

---

Wednesday, May 31

---

- 09:00 – 09:55 Registration and coffee-break  
09:55 – 10:00 Opening remarks  
10:00 – 10:50 **Emmanuel Candes** (California Institute of Technology)  
*The Dantzig Selector: Statistical Estimation when  $p$  is Larger than  $n$*   
11:00 – 11:30 Coffee-break  
11:30 – 12:20 **Franck Barthe** (Université Toulouse III)  
*About Talagrand's Concentration Inequality for Exponential Measures*  
12:30 – 14:30 Lunch break  
14:30 – 15:20 **Dean Foster** (University of Pennsylvania)  
*Deterministic Calibration and Nash Equilibrium*  
15:30 – 16:20 **Neri Merhav** (Technion, I.I.T.)  
*On Context-tree Prediction of Individual Sequences*  
16:30 – 17:00 Coffee-break  
17:00 – 17:50 **Peter Grünwald** (CWI & Eurandom)  
*Suboptimality of MDL and Bayes in Classification under Misspecification*  
18:00 – 19:30 Poster session  
19:30 – 21:30 Light welcoming buffet

---

Thursday, June 1

---

- 09:00 – 09:50 **Ofer Zeitouni** (University of Minnesota)  
*A Correlation Inequality for Nonlinear Reconstruction*  
10:00 – 10:30 Coffee-break  
10:30 – 11:20 **Ehud Lehrer** (Tel-Aviv University)  
*Bayesian and Non-Bayesian Learning in Games*  
11:30 – 12:20 **Bernhard Schölkopf** (Max Planck Institut, Tübingen)  
*Applications of Kernel Methods*  
12:30 – 14:30 Lunch break  
14:30 – 15:20 **Z. D. Bai** (National University of Singapore)  
*Statistical Analysis for Rounding Data*  
15:30 – 16:20 **Shie Mannor** (McGill University)  
*An Isoperimetric Inequality with Applications to Learning*  
16:30 – 17:00 Coffee-break  
17:00 – 17:50 **Ingo Steinwart** (Los Alamos National Laboratory)  
*Learning from Dependent Observations*

---

Friday, June 2

---

- 09:00 – 09:50 **John Shawe-Taylor** (University of Southampton)  
*Statistical Analysis of Subspace Methods and Associated Learning Algorithms*
- 10:00 – 10:30 Coffee-break
- 10:30 – 11:20 **Tong Zhang** (Yahoo Inc.)  
*Theory and Algorithms for Large Scaling Ranking Problems*
- 11:30 – 12:20 **Vladimir Koltchinskii** (Georgia Institute of Technology)  
*Sparsity in High-Dimensional Learning Problems*
- 12:30 – 14:30 Lunch break
- 14:30 – 15:20 **Jean-Philippe Vert** (Ecole des Mines de Paris)  
*Regularization of Kernel Methods by Decreasing the Bandwidth of the Gaussian Kernel*
- 15:30 – 16:20 **Nicolò Cesa-Bianchi** (Università degli Studi di Milano)  
*Learning and Randomization*
- 16:30 – 17:00 Coffee-break
- 17:00 – 17:50 **Vladimir Temlyakov** (University of South Carolina)  
*On Optimal and Universal Estimators in Learning Theory*

---

Saturday, June 3

---

- 9:00 – 9:50 **Santosh Vempala** (Massachusetts Institute of Technology)  
*Sampling, Integration and Optimization of High-dimensional Log-concave Functions*
- 10:00 – 10:30 Coffee-break
- 10:30 – 12:00 **Nathan Linial** (The Hebrew University of Jerusalem)  
*Complexity of Sign Matrices and its Many Aspects*
- 12:00 – 14:00 Buffet lunch
- 14:00 – 14:50 **Nicolas Vayatis** (Université Paris VI)  
*Is There Life beyond the Classification Problem?*
- 15:00 – 15:50 **Peter Bartlett** (UC Berkeley)  
*Asymptotic Properties of Convex Optimization Methods for Multiclass Classification*
- 16:00 – 17:00 Final coffee-break

**Emmanuel Candes (California Institute of Technology)**  
**The Dantzig Selector: Statistical Estimation when  $p$  is Larger than  $n$**

In many important statistical applications, the number of variables or parameters is much larger than the number of observations. In radiology and biomedical imaging for instance, one is typically able to collect far fewer measurements about an image of interest than the unknown number of pixels. Examples in functional MRI and tomography immediately come to mind. Other examples of high-dimensional data in genomics, signal processing and many other fields abound. In the context of multiple linear regression for instance, a fundamental question is whether it is possible to estimate a vector of parameters of size  $p$  from a vector of observations of size  $n$  when  $n \ll p$ . This seems a priori hopeless.

This talk introduces a new estimator, dubbed the “Dantzig selector” in honor of the late George Dantzig as it invokes linear programming, and which enjoys remarkable statistical properties. Suppose that the data or design matrix obeys a uniform uncertainty principle and that the true parameter vector is sufficiently sparse or compressible which roughly guarantees that the model is identifiable. Then the estimator achieves an accuracy which nearly equals that one would achieve with an *oracle* that would supply perfect information about which coordinates of the unknown parameter vector are nonzero and which were above the noise level. Our results connect with the important model selection problem. In effect, the Dantzig Selector automatically selects the subset of covariates with nearly the best predictive power, by solving a convenient linear program.

Our results are also inspired by a recent body of work perhaps now best known under the name of “Compressive Sampling,” a new sampling theory we introduced very recently. If time allows, I will discuss applications of Compressive Sampling in other fields such as coding theory.

**Further references:** The main paper is

<http://www.acm.caltech.edu/~emmanuel/papers/DantzigSelector.pdf>,

but there are also many other papers on a related subject, for instance, “Decoding by linear programming” (<http://www.acm.caltech.edu/~emmanuel/papers/DecodingLP.pdf>)

and “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information” (<http://www.acm.caltech.edu/~emmanuel/papers/ExactRecovery.pdf>).

**Franck Barthe (Université Toulouse III)**  
**About Talagrand’s Concentration Inequality for Exponential Measures**

This lecture provides a survey on the recent works extending Talagrand’s concentration inequality for the exponential measures, or more generally product log-concave measures. Among others, transportation cost inequalities and modified logarithmic Sobolev inequalities will be presented and studied.

**Dean Foster (University of Pennsylvania)**  
**Deterministic Calibration and Nash Equilibrium**

We provide a natural learning process in which the joint frequency of empirical play converges into the set of convex combinations of Nash equilibria. In this process, all players rationally choose their actions using a public prediction made by a deterministic, weakly calibrated algorithm. Furthermore, the public predictions used in any given round of play are frequently close to some Nash equilibrium of the game. (Joint work with Sham M. Kakade.)

**Neri Merhav (Technion, I.I.T.)**  
**On Context-tree Prediction of Individual Sequences**

Motivated by the evident success of context–tree based methods in lossless data compression, we explore, in this talk, methods of the same spirit in universal prediction of individual sequences. By context–tree prediction, we refer to a family of prediction schemes, where at each time instant  $t$ , after having observed all outcomes of the data sequence  $x_1, \dots, x_{t-1}$ , but not yet  $x_t$ , the prediction is based on a “context” (or a state) that consists of the  $k$  most recent past outcomes  $x_{t-k}, \dots, x_{t-1}$ , where the choice of  $k$  may depend on the contents of a possibly longer, though limited, portion of the observed past,  $x_{t-k_{\max}}, \dots, x_{t-1}$ . This is different from the study reported in Feder, Merhav, and Gutman (2002), where general finite–state predictors as well as “Markov” (finite–memory) predictors of fixed order, were studied in the regime of individual sequences.

Another important difference between this study and Feder, Merhav, and Gutman (2002) is the asymptotic regime. While in Feder, Merhav, and Gutman (2002), the resources of the predictor (i.e., the number of states or the memory size) were kept fixed regardless of the length  $N$  of the data sequence, here we investigate situations where the number of contexts, or states, is allowed to grow concurrently with  $N$ . We are primarily interested in the following fundamental question: What is the critical growth rate of the number of contexts, below which the performance of the best context–tree predictor is still universally achievable, but above which it is not? We show that this critical growth rate is linear in  $N$ . In particular, we propose a universal context–tree algorithm that essentially achieves optimum performance as long as the growth rate is sublinear, and show that, on the other hand, this is impossible in the linear case.

**Further references:**

<http://www.ee.technion.ac.il/people/merhav/papers/p100.pdf>

**Peter Grünwald (Centrum voor Wiskunde en Informatica & Eurandom)**  
**Suboptimality of MDL and Bayes in Classification under Misspecification**

We show that forms of Bayesian and MDL learning that are often applied to classification problems can be “statistically inconsistent”. We present a classification model (a large family of classifiers) and a distribution such that the best classifier within the model has classification risk  $r$ , where  $r$  can be taken arbitrarily close to 0. Nevertheless, no matter how many data are observed, both the classifier inferred by MDL and the classifier based on the Bayesian posterior will make predictions with error much larger than  $r$ . If  $r$  is chosen not too small, predictions based on the Bayesian posterior can even perform substantially worse than random guessing, no matter how many data are observed. Our result can be re-interpreted as showing that, if a probabilistic model does not contain the data generating distribution, then Bayes and MDL do not always converge to the distribution in the model that is closest in KL divergence to the data generating distribution. We compare this result with earlier results on Bayesian inconsistency by Diaconis, Freedman and Barron.

This work is a follow-up on joint work with John Langford of the Toyota Technological Institute, Chicago, published at COLT 2004, available at [www.grunwald.nl](http://www.grunwald.nl).

**Ofer Zeitouni (University of Minnesota)**  
**A Correlation Inequality for Nonlinear Reconstruction**

Consider the problem of reconstructing a Gaussian vector based on the maximum of its projections on the elements of an orthogonal basis. S. Mallat and myself showed that the optimal basis for this problem is the Karhunen-Loeve one. I will discuss the proof and conjectured generalizations.

**Ehud Lehrer (Tel-Aviv University)**  
**Bayesian and Non-Bayesian Learning in Games**

I will contrast Bayesian with non-Bayesian learning to play an equilibrium. I will primarily refer to the mathematics involved in the two corresponding research directions.

**Bernhard Schölkopf (Max Planck Institut, Tübingen)**  
**Applications of Kernel Methods**

We discuss several new applications of kernel methods, including algorithms developed for the tasks at hand. The problems we are working on range from aspects of biomedical signal processing (analysis of neural data, brain computer interfacing) to applications in computer graphics (surface modeling and morphing).

**Z. D. Bai (National University of Singapore)**  
**Statistical Analysis for Rounding Data**

Unless the model is discrete, data rounding is unavoidable in practical measurement. However, the errors caused by rounding of data are almost ignored by all classical statistical theories. Although some pioneers have noticed this problem, few suitable approaches were proposed to deal with this error. In this work, both by simulations as well as by theoretical analysis, we demonstrate that the traditionally used sample mean and sample variance, covariance are no longer consistent nor asymptotically normal, when rounding errors are present. Also, by some concrete examples when measurements are rounded to some extent, we propose to use MLE or approximated MLE (AMLE) to estimate the parameters and discuss the properties of them and tests based on the new estimators. In particular, as an example, we shall discuss the limiting properties of the new estimator of parameters in an  $AR(p)$  model and  $MA(q)$  model when the observations are rounded.

(Joint work with Shurong Zheng and Baoxue Zhang.)

**Shie Mannor (McGill University)**  
**An Isoperimetric Inequality with Applications to Learning**

An issue of central importance is learning in the presence of data corruption, or noise. In this talk, we consider the case where data corruption has produced a data sample with a large margin. The essential question is “what is the cost of this margin?” in terms of generalization error. We provide an answer for the case where the underlying distribution has a nearly log-concave density.

First, we prove that given such a nearly log-concave density, in any partition of the space into two well separated sets, the measure of the points that do not belong to these sets is large. Next, we apply this isoperimetric inequality to derive lower bounds on the generalization error in classification. We further consider regression problems and show that if the inputs and outputs are sampled from a nearly log-concave distribution, the measure of points for which the prediction is wrong by more than  $\epsilon_0$  and less than  $\epsilon_1$  is (roughly) linear in  $\epsilon_1 - \epsilon_0$ , as long as  $\epsilon_0$  is not too small, and  $\epsilon_1$  not too large. We also show that when the data are sampled from a nearly log-concave distribution, the margin cannot be large in a strong probabilistic sense.

**Ingo Steinwart (Los Alamos National Laboratory)**  
**Learning from Dependent Observations**

The standard assumption in statistical learning theory is that the available samples are realizations of i.i.d. random variables. However, in many applications this assumption cannot be rigorously justified, in particular if the observations are intrinsically temporal. In this talk I will present some recent results on the learnability of rather general observation-generating random processes. In particular, I will establish a weak consistency result for support vector machine classification and regression. In addition, refined results for e.g.  $\alpha$ -mixing processes will be presented. If time permits I will finally discuss whether the behaviour of certain dynamical systems can be learned.

**John Shawe-Taylor (University of Southampton)**  
**Statistical Analysis of Subspace Methods and Associated Learning Algorithms**

Subspace inference is a critical component in many practical applications of learning from data, yet very little analysis has been made of the performance of these algorithms. The talk considers the question of providing a statistical analysis of subspace methods and of learning using the associated representations. We begin with considering principal components analysis and the relation between process and empirical eigenvalues. We go on to consider more advanced techniques such as canonical correlation analysis and linear functions learned in the inferred representation. Sparse analogies of these techniques will be discussed with associated bounds.

**Tong Zhang (Yahoo Inc.)**  
**Theory and Algorithms for Large Scaling Ranking Problems**

I will discuss machine learning problems encountered in web search and advertising, and then focus on ranking. In the web search setting, I will talk about training relevance models based on DCG (discounted cumulated gain) optimization. Under this metric, the system output quality is naturally determined by the performance near the top of its rank-list. I will mainly focus on various theoretical issues in this learning problem.

As a related practical illustration, I will talk about optimizing the ranking function of a statistical machine translation system according to the BLEU metric (standard measure of translation quality). Our approach treats machine translation as a black-box, and can optimize millions of system parameters automatically. This has never been attempted before. I will present our method and some results. (Joint work with David Cossock, Yahoo, and Christoph Tillmann, IBM.)

**Vladimir Koltchinskii (Georgia Institute of Technology)**  
**Sparsity in High-Dimensional Learning Problems**

We study penalized empirical risk minimization with convex loss over the linear span of a large finite set  $\mathcal{H}$  of base functions. The penalty is based on the  $\ell_p$ -norm of the vector of coefficients with  $p = 1 + c/\log N$ , where  $N$  is the cardinality of  $\mathcal{H}$ . We prove several inequalities that directly relate "the degrees of sparsity" of empirical and true solutions of such problems and show what impact the sparsity has on the excess risk bounds and on the accuracy of estimation of the vector of coefficients. We discuss several other problems, such as data-driven choice of regularization parameter that provides adaptation to unknown sparsity of the true solution as well as the problem of adaptation to linear dependencies in the set  $\mathcal{H}$ . We also discuss the connections of these results to recent work on aggregation of statistical estimators (Tsybakov and coauthors) and to sparse recovery problems in computational harmonic analysis (Donoho, Candes, Tao among others).

**Jean-Philippe Vert (Ecole des Mines de Paris)**  
**Regularization of Kernel Methods by Decreasing the Bandwidth of the Gaussian Kernel**

We consider learning algorithms that minimize an empirical risk regularized by the norm in the reproducing kernel Hilbert space of the Gaussian kernel. The conditions on the loss function for Bayes consistency of such methods have been studied recently when the regularization term asymptotically vanishes as the sample size increases. Here we study the different situation where the regularization term does not vanish, but the bandwidth of the Gaussian kernel instead decreases with the sample size. We will explicit the asymptotic limit of the function selected by the algorithm, give conditions on the loss function to ensure Bayes consistency, and provide non-asymptotic learning bounds in this case. We will deduce in particular the consistency of the one-class support vector machine algorithm as a density level set estimator. (Joint work with Régis Vert.)

**Nicolò Cesa-Bianchi (Università degli Studi di Milano)**  
**Learning and Randomization**

Randomization is a fundamental tool in learning. In this talk we illustrate some interesting applications of randomized algorithms to the solution of various problems in the areas of individual sequence prediction and pattern classification.

**Vladimir Temlyakov (University of South Carolina)**  
**On Optimal and Universal Estimators in Learning Theory**

This talk addresses some problems of supervised learning. Supervised learning, or learning-from-examples, refers to a process that builds on the base of available data of inputs  $x_i$  and outputs  $y_i$ ,  $i = 1, \dots, m$ , a function that best represents the relation between the inputs  $x \in X$  and the corresponding outputs  $y \in Y$ . The goal is to find an estimator  $f_z$  on the base of given data  $\mathbf{z} := ((x_1, y_1), \dots, (x_m, y_m))$  that approximates well the regression function  $f_\rho$  (or its projection) of an unknown Borel probability measure  $\rho$  defined on  $Z = X \times Y$ . We assume that  $(x_i, y_i)$ ,  $i = 1, \dots, m$ , are independent and distributed according to  $\rho$ .

There are several important ingredients in mathematical formulation of this problem. We follow the way that has become standard in approximation theory and has been used in recent papers. In this approach we first choose a function class  $W$  (a hypothesis space  $\mathcal{H}$ ) to work with. After selecting a class  $W$  we have the following two ways to go. The first one is based on the idea of studying approximation of the  $L_2(\rho_X)$  projection  $f_W := (f_\rho)_W$  of  $f_\rho$  onto  $W$ . Here,  $\rho_X$  is the marginal probability measure. This setting is known as the *improper function learning problem* or the *projection learning problem*. In this case we do not assume that the regression function  $f_\rho$  comes from a specific (say, smoothness) class of functions. The second way is based on the assumption  $f_\rho \in W$ . This setting is known as the *proper function learning problem*. For instance, we may assume that  $f_\rho$  has some smoothness. We will give some upper and lower estimates in both settings.

In the problem of universal estimators we assume that an unknown measure  $\rho$  satisfies some conditions. Following the standard way from nonparametric statistics we formulate these conditions in the form  $f_\rho \in \Theta$ . Next, we assume that the only a priori information available is that  $f_\rho$  belongs to a class  $\Theta$  (unknown) from a known collection  $\{\Theta\}$  of classes. We want to build an estimator that provides approximation of  $f_\rho$  close to the optimal for the class  $\Theta$ . We use a standard method of penalized least squares estimators for construction of universal estimators.

**Santosh Vempala (Massachusetts Institute of Technology)**  
**Sampling, Integration and Optimization of High-dimensional Log-concave Functions**

Logconcave functions are a common generalization of Gaussians and indicator functions of convex bodies; they appear in many areas. In this talk, we survey the algorithmic and geometric ideas behind the most recent developments in sampling, integration and optimization of logconcave functions. In particular, we will discuss the analysis of the random walk called “hit-and-run”, and a general method called simulated annealing which is used in the current best algorithms for both integration and optimization.

**Nathan Linial (The Hebrew University of Jerusalem)**  
**Complexity of Sign Matrices and its Many Aspects**

Consider a matrix of  $+1/-1$  as a family of concepts to be learned. Various measures can be associated with this matrix in an attempt to quantify how hard it is to learn this concept class. Among the better known measures are the VC dimension and the margin. In joint work with Adi Shraibman we are putting these notions in a broader framework of complexity measures of sign matrices. The simplest complexity measure is the rank, and many other natural concepts arise which are related to various other fields such as Banach Space Theory, communication complexity and discrepancy theory. We are investigating these different concepts and their mutual relationships.

**Nicolas Vayatis (Université Paris VI)**  
**Is There Life beyond the Classification Problem?**

In the recent years, significant progress has been achieved on the statistical understanding of celebrated classification algorithms such as boosting and SVM. The key for proceeding to a statistical analysis was to interpret these algorithms as optimization procedures minimizing a penalized convex risk functional. From there it was possible: first, to relate the convex criterion to the standard performance measure -the classification error- and then, to adapt the flourishing theory of empirical risk minimization in order to provide generalization error bounds and oracle inequalities for convex risk minimization procedures. In the talk, I will discuss whether this programme can be applied to another problem: the ranking/scoring problem. Indeed, in applications such as Information Retrieval or Credit Risk screening, the goal is to rank/score webpages or individuals, rather than simply assigning them to a specified category. In this perspective, standard performance measures lead to statistical functionals of order two for which classification theory does not apply straightforwardly. In the talk, I will give some insights and results on these new challenging issues. (Joint work with Stephan Cléménçon and Gábor Lugosi.)

**Peter Bartlett (UC Berkeley)**  
**Asymptotic Properties of Convex Optimization Methods for Multiclass Classification**

We consider the following pattern classification problem: given a sample of i.i.d. pairs  $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ , where  $\mathcal{Y}$  is finite, find a function  $f : \mathcal{X} \rightarrow \mathcal{Y}$  that has small misclassification probability. Many successful algorithms for binary classification (with  $|\mathcal{Y}| = 2$ ) involve optimization of a convex criterion. These methods can be generalized in many ways to handle the multiclass case. It turns out that the study of multiclass methods is not a simple extension of results for the binary case. For instance, many apparently natural generalizations of binary methods do not preserve the attractive property of universal consistency (that is, for any probability distribution, the risk of the classifier approaches the best possible). We consider methods that choose a vector-valued function  $f$  to optimize a convex criterion of the form  $\sum_i \Psi(f(X_i), Y_i)$ , where  $\Psi(\cdot, y)$  is convex. We give a characterization of such criteria that allow the universal consistency property, in terms of geometric properties of the convex hull of the image of  $\Psi$ . We describe the implications for several multiclass methods from the literature. (Joint work with Ambuj Tewari.)

**Further references:** See

<http://www.cs.berkeley.edu/~ambuj/research/tewari05consistency.pdf>

Poster session	Wednesday, May 31, 18:00 – 19:30
Felix Agakov	Variational Information-Maximization and Probabilistic Self-Supervised Training
Pierre Alquier	Iterative Feature Selection in Regression and Density Estimation
Maria-Florina Balcan	On a Theory of Kernels as Similarity Functions
Matthieu Cornec	Distribution-free Finite Sample Results for Generalized Cross-validation
András György	The Shortest Path Problem with Limited Feedback
Matthias Hein	Uniform Convergence of Adaptive Graph-based Regularization
H. X. Liu	Making Markowitz's Portfolio Optimization Theory Practically Useful
Sébastien Loustau	Learning Rates for Support Vector Machines Using Sobolev Spaces
Leila Mohammadi	The Nonnegative Garrote Estimator in Classification
Tsuyoshi Okita	Kernel Methods for Conditional Quantities
György Ottucsák	Hannan Consistency in On-line Learning in Case of Unbounded Losses under Partial Monitoring
Kristiaan Pelckmans	Risk Scores and its Use in Censored Regression
Evgeniy Rafikov	About Universal Estimators in the Case of Unbounded Responses
Nima Reyhani	Noise Variance Estimation, Difficulties and Applications
Manuel Samuelides	Learning Surrogate Models for Optimization and Applications to Structure Optimization
Dina Anna Sudarsky-Guez	Geometry of Excursion Sets of the Non-stationary Elliptic Gaussian Parabolic Bending Non-isotropic Scale Space Random Fields
Csaba Szepesvári	Learning Near-optimal Policies with Bellman-Residual Minimization Based Fitted Policy Iteration and a Single Sample Path
Yiming Ying	Online Gradient Descent Learning Algorithms
Gilbert Young	Complexity of Tetris-Packing Problem

---

## Variational Information-Maximization and Probabilistic Self-Supervised Training Felix Agakov (University of Edinburgh)

In this work we investigate a relation between the variational Arimoto-Blahut (Information Maximizing) algorithm for the channel capacity of *encoder* models and the variational EM for generative models and *stochastic* autoencoders. Much of the previous work on relating maximization of the mutual information, likelihood, and conditional likelihood in such models focused primarily on specifically constrained invertible mappings (e.g. Cardoso, 1997; MacKay, 1999) or specific noiseless autoencoders (Oja, 1989), where the computations were exact. Our goal here was to investigate relations between these learning paradigms for more general graphical models, which could arguably be more practical for describing real-world communication channels or data-generating processes. Since in our case the optimized objectives were generally computationally intractable, we considered their common variational relaxations. Our focus on the variational EM and IM was due to popularity and simplicity of these approaches for approximate training of generally intractable graphical models.

Our study here was motivated by a simple observation that independently of a specific (generative, self-supervised, or encoding) modeling approach, we may often be interested in finding latent variable representations  $y$  which are somewhat informative about the observations  $x$ . One possible principled way to relate such approaches could be by comparing the induced optimization surfaces with (the bounds on) the information content between  $x$  and  $y$ , under specific modeling assumptions. Our assumption here is that the (exact or approximate) posteriors of the considered models lie in the same parametric families; i.e. for the fixed parameters of the posteriors, the models should lead to identical inferences.

For example, in the first part of our study we assume that the encoding distribution  $p(y|x)$  of the encoder model  $\mathcal{M}_I \stackrel{\text{def}}{=} p(y|x)\tilde{p}(x)$  is constrained to be equivalent to the posterior of a generative model  $\mathcal{M}_L \stackrel{\text{def}}{=} p(y)p(x|y)$ , or its variational approximation  $q(y|x)$  (in the latter case,  $q(y|x)$  is the variational parameter of the Jensen's lower bound on the likelihood of  $\mathcal{M}_L$ ). A simple comparison of optimization surfaces of the variational EM for  $\mathcal{M}_L$  and the variational IM for  $\mathcal{M}_I$  gives rise to a sufficient condition for equivalence of both learning approaches. Another simple outcome of this comparison is the observation that under the considered parameterization, the Jensen's lower bound on the likelihood in  $\mathcal{M}_L$  defines a (potentially loose) lower bound on the mutual information  $I(x, y)$  in  $\mathcal{M}_I$ , which may be strengthened by a specific tractable instance of the IM bound. In the second part of our study we analyze the fixed points of the variational EM for a stochastic autoencoder  $\mathcal{M}_C$ , and show that the variational EM learning in  $\mathcal{M}_C$  is identical to a special instance of the IM learning in the corresponding encoder model (again, the encoder is assumed to satisfy the equivalent inference constraint). One of the practical results of the study is a simplification of the conditional self-supervised training of noisy autoencoders, as the equivalent IM formulation has a simpler form with the lower cardinality of the variational parameters.

Finally, we note that availability of simple and general lower bounds on  $I(x, y)$  (such as those based on conditional likelihood of autoencoders) may alleviate a need of its approximations, which may often be accurate only under strong assumptions about the encoder models. We briefly demonstrate that although it is easy to strengthen Jensen's lower bounds on  $I(x, y)$  (as it is on the marginal or conditional likelihood), even simple choices of the variational distributions may lead to better estimates of  $I(x, y)$  than some of the common (Brunel and Nadal, 1998) or more recent (Corduneanu and Jaakkola, 2003) approximations.

### Further references:

The extended version is available at [http://homepages.inf.ed.ac.uk/felixa/Papers/info\\_t03.pdf](http://homepages.inf.ed.ac.uk/felixa/Papers/info_t03.pdf).

**Iterative Feature Selection in Regression and Density Estimation**  
**Pierre Alquier (University Paris 6 & Crest)**

In this work, we give an iterative algorithm for regression estimation in the transductive setting (the technique described also covers regression estimation in the inductive setting and density estimation, both with quadratic loss). We assume that we observe a learning sample  $(X_1, Y_1), \dots, (X_N, Y_N)$  (i.i.d. from a distribution  $P$  on  $\mathcal{X} \times \mathbf{R}$ , there is no assumption on  $\mathcal{X}$ ) and the design of a test sample,  $X_{N+1}, \dots, X_{2N}$ , our aim being to estimate  $Y_{N+1}, \dots, Y_{2N}$ . We assume that we are given a large set of feature, more precisely  $m$  functions  $f_1, \dots, f_m$  from  $\mathcal{X}$  to  $\mathbf{R}$ , with  $m \geq N$ . The aim of the method is to select functions in this set that are relevant to estimate  $Y_{N+1}, \dots, Y_{2N}$  and to aggregate them - estimating the  $Y_{N+i}$  by  $f(X_{N+i})$  where:

$$f(\cdot) = \sum_{k \in I} \alpha_k f_k(\cdot). \quad (1)$$

Focusing on the least square error:

$$r(f) = \frac{1}{N} \sum_{i=N+1}^{2N} [Y_i - f(X_i)]^2$$

we give in a first time a deviation inequality providing control for  $r(\hat{\alpha}_k f_k)$ , the risk of an estimator in a unidimensional model defined by the function  $f_k$ . Actually, each of these inequalities provides a confidence region for  $\bar{f}$ , the minimizer of  $r$  under the form given by equation (1). We propose to perform successive projections on such confidence regions using a suitable scalar product, and describe a practical algorithm to do it. Every projection adds (at most) one feature to the estimator.

The result we prove is that with large probability, at each projection step, the performance of the current estimator (measured by  $r$ ) is actually improved, providing a guarantee against overlearning.

We focus on two particular examples. In the case where  $m = 2N$  and  $f_i = K(X_i, \cdot)$  for some kernel  $K$  we obtain a SVM estimator. In the inductive setting (for regression or density estimation), when the features are actually an orthogonal basis of functions, this algorithm is equivalent to a soft-thresholding procedure. If the true regression function belongs to a space of unknown regularity  $\beta$ , this method adapts to this regularity and reaches the right speed of convergence - up to a log factor:  $(\log N/N)^{2\beta/(2\beta+1)}$ .

**Further references:**

The extended version of this work is available on

[http://www.crest.fr/pageperso/alquier/alquier\\_eng.htm](http://www.crest.fr/pageperso/alquier/alquier_eng.htm).

**On a Theory of Kernels as Similarity Functions**  
**Maria-Florina Balcan (Carnegie Mellon University)**

Kernel functions have become an extremely popular tool in machine learning, with an attractive theory as well. This theory views a kernel as performing an implicit mapping of data points into a possibly very high dimensional space, and describes a kernel function as being good for a given learning problem if data is separable by a large margin in that implicit space. However, while quite elegant, this theory does not directly correspond to one's intuition of a good kernel as a good similarity function. Furthermore, it may be difficult for a domain expert to use the theory to help design an appropriate kernel for the learning task at hand since the implicit mapping may not be easy to calculate. Finally, the requirement of positive semi-definiteness may rule out the most natural pairwise similarity functions for the given problem domain.

In this work we develop an alternative, more general, theory of learning with similarity functions (i.e., sufficient conditions for a similarity function to allow one to learn well) that does not require reference to implicit spaces, and does not require the function to be positive semi-definite (or even symmetric). Our results also generalize the standard theory in the sense that any good kernel function under the usual definition can be shown to also be a good similarity function under our definition (though with some loss in the parameters). In this way, we provide the first steps towards a theory of kernels that describes the effectiveness of a given kernel function in terms of natural similarity-based properties.

**Further references:**

The extended version of this work is available on  
<http://www.cs.cmu.edu/~ninamf/papers/similarity.ps>.

## Distribution-free Finite Sample Results for Generalized Cross-validation Matthieu Cornec (Université Paris X)

In this article, we derive finite-sample results for the cross-validation estimate of the generalisation error in the context of risk assessment and model selection. In the general setting, we prove sanity-check bounds in the spirit of Kearns (1997) “bounds showing that the worst-case error of this estimate is not much worse than that of training error estimate”. For special algorithms (namely subbagging), our bounds can be much tighter than Vapnik’s bounds especially in the case of small sample sizes.

General loss functions and class of predictors with both finite and infinite VC dimension are considered. We generalize slightly the formalism introduced by Dudoit and Van der Laan (2003) to cover a large variety of cross-validation procedures including leave-one-out cross-validation,  $k$ -fold cross-validation, hold-out cross-validation (or split sample), leave- $v$ -out cross-validation and the resubstitution error.

In particular, we focus on :

- proving the accuracy of the various cross-validation procedures,
- pointing out the interest of each cross-validation procedure in terms of rate of convergence. In particular, the special interest of the  $k$ -fold cross-validation allowing both a low bias and a low variance is emphasized. An estimation curve with transition phases depending on the cross-validation procedure is derived. It gives a simple rule on how to choose the cross-validation method knowing the loss function, the class of predictors and the sample size. The conclusions about the optimal splitting procedure are different from previous ones (Kearns, 1998).
- showing when the cross-validation estimates can outperform the training estimate,
- proving that cross-validation can work out with infinite VC-dimension predictor,
- at last, providing simulation studies to illustrate our results.

**The Shortest Path Problem with Limited Feedback****András György (Computer and Automation Research Institute of the Hungarian Academy of Sciences)**

(Joint work with Tamás Linder, Queen's University, Gábor Lugosi, Pompeu Fabra University, György Ottucsák, Budapest University of Technology and Economics)

The on-line shortest path problem is considered with limited feedback. At each round, a decision maker has to choose a path between two distinguished vertices of a weighted directed acyclic graph whose edge weights can change in an arbitrary (adaptive adversarial) way such that the loss of the chosen path (defined as the sum of the weights of its composing edges) be small. In the multi-armed bandit setting, after choosing a path, the decision maker learns only the weights of those edges that belong to the chosen path. For this scenario, an algorithm is given whose average cumulative loss in  $n$  rounds exceeds that of the best path, matched off-line to the entire sequence of the edge weights, by a quantity that is proportional to  $1/\sqrt{n}$  and depends only polynomially on the number of edges of the graph. The algorithm can be implemented with linear complexity in the number of rounds  $n$  and in the number of edges. The main idea in constructing the algorithm is to modify the bandit algorithm of Auer et al. [1] such that one needs to keep weights for the edges of the graph instead of the paths, and these weights are combined efficiently to choose a path in each round. This result improves earlier bandit-algorithms which have performance bounds that either depend exponentially on the number of edges or converge to zero at a slower rate than  $O(1/\sqrt{n})$ . An extension is given for the tracking problem, where the performance of the algorithm is compared to the performance of dynamic paths that can switch between fixed paths several times. Another extension to the so-called label efficient setting is also given, where the decision maker is informed about the edge weights of the chosen path only with probability  $\varepsilon < 1$ . Applications to routing in packet switched networks along with simulation results are also presented.

**Further references:**

The extended version of this work is available at

<http://www.szit.bme.hu/~gya/publications/GyLiLu0t06.pdf>.**References**

- [1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

**Uniform Convergence of Adaptive Graph-based Regularization**  
**Matthias Hein (Max Planck Institute for Biological Cybernetics)**

The regularization functional induced by the graph Laplacian of a random neighborhood graph based on the data is adaptive in two ways. First it adapts to an underlying manifold structure and second to the density of the data-generating probability measure. We identify in this paper the limit of the regularizer and show uniform convergence over the space of Hölder functions. As an intermediate step we derive upper bounds on the covering numbers of Hölder functions on compact Riemannian manifolds, which are of independent interest for the theoretical analysis of manifold-based learning methods. We include also an extensive discussion of the properties of the limit smoothness functional and why and how it can be interesting in different learning algorithms such as regression, semi-supervised learning and clustering. In particular the adaptation to the two independent structures inherent to the data, the geometry of the data manifold  $M$  and the density of the data generating probability measure, is discussed.

**Further references:**

The extended version of this work is available on

[http://www.kyb.mpg.de/publications/attachments/Hein-Unif\\_Conv\\_Graph\\_Reg\(2006\)\\_3893\[1\].pdf](http://www.kyb.mpg.de/publications/attachments/Hein-Unif_Conv_Graph_Reg(2006)_3893[1].pdf).

## **Making Markowitz's Portfolio Optimization Theory Practically Useful**

**H. X. Liu (National University of Singapore)**

(Joint work with Zhidong Bai and Wing-Keung Wong)

The Markowitz mean-variance optimization procedure to compute the optimal return is highly appreciated as a one of the most important cornerstones in modern finance theory. However, the traditional estimated return has been demonstrated not to be applicable in practice due to its serious departure from its theoretic optimal return, attributed to the substantial measurement error. Applying the theory of large dimensional data analysis, we first theoretically explain this phenomenon is natural when the number of assets is large. We also show that the huge measurement error is due to the serious departure of the estimated portfolio from its theoretic counterpart. Thereafter, we prove that the estimated optimal return is always larger than its theoretic parameter when the number of assets is large. To circumvent this problem, we utilize both the large dimensional random matrix theory and the parametric bootstrap method to develop new bootstrap estimators for the optimal return and its asset allocation. We further theoretically prove that these bootstrap estimates are consistent to their counterpart parameters. Our simulation confirms the consistency and shows that, comparing with the traditional estimate, our proposed estimate improves the estimation accuracy so substantial that its relative efficiency is as high as 220 times for sample size of 500; implying that the essence of the portfolio analysis problem could be adequately captured by our proposed estimates. The improvement of our proposed estimates are so big that there is a sound basis for believing our proposed estimates to be the best estimates to date that they greatly enhance the Markowitz mean-variance optimization procedure to be implementable and practically useful.

**Learning Rates for Support Vector Machines Using Sobolev Spaces**  
Sébastien Loustau (C.M.I., Marseille)

The goal of the present poster is to rely ideas from linear algebra and approximation theory with Learning. We state learning rates to the Bayes risk approaching  $O(1/n)$  for SVMs using a new type of kernel. This kernel will be described in terms of the spectrum of his integral operator.

For the kernel approximation, we establish polynomial rates depending on the regularity of the kernel function. We deduce some theoretical estimates for the approximation error.

For the stochastic part of this analysis, we use tools of concentration theory and local Rademacher averages. Recent advances on concentration inequalities improve Vapnik's structural risk minimization for pattern recognition. More precisely, we related the eigenvalues of the integral operator to his entropy numbers to deduce a structural control of the RKHS considered. Gathering this with the now standard Tsybakov's noise assumption, our results shows that it is possible to obtain fast rates of convergence depending on our eigenspectrum associated to the kernel. The dependence is explicitly given and is compared with recent results on this topic.

**The Nonnegative Garrote Estimator in Classification**  
Leila Mohammadi (EURANDOM, Technical University of Eindhoven)

Subset selection regression is a frequently used statistical method. Suppose we are given data of the form  $\{(y_n, x_{1n}, \dots, x_{Mn}), n = 1, \dots, N\}$ . Subset selection waives some of the predictor variables  $x_1, \dots, x_M$  and then the prediction equation for  $y$  is based on the remaining set of variables. Subset selection is simple and it clearly reduces the variance if  $M$  is large. An other method for reducing the variance is ridge regression. In this method we assume  $\lambda$  to be a positive value (shrinkage parameter) and the coefficients are estimated by  $(X^T X + \lambda I)^{-1} X^T Y$ . Let  $y = \sum_k \beta_k x_k + \varepsilon$ . If a few of the  $\{\beta_k\}$  are nearly zero and the rest are large, then subset selection gives more accurate prediction than ridge regression. If it is not the case, then ridge regression acts better. Thus usually, subset selection is not as accurate as ridge. The problems with ridge regression are for example: 1) it is not scale invariant 2) it does not give a simple equation. As it is known, we need an intermediate method which selects subsets, is stable and gains its accuracy by selective shrinking. The nonnegative garrote estimator in a linear regression model was introduced by Breiman (1995). This is an intermediate method.

We define the nonnegative garrote estimator in a binary classification problem. We obtain the rate of convergence of the risk of the estimator which is  $(\log n)/n$ . It shows that the same asymptotic convergence rate as the hard and soft shrinkage estimates holds for the nonnegative garrote estimator.

**Further references:**

A paper related to this work is available on

<http://www.math.leidenuniv.nl/%7Eleila/publicat.html>.

**Kernel Methods for Conditional Quantities**  
**Tsuyoshi Okita (Vrije Universiteit Brussel)**  
(Joint work with Bernard Manderick)

Kernel methods have rapidly become popular in the last decades. Despite the easy usage without prior knowledge about data, they can handle vast categories of learning problems with high accuracies (Shawe-Taylor and Cristianini, 2004; Scholkopf and Smola, 2002; Herbrich, 2001; Joachims, 2002; Vapnik, 1998). Easy usage is due to the IID assumption on the data generation process which lets us free from annoying about which assumptions we need to make on a data generation process which is often required for a Bayesian approach (MacKay, 2003; Neal, 2005; Jordan and Bishop, 2000; Ghahramani, 2005; Korb and Nicholson, 2004; Neapolitan, 2003; Koller and Friedman, 2003). An intuitive concern is that if we may employ some prior knowledge about data which contradicts the IID assumption, we could get better test accuracies: indeed this is the Bayesian argument towards kernel methods. As a primal source of prior knowledge, Bayesian incorporates the correlations among random variables, which can be described by the conditional probabilities  $P(X_i|X_j)$ , and uses the priors (prior belief), both of which bases on the Bayes rule. The similar situation occurs as well in a different context: the context of reduction approach (based on information theory) (Dietterich and Bakiri, 1995; Langford and Beygelzimer, 2005) and that of an on-line learning (Cesa-Bianchi et.al., 2005). In a reduction approach after reducing a problem into subproblems, the target bits of error correction is the bits which are likely to be modified by a noisy channel, which can be discribed by the conditional entropy  $H(X_R|Y_S)$ . Similarly in on-line learning, if we see the random variables along a time sequence, the natural strategy how to choose the next time step can be described using the conditional expectation  $E(X_{n+1}|X_n, \dots, X_1)$ , which is related to the martingale for gambling. The interesting similarities among Bayesian, on-line learning, and reduction approach are in their common usage of conditional quantities (conditional probability / expectation / entropy).

This poster shows one approach to handle those conditional quantities by kernel methods. The key ingredient is to incorporate an order statistics to represent the exchangeability principle (de Finetti, 1939), where this principle states that the conditionally independent random variables are infinitely exchangeable. The advantage of this kernel method over the Bayesian approach is that the kernel method can guarantee the existence and uniqueness of solutions by RKHS. It is noted that the non IID assumption is half explained by this preliminary / ongoing study in terms of conditional quantities (other factor is the nonidentical assumption).

**Hannan Consistency in On-line Learning**  
**in Case of Unbounded Losses under Partial Monitoring**  
György Ottucsák (Budapest University of Technology and Economics)  
(Joint work with László Györfi)

In most of the machine learning literature, one assumes that the losses are bounded, and such a bound is known in advance, when designing an algorithm. In many applications, including regression problem or routing in communication networks the bound of the loss function is not known beforehand or the loss function is unbounded.

In this paper the sequential prediction problem with expert advice is considered when the bound of the loss function is unknown in advance, or when the loss function is unbounded. We analyze a simple modification of Allenberg and Auer's "bandit" algorithm. The modified algorithm is able to handle a wider class of the partial monitoring problems: the combination of the label efficient and multi-armed bandit problem, that is, where the algorithm is informed about the performance of the chosen action only with probability  $\varepsilon < 1$ .

We prove that Hannan consistency, a fundamental property in game-theoretic prediction models, can be achieved by the algorithm in full monitoring and label efficient case if the maximum of the normalized average of the square of the loss to grow infinity slightly slower than linearly in number of prediction rounds. We also determine the sufficient conditions of Hannan consistency under other partial monitoring cases: in the multi-armed bandit problem and in the combined version with label efficient setting. In the proof we avoid using doubling trick (epochs) although we are still able to trace the range of the value of the loss function and to handle the infinite time-horizon.

**Further references:**

The extended version of this work is available on <http://www.szit.bme.hu/~oti>.

**Risk Scores and its Use in Censored Regression**  
**Kristiaan Pelckmans (K.U. Leuven, ESAT/SCD-SISTA, Belgium)**  
 (Joint work with J.A.K. Suykens and B. De Moor)

This work explores the use of risk scores into learning theory and for the design of learning machines in the context of censored observations. The notion of risk scores relates to the classical notion of likelihood scores as lying on the basis of maximum likelihood inference [3], but is conceived instead in a context of empirical risk minimization theory. Given a loss function  $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$  belonging to  $\mathcal{C}^1(\mathbb{R}^d)$  which measures the appropriateness of a certain parameter vector  $\theta \in \mathbb{R}^d$  (for  $d \in \mathbb{N}_0$ ) in the context of a fixed but unknown distribution  $F_Z$ . The corresponding risk score  $r : \mathbb{R}^d \rightarrow \mathbb{R}$  in  $\mathcal{C}^0(\mathbb{R}^d)$  can be formalized as follows

$$r(\theta_0; F_Z) = E \left[ \frac{\partial \ell(\theta; F_Z)}{\partial \theta} \Big|_{\theta=\theta_0} \right].$$

Theoretical properties as e.g. existence and continuity are derived, while the finite sample properties of the empirical counterpart can be quantified exploiting the linearity of the involved operators. Risk minimization then amounts to finding a parameter vector  $\theta$  such that  $r(\theta; F_Z) = 0$ . We investigate properties of the empirical counterpart of the risk score in case of a number of loss functions. In our case the risk score is the random variable of interest, instead of the risk terms itself.

A first result motivating the use of risk scores in a distribution-free setting is obtained by the construction of a nontrivial confidence set based on an i.i.d. sample  $Z_i \sim F_Z$  with  $i = 1, \dots, n$ :

$$\mathcal{S}_{\alpha, n} \subset \mathbb{R}^d : Pr(\theta^* \in \mathcal{S}_{\alpha, n}) \geq \alpha \quad \text{s.t.} \quad r(\theta^*, F_Z) = 0,$$

where  $0 < \alpha < 1$  is a prespecified confidence level and  $n$  denotes the sample setting. This derivation is based on large deviation bounds and application of the union bound technique. For convex loss-functions  $\ell$ , an analytical expression can be derived for the distribution-free confidence set of a location estimator. This construction of confidence sets relates closely to the likelihood based counterparts of score tests [3]. The close links with the Fisher information, recent stability based results [1], perturbation theory and topics in robust inference [2] are highlighted.

Risk scores prove to be useful in order to develop a theoretical motivated strategy to deal with censored observations in location estimation or function approximation. Estimation problems involving censored observations occur often in econometrics or in survival analysis for clinical studies. This problem is classically approached with a Tobit model based on parametric likelihood, Powell's LAD estimator, and it is also the subject of the empirical likelihood approach of the Kaplan-Meier product limit estimator, see e.g. [4]. Empirical risk scores can be used to approach the problem of an appropriate estimation technique in the context of censoring from a different angle. Herefor, we exploit the property of bounded empirical risk scores in case a Lipschitz smooth loss function is used. The convexity of the resulting problem is exploited to construct a corresponding confidence set.

## References

- [1] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.
- [2] A. Christmann and I. Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *JMLR*, 5:1007–1034, 2004.
- [3] R.V. Hogg, A.T. Allen. *Introduction to Mathematical Statistics*. Macmillan, New York, 5th edition, 1995.
- [4] J.D. Kalbfleisch, R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley series in probability and statistics. Wiley, 2002.

## About Universal Estimators in the Case of Unbounded Responses Evgeniy Rafikov (Moscow State University)

Learning from examples refers to a process that builds on the base of available data  $x_i$  and outputs  $y_i, i = 1, \dots, m$  a function that best represents the relation between the inputs  $x \in X$  and the corresponding outputs  $y \in Y$ . Let  $X \in \mathbb{R}^d, Y \in \mathbb{R}$  be Borel sets and  $\rho$  be a Borel probability measure on  $Z = X \times Y$ . Consider  $\rho(y|x)$  – conditional (with respect to  $x$ ) probability measure on  $Y$  and define the regression function of  $\rho$  as usual

$$f_\rho(x) = \int_Y y d\rho(y|x).$$

A lot of authors studied the problem of estimating  $f_\rho(x)$ , see, e.g., [1, 2, 3, 4]. The typical assumption for the unknown measure  $\rho$  is that  $f_\rho(x) \in W$  for some functional class  $W$  and  $|y| < M$  a.s. for some constant  $M$ . We discuss several aspects of estimating  $f_\rho(x)$  in the case of unbounded outputs, i.e. when the condition  $|y| < M$  is replaced by some rates of convergence of  $\rho(y|x)$ -probability tails of  $y$ . For example we study the case of uniform (with respect to  $x$ ) subgaussianness of  $y$ .

## References

- [1] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics, 2002.
- [2] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1-49, 2001.
- [3] R. DeVore, G. Kerkycharian, D. Picard, V. Temlyakov. Mathematical methods for supervised learning. *IMI Preprints* 22:1-24, 2004.
- [4] V. Temlyakov. Approximation in learning theory. *IMI Preprints* 05:1-49, 2005.

**Noise Variance Estimation, Difficulties and Applications**  
**Nima Reyhani (Helsinki University of Technology)**  
(Joint work with Amaury Lendasse)

In field of machine learning or signal processing, usually we assume that the data set is contaminated by some additive noise where the noise is independent to the design points. Let's consider input variable  $\mathbf{x} \in [0, 1]^M$ , for some fixed  $M$ , and the output or response  $y \in \mathbb{R}$ . Suppose the relation between  $\mathbf{x}$  and  $y$  satisfies  $y_i = g(\mathbf{x}_i) + \varepsilon$ , where  $\varepsilon$  is the noise variable. We also assume  $|\mathbf{x}_i - \mathbf{x}_j| = \mathcal{O}(\frac{1}{N})$ ,  $\forall i, j = 1, \dots, N$ . The problem is estimation of the  $\text{Var}\{\varepsilon\}$ .

Referring to the law of large numbers, we can state that, in general, the noisy signal,  $y$ , is more gaussian/less smooth than the original signal  $g(\cdot)$ . Now, let's consider the regression function  $g(\cdot) \in \mathcal{A}(\lambda)$ , where  $\mathcal{A}(\lambda)$  is the class of functions having smoothness not less than  $\lambda$ . Also, assume that this class of functions can be estimated without error by a class of regression functions belonging to  $\mathcal{G}(\lambda)$ , e.g. the class of polynomials of some fixed order  $p$ . Then, the local residuals between the approximator belonging to  $\mathcal{G}(\lambda)$  show the noise at that point. Thus, a priori on the smoothness of the regression function can be used to determine the noise signal.

Furthermore, we assume that the underlying function in an open ball  $\mathcal{B}(\varepsilon)$  with  $\varepsilon \rightarrow 0$  is locally smooth (the second term in Taylor expansion is negligible). Therefore, again, given that the design points  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are sufficiently close to each other, we have  $\frac{1}{2}(g(\mathbf{x}_i) - g(\mathbf{x}_j))^2 \rightarrow \text{Var}\{\varepsilon\}$ . The expression converges to the noise variance under the condition mentioned in the beginning, if a similarity measure for the input space, like kernels, be involved in the computations as well. Now, let's turn into applications of such estimates.

In function approximation, the goal is to find the function  $F(\mathbf{x}; \mathbf{w})$  which can model the data set  $\{\mathbf{x}_i, y_i\}_{i=1}^N$  by selecting the proper values of  $\mathbf{w}$ . The goal is to minimize the functional Mean Square Error (MSE), i.e.  $\min_{\mathbf{w}} \sum_{i=1}^N (g(\mathbf{x}_i) - F(\mathbf{x}_i; \mathbf{w}))^2$ . The MSE can be decomposed to  $\varepsilon^2 + (g(\mathbf{x}) - F(\mathbf{x}; \mathbf{w}))^2$ . Therefore, the noise variance estimate indicates to the minimum MSE that can be achieved from a given data set without overfitting. We can apply the noise variance estimate for the purpose of model selection in such a way that the MSE of trained model with fixed hyper-parameters, e.g. regularization parameter in support vector regression (SVR) or the number of hidden neurons in Multi Layer Perceptron (MLP), should not get below the noise variance. Moreover, one can apply the noise variance in order to implement variable selection or variable scaling. Also, the noise variance provides a measure to define the predictability of a data set. In other words, the noise variance indicates the possibility of finding an efficient model which best fits to the given data set. Moreover, it can be applied as indicator for stopping the training/optimization procedure.

In summary, first of all, based on the smoothness assumptions and independency between the noise and design distribution, it is possible to derive estimators of the noise variance. Obviously, due to the assumptions about the smoothness of the regression function, the estimation is biased, e.g. bias of kernel width or, in general, bias of the set  $\mathcal{G}(\lambda)$ . Without any assumption about the regression function smoothness it is impossible to estimate the noise variance. This makes the major difficulty in deriving the noise variance estimations.

**Further references:**

The extended version of this work will be on <http://www.cis.hut.fi/nreyhani/MFLT2.pdf> by June 2006.

## Learning Surrogate Models for Optimization and Applications to Structure Optimization

Manuel Samuelides (SUPAERO and ONERA/DTIM)  
(Joint work with Antoine Merval, Transiciel and ONERA/DTIM)

Optimization is a field of great interest in aeronautical world, since a decrease of a plane weight both allows a decrease in fuel consumption and an increase of fly range. Actually, such an optimization process is often hard to solve as it possibly handles a very large number of variables. For instance, optimization of a whole fuselage is an iterative, time consuming process solved as a bilevel optimization: indeed, as a direct complete optimization process is out of scope as it would handle too many design variables, the structure to be optimized is divided into elementary parts that are optimized independently under mechanical constraints considering current internal loads and then put together to compute the load distribution in the new complete structure. Nowadays, this process is run using analytical dedicated software to compute stability criteria at local level and finite element analysis to compute load distribution in the structure at global level. Moreover, as the constraints values are supplied by a black box model, their gradients have to be computed by finite difference, increasing the number of calls to the software and consequently the time spent in local optimization processes. The reduction of computational complexity is a key issue of engineering research. A possible way to achieve this goal is to use surrogate models. Then computation time is significantly reduced as constraint values are expressed as analytical approximate formulations instead of software computation. Furthermore, gradients can be easily access by differentiating metamodel formulation.

In that framework, we used non linear response surface methodology (RSP) combined with DACE (Design of Assisted Computer Experiments) to perform regression of static mechanical criteria that are constraints of the local optimization problems. We chose Artificial Neural Networks as non linear models for their universal approximation property. We encountered some difficulties to get a precise approximation on the whole design space. Then, we show how the use of Mixture of Experts (MoE) methodology, which is based upon design variable smoothness prior knowledge allows us to cope with this problem. So we are able to build efficient surrogate models of local optimization constraints on two application cases that are buckling and collapse of metallic and composite stiffened panels. We then embed successfully these response surface models in local optimization benchmarks.

The drawbacks of our study is the prior knowledge that was used to divide the space into regions of interest to perform elementary non linear regression (individual expert). For more complex optimization problems with poor prior knowledge about constraints switching boundaries, an automatic method is expected to achieve the localization of experts. Several algorithms exist: fuzzy decision trees, local estimation of approximation error using kriging or bootstrapped regressors. Future work will be dedicated to test these methods on more complex design state (20 variables and more).

### Further references:

The extended version of this work will be published in proceedings of 2nd AIAA Multidisciplinary Design Optimization Specialist Conference.

This work is done in collaboration with Stéphane Grihon (Airbus, Engineering Structure Analysis). It is supported by Vivace Project of European community.

## Geometry of Excursion Sets of the Non-stationary Elliptic Gaussian Parabolic Bending Non-isotropic Scale Space Random Fields

Dina Anna Sudarsky-Guez (Universités Paris 13 et Paris 7)

I am interested in the geometric properties of real-valued nonstationary and nonisotropic  $(N, d)$  random fields  $X = \{X(t, \Upsilon) : t \in S_{loc} \subset \mathbb{R}^D, \Upsilon \in S_p \subset \mathbb{R}^{N-D}\}$  defined on a parameter space  $S_p$ : a subset of  $D$ -dimensional Euclidean space, for e.g.;  $D \geq 2, d = 1$ . My first interest in these fields focuses on their excursion sets: The set of points of the field with a value which exceeds a fixed threshold  $\mathcal{T} \in \mathbb{R}$ ,  $\mathcal{A}_{\mathcal{T}} = \mathcal{A}_{\mathcal{T}}(X(t, \Upsilon), S_{loc} \times S_p) = \{(t, \Upsilon) \in S_{loc} \times S_p \subset \mathbb{R}^N : X(t, \Upsilon) \geq \mathcal{T}\} = X^{-1}[\mathcal{T}, \infty)$  and their geometric characteristics which are amenable to statistical analysis [4]. In particular the Differential Topology (DT) or absolute Euler characteristics:  $\Psi = \Psi(\mathcal{A}_{\mathcal{T}}(X(t, \Upsilon), S_{loc} \times S_p)) = \sum_{i=0}^N (-1)^i \mu_i$  with  $\mu_i(X, \mathcal{A}_{\mathcal{T}}) = \#C_i((X, \mathcal{A}_{\mathcal{T}}))$  and  $C_i$  = the critical points of  $X$  of index  $i$ ; as well as the relative Euler characteristics,  $\Psi = \sum_{k=0}^N \sum_{J \in \mathcal{J}_k} \sum_{i=0}^k (-1)^i \mu_i(J)$  with  $\mathcal{J}_k = \partial_k S_p, \mu_i(J) = \#\{t \in J, \nabla X(t) = 0\}$ . The DT characteristics cannot see boundary events and are identical to the EC if  $\mathcal{A}_{\mathcal{T}} \cap \partial S_p = \emptyset$ . The field  $X(t, \Upsilon)$  is taken as a real-valued function of class  $c^2$ , admissible relative to a regular compact  $c^2$  domain with  $c^2$  boundary also called Morse function. The geometry of the stationary and isotropic Gaussian random fields on Euclidean space and on manifolds has been well developed by Adler [1, 2], and Adler and Taylor [3, 4] through the relation  $P\{\sup_{t \in S_{loc}} X(t) \geq \mathcal{T}\} \approx E\{\Psi[X^{-1}(\mathcal{T}, \infty)]\}$  also valid for nonstationary and nonisotropic NSNI random fields and for particular situations by Worsley [5].

The aim of the present work is to extend these results to the statistical analysis of  $\Psi$  for a NSNI smooth Gaussian and related random field excursion set on  $S_{loc} \times S_p$  and especially to develop an *explicit* expression for the expected Euler characteristic,  $E\{\psi(\mathcal{A}_{\mathcal{T}}(X(t, \Upsilon), S_{loc} \times S_p))\}$ , to approximate the number of local maxima of the NSNI random field with  $\psi(\mathcal{A}_{\mathcal{T}}(X(t, \Upsilon), S_{loc} \times S_p))$ , and the distribution of the global maximum of  $X(t, \Upsilon)$ ,  $P\{\sup_{(t, \Upsilon) \in S_{loc} \times S_p} X(t, \Upsilon) \geq \mathcal{T}\}$  with  $E\{\psi(\mathcal{A}_{\mathcal{T}}(X(t, \Upsilon), S_{loc} \times S_p))\}$  above high threshold  $\mathcal{T}$ .

Part of the motivations for this comes fundamentally from applications to geo-statistics, astrophysics (Guttorp and Sampson, 1992), statistics of medical images (Cao and Worsley, 1997-1999), to image processing and analysis; as signal detection and extraction (Siegmund and Worsley, 1995; Shafie and Worsley, 2003) which is in fact a well known statistical problem.

The purposes of this and pursuing researches are to establish the relation between the expected Euler characteristic of the NSNI elliptic-Gaussian parabolic-bending non-isotropic scale space  $(\frac{D}{2}(D+5) - 1, d)$  random field excursion sets above high  $\mathcal{T}$  and  $P\{\sup_{(t, \Upsilon) \in S_{loc} \times S_p} X(t, \Upsilon) \geq \mathcal{T}\}$  and to a non lesser extent to apply the method to the problem of searching for multiple sclerosis lesion (MS), evolution and structural changes in brain white matter image sequences obtained by magnetic resonance imaging (MRI).

## References

- [1] R.J. Adler. *The Geometry of Random Fields*. Wiley, New York, 1981.
- [2] R.J. Adler. On excursion sets, tube formula and maxima of random fields. *The Annals of Applied Probability*, 10(1):1–74, 2000.
- [3] J.E. Taylor and R.J. Adler. Euler characteristics for Gaussian fields on manifolds. *The Annals of Probability*, 31(3):533, 2003.
- [4] R.J. Adler and J.E. Taylor. *Random Fields and Geometry*. Birkhauser, 2005.
- [5] K.J. Worsley. Boundary correction for the expected euler characteristic of excursion sets of random fields fields, with an application to astrophysics. *Advances in Applied Probability*, 27:943–959, 1995.

## Learning Near-optimal Policies with Bellman-Residual Minimization Based Fitted Policy Iteration and a Single Sample Path

Csaba Szepesvári (MTA SZTAKI, Hungary)

(Joint work with András Antos, MTA SZTAKI, and Remi Munos, Ecole Polytechnique)

Consider the problem of optimizing a controller for an industrial environment. In many cases the data is collected on the field by running a fixed controller and then taken to the laboratory for optimization. The goal is to derive an optimized controller that improves upon the performance of the controller generating the data.

Here we are interested in the performance improvement that can be guaranteed given a finite amount of data. In particular, we are interested in how performance scales as a function of the amount of data available. We study Bellman-residual minimization based policy iteration assuming that the environment is stochastic and the state is observable and continuous valued. The algorithm considered is an iterative procedure where each iteration involves solving a least-squares problem, similar to the Least-Squares Policy Iteration algorithm of Lagoudakis and Parr [1]. However, whilst Lagoudakis and Parr considered the so-called least-squares fixed-point approximation to avoid problems with Bellman-residual minimization in the case of correlated samples, we modify the original Bellman-residual objective.

The main conditions of our results can be grouped into three parts: Conditions on the system, conditions on the trajectory (and the behaviour policy used to generate the trajectory) and conditions on the algorithm. The most important conditions on the system are that the state space should be compact, the action space should be finite and the dynamics should be smooth in a sense to be defined later. The major condition on the trajectory is that it should be rapidly mixing. This mixing property plays a crucial role in deriving a PAC-bound on the probability of obtaining suboptimal solutions in the proposed Bellman-residual minimization subroutine. The major conditions on the algorithm are that an appropriate number of iterations should be used and the function space used should have a finite capacity and be sufficiently rich at the same time. It follows that these conditions, as usual, require a good balance between the power of the approximation architecture (we want large power to get good approximation of the action-value functions of the policies encountered during the algorithm) and the number of samples: If the power of the approximation architecture is increased the algorithm will suffer from overfitting, as it also happens in supervised learning. Although the presence of the tradeoff between generalization error and model complexity should be of no surprise, this tradeoff is somewhat underrepresented in the reinforcement literature, presumably because most results where function approximators are involved are asymptotic.

### Further references:

The extended version of this work is available on

<http://www.sztaki.hu/~szcsaba/research/onlinepubs.htm>.

## References

- [1] M. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

**Online Gradient Descent Learning Algorithms**  
**Yiming Ying (University College London)**  
(Joint work with Massimiliano Pontil)

Let  $X$  be a set in  $\mathbb{R}^d$ ,  $Y \subseteq \mathbb{R}$ ,  $\mathbb{N}_T := \{1, \dots, T\}$  and  $\mathbf{z} := \{z_t = (x_t, y_t) : t \in \mathbb{N}_T\}$  be a set of random samples independently distributed according to an unknown probability  $\rho$  on  $X \times Y$ . We consider the least-square online gradient descent algorithm in a real reproducing kernel Hilbert space  $\mathcal{H}_K$  without regularization, that is,

$$f_{t+1} = f_t - \eta_t(f_t(x_t) - y_t)K_{x_t}, \text{ for } t \in \mathbb{N}_T$$

where  $f_1 \in \mathcal{H}_K$  is a given function (typically  $f_1 = 0$ ). Our primary goal is to understand the statistical behavior of the last output  $f_{T+1}$ . In particular, we show how the choice of the step sizes in the algorithm affects the error rates.

We present a novel capacity independent approach which allows us to derive error bounds and convergence results for the above algorithm. Explicit error rates are given in terms of the choice of the step sizes which turn out to be competitive with the state-of-art rates for both offline and online regularized learning algorithms.

**Further references:**

The extended version of this work is available on

<http://www.cs.ucl.ac.uk/staff/Y.Ying/publication.html>.

## References

- [1] N. Cesa-Bianchi, P. Long, and M. K. Warmuth, Worst-case quadratic loss bounds for prediction using linear functions and gradient descent, *IEEE Trans. Neural Networks* **7** (1996), 604–619.
- [2] J. Kivinen, A.J. Smola, and R.C. Williamson, Online learning with kernels, *IEEE Trans. Signal Processing* **52** (2004), 2165–2176.
- [3] S. Smale and Y. Yao, Online learning algorithms, *Found. Comp. Math.*, online version, September 2005.
- [4] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.
- [5] V. Vovk, On-line regression competitive with reproducing kernel Hilbert spaces, Technical report, November 2005. Available at <http://arxiv.org/abs/cs.LG/0511058>.
- [6] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, ICML, 2004.

**Complexity of Tetris-Packing Problem**  
**Gilbert Young (California State Polytechnic University)**  
(Joint work with P. To)

Tetris is a classic video game invented by Alexey Pazhitnov, a Russian mathematician, two decades ago. A player starts out with an empty vertical game board divided into unit squares. The game pieces consist of an endless random sequence of tetrominoes, or shapes made up of four unit squares arranged in various ways. As each piece falls, players must slide them left or right or rotate them, in order to form complete rows of unit blocks. Completed rows of unit grid squares are eliminated from the stack and a new empty row is created on top. The game ends when the height of the block stack prevents placing new pieces. Scoring is based on the number of rows eliminated. Breukelaar et al. recently formalized the game into the Tetris Problem, which they proved NP-hard. Their version allowed for game boards of arbitrary initial configuration, width, and height.

We present a new class of packing problems called the Tetris-Packing Problem (TPP). A problem instance is a 5-tuple  $G = (\beta, (w, h), (P_1, P_2, \dots, P_p), (S_1, S_2, \dots, S_s), u)$ , where  $\beta$  is the initial configuration of the game board (empty or arbitrary),  $(w, h)$  specifies the dimensions of the game board,  $(P_1, P_2, \dots, P_p)$  is a sequence of  $p$  game pieces each belonging to the shape set  $S_1, S_2, \dots, S_s$ , and  $u$  specifies if the top of the board is open or closed (whether pieces can be extended partially beyond its top). Unlike the game of Tetris, TPP has no line elimination. We prove that TPP is generally NP-hard with the objective function of number of filled grid squares.

**Further references:**

The extended version of this work is available on the following URL:  
<http://www.csupomona.edu/~gsyoung/MFLT06.htm>.