting that provably can improve over passive learning. Unlike other learning settings (such as classification), in regression the passive learning rate of $O(1/\epsilon)$ cannot in general be improved upon. Nonetheless, the so-called 'constant' in the rate of convergence, which is characterized by a distribution- dependent *risk*, can be improved in many cases. For a given distribution, achieving the optimal risk requires prior knowledge of the distribution. Following the stratification technique advocated in Monte-Carlo function integration, our active learner approaches the optimal risk using piecewise constant approximations.

Sivan Sabato is supported by the Lynne and William Frankel Center for Computer Science.

3.20 Dictionary learning – fast and dirty

Karin Schnass (Universität Innsbruck, AT)

Main reference K. Schnass, "Convergence radius and sample complexity of ITKM algorithms for dictionary learning," arXiv:1503.07027v2 [cs.LG], 2015.
URL http://arxiv.org/abs/1503.07027v2

In this talk we give a short introduction to fast dictionary learning algorithms with local convergence guarantees. Using the classic optimization principle underlying K-SVD as starting point we motivate a response maximization principle and and the associated algorithm ITKM (Iterative Thresholding and K Means). We then progress to a variant using residual means ITKrM (Iterative Thresholding and K residual Means), which can be seen as hybrid between K-SVD and ITKrM and as such inherits the best of both worlds. Experimental global convergence from K-SVD, and computational efficiency, sequentiality/parallelizability and local convergence guarantees under low sample complexity from ITKM.

3.21 Variational approach to consistency of clustering of point clouds

Dejan Slepcev (Carnegie Mellon University, US)

Joint work of Garcia Trillos, Nicolas; Laurent, Thomas; von Brecht, James; Bresson, Xavier; Slepcev, Dejan
Main reference N. Garcia Trillos, D. Slepcev, J. van Brecht, T. Laurent, X. Bresson, "Consistency of Cheeger and Ratio Graph Cuts," arXiv:1411.6590v1 [stat.ML], 2014.
URL http://arxiv.org/abs/1411.6590v1

The talk discussed variational problems arising in machine learning and their consistency as the number of data points goes to infinity. Consider point clouds obtained as random samples of an underlying "ground-truth" measure on a Euclidean domain. Graph representing the point cloud is obtained by assigning weights to edges based on the distance between the points. We discussed approaches to clustering based on minimizing objective functionals defined on these graphs. We focused is on functionals based on graph cuts like the Cheeger and ratio cuts. We showed that minimizers of the these cuts converge as the sample size increases to a minimizer of a corresponding continuum cut (which partitions the ground truth measure). A setup based on Gamma-convergence and optimal transportation to study such questions was introduced. Sharp conditions on how the connectivity radius can be scaled with respect to the number of sample points for the consistency to hold were obtained.

3.22 Optimization, Regularization and Generalization in Multilayer Networks

Nathan Srebro (TTIC – Chicago, US)

Joint work of Srebro, Nathan; Neyshabur, Behnam; Tomioka, Ryota; Salakhutdinov, Russ

What is it that enables learning with multi-layer networks? What causes the network to generalize well? What makes it possible to optimize the error, despite the problem being hard in the worst case? In this talk I will attempt to address these questions and relate between them, highlighting the important role of optimization in deep learning. I will then use the insight to suggest studying novel optimization methods, and will present Path-SGD, a novel optimization approach for multi-layer RELU networks that yields better optimization and better generalization.

3.23 Oracle inequalities for network models and sparse graphon estimation

Alexandre Tsybakov (UPMC – Paris, FR)

Joint work of Klopp, Olga; Tsybakov, Alexandre B.; Verzelen, Nicolas
Main reference O. Klopp, A. B. Tsybakov, N. Verzelen, "Oracle inequalities for network models and sparse graphon estimation," arXiv:1507.04118v1 [math.ST], 2015.
URL http://arxiv.org/abs/1507.04118v1

Inhomogeneous random graph models encompass many network models such as stochastic block models and latent position models. In this paper, we study two estimators: the ordinary block constant least squares estimator, and its restricted version. We show that they satisfy oracle inequalities with respect to the block constant oracle. As a consequence, we derive optimal rates of estimation of the probability matrix. Our results cover the important setting of sparse networks. Nonparametric rates for graphon estimation in the L_2 norm are also derived when the probability matrix is sampled according to a graphon model. The results shed light on the differences between estimation under the empirical loss (the probability matrix estimation) and under the integrated loss (the graphon estimation).

3.24 Learning Economic Parameters from Revealed Preferences

Ruth Urner (MPI für Intelligente Systeme – Tübingen, DE)

Joint work of Balcan, Maria-Florina; Daniely, Amit; Mehta, Ruta; Urner, Ruth; Vazirani, Vijay V.
Main reference M.-F. Balcan, A. Daniely, R. Mehta, R. Urner, V. V. Vazirani, "Learning Economic Parameters from Revealed Preferences," in Proc. of the 10th Int'l Conf. on Web and Internet Economics (WINE'14), LNCS, Vol. 8877, pp. 338–353, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-13129-0_28

A recent line of work, starting with Beigman and Vohra and Zadimoghaddam and Roth, has addressed the problem of learning a utility function from revealed preference data. The goal here is to make use of past data describing the purchases of a utility maximizing agent when faced with certain prices and budget constraints in order to produce a hypothesis function that can accurately forecast the future behavior of the agent.

In this work we advance this line of work by providing sample complexity guarantees and efficient algorithms for a number of important classes. By drawing a connection to recent advances in multi-class learning, we provide a computationally efficient algorithm with tight sample complexity guarantees ($\Theta(d/\epsilon)$ for the case of d goods) for learning linear utility functions under a linear price model. This solves an open question in Zadimoghaddam and Roth. Our technique yields numerous generalizations including the ability to learn other well-studied classes of utility functions, to deal with a misspecified model, and with non-linear prices.

References

1 Maria-Florina Balcan, Amit Daniely, Ruta Mehta, Ruth Urner and Vijay V. Vazirani. *Learning Economic Parameters from Revealed Preferences*. Web and Internet Economics – 10th International Conference (WINE) 2014, Beijing, China, December 14–17, 2014. Proceedings.

3.25 Stochastic Forward-Backward Splitting

Silvia Villa (Italian Institute of Technology – Genova, IT)

Joint work of Rosasco, Lorenzo; Vu, Cong Bang; Villa, Silvia
Main reference L. Rosasco, S. Villa, B. C. Vu, "Convergence of stochastic proximal gradient algorithm," arXiv:1403.5074v3 [math.OC], 2014.
URL http://arxiv.org/abs/1403.5074v3

I analyzed the convergence of a novel stochastic forward-backward splitting algorithm for solving monotone inclusions given by the sum of a maximal monotone operator and a single-valued maximal monotone cocoercive operator. This latter framework has a number of interesting special cases, including variational inequalities and convex minimization problems, while stochastic approaches are practically relevant to account for perturbations in the data. The algorithm I discussed is a stochastic extension of the classical deterministic forward-backward method, and is obtained considering the composition of the resolvent of the maximal monotone operator with a forward step based on a stochastic estimate of the single-valued operator.

The talk was based on the following papers:

References

1 L. Rosasco, S. Villa, and B. C. Vu. *Convergence of stochastic proximal gradient algorithm*. arxiv 1403.5074

2 L. Rosasco, S. Villa, and B. C. Vu. *Stochastic forward-backward splitting for monotone inclusions*. arxiv 1403.7999

3 L. Rosasco, S. Villa, and B. C. Vu. *A stochastic inertial forward-backward splitting algorithm for multivariate monotone inclusions*. arXiv:1507.00848

3.26 Finding global k-means clustering solutions

Rachel Ward (University of Texas – Austin, US)

K-means clustering aims to partition a set of n points into k clusters in such a way that each observation belongs to the cluster with the nearest mean, and such that the sum of squared distances from each point to its nearest mean is minimal. In general, this is a hard optimization problem, requiring an exhaustive search over all possible partitions of the data into k clusters in order to find the optimal clustering. At the same time, fast heuristic algorithms for the k-means optimization problem are often applied in many data processing applications, despite having few guarantees on the clusters they produce. In this talk, we will introduce a semidefinite programming relaxation of the k-means optimization problem, along with geometric conditions on a set of data such that the algorithm is guaranteed to find the optimal k-means clustering for the data. For points drawn randomly within separated balls, the important quantities are the distances between the centers of the balls compared to the relative densities of points within them, and at sufficient density, the SDP relaxation is guaranteed to resolve such clusters at arbitrarily small separation distance. We will also discuss certain convex relaxations and recovery guarantees for another geometric clustering objective, k-median clustering. We will conclude by discussing several open questions related to this work.

3.27 Symmetric and Asymmetric k-Center Clustering under Stability

Colin White (Carnegie Mellon University, US)

Joint work of Balcan, Maria-Florina; Haghtalab, Nika; White, Colin
Main reference M.-F. Balcan, N. Haghtalab, C. White, "Symmetric and Asymmetric k-center Clustering under Stability," arXiv:1505.03924v2 [cs.DS], 2015.
URL http://arxiv.org/abs/1505.03924v2

In this work, we take a beyond the worst case approach to asymmetric and symmetric k-center problems under two very natural input stability (promise) conditions. We consider both the α-perturbation resilience notion of Bilu and Linial [BL12], which states that the optimal solution does not change under any α-factor perturbation to the input distances, and the (α,ϵ)-approximation stability notion of Balcan et al. [BBG09], which states that any α-approximation to the cost of the optimal solution should be ϵ-close in the solution space (i.e., the partitioning) to the optimal solution. We show that by merely assuming 3-perturbation resilience or $(2, 0)$-approximation stability, the exact solution for the asymmetric k-center problem can be found in polynomial time. To our knowledge, this is the first problem that is hard to approximate to any constant factor in the worst case, yet can be optimally solved in polynomial time under perturbation resilience for a constant value of α. In the case of 2-approximation stability, we prove our result is tight by showing k-center under (2-ϵ)-approximation stability is hard unless $NP = RP$. For the case of symmetric k-center, we give an efficient algorithm to cluster 2-perturbation resilient instances. Our results illustrate a surprising relation between symmetric and asymmetric k-center instances under these stability conditions. Unlike approximation ratio, for which symmetric k-center is easily solved to a factor of 2 but asymmetric k-center cannot be approximated to any constant factor, both symmetric and asymmetric k-center can be solved optimally under resilience to small constant-factor perturbations.

3.28 A Dynamic Approach to Variable Selection and Sparse Recovery: Differential Inclusions with Early Stopping

Yuan Yao (Peking University, CN)

Sparse signal recovery from linear noisy measurements has been a classical topic in compressed sensing and high dimensional statistics. There has been a large volume of literature around l1-regularization or LASSO approach and it is well-known that the convex relaxation in LASSO leads to biased solutions. So in practice, people compute LASSO regularization paths for model selection, followed by a subset least square to remove the bias. Here we discuss an alternative approach to sparse recovery via differential equations with inclusion constraints, which we call Bregman ISS (Inverse Scale Space) or Linearized Bregman ISS. We shall see that the new approach has great advantages over LASSO in its algorithmic simplicity and estimate quality. Its dynamics naturally induces a solution path for regularization and the points on the paths can be unbiased or less biased than LASSO. We show that under nearly the same conditions for LASSO's sign consistency, there exists a bias-free and sign-consistent point on the solution paths, where early stopping is crucial for regularization.

3.29 Minimum Error Entropy and Related Problems

Ding-Xuan Zhou (City University – Hong Kong, HK)

Minimum error entropy principle has been widely used in the community of signal processing and is closely related to kernel methods in learning theory. Its idea is to seek as much information as possible from data by minimizing various entropies of the error random variable. A minimum error entropy method takes moments of all orders into consideration and may perform well in dealing with heavy-tailed noise. Compared with its practical developments within the last decade, its rigorous theoretical consistency analysis is unknown. This talk demonstrates some rigorous consistency analysis of the minimum error entropy principle in the framework of regression. Some new methods arise from the study and might be used for investigating other related problems: Fourier analysis of the generalization error associated with pairwise loss functions, minimax rates of convergence achieved by the least squares regularization scheme, and the choice of step sizes for online or gradient descent algorithms.

Participants

- Shivani Agarwal
  Indian Institute of Science – Bangalore, IN
- Animashree Anandkumar
  Univ. of California – Irvine, US
- Peter L. Bartlett
  University of California – Berkeley, US
- Shai Ben-David
  University of Waterloo, CA
- Gilles Blanchard
  Universität Potsdam, DE
- Stephane Boucheron
  Paris Diderot University, FR
- Sebastien Bubeck
  Microsoft Res. – Redmond, US
- Joachim M. Buhmann
  ETH Zürich, CH
- Constantine Caramanis
  University of Texas at Austin, US
- Sou-Cheng Choi
  NORC – Chicago, US
- Luc Devroye
  McGill Univ. – Montreal, CA
- Jack Fitzsimons
  University of Oxford, GB
- Antoine Gautier
  Universität des Saarlandes, DE
- Remi Gribonval
  INRIA Rennes – Bretagne Atlantique, FR
- László Györfi
  Budapest University of Technology & Economics, HU
- Moritz Hardt
  Google Research – Mountain View, US
- Matthias Hein
  Universität des Saarlandes, DE
- Prateek Jain
  Microsoft Research India – Bangalore, IN
- Stefanie Jegelka
  MIT – Cambridge, US
- Felix Krahmer
  TU München, DE
- Andreas Krause
  ETH Zürich, CH
- Lek-Heng Lim
  University of Chicago, US
- Gabor Lugosi
  UPF – Barcelona, ES
- Robert D. Nowak
  University of Wisconsin – Madison, US
- Guillaume Obozinski
  ENPC – Marne-la-Vallée, FR
- Duy Khanh Pham
  Ho Chi Minh City University of Pedagogy, VN
- Lorenzo Rosasco
  MIT – Cambridge, US
- Alessandro Rudi
  MIT – Cambridge, US
- Sivan Sabato
  Ben Gurion University – Beer Sheva, IL
- Karin Schnass
  Universität Innsbruck, AT
- Dejan Slepcev
  Carnegie Mellon University, US
- Nathan Srebro
  TTIC – Chicago, US
- Yannik Stein
  FU Berlin, DE
- Alexandre Tsybakov
  UPMC – Paris, FR
- Ruth Urner
  MPI für Intelligente Systeme – Tübingen, DE
- Silvia Villa
  Italian Institute of Technology – Genova, IT
- Rachel Ward
  University of Texas – Austin, US
- Colin White
  Carnegie Mellon University, US
- Robert C. Williamson
  Australian National Univ., AU
- Yuan Yao
  Peking University, CN
- Ding-Xuan Zhou
  City University – Hong Kong, HK