Report from Dagstuhl Seminar 11291

Mathematical and Computational Foundations of Learning Theory

Edited by

Matthias Hein[1], Gabor Lugosi[2], Lorenzo Rosasco[3], and Steve Smale[4]

1 **Universität des Saarlandes, DE,** hein@cs.uni-sb.de
2 **Univ. Pompeu Fabra – Barcelona, ES,** gabor.lugosi@gmail.com
3 **MIT – Cambridge, US, and IIT, Italy,** lrosasco@mit.edu
4 **City University – Hong Kong, CN,** smale@cityu.edu.hk

Abstract

The main goal of the seminar "Mathematical and Computational Foundations of Learning Theory" was to bring together experts from computer science, mathematics and statistics to discuss the state of the art in machine learning broadly construed and identify and formulate the key challenges in learning which have to be addressed in the future. This Dagstuhl seminar was one of the first meetings to cover the full broad range of facets of modern learning theory. The meeting was very successful and all participants agreed that such a meeting should take place on a regular basis.

Seminar 17.–22. July, 2011 – www.dagstuhl.de/11291
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, machine learning, sparsity, high-dimensional geometry, manifold learning, online learning
Digital Object Identifier 10.4230/DagRep.1.7.53

1 Executive Summary

Matthias Hein
Gabor Lugosi
Lorenzo Rosasco
Steve Smale

The study of learning is at the very core of the problem of intelligence both in humans and machines. We have witnessed an exciting success story of machine learning in recent years. Among other examples, we now have cars that detect pedestrians, and smart-phones that can be controlled simply by our voices. Indeed, aside from the increase in computational power and availability of large amount of data, the key to these successes has been the development of efficient learning algorithms based on solid theoretical foundations. As the science and engineering of learning move forward to understand and solve richer and more articulated classes of problems, broadening the mathematical and computational foundations of learning becomes essential for future achievements.

Mathematical and Computational Foundations of Learning Theory, *Dagstuhl Reports*, Vol. 1, Issue 7, pp. 53–69
Editors: Matthias Hein, Gabor Lugosi, Lorenzo Rosasco, and Steve Smale
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The main goal of our seminar was to account for the newest developments in the field of learning theory and machine learning as well as to indicate challenges for the future. This seminar was in the same spirit of two very successful conferences titled "Mathematical Foundations of Learning Theory", organized in 2004 in Barcelona and 2006 in Paris. The seminar brought together leading researchers from computer science and mathematics to discuss the state of the art in learning and generate synergy effects between the different usually disconnected communities. This Dagstuhl seminar has been the first to cover the full range of facets of modern learning theory.

The seminar has focused on three main topics, while trying to keep a broader view on all recent advances. The three main topics were: 1) the role of sparsity in learning, 2) the role of geometry in learning, and 3) sequential learning and game theory. Experts in each field gave tutorials on each topic, covering basic concepts as well as recent results.

The meeting was hold in a very informal and stimulating atmosphere. The participants all agreed that such a seminar should be come a regular meeting.

Acknowledgements. We thank Annette Beyer and Claudia Thiele for their continuous support and help in organizing this workshop. Moreover, we would like to thank the staff at Schloss Dagstuhl for making this seminar such a remarkably enjoyable event. Special thanks go to Elisabeth Chaverdian for her wonderful piano concert with excerpts from her current program of the works of Liszt.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

The seminar has been structured to have in the first part of the meeting three segments covering the main topics listed in the previous section. Each segment has been introduced by a tutorial on the corresponding topic. Each tutorial has been 45 minutes long while we kept other contributions to 30 minutes. The tutorials were given by Vladimir Koltchinskii (sparsity in learning), Steve Smale (geometry in learning) and Nicolo Cesa Bianchi (game theory and sequential prediction).

3.1 Toward understanding (more) complex data. Graph Laplacians on manifolds with singularities and boundaries

Misha Belkin (Ohio State University, US)

Joint work of Qichao Que, Yusu Wang and Xueyuan Zhou

In this talk I will discuss our recent work on understanding geometry of the space using graph Laplacians. In particular, I will talk about how singularities and boundaries of the space change the behavior of the limiting Laplace operator, a phenomenon somewhat reminiscent of the Gibbs effect in Fourier analysis.

3.2 New (and old) estimators for the mean

Sébastien Bubeck (Universitat Autonoma de Barcelona, ES)

In this talk we discuss the basic problem of estimating the mean of a distribution based on an i.i.d sequence of random variables from this distribution. For subgaussian distributions the empirical mean estimator has all the guarantees that one could hope for: it has exponential tails, it is easy to update with a new observation, and it does not require any parameter tuning. Unfortunately for distributions with only a finite variance, the empirical mean has heavy tails (polynomial size). We discuss here three alternatives, the truncated empirical mean, the median of means, and Catoni's estimator. These three estimators has exponential tails for distributions with finite variances, but each has different advantages and disadvantages regarding computational complexity and parameters tuning.

3.3 Context sensitive information: Model validation by information theory

Joachim M. Buhmann (ETH Zürich, CH)

Joint work of A. Busetto, M. Chehreghani, M. Frank, A. Streich

Model selection in pattern recognition requires (i) to specify a suitable cost function for the data interpretation and (ii) to control the degrees of freedom depending on the noise level in the data. We advocate an information theoretic perspective where the uncertainty in the measurements quantizes the solution space of the underlying optimization problem, thereby adaptively regularizing the cost function. A pattern recognition model, which can tolerate a higher level of fluctuations in the measurements than alternative models, is considered to be superior provided that the solution is equally informative. The optimal tradeoff between "informativeness" and "robustness" is quantified by the approximation capacity of the selected cost function.

Empirical evidence for this model selection concept is provided by cluster validation in computer security, i.e., multilabel clustering of Boolean data for role based access control, but also in high dimensional Gaussian mixture models and the analysis of microarray data. The principle also allows us to rank different spectral clustering models w.r.t. information content. Furthermore, the approximation capacity of the SVD cost function suggests an optimal cutoff value for the SVD spectrum.

3.4 Active Learning and Adaptive Sensing for Sparse Signal Estimation and Testing

Rui M. Castro (TU Eindhoven, NL)

Joint work of Jarvis Haupt and Robert Nowak

Many traditional approaches to statistical inference and machine learning are passive, in the sense that all data are passively collected prior to analysis. However, in many practical scenarios it is possible to adjust the data collection process based on information gleaned from previous observations, closing the loop between data analysis and acquisition. Inference under such scenarios is often referred to as active learning, adaptive sensing, or inference using sequential experimental designs. In this talk I'll focus in particular on estimation and testing problems when the objects of interest are sparse vectors. I'll present a simple but powerful adaptive sensing procedure - Distilled Sensing - which is highly effective for detection and estimation of high-dimensional sparse signals in noise. Large-sample analysis shows that this procedure provably outperforms the best possible inference methods based on non-adaptive data collection methods, allowing for both detection and estimation of extremely weak signals, imperceptible without adaptive sensing. Furthermore, it can be shown that this procedure is essentially optimal for a wide range of scenarios, meaning no other adaptive sensing procedure can yield a significant performance improvement.

3.5 The Game-Theoretic Approach to Machine Learning and Adaptation

Nicoló Cesa-Bianchi (Universitá di Milano, IT)

Joint work of Ohard Shamir

In the first part of the talk, we trace the roots of the game-theoretic approach in learning theory mentioning some of the key results in prediction with expert advice and online learning. In the second part, we describe the first computationally efficient online algorithm for collaborative filtering with norm-constrained matrices. The algorithm combines "random playout" and randomized rounding of loss subgradients.

3.6 On Stability and Bootstrap of Support Vector Machines

Andreas Christmann (Universität Bayreuth, DE)

Joint work of Robert Hable, Ingo Steinwart, and Arnout Van Messem

Support Vector Machines (SVMs) play an important role in statistical machine learning. The talk will focus on some recent results on SVMs: modeling heteroscedasticity by SVMs, a consistency result of SVMs for dependent data, statistical stability (=robustness) of SVMs, and stability of bootstrap estimators of SVMs.

3.7 Nonlinear Eigenproblems in Machine Learning

Matthias Hein (Universität des Saarlandes , DE)

Many problems in data analysis can be formulated as (generalized) eigenproblems. In this work I discuss nonlinear eigenproblems, which allow extended modeling freedom compared to linear eigenproblems in particular concerning robustness and sparsity. After an introduction of the framework and the discussion of an efficient generalization of the inverse power method, two examples of nonlinear eigenproblems are discussed in more detail: the tight relaxation of a large family of balanced graph cuts based on the nonlinear 1-graph Laplacian and sparse PCA.

3.8 Well localized frames, representation of function spaces, and heat kernel estimates

Gerard Kerkyacharian (University Paris-Diderot, FR)

Joint work of T. Coulhon and P.Petrushev

Since during the last twenty years, wavelet theory has proved to be a very useful tool for theoretical purposes as well as for applications, in this talk, we will revisit and provide an extension of this theory in a general geometric framework. Our object here, will be a metric space (M, ρ) equipped with a positive Radon measure, such that (M, ρ, μ) is a homogeneous space in the sense of Harmonic Analysis (there exists $d > 0$,which plays the role of a dimension, such that for all $x \in M$, and $r > 0$, $\mu(B(x, 2r)) \leq 2^d \mu(B(x, r))$). Moreover, the geometry of the space is related to a positive self-adjoint operator L and to the associated semi-group e^{-tL}. We suppose in addition that e^{-tL} is markovian. Here is the main hypothesis : e^{-tL} is a kernel operator, and this kernel $P_t(x, y)$ has the following Gaussian estimate : for all x, y in M, $t > 0$,

$$P_t(x, y) \leq \frac{C_2 \, e^{-c \frac{\rho^2(x,y)}{t}}}{\sqrt{\mu(B(x, \sqrt{t}))\mu(B(y, \sqrt{t}))}}.$$

It is well known that this property is verified for the Laplacian of a Riemannian manifold with non negative Ricci curvature, for Nilpotent Lie Groups, compact Lie Groups and their homogeneous spaces The Besov spaces $B^s_{p,q}$, $0 < s$, $1 \leq p \leq \infty$, $0 \leq q \leq \infty$ could be defined in several equivalent way: as spaces of approximation, or interpolations spaces. The main results are the following :

1. One can build an efficient Littlewood-Paley decomposition and give a charaterization of the Besov spaces.
2. It is possible to build localized frames in duality : $\psi_{j,\xi}, \tilde{\psi}_{j,\xi}$, $j \in \mathbb{N}$, $\xi \in A_j$

   a.

   $$\forall f \in L^p,\ 1 \leq p \leq \infty, \quad f(x) = \sum_j \sum_{\xi \in A_j} \langle f, \psi_{j,\xi} \rangle \tilde{\psi}_{j,\xi}$$

   b.

   $$\exists c > 0,\ \forall j \in \mathbb{N},\ \xi \in A_j, \qquad \psi_{j,\xi} \text{ and } \tilde{\psi}_{j,\xi} \in \Sigma_{cb^j}.$$

   c. and we have the following characterization:

   $$f \in B^s_{p,q} \Leftrightarrow \forall j \in \mathbb{N}, \Big(\sum_{\xi \in A_j} |\langle f, \psi_{j,\xi} \rangle|^p \|\tilde{\psi}_{j,\xi}\|_p \Big)^{\frac{1}{p}} = \alpha_j 2^{-js}, \qquad \alpha \in l_q$$

   (with the usual modification for $p = \infty$).

3.9 Classifying Clustering Schemes

Facundo Mémoli (Stanford University, US)

Many clustering schemes are defined by optimizing an objective function defined on the partitions of the underlying set of a finite metric space. In this paper, we construct a framework for studying what happens when we instead impose various structural conditions on the clustering schemes, under the general heading of functoriality. Functoriality refers to the idea that one should be able to compare the results of clustering algorithms as one varies the data set, for example by adding points or by applying functions to it. We show that within this framework, one can prove a theorem analogous to one of J. Kleinberg, in which for example one obtains an existence and uniqueness theorem instead of a non-existence result. We obtain a full classification of all clustering schemes satisfying a condition we refer to as excisiveness. The classification can be changed by varying the notion of maps of finite metric spaces. The conditions occur naturally when one considers clustering as the statistical version of the geometric notion of connected components. By varying the degree of functoriality that one requires from the schemes it is possible to construct richer families of clustering schemes that exhibit sensitivity to density.

3.10 Testing the Manifold Hypothesis

Hari Narayanan (MIT – Cambridge, US)

Increasingly, we are confronted with very high dimensional data sets in areas like computational biology and medical imaging. As a result, methods of avoiding the curse of dimensionality have come to the forefront of machine learning research. One approach, which relies on exploiting the geometry of the data, has evolved into a subfield called manifold learning.

The underlying hypothesis of this field is that data tend to lie near a low dimensional submanifold, due to constraints that limit the degrees of freedom of the process generating them. This has been empirically observed to be the case, for example, in speech and video data. Although there are many widely used algorithms which assume this hypothesis, the basic question of testing this hypothesis is poorly understood.

I will discuss forthcoming work with Charles Fefferman and Sanjoy Mitter towards developing a provably correct, efficient algorithm to test this hypothesis from random data.

3.11 Active Clustering

Rob Nowak (University of Wisconsin – Madison, US)

Joint work of Brian Eriksson, Gautam Dasarathy, and Aarti Singh

Hierarchical clustering is a common tool used in a broad range of scientific applications. However, in many problems it may be expensive to obtain or compute similarities between the items to be clustered. If we choose to randomly subsample similarities, we cannot hope to recover small clusters with size that is sublinear in the number of objects. This necessitates an active procedure that sequentially selects which similarities to obtain in an adaptive fashion. I will describe such an active clustering procedure that generates a hierarchical clustering of N objects using only N log N similarities, instead of all N(N-1)/2 similarities. The method can recover all clusters of size larger than log N, even in the presence of a limited fraction of arbitrarily corrupted or noisy similarities. I will also discuss potential applications to network tomography and genomic data analysis.

3.12 Convex relaxations for Combinatorial Penalties

Guillaume Obozinski (ENS – Paris, FR)

In structured sparsity, one attempts to estimate a function which, in a appropriate parameterization, is encoded by a sparse vector; the support (or set of non-zero elements) of this sparse vector is furthermore assumed to present a type of structure which is known a priori. A common approach to the problem is to penalize implicitly or explicitly the structure of the support of the estimated parameter vector. In this talk, I will present a generic convex relaxation for a family of functions that penalize simultaneously the structure of the support through a general set function, and the L_p norm of the parameter vector for an arbitrary fixed p.

The formulation considered allows to treat in a unified framework several a priori disconnected approaches such as block-coding and submodular functions, and extend in the latter case theoretical results obtained by Bach.

3.13 A Meta-Learning Approach to the Regularized Learning - Case Study: Blood Glucose Prediction

Sergei Pereverzev (RICAM – Linz, AT)

The motivation for this research appeared in the course of the project “DIAdvisor” funded by the European Commission with the aim to improve the diabetes therapy.

The massive increase in the incidence of diabetes is now a major global healthcare challenge, and the treatment of diabetes is one of the most difficult therapies to manage, because of the difficulty in actively predicting blood glucose levels.

From the literature we know that nowadays there are mainly two approaches to predict the future blood glucose based upon the patient's current and past blood glucose values. One of them uses the time-series methodology, while another one employes artificial neural networks techniques. But time-series predictors seem to be too sensitive to gaps in the data, which may frequently appear when available blood glucose meters are used. As to neural networks predictors, they need long training periods and much more information to be set up.

Therefore, we join the "DIAdvisor" consortium with the idea to use regularized learning algorithms in predicting blood glucose. These algorithms are well understood now, and it is known that their performance essentially depends on the choice of regularization parameters and, which is even more important, on the choice of kernels generating Reproducing Kernel Hilbert Spaces, in which the regularization is performed. As it was quickly realized, in the context of blood glucose prediction these algorithmic instances cannot be a priori fixed, but need to be adjusted to each particular prediction input. Thus, a regularized learning based predictor should learn how to learn kernels and regularization parameters from the inputs. Such a predictor is constructed as a result of a process of learning to learn, or "meta-learning". In this way we have developed the Fully Adaptive Regularized Learning (FARL) approach to the blood glucose prediction.

The developed approach allows the construction of blood glucose predictors which, as it has been demonstrated in the extensive clinical trials, outperform the state-of-art algorithms. Moreover, it turns out that in the context of the blood glucose prediction the FARL-approach is more advanced than other meta-learning technologies such as k-NN ranking.

In this talk we are going to present a theoretical justification of the FARL-approach, as well as performance assessment results from clinical trials. The approach is described in the patent application EP 11163219.6 filed jointly by Austrian Academy of Sciences and Novo Nordisk A/S (Denmark).

3.14 The computational magic of the ventral stream: towards a theory?

Tomaso Poggio (MIT – Cambridge , US)

I conjecture that the sample complexity of object recognition is mostly due to geometric image transformations and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations. The most surprising implication of the theory emerging from these assumptions is that the computational goals and detailed properties of cells in the ventral stream follow from *symmetry properties* of the visual world through a process of unsupervised correlational learning.

From the assumption of a hierarchy of areas with receptive fields of increasing size the theory predicts that the size of the receptive fields determines which transformations are learned during development and then factored out during normal processing; that the transformation represented in each area determines the tuning of the neurons in the area, independently of the statistics of natural images; and that class-specific transformations are learned and represented at the top of the ventral stream hierarchy.

Some of the main predictions of this theory-in-fieri are:

- the type of transformation that are learned from visual experience depend on the size (measured in terms of wavelength) and thus on the area (layer in the models) – assuming that the aperture size increases with layers;
- the mix of transformations learned determine the properties of the receptive fields – oriented bars in V1+V2, radial and spiral patterns in V4 up to class specific tuning in AIT (e.g. face tuned cells);
- class-specific modules – such as faces, places and possibly body areas – should exist in IT to process images of object classes.

3.15 Learning Theory: A Minimax Analysis

Alexander Rakhlin (University of Pennsylvania, USA)

Statistical Learning Theory studies the problem of estimating (learning) an unknown function given a class of hypotheses and an i.i.d. sample of data. Classical results show that combinatorial parameters (such as Vapnik-Chervonenkis and scale-sensitive dimensions) and complexity measures (such as covering numbers, Rademacher averages) govern learnability and rates of convergence. Further, it is known that learnability is closely related to the uniform Law of Large Numbers for function classes.

In contrast to the i.i.d. case, in the online learning framework the learner is faced with a sequence of data appearing at discrete time intervals, where the data is chosen by the adversary. Unlike statistical learning, where the focus has been on complexity measures, the online learning research has been predominantly algorithm-based. That is, an algorithm with a non-trivial guarantee provides a certificate of learnability.

We develop tools for analyzing learnability in the game-theoretic setting of online learning without necessarily providing a computationally feasible algorithm. We define complexity measures which capture the difficulty of learning in a sequential manner. Among these measures are analogues of Rademacher complexity, covering numbers and fat shattering dimension from statistical learning theory. These can be seen as temporal generalizations of classical results. The complexities we define also ensure uniform convergence for non-i.i.d. data, extending the Glivenko-Cantelli type results. A further generalization beyond external regret covers a vast array of known frameworks, such as internal and Phi-regret, Blackwell's Approachability, calibration of forecasters, global non-additive notions of cumulative loss, and more

3.16 Sparse Recovery and Structured Random Matrices

Holger Rauhut (Universität Bonn, DE)

Compressive sensing (sparse recovery) is a recent paradigm in signal processing and sampling theory that predicts that sparse signals can be recovered from a small number of linear and non- adaptive measurements using convex optimization or greedy algorithms. Quite remarkably, all good constructions of the so called measurement matrix known so far are based on randomness. While Gaussian random matrices provide optimal recovery guarantees, such unstructured matrices are of limited use in applications. Indeed, structure often allows to have fast matrix vector multiplies. This is crucial in order to speed up recovery algorithms and to deal with large scale problems. The talk discusses models of structured random matrices that are useful in certain applications, and presents corresponding recovery guarantees. An important type of structured random matrix arises in connection with sampling sparse expansions in terms of bounded orthogonal systems (such as the Fourier system). The second type of structured random matrices to be discussed are partial random circulant matrices, that is, from convolution. In particular, we present recent results with J. Romberg and J. Tropp on the restricted isometry property of such matrices. The third type of measurement matrices arises in kernel-based semisupervised learning, for which preliminary results are reported.

3.17 Nonparametric Bandits with Covariates v.2.0

Philippe Rigollet (Princeton University, US)

Joint work of Vianney Perchet (Univ. Paris 7)

We consider a multi-armed bandit problem in a setting where each arm produces a noisy reward realization which depends on an observable random covariate. As opposed to the traditional static multi-armed bandit problem, this setting allows for dynamically changing rewards that better describe applications where side information is available. We adopt a nonparametric model where the expected rewards are smooth functions of the covariate and where the hardness of the problem is captured by a margin parameter. To maximize the expected cumulative reward, we introduced in Rigollet and Zeevi (2010) a policy based on a fixed partitioning of the covariate space. While it achieved optimal regret for a certain class of "difficult" problems, it failed to adapt the the complexity of "easy" problems. In this second attempt, we rely on a dynamic, adaptive partition that is implemented in a policy called Adaptively Binned Successive Elimination (abse) . It is proved to achieve optimal bounds on the regret an reveals an interesting phenomenon: the effect of the exploration vs. exploitation dilemma is washed away by the difficulty of nonparametric estimation. To derive these bounds, we developed a modification of the Successive Elimination policy in the static framework that achieves the sharper regret bounds necessary to prove our main theorem.

3.18 Nonparametric Sparsity Based Regularization

Lorenzo Rosasco (MIT – Cambridge, US)

In this work we are interested in the problems of supervised learning and variable selection when the input-output dependence is described by a nonlinear function depending on a few variables. Our goal is to consider a sparse nonparametric model, hence avoiding linear or additive models. The key idea is to measure the importance of each variable in the model using partial derivatives. Based on this intuition we propose and study a new regularizer and a corresponding least squares regularization scheme. Using concepts and results from the theory of reproducing kernel Hilbert spaces and proximal methods, we show that the proposed learning algorithm corresponds to a minimization problem which can be provably solved by an iterative procedure. The consistency properties of the obtained estimator are studied both in terms of prediction and selection performance.

3.19 Learnability Beyond Uniform Convergence

Shai Shalev-Shwartz (The Hebrew University of Jerusalem, IL)

The problem of characterizing learnability is the most basic question of statistical learning theory. A fundamental result is that learnability is equivalent to uniform convergence of the empirical risk to the population risk, and that if a problem is learnable, it is learnable via empirical risk minimization. The equivalence of uniform convergence and learnability was formally established only in the supervised classification and regression setting. We show that in (even slightly) more complex prediction problems learnability does not imply uniform convergence. We discuss several alternative attempts to characterize learnability.

3.20 Entire Relaxation Path for Maximum Entropy Models

Yoram Singer (Google Inc. – Mountain View, US)

We describe a relaxed and generalized notion of maximum entropy problems for multinomial distributions. By introducing a simple re-parametrization we are able to derive an efficient homotopy tracking for the entire relaxation path. The end result is an algorithm that can provide optimal probabilistic estimates for any relaxation parameter using linear space and sub-linear time. We also show that the Legendre dual of the relaxed maximum entropy problem is the task of finding the maximum-likelihood estimator for an exponential distribution with L_1 regularization. Hence, our solution can be used for problems such as language modeling with sparse parameter representation.

3.21 Robust approachability with applications to regret minimization in games with partial monitoring

Gilles Stoltz (ENS – Paris, FR)

Approachability has become a standard tool in analyzing learning algorithms in the adversarial online learning setup. We first define this notion and recall Blackwell's early characterization [1, 2]. We then develop a variant of approachability for games where there is ambiguity in the obtained reward that belongs to a set, rather than being a single vector. Using this variant we tackle the problem of approachability in games with partial monitoring and develop simple and efficient algorithms (i.e., with constant per-step complexity) for this setup, where the characterization of approachability has been obtained by Perchet [4] but came without an efficient algorithm. This talk is based on [3].

References

1 D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

2 D. Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians, 1954, Amsterdam, vol. III*, pages 336–338, 1956.

3 S. Mannor, V. Perchet, and G. Stoltz. Robust approachability with applications to regret minimization in games with partial monitoring. In *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory (COLT'11)*. Omnipress, 2011.

4 V. Perchet. Internal regret with partial monitoring calibration-based optimal algorithms. *Journal of Machine Learning Research*, 12(Jun):1893–1921, 2011.

3.22 The Lasso, correlated design, and improved oracle inequalities

Sara van de Geer (ETH Zürich, CH)

We study high-dimensional linear models and the l_1-penalized least squares estimator, also known as the Lasso estimator. In literature, oracle inequalities have been derived under restricted eigenvalue or compatibility conditions. In this paper, we complement this with entropy conditions which allow one to improve the dual norm bound, and demonstrate how this leads to new oracle inequalities. The new oracle inequalities show that a smaller choice for the tuning parameter and a trade-off between l_1-norms and small compatibility constants are possible. This implies, in particular for correlated design, improved bounds for the prediction error of the Lasso estimator as compared to the methods based on restricted eigenvalue or compatibility conditions only.

3.23 Dictionary learning: theme and variations

Alessandro Verri (University of Genova, IT)

In this talk i will illustrate our recent work on dictionary learning. The motivation of our work is rooted in the typical setting of biomedical imaging in which image annotation is expensive, while unlabeled or weakly labeled data abound. Starting from a rather conventional approach i discuss several developments including a scheme in which the analysis matrix is learnt during training and a scheme in which slowness plays a role in enforcing sparsity in the encoding stage.

3.24 Phase-transition in the family of p-resistances

Ulrike von Luxburg (Universität Hamburg

We study the family of p-resistances on graphs for $p \geq 1$. This family generalizes the standard resistance distance. We prove that for any fixed graph, for $p = 1$ the p-resistance coincides with the shortest path distance, for $p = 2$ it coincides with the standard resistance distance, and for $p \to \infty$ it converges to the inverse of the minimal s-t-cut in the graph. Secondly, we consider the special case of random geometric graphs (such as k-nearest neighbor graphs) when the number n of vertices in the graph tends to infinity. We prove that an interesting phase-transition takes place. There exists a critical threshold p^* such that if $p < p^*$, then the p-resistance depends on meaningful global properties of the graph, whereas if $p > p^*$, it only depends on trivial local quantities and does not convey any useful information. We can explicitly compute the critical value: $p^* = 1 + 1/(d - 1)$ where d is the dimension of the underlying space. We also relate our findings to Laplacian regularization and suggest to use q-Laplacians as regularizers, where q satisfies $1/p^* + 1/q = 1$.

3.25 Loss Functions, and Relations Between Machine Learning Problems

Robert Williamson(Australian National University – Canberra, AU)

Loss functions are central to supervised machine learning problems, but there has been little work in the recent machine learning literature in systematically understanding the effect of choice of loss functions. In this talk I will summarize some recent work starting with consideration of proper losses for classification problems (binary and multiclass). I will consider relationships to divergences (f-divergences and Bregman), surrogate regret bounds, composite losses (the composition of a proper loss with an invertible link function), existence and uniqueness results for such representations, integral representations, and characterization of mixability and convexity.

I will conclude by situating the work as part of a larger project on relating machine learning problems.

3.26 Some Learning Algorithms Producing Sparse Approximations

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

We shall discuss two classes of kernel-based learning algorithms which produce sparse approximations for regression. The first class is of kernel projection machine type and generated by least squares regularization schemes with ℓ^q-regularizer $(0 < q \leq 1)$ in a data dependent hypothesis space based on empirical features (constructed by reproducing kernels and samples). The second class is spectral algorithms associated with high-pass filter functions. Learning rates and sparsity estimations will be provided based on properties of the kernel, the regression function, and the probability measure.

Participants

- Misha Belkin
Ohio State University, US
- Robert J. Bonneau
AFOSR – Arlington, US
- Sebastien Bubeck
Princeton University, US
- Joachim Buhmann
ETH Zürich, CH
- Rui Castro
Eindhoven University of Technology, NL
- Nicoló Cesa-Bianchi
Univ. degli Stugli de Milano, IT
- Andreas Christmann
Universität Bayreuth, DE
- Matthias Hein
Saarland University, DE
- Jürgen Jost
MPI for Mathematics in the Sciences – Leipzig, DE
- Gerard Kerkyacharian
University Paris-Diderot, FR
- Vladimir Koltchinskiii
Georgia Tech, US
- Lek-Heng Lim
University of Chicago, US
- Gabor Lugosi
Pompeu Fabra University – Barcelona, ES
- Stephane Mallat
Ecole Polytechnique – Paris, FR
- Facundo Memoli
University of Adelaide, AU
- Hari Narayanan
MIT – Cambridge, US
- Robert Nowak
Univ. of Wisconsin-Madison, US
- Guillaume Obozinski
Ecole Normale Superieure – Paris, FR
- Sergei Pereverzyev
RICAM – Linz, AT
- Tomaso Poggio
MIT – Cambridge, US
- Massimiliano Pontil
University College London, UK
- Alexander Rahklin
University of Pennsylvania, US
- Philippe Rigollet
Princeton University, US
- Holger Rauhut
Universität Bonn, DE
- Lorenzo Rosasco
MIT – Cambridge, US, and IIT, IT
- Stephen Smale
City University, HK
- Bernhard Schölkopf
MPI for Intelligent Systems – Tübingen, DE
- Shai Shalev-Shwartz
Hebrew Univ. – Jerusalem, IL
- Yoram Singer
Google Research, US
- Ingo Steinwart
Universität Stuttgart, DE
- Gilles Stoltz
Ecole Normale Superieure – Paris, FR
- Alexandre Tsybakov
Université Paris VI, FR
- Sara van de Geer
ETH Zürich, CH
- Ulrike von Luxburg
Universität Hamburg, DE
- Alessandro Verri
Univ. degli Studi di Genova, IT
- Martin Wainwright
UC Berkeley, US
- Bob Williamson
ANU – Canberra, AU
- Ding-Xuan Zhou
City University, HK
- Tong Zhang
Rutgers University, US