Mathematisches Forschungsinstitut Oberwolfach

Report No. 42/2011

DOI: 10.4171/OWR/2011/42

Mini-Workshop: Mathematics of Machine Learning

Organised by
Laszlo Györfi, Budapest
Gabor Lugosi, Barcelona
Ingo Steinwart, Stuttgart
Sara van de Geer, Zürich

August 21st – August 27th, 2011

Abstract. This is a report for a mini-workshop on the mathematical theory of learning. The purpose of the workshop was to bring together internationally recognized experts and young researchers to discuss new approaches, trends, and problems of the area.

Mathematics Subject Classification (2000): 68T05, 68Q32, 68T10, 62H30, 62G08, 62L12, 62M20, 47G10; International Mathematical Union: 10, 15.

Introduction by the Organisers

Learning theory encompasses the mathematical, statistical, and algorithmic aspects arising from problems that appear frequently when one aims at deducing prediction and classification rules from massive and often high-dimensional data. Such problems abound in an increasing number of diverse areas such as bioinformatics, computer vision, data mining, speech processing, and finance, in which learning methods have successfully been applied. As a result of this development, there is now a widely interdisciplinary and highly international community interested in all aspects of learning.

During the last decade, the field of learning theory has witnessed an enormous advance and growth. This progress was both triggered and made possible by successfully merging quite different communities, such as the machine learning community, which traditionally resides in computer science and engineering, on the one hand and the mathematicians coming with diverse background as nonparametric statistics, high-dimensional geometry, theoretical computer science, etc. on the other hand.

While there are many conferences aiming at either applications of learning methods or the development of new learning methods, only very few conferences have focused on the mathematical aspects of learning, so far. The most prominent ones in this direction were probably the conferences "Mathematical Learning Theory" held in Barcelona (2004), Paris (2006), and Dagstuhl (2011). Furthermore, there have been only a few workshops that focused on one or two aspects of mathematical learning theory. Among the most important ones are probably the Oberwolfach workshops "Learning Theory and Approximation" (2008, 2012), "Sparse Recovery Problems in High Dimensions: Statistical Inference and Learning Theory" (2009). We believe that the proposed mini-workshop has an important role in this quickly evolving field. Moreover, the mini-workshop aims in a direction, the above mentioned workshops did or will not cover, and therefore, it nicely complements the MFO program in this field.

Nonparametric classification and regression are probably the mathematically best understood learning problems since their rigorous statistical investigation dates back to the 1970's. While at the beginning, mostly relatively simple statistical procedures were considered, this later changed, e.g. with the development of neural networks and kernel methods. At their beginning, the latter were crudely motivated by heuristics from computer science, but during the last decade they have been intensively studied, so that nowadays their mathematical theory, in particular for classification and regression, is quite mature, and the same is true for some other classification and regression methods. Therefore, the proposed mini-workshop aims at discussing learning problems and successful learning methods of machine learning that have not attracted as much attention from the mathematical community, yet. In particular, the following topics will be covered:

Spectral methods for clustering and semi-supervised learning.: During the last decade, a new class of algorithms for cluster analysis and semi-supervised learning have been developed, which on the one hand, show promising results in applications, and on the other hand, have not been mathematically well-understood. These methods use a similarity measure between data points to construct a similarity graph, which is then used to compute the first eigenvectors of the associated graph Laplacian. While some progress has already been made in understanding these methods, the large-sample behavior of these methods is, so far, mostly unknown. However, recent results indicate that at least for some of the methods, spectral techniques that have already been developed for the analysis of kernel methods may be crucial for a deeper understanding. We will invite experts from both domains to discuss this question.

Merging online and kernel methods.: While kernel methods often produce state-of-the-art results, their computation is rather expensive, which prohibits their use for data sets consisting of millions of data points, a quite common situation in recent practice. On the other hand, online algorithms, which are inherently fast, do not possess strong statistical guarantees such as universal consistency or minmax learning rates, as they are typically investigated in a conditional worst-case scenario. However, in recent years, some progress has been made in merging the best of the two worlds, by either designing online inspired optimization algorithms for kernel methods, or "kernelizing" online algorithms. Our preliminary list of participants, contains experts for online learning and kernel methods, to achieve synergies.

Prediction based on structured classes of experts: Online learning has been an increasingly active area of learning theory. In online learning one is interested in prediction algorithms that act as well as the best among a given class of reference forecasters (the so-called experts). Recently there a lot of research has been carried out in understanding large and structured classes of experts. Such problems give rise to interesting combinatorial, probabilistic, and algorithmic questions. Even though some interesting progress has been achieved recently, many important questions remain unanswered and one of the aims of the workshop is to discuss these problems.

Prediction of stationary time series: The existing prediction algorithms are mainly analyzed for special parametric stochastic processes, where the optimality means mean square optimality. Here the problem is how to construct prediction rules, which are optimal with probability 1, i.e., for any stationary time series, the average of squared errors converges, almost surely, to that of the optimum, given by the Bayes predictor. Such algorithms can be introduced and studied by combining the principles of nonparametric statistics and machine learning.

Mini-Workshop: Mathematics of Machine Learning

Table of Contents

Abstracts

Learning Mixtures of Gaussians

Mikhail Belkin

The study of Gaussian mixture distributions goes back to the late 19th century, when Pearson introduced the method of moments to analyze the statistics of a crab population. They have since become one of the most popular tools of modeling and data analysis, extensively used in speech recognition, computer vision and other fields. Yet their properties are still not well understood. Widely used algorithms, such as Expectation Maximization (EM), often fail even on simple generated data and their theoretical properties are often unclear.

In my talk I will discuss some theoretical aspects of the problem of learning Gaussian mixtures. In particular, I will discuss our recent result with Kaushik Sinha, which, in a certain sense, completes work on an active recent topic in theoretical computer science by establishing general conditions for polynomial learnability of mixture distributions using methods of semi-algebraic geometry.

An open problem on strongly consistent learning of the best prediction for Gaussian processes

László Györfi

Let $\{Y_n\}_{-\infty}^{\infty}$ be a stationary, ergodic, Gaussian process. The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$. It is an open problem whether it is possible to learn the best predictor from the past data in a strongly consistent way, i.e., whether there exists a prediction rule g such that

$$\lim_{n\to\infty} \left(\mathbf{E}\{Y_n \mid Y_1^{n-1}\} - g_n(Y_1^{n-1})\right) = 0 \quad \text{almost surely} \tag{1}$$

for all stationary and ergodic Gaussian processes. (Here Y_1^{n-1} denotes the string $Y_1, \ldots, Y_{n-1}$.)

Bailey [1] and Ryabko [3] proved that just the stationarity and ergodicity is not enough, i.e., for any predictor g, there is a binary valued stationary ergodic process such that

$$\mathbf{P}\left\{\limsup_{n\to\infty} |g_n(Y_1^{n-1}) - \mathbf{E}\{Y_n \mid Y_1^{n-1}\}| \geq 1/2\right\} \geq 1/8.$$

Schäfer [4] proved that, under some conditions on the Gaussian process, we have that

$$\lim_{n\to\infty} \left(\mathbf{E}\{Y_n \mid Y_{n-k_n}^{n-1}\} - g_n(Y_1^{n-1})\right) = 0 \quad \text{almost surely.}$$

His conditions include that the process is purely nondeterministic and the spectral density exists. For example, he proved the strong consistency with $k_n = n^{1/4}$ if the spectral density is bounded away from zero. His proof is based on the fact that under the conditions above the covariance matrix has an inverse. The question left is how to avoid these conditions. Maybe using the recent techniques of machine learning it is doable.

For Gaussian process, Bleakley et al. [2] defined an infinite array of elementary predictors $\tilde{h}^{(k)}$, $k = 1, 2, \ldots$ as follows:

$$\tilde{h}^{(k)}(Y_1^{n-1}) = \sum_{j=1}^{k} c_{n,j}^{(k)} Y_{n-j}$$

such that the coefficients $c_{n,j}^{(k)}$ minimize

$$\sum_{i=k+1}^{n-1} \left(\sum_{j=1}^{k} c_j Y_{i-j} - Y_i \right)^2$$

if $n > k$, and the all-zero vector otherwise. The minimum always exists, it is not unique in general, it can be uniquely defined choosing the minimizer vector with minimal Euclidean norm Set

$$h_n^{(k)}(Y_1^{n-1}) = T_{\min\{n^\delta, k\}} \left(\tilde{h}_n^{(k)}(Y_1^{n-1}) \right),$$

where the truncation function is

$$T_a(z) = \begin{cases} a & \text{if } z > a; \\ z & \text{if } |z| < a; \\ -a & \text{if } z < -a, \end{cases}$$

and $0 < \delta < \frac{1}{8}$. Combine these experts as follows. Let $\{q_k\}$ be an arbitrarily probability distribution over the positive integers such that for all k, $q_k > 0$, and define the weights

$$w_{k,n} = q_k e^{-(n-1)L_{n-1}(h_n^{(k)})/\sqrt{n}} = q_k e^{-\sum_{i=1}^{n-1}(h_i^{(k)}(Y_1^{i-1}) - Y_i)^2/\sqrt{n}}$$

and their normalized values

$$p_{k,n} = \frac{w_{k,n}}{\sum_{i=1}^{\infty} w_{i,n}}.$$

The prediction strategy g at time n is defined by

$$g_n(Y_1^{n-1}) = \sum_{k=1}^{\infty} p_{k,n} h_n^{(k)}(Y_1^{n-1}), \qquad n = 1, 2, \ldots$$

Bleakley et al. [2] proved that the prediction strategy g defined above is universally consistent with respect to the class of all jointly stationary and ergodic zero-mean Gaussian processes, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{E}\{Y_i \mid Y_{-\infty}^{i-1}\} - g_i(Y_1^{i-1}) \right)^2 = 0 \quad \text{almost surely.}$$

This later convergence is expressed in terms of an almost sure Cesáro consistency. I guess that even the almost sure consistency (1) holds. In order to support this conjecture mention that

$$g_n(Y_1^{n-1}) = \sum_{k=1}^{\infty} p_{k,n} h_n^{(k)}(Y_1^{n-1}) \approx \sum_{k=1}^{\infty} p_{k,n} \tilde{h}_n^{(k)}(Y_1^{n-1}) = \sum_{j=1}^{\infty} c_{n,j} Y_{n-j},$$

where $c_{n,j} = \sum_{k=j}^{\infty} p_{k,n} c_{n,j}^{(k)}$. It is well known for Gaussian time series that the best predictor is a linear function of the past:

$$\mathbf{E}\{Y_n \mid Y_{n-1}, Y_{n-2}, \dots\} = \sum_{j=1}^{\infty} c_j^* Y_{n-j},$$

where the c_j^*'s minimize the criterion

$$\mathbf{E}\left\{ \left(\sum_{j=1}^{\infty} c_j Y_{1-j} - Y_1 \right)^2 \right\}.$$

Again, the vector $c^* = (c_1^*, c_2^*, \dots)$ can be uniquely defined choosing the minimizer vector with minimal Euclidean norm. For the notation, $c_n = (c_{n,1}, c_{n,2}, \dots)$, we need that $c_n \to c^*$ almost surely in an appropriate topology such that

$$\mathbf{E}\{Y_n \mid Y_{n-1}, Y_{n-2}, \dots\} - g_n(Y_1^{n-1}) = \sum_{j=1}^{\infty} (c_j^* - c_{n,j}) Y_{n-j} \to 0$$

almost surely.

References

[1] D. H. Bailey. *Sequential schemes for classifying and predicting ergodic processes.* PhD thesis, Stanford University, 1976.

[2] K. Bleakley, G. Biau, L. Györfi, G. Ottucsák (2010) "Nonparametric sequential prediction of time series", *Journal of Nonparametric Statistics*, 22, pp. 297-317.

[3] B.Ya. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24:87–96, 1988.

[4] D. Schäfer. Strongly consistent online forecasting of centered Gaussian processes, *IEEE Trans. Inform. Theory*, Vol. 48, pp. 791–799, 2002.

Information Optimal Algorithms in Machine Learning

Elad Hazan

In this talk we survey recent advancement in the design of efficient algorithms in machine learning, whose running time is best possible. To illustrate these advancements we start with the fundamental problem of Linear Classification.

In the problem of Linear Classification, labeled examples of a concept are represented by in Euclidean space by pairs of their feature vectors and labels, denoted $\{(x_i, y_i) \in (\mathbb{R}^d, \pm 1)\}$. The goal is to find a hyperplane $h \in \mathbb{R}^d$ separating the two classes of vectors, those with positive and negative labels. First consider the separable case, i.e. the case in which a hyperplane exists such that all example are correctly classified. In this case, the objective function can be re-written as

$$\max_{\|h\|\leq 1} \min_{i \in [n]} y_i \cdot (h^\top x_i)$$

The quantity $\omega = \min_{i \in [n]} y_i \cdot (h^\top x_i)$, where h is the solution to the above optimization problem, is called the *margin* of the instance, and is a quantity of significance beyond this talk (for example in the analysis of generalization error in statistical learning theory). When the instance is separable, this margin is non-negative, and indeed a hyperplane can be found which correctly classifies all examples.

The Perceptron Algorithm for linear classification is one of the oldest algorithms studied in machine learning [Nov62, Min88]. It has been the best known algorithm for the linear classification problem for over fifty years. The Perceptron algorithm is extremely simple to describe: iteratively, the Perceotron searcher for an example which is not yet correctly classified, and adds it (as a vector) modulo a sign change to the current candidate hyperplane. It can be shown that when a hyperplane exists that classifies all examples correctly, this process terminates quickly. This classical Perceptron Algorithm, when applied to n vectors in d dimensions, returns an ε-approximate solution to this problem in total time $O(\varepsilon^{-2}nd)$. [1].

This running time is $l$inear in the input representation, since to represent an instance of linear classification one needs to represent n vectors in d dimensions, for a total of $n \times d$ data entries in the unit RAM model. An algorithm that runs in less time then $n \times d$ time does not view the entire data even once, and hence called $s$ublinear.

It might seem surprising that a sublinear algorithm is at all possible, given that the correct classifier might be determined by very few examples, as shown in figure 1. It thus seems necessary to go over all examples at least once.

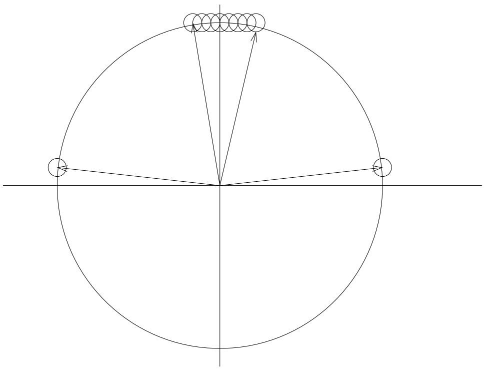

Figure 1. The optimum is determined by very few examples.

However, in recent work with colleagues at IBM ARC [CHW10], we have proved the following result. For given $\delta \in (0, 1)$, our new algorithm takes $O(\varepsilon^{-2}(n + d)(\log n) \log(n/\delta))$ time to return an ε-approximate solution with probability at

[1] An ε-approximate solution in this context is a hyperplane with margin at least $\omega - \varepsilon$. Here ω is the maximum margin attainable by any hyperplane of norm at most one

least $1 - \delta$. Instead of observing all examples at least once, we show it suffices to look at noisy estimates, and plug that into a primal-dual game theoretic optimization framework, along with a novel multiplicative updates algorithm. Further, we show this is optimal in the unit-cost RAM model, up to poly-logarithmic factors.

Ignoring poly-logarithmic factors, this running time, $\tilde{O}(\varepsilon^{-2}(n+d))$, improves by leading order term over the state of the art. In super-scale data analysis problems for which both the dimension and number of examples are huge, this improvement may be of practical significance.

We survey applications of this machinery to other problems in machine learning, such as linear classification with kernels, semi-definite programming [GarHaz11] and soft-margin SVM [HazKorSre11].

References

[CHW10] Ken Clarkson, Elad Hazan and David Woodruff. Sublinear Optimization for Machine Learning *The 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS 2010).*

[Val84] L. G. Valiant. A theory of the learnable. Communications of the ACM, 27(11):1134-1142, November 1984

[Vap98] V. N. Vapnik. Statistical Learning Theory. Wiley, 1998

[CBL06] Nicolo Cesa-Bianchi and Gabor Lugosi. Prediction, learning, and games. Cambridge University Press, 2006 ISBN 0521841089

[BTN01] Ben-Tal, A. and Nemirovski, A. Lectures on Modern Convex Optimization: Analysis, Algorithms; Engineering Applications. SIAM-MPS Series in Optimization, (2001).

[ScSm02] Bernhard Scholkopf and Alexander J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2002).

[GriKha95] Michael D. Grigoriadis and Leonid G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. In *Operations Research Letters*, Vol. 18, No. 2, pp. 53-58, (1995).

[HazKorSre11] Elad Hazan, Tomer Koren and Nati Srebro. Beating SGD: Learning SVMs in Sublinear Time. To appear in *NIPS 2011*, (2011).

[GarHaz11] Dan Garber and Elad Hazan. Approximating Semidefinite Programs in Sublinear Time. To appear in *NIPS 2011*, (2011).

[Nov62] A.B.J. Novikoff. On convergence proofs on perceptrons. Proceedings of the Symposium on the Mathematical Theory of Automata, Vol XII, pages 615-622 (1962).

[Min88] Minsky, M. L. and Papert, S. Perceptrons: An introduction to computational geometry. (1988) publisher: MIT press Cambridge, Mass

Structured Sparsity and Generalization

Massimiliano Pontil
(joint work with Andreas Maurer)

We present a data dependent generalization bound for a large class of regularized algorithms which implement structured sparsity constraints. A novel feature of our bound is that it can be applied in an infinite dimensional setting such as the Lasso in a separable Hilbert space or multiple kernel learning with a countable number of kernels.

We study a class of regularization methods used to learn a linear function from a finite set of examples. The regularizer is expressed as an infimum convolution which involves a set $\mathcal{M}$ of linear transformations (see equation (1) below). This regularizer generalizes, depending on the choice of the set $\mathcal{M}$, the regularizers used by several learning algorithms, such as ridge regression, the Lasso, the group Lasso, multiple kernel learning and others, see [Maurer and Pontil(2006)] for a discussion. Our study was originally motivated by the the methods described in [Micchelli et al.(2010)].

We give a bound on the Rademacher average of the linear function class associated with this regularizer. The result matches existing bounds in the above mentioned cases but also admits a novel, dimension free interpretation. In particular, the bound applies to the Lasso in a separable Hilbert space or to multiple kernel learning with a countable number of kernels, see e.g. [Micchelli and Pontil(2005)] and references therein.

Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $\mathcal{M}$ be a finite or countably infinite set of symmetric bounded linear operators on H such that for every $x \in H$, $x \neq 0$, there is some linear operator $M \in \mathcal{M}$ with $Mx \neq 0$ and that $\sup_{M \in \mathcal{M}} |||M||| < \infty$, where $||| \cdot |||$ is the operator norm. Define the function $\|\cdot\|_{\mathcal{M}} : H \to \mathbb{R}_+ \cup \{\infty\}$ by

$$(1) \qquad \|\beta\|_{\mathcal{M}} = \inf \left\{ \sum_{M \in \mathcal{M}} \|v_M\| : v_M \in H, \ \sum_{M \in \mathcal{M}} M v_M = \beta \right\}.$$

The notation is justified, because $\|\cdot\|_{\mathcal{M}}$ is indeed a norm [Maurer and Pontil(2006)] on the subspace of H where it is finite, and the dual norm is, for every $z \in H$, given by

$$\|z\|_{\mathcal{M}*} = \sup_{M \in \mathcal{M}} \|Mz\|.$$

The somewhat complicated definition of $\|\cdot\|_{\mathcal{M}}$ is contrasted by the simple form of the dual norm. As an example, if $H = \mathbb{R}^d$ and $M = \{P_1, \dots, P_d\}$, where P_i is the orthogonal projection on the i-th coordinate, then the function (1) reduces to the ℓ_1 norm.

Using well known techniques, as described in [Koltchinskii and Panchenko(2002)] and [Bartlett and Mendelson(2002)], our study of generalization reduces to the search for a good bound on the empirical Rademacher complexity of a set of linear functionals with $\|\cdot\|_{\mathcal{M}}$-bounded weight vectors

$$(2) \qquad \mathcal{R}_{\mathcal{M}}(\mathbf{x}) = \frac{2}{n} \mathbb{E} \sup_{\beta : \ \|\beta\|_{\mathcal{M}} \le 1} \sum_{i=1}^{n} \epsilon_i \langle \beta, x_i \rangle,$$

where $\mathbf{x} = (x_1, \dots, x_n) \in H^n$ is a sample vector representing observations, and $\epsilon_1, \dots, \epsilon_n$ are Rademacher variables, mutually independent and each uniformly distributed on $\{-1, 1\}$. Given a bound on $\mathcal{R}_{\mathcal{M}}(\mathbf{x})$ we obtain uniform bounds on the estimation error, for example using the following standard result (adapted from [Bartlett and Mendelson(2002)]), where the Lipschitz function ϕ is to be interpreted as a loss function.

Theorem 1. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of iid random variables with values in H, let X be iid to X_1, let $\phi : \mathbb{R} \to [0,1]$ have Lipschitz constant L and $\delta \in (0,1)$. Then with probability at least $1 - \delta$ in the draw of $\mathbf{X}$ it holds, for every $\beta \in \mathbb{R}^d$ with $\|\beta\|_{\mathcal{M}} \leq 1$, that*

$$\mathbb{E}\phi\left(\langle \beta, X \rangle\right) \leq \frac{1}{n}\sum_{i=1}^{n} \phi\left(\langle \beta, X_i \rangle\right) + L\ \mathcal{R}_{\mathcal{M}}\left(\mathbf{X}\right) + \sqrt{\frac{9 \ln 2/\delta}{2n}}.$$

A similar (slightly better) bound is obtained if $\mathcal{R}_{\mathcal{M}}\left(\mathbf{X}\right)$ is replaced by its expectation $\mathcal{R}_{\mathcal{M}} = \mathbb{E}\mathcal{R}_{\mathcal{M}}\left(\mathbf{X}\right)$ (see [Bartlett and Mendelson(2002)]).

The following result leads to consistency proofs and finite sample generalization guarantees for all algorithms which use a regularizer of the form (1). Its proof can be found in [Maurer and Pontil(2006)].

Theorem 2. *Let $\mathbf{x} = (x_1, \ldots, x_n) \in H^n$ and $\mathcal{R}_{\mathcal{M}}\left(\mathbf{x}\right)$ be defined as in (2). Then*

$$\begin{aligned} \mathcal{R}_{\mathcal{M}}\left(\mathbf{x}\right) &\leq \frac{2^{3/2}}{n}\sqrt{\sup_{M \in \mathcal{M}} \sum_{i=1}^{n} \left\|Mx_i\right\|^2}\left(2 + \sqrt{\ln\left(\sum_{M \in \mathcal{M}} \frac{\sum_i \left\|Mx_i\right\|^2}{\sup_{N \in \mathcal{M}} \sum_j \left\|Nx_j\right\|^2}\right)}\right) \\ &\leq \frac{2^{3/2}}{n}\sqrt{\sum_{i=1}^{n} \left\|x_i\right\|_{\mathcal{M}*}^2}\left(2 + \sqrt{\ln\left|\mathcal{M}\right|}\right). \end{aligned}$$

The second inequality follows from the first one, the inequality

$$\sup_{M \in \mathcal{M}} \sum_{i=1}^{n} \left\|Mx_i\right\|^2 \leq \sum_{i=1}^{n} \left\|x_i\right\|_{\mathcal{M}*}^2,$$

a fact which will be tacitly used in the sequel, and the observation that every summand in the logarithm appearing in the first inequality is bounded by 1. Of course the second inequality is relevant only if $\mathcal{M}$ is finite. In this case we can draw the following conclusion: If we have an a priori bound on $\|X\|_{\mathcal{M}*}$ for some data distribution, say $\|X\|_{\mathcal{M}*} \leq C$, and $\mathbf{X} = (X_1, \ldots, X_n)$, with X_i iid to X, then

$$\mathcal{R}_{\mathcal{M}}\left(\mathbf{X}\right) \leq \frac{2^{3/2} C}{\sqrt{n}}\left(2 + \sqrt{\ln\left|\mathcal{M}\right|}\right),$$

thus passing from a data-dependent to a distribution dependent bound.

But the first bound in Theorem 2 can be considerably smaller than the second and may be finite even if $\mathcal{M}$ is infinite. This gives rise to some novel features, even in the well studied case of the Lasso, when there is a (finite but potentially large) ℓ_2-bound on the data.

Corollary 3. *Under the conditions of Theorem 2 we have*

$$\mathcal{R}_{\mathcal{M}}\left(\mathbf{x}\right) \leq \frac{2^{3/2}}{n}\sqrt{\sup_{M \in \mathcal{M}} \sum_{i} \left\|Mx_i\right\|^2}\left(2 + \sqrt{\ln \frac{1}{n}\sum_{i}\sum_{M \in \mathcal{M}} \left\|Mx_i\right\|^2}\right) + \frac{2}{\sqrt{n}}.$$

To obtain a distribution dependent bound we retain the condition $\|X\|_{\mathcal{M}*} \leq C$ and replace finiteness of $\mathcal{M}$ by the condition that

$$R^2 := \mathbb{E} \sum_{M \in \mathcal{M}} \|MX\|^2 < \infty. \tag{3}$$

Taking the expectation in Corollary 3 and using Jensen's inequality then gives a bound on the expected Rademacher complexity

$$\mathcal{R}_{\mathcal{M}} \leq \frac{2^{3/2}C}{\sqrt{n}} \left(2 + \sqrt{\ln R^2}\right) + \frac{2}{\sqrt{n}}. \tag{4}$$

The key features of this result are the dimension-independence and the only logarithmic dependence on R^2, which in many applications turns out to be simply $R^2 = \mathbb{E}\|X\|^2$.

References

[Bartlett and Mendelson(2002)] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[Koltchinskii and Panchenko(2002)] V. Koltchinskii and D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *Annals of Statistics*, 30(1):1–50, 2002.

[Maurer and Pontil(2006)] A. Maurer and M. Pontil (2011). Structured Sparsity and Generalization. arXiv:1108.3476.

[Micchelli et al.(2010)] C.A. Micchelli, J.M. Morales, M. Pontil. A family of penalty functions for structured sparsity. In *Advances in Neural Information Processing Systems 23*, pages 1612–1623, 2010.

[Micchelli and Pontil(2005)] C.A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, 66:297–319, 2007.

Convex relaxation for combinatorial penalties

Guillaume Obozinski

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 best convex relaxation for a family of functions that penalize simultaneously (a) the structure of the support through a general combinatorial set function, and (b) the L_p norm of the parameter vector for an arbitrary fixed p.

The convex relaxation only certain characteristics of the original set- function are kept, and we introduce the notion of lower combinatorial envelope which characterizes the retained properties.

The formulation considered allows to treat in a unified framework several a priori disconnected approaches such as norms based on overlapping groups, norms based on latent representations such as block-coding and submodular functions.

Learning Theory: A Minimax Analysis

Alexander Rakhlin

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.

Multi-class Learning: Simplex Coding and Relaxation Error

Lorenzo Rosasco
(joint work with Youssef Mroueh, Tomaso Poggio, Jean-Jacques E. Slotine)

We study multi-category classification in the framework of computational learning theory. We show how a relaxation approach, which is commonly used in binary classification, can be generalized to the multi-class setting. We propose a vector coding, namely the simplex coding, that allows to introduce a new notion of multi-class margin and cast multi-category classification into a vector valued regression problem. The analysis of the relaxation error be quantified and the binary case is recovered as a special case of our theory. From a computational point of view we can show that using the simplex coding we can design regularized learning algorithms for multi-category classification that can be trained at a complexity which is independent to the number of classes.

1. Problem Setting

We consider an input space $\mathcal{X} \subset \mathbb{R}^d$, and output space $\mathcal{Y} = \{1, \ldots, T\}$. Given a probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$ we let $\rho_{\mathcal{X}}$ be the marginal probability on $\mathcal{X}$ and $\rho_j(x) = \rho(j|x)$ the conditional distribution of class j given x, for each $j = 1, \ldots, T$ and $x \in \mathcal{X}$. A training set is a sequence $(x_i, y_i)_{i=1}^n$ sampled i.i.d. with respect to ρ. A classification rule is a map $c : \mathcal{X} \to \mathcal{Y}$ and its properties can be measured via the misclassification error (or misclassification risk),

$$R(c) = \mathbb{P}(c(x) \neq y),$$

which is minimized, by the Bayes rule $b_\rho(x) = \operatorname{argmax}_{j=\{1,\ldots,T\}} \rho_j(x)$. This risk functional cannot be directly minimized for two reasons: 1) the true probability distribution is unknown, 2) it requires optimizing a non convex functional over a set of discrete valued functions, in fact the risk can be written as $R(c) = \int \Theta(yc(x)) d\rho(x, y)$ where $\Theta(x) = 1$ if $x < 0$ and 0 otherwise. While we can tackle the first issue looking at the empirical error on the data– rather than the risk, in this work we consider the second issue.

The typical approach in binary classification, i.e. $T = 2$, is based on the following steps. First real valued functions are considered in place of binary valued ones so that a classification rule is defined defined by the sign of a function. Second, the *margin* of a function is defined to be the quantity $m = yf(x)$ and $\Theta(m)$ is replaced by a *margin loss* function $V(m)$ where V is a non-negative and convex. This *relaxation* approach introduces an error which can be quantified. In fact, if we define $\mathcal{E}(f) = \int V(yf(x)) d\rho(x, y)$, and let f_ρ be its minimizer, it is possible to prove [2] that if V is decreasing in a neighborhood of 0, and differentiable in 0, then $b_\rho(x) = \operatorname{sign}(f_\rho)(x)$, namely the loss is *classification calibrated.* Moreover, for any measurable function $f : \mathcal{X} \longmapsto \mathbb{R}$ and probability distribution ρ we can derive a so called *comparison theorem*, that is, there exits a function $\psi_V : [0, 1] \mapsto [0, \infty)$

$$\psi_V(R(\operatorname{sign}(f)) - R(\operatorname{sign}(f_\rho))) \leq \mathcal{E}(f) - \mathcal{E}(f_\rho).$$

For example for the the square loss $V(m) = (1 - m)^2$ we have $\psi_V(t) = t^2$ and for the hinge loss $V(m) = |1 - m|_+$ we have $\psi_V(t) = t$. In this note we discuss how the above approach can be extended to $T \geq 2$.

1.1. Simplex Coding and Relaxation Error.

The following definition is at the core of our approach.

Definition 1. *The simplex coding is a map $C : \{1, \ldots, T\} \to \mathbb{R}^{T-1}$ such that for $i = 1, \ldots, T$, $C(i) = a_i$, where the code vectors $\mathcal{A} = \{a_1, \ldots, a_T\} \subset \mathbb{R}^{T-1}$ satisfy*

$$\|a_i\|^2 = 1, \quad \forall i = 1, \ldots, T, \quad \langle a_i, a_j \rangle = -\frac{1}{T-1}, \quad \forall i, j = 1, \ldots, T,\ i \neq j,$$

and $\sum_{i=1}^T a_i = 0$. The corresponding decoding is the map $D : \mathbb{R}^{T-1} \to \{1, \ldots, T\}$ $D(\alpha) = \operatorname{argmax}_{i=1,\ldots,T} \langle \alpha, a_i \rangle$, $\forall \alpha \in \mathbb{R}^{T-1}$.

The simplex coding corresponds to the T most separated vectors on the hypersphere $\mathbb{S}^{T-2}$ in $\mathbb{R}^{T-1}$, which are the vertices of the simplex. For binary classification it reduces to the ± 1 coding. The decoding map has a natural geometric interpretation: an input point is mapped to a vector $f(x)$ by a vector regressor and hence assigned to the class having closer code vector. The projection $\langle f(x), a_j \rangle$ is precisely the multi-class extension of the notion of margin that we discussed in binary classification and allows to extend the relaxation approach. Using the simplex coding the misclassification risk can be written as

$$R(D(f)) = \int \Theta(\langle f(x), a \rangle) d\rho(a, x) = \sum_{j=1}^{T} \int \Theta(\langle f(x), a_j \rangle) \rho_j(x) d\rho_{\mathcal{X}}(x).$$

where $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^{T-1}$. Then, we can simply consider any margin loss, e.g. hinge or logistic loss, and can replace the misclassification risk by the expected risk $\mathcal{E}(f) = \int V(\langle f(x), y \rangle) d\rho(x, y)$. Note that the square loss can be seen as margin loss if f is on the sphere.

1.2. **Relaxation error analysis.** As in the binary case, it is natural to ask what is the error we incur into by considering a convex relaxation of the classification problem. Interestingly, the results in the binary case can be now extended to the multiclass setting. In fact, also in this case if V is decreasing in a neighborhood of 0, and differentiable in 0, then $b_\rho(x) = D(f_\rho)(x)$, where the sign is replaced by the decoding map. Comparison theorems can also be proved. For example, for the the square loss $V(m) = (1-m)^2$ we have $\psi_V(t) = t^2/(T-1)^2$ and for the hinge loss $V(m) = |1-m|_+$ we have $\psi_V(t) = t/(T-1)$, where we see the price to pass from $T = 2$ to $T \geq 2$. While we omit further details we mention here that a notion of (multi) classification noise related to the one used in binary classification [2] can also be defined, which allows to improve the above results. Compared to previous works [7, 8] we see that the simplex coding allows to avoid any further constraint to the function class.

1.3. **Computing the simplex coding.** The simplex coding can be easily implemented and can induce regularized learning methods for multi-category classification that can be trained at the same computational complexity of a binary classification problem, hence independently to the number of classes.

We start discussing a simple algorithm for the generation of the simplex coding. We use a recursive projection of simpleces, by observing that the simplex in $\mathbb{R}^{T-1}$, can be obtained projecting the simplex in $\mathbb{R}^T$ on the hyperplane orthogonal to the element $(1, \dots, 0)$ of the canonical basis in $\mathbb{R}^T$. Let $C[T-1]$ be the simplex code associated to T classes, $C[T-1]$ is a matrix of dimension $T \times (T-1)$. We have the following recursion, where at each step we add a dimension, and backproject:

$$C[T] = \begin{pmatrix} 1 & u \\ v & C[T-1] \times \sqrt{1 - \frac{1}{T^2}} \end{pmatrix} \tag{1}$$

$$C[1] = [1; -1]$$

Where u is row vector of dimension T, $u = (-\frac{1}{T} \cdots - \frac{1}{T})$, and v a column vector of dimension T, $v = (0, \dots, 0)$.
Kernels and Regularization Algorithms. Next we need to recall some basic concepts in the theory of reproducing kernel Hilbert spaces (RKHS) of vector valued functions. The definition of RKHS for vector valued functions parallels the one in the scalar [1], with the main difference that the reproducing kernel is now *matrix* valued – see [3] and references therein. A reproducing kernel is a symmetric function $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{D \times D}$, such that for any $x, x' \in \mathcal{X}$ $\Gamma(x, x')$ is a positive semi-definite *matrix*. A vector valued RKHS is a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}^D$, such that for very $c \in \mathbb{R}^D$, and $x \in \mathcal{X}$, $\Gamma(x, \cdot)c$ belongs to h and moreover Γ has the reproducing property $\langle f, \Gamma(\cdot, x)c \rangle_{\mathsf{h}} = \langle f(x), c \rangle$, where $\langle \cdot, \cdot \rangle_{\mathsf{h}}$ is the inner product in $\mathcal{H}$. The choice of the kernel corresponds to the choice of the representation (parameterization) for the functions of interest. In fact it can be shown that any function in a RKHS with kernel Γ, is in the completion of the span of $\Gamma(x_i, \cdot)$ with $c_j \in \mathbb{R}^D$. Given the reproducing property, the norm of f can be written as $\|f\|_{\mathsf{h}}^2 = \sum_{i,j=1}^{\infty} \langle c_j, \Gamma(x_i, x_j) c_j \rangle$. Note that for $D = 1$ we recover the classic theory of scalar valued RKHS. In the following we restrict our attention to kernels of the form $\Gamma(x, x') = k(x, x')A, \quad A = I$, where k is a scalar reproducing kernel. As we discuss elsewhere [6] the choice of A corresponds to a prior belief that different components can be related. In fact, if we let $f = (f_1, \dots, f_D)$ it is possible to see that the entry $A_{t,t'}$ defines the relation between f_t and $f_{t'}$. For the sake of simplicity we restrict ourselves to $A = I$, hence treating each component as independent. This case is directly comparable to the one-vs-all approach.

Next, we discuss the properties of different learning algorithms using the simplex coding. We use the following notation, $Y \in \mathbb{R}^{n \times (T-1)}, Y = (y_1, ..., y_n)$, $y_i \in \mathcal{A}$, $i = 1, \dots, n$; $K \in \mathbb{R}^{n \times n}, K_{ij} = k(x_i, x_j)$; $C \in \mathbb{R}^{n \times (T-1)}, C = (c_1, c_2, ..., c_n)$. We consider algorithms defined by the minimization of a Tikhonov functional

$$\mathcal{E}_n^{\lambda}(f) = \frac{1}{n} \sum_{i=1}^{n} V(\langle y_i, f(x_i) \rangle) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2,$$

where in particular $V(\langle y_i, f(x_i) \rangle)$ will be either the square loss or a margin loss (in particular the SVM's hinge loss). It is well known [5] that the representer theorem [4] can be easily extended to a vector valued setting to show that minimizer of the above functional over h can be written as $f(x) = \sum_{j=1}^{n} k(x, x_j) c_j, \quad c_j \in \mathbb{R}^{T-1}$. The choice of different loss functions induce different strategy to compute C.

If we choose let the loss to be $\|y - f(x)\|^2$ it is easy to see that, $(K + \frac{\lambda}{2} nI)C = Y$. If we want to compute a solution for N values of λ, by using SVD to perform the matrix inversion, we can still compute a regularized inverse in essentially $O(n^3)$ but the multiplication $(K + \frac{\lambda}{2} nI)^{-1} Y$ is going to be $O(n^2 TN)$, which is linear in T. Note that this complexity is still much better than the one-vs-all approach that would give a $O(n^3 TN)$. If we choose let the loss to be $|1 - \langle y, f(x) \rangle|_+$, following standard reasonings from the binary case [9] to see that we have to solve the problem

$$\max_{\alpha\in\mathbb{R}^n}\left\{\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\alpha^\top Q\alpha\right\},\qquad 0\le\alpha_i\le\frac{1}{n\lambda}, i=1,\dots,n$$

where $Q_{ij}=K(x_i,x_j)y_i^Ty_j$ and $c_k=\alpha_k y_k$ where $\alpha_k\in\mathbb{R}$, for $k=1,\dots,n$. Note that the optimization is now only over the n dimensional vector α and T appears only in the computation of the matrix Q. Training for fixed C is hence independent of the number of classes and is essentially $O(n^3)$ in the worst case. If we are interested into N different values of λ we would get a complexity $O(n^3N)$. Note that more sophisticated strategy to compute the whole regularization path could be coupled with the use of simplex coding.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] P.L. Bartlett, M.I. Jordan, and J.D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2005. To appear. (Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

[3] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Anal. Appl. (Singap.)*, 4(4):377–408, 2006.

[4] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1970.

[5] C.A. Micchelli and M. Pontil. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.

[6] Y. Mroueh and L. Poggio, T.and Rosasco. Regularization predicts while discovering taxonomy. Technical Report MIT-CSAIL-TR-2011-029/CBCL-299, Massachusetts Institute of Technology,cambridge,MA, june 2011.

[7] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, May 2007.

[8] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 2004.

[9] V. N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

Some Applications of Scaled Bregman distances to Analyses of Random Data

Wolfgang Stummer

(joint work with Igor Vajda)

It is well known that the Csiszar divergences (including the Kullback-Leibler information divergence) as well as the “classical” Bregman divergences are very useful tools for machine learning purposes, see e.g. Laferty (1999), Banerjee et al. (2005), Amari (2007), Teboulle (2007), Nock & Nielsen (2009), and the comprehensive exposition in Cesa-Bianchi & Lugosi (2006), as well as the corresponding references therein.

In this talk we present the new concept of *scaled* Bregman divergences, cf. Stummer & Vajda (2011) (see also Stummer (2007)), which generalizes both the Csiszar divergences as well as the classical Bregman divergences.
Definition 1. Let $\phi : (0, \infty) \mapsto \mathbb{R}$ be a continuous convex function, P, Q be two probability measures and M a finite measure having densities

$$p = \frac{\mathrm{d}P}{\mathrm{d}\lambda}, \quad q = \frac{\mathrm{d}Q}{\mathrm{d}\lambda} \qquad \text{and} \qquad m = \frac{\mathrm{d}M}{\mathrm{d}\lambda}$$

with respect to a σ-finite measure λ. Then the *Bregman divergence* of P, Q *scaled* by M is defined by the formula

$$\begin{aligned} B_\phi(P, Q \,|\, M) &= \int_{\mathcal{X}} \left[\phi\left(\frac{p}{m}\right) - \phi\left(\frac{q}{m}\right) - \phi'_+\left(\frac{q}{m}\right)\left(\frac{p}{m} - \frac{q}{m}\right)\right] \mathrm{d}M \\ &= \int_{\mathcal{X}} \left[m\phi\left(\frac{p}{m}\right) - m\phi\left(\frac{q}{m}\right) - \phi'_+\left(\frac{q}{m}\right)(p - q)\right] \mathrm{d}\lambda. \end{aligned}$$

The convex ϕ under consideration can be interpreted as a generating function of the divergence.

We illuminate the connections of scaled Bregman divergences with Csiszar divergences and discuss several important special cases such as e.g. the scaled *power* Bregman divergences. For an exemplary highly complex financial-diffusion-process setup, we show how these divergences can be used for sequential learning and decision making. As another field of application, we present a new concept of robust graphical 3D goodness-of-fit tests which is discussed for the very important and widely used exponential-families class of probability laws, as well as mixtures thereof.

References

[1] Amari, S.-I. (2007): Integration of stochastic models by minimizing α-divergence. *Neural Computation*, **19**(10), pp. 2780-2796.

[2] Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005): Clustering with Bregman divergences. *J. Machine Learning Research*, **6**, pp. 1705-1749.

[3] Cesa-Bianchi, N. and Lugosi, G. (2006): *Prediction, Learning, Games.* Cambridge, Cambridge University Press.

[4] Lafferty, J.D. (1999): Additive models, boosting, and inference for generalized divergences. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory.* New York, ACM Press, pp. 125-133.

[5] Nock, R. and Nielsen, F. (2009): Bregman divergences and surrogates for learning. *IEEE Transactions on PAMI*, **31**(11), pp. 2048 - 2059.

[6] Stummer, W. (2007): Some Bregman distances between financial diffusion processes. *Proc. Appl. Math. Mech.*, **7**(1), pp. 1050503 - 1050504.

[7] Stummer, W. and Vajda, I. (2011): On Bregman Distances and Divergences of Probability Measures. *To appear in IEEE Transaction on Information theory.*

[8] Teboulle, M. (2007): A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, **8**, pp. 65-102.

Least Squares Temporal Difference Learning and Galerkin's Method

Csaba Szepesvári

The problem of estimating the value function underlying a Markovian reward process is considered. As it is well known, the value function underlying a Markovian reward process satisfied a linear fixed point equation. One approach to learning the value function from finite data is to find a good approximation to the value function in a given (linear) subspace of the space of value functions. We review some of the issues that arise when following this approach, as well as some results that characterize the finite-sample performance of some of the algorithms.

1. Markovian Reward Processes

Let $\mathcal{X}$ be a measurable space and consider a stochastic process

$$(X_0, R_1, X_1, R_2, X_2, \ldots),$$

where $X_t \in \mathcal{X}$ and $R_{t+1} \in \mathbb{R}$, $t = 0, 1, 2, \ldots$. The process is called a *Markovian Reward process* if

- $(X_0, X_1, \ldots)$ is a Markov process, and
- for any $t \geq 0$, given X_t, X_{t+1} the distribution of R_{t+1} is independent of the history of the process.

Here, X_t is called the *state* of the system at time t, while R_{t+1} is the *reward* associated to transitioning from X_t to X_{t+1}. We shall denote by $\mathcal{P}$ the Markovian kernel underlying the process: Thus, the distribution of (X_{t+1}, R_{t+1}) given X_t is given by $\mathcal{P}(\cdot, \cdot | X_t)$, $t = 0, 1, \ldots$.

Fix the so-called *discount factor* $0 \leq \gamma \leq 1$ and define the (total discounted) *return* associated to the process

$$\mathcal{R} = \sum_{t=0}^{\infty} \gamma^t R_{t+1}$$

and the value function

$$V^*(x) = \mathcal{E}\mathcal{R} \,|\, X_0 = x, \quad x \in \mathcal{X}\,.$$

For simplicity, assume that the rewards are bounded with probability one, say, $\sup_{t \geq 0} |R_{t+1}| \leq 1$ a.s. Further, assume that the support of the distribution of X_0 is the entire state space $\mathcal{X}$. Under these conditions and if $\gamma < 1$, $V^* : \mathcal{X} \to \mathbb{R}$ is well-defined and depends solely on the transition probability kernel $\mathcal{P}$. When $\gamma = 1$, further conditions are necessary to ensure that V is well-defined. Hence, for simplicity in the theoretical developments below we shall assume that $\gamma < 1$.

2. Markovian Value Prediction Problems

The *Markovian value prediction problem* is to estimate V given data that follows the law of the underlying Markovian Reward Process. This problem arises in a number of applications (e.g., predicting long term values in financial applications, predicting whether a biped robot is going to fall, or as a subproblem of approximate dynamic programming algorithms).

By overloading some of the letters, let $D_n = ((X_1, R_2, X_2'), \ldots, (X_n, R_{n_1}, X_n'))$ be the data available for the statistician, where the distribution of (X_t', R_{t+1}) given the history $((X_1, R_2, X_2'), \ldots, (X_{t-1}, R_t, X_t'))$ and X_t is $\mathcal{P}(\cdot|X_t)$. Two important special cases are the following:

(i) The process $((X_t, R_{t+1}, X_{t+1}'))_{1 \le t \le n}$ is i.i.d.
(ii) For all i, $X_{t+1} = X_{t+1}'$.

The second case arises in applications where only a single *trajectory* is available for learning, while the first is mainly used as a convenient simplifying assumption when studying the performance of learning methods (a situation, when the first assumption is met in practice is when a simulator of the Markovian Reward Process is available). In practice, one might also have a dataset that consists of a number of trajectories whose start states are independently sampled from each other.

If the training data was an infinitely long trajectory, one could approach estimating V^* by casting it as a regression problem. Indeed, when $n = \infty$, $\mathcal{R}_t = \sum_{j=t}^{\infty} \gamma^{j-t} R_{j+1}$ provides an unbiased estimate of $V^*(X_t)$: $V^*(X_t) = \mathcal{E}\mathcal{R}_t|X_t$. Of course, in practice n is finite. Then, one is forced to truncate the estimates $\mathcal{R}_t$ and use $\mathcal{R}_t^{(n)} = \sum_{j=t}^{n} \gamma^{j-t} R_{j+1}$. This way, however, a bias of size $O(\gamma^{n-t})$ is introduced at sample t. Thus, for t large, the bias introduced is considerable, whereas for t small, the variance of $\mathcal{R}_t^{(n)}$ can be large. The approach can be extended to the case when the data consists of a set of long trajectories. However, this approach will not work when the data consists of short trajectories (i.e., case i above).

3. Least squares temporal difference learning: A Statistical Galerkin Method

The approach proposed by [Bradtke and Barto, 1996] allows one to deal with such sequences of short trajectories. This approach was inspired by earlier work by [Sutton, 1988]. In fact, it can be viewed as the least-squares analogue of the stochastic approximation algorithm proposed by [Sutton, 1988]. Here, we show how this approach can be viewed as a "statistical Galerkin method". Thus our approach to introduce this method will be different than the usual approach. Although the connection to Galerkin methods has been acknowledged beforehand (e.g., [Yu and Bertsekas, 2010]), the connection is not very well known amongst reinforcement learning researchers and hence we find it useful to explain it. For yet another alternative derivation of the method, as well as a statistical analysis, see [Antos et al., 2008]. For the analysis of non-parametric, regularized versions of the method, see, e.g., [Farahmand et al., 2009].

As it is well known, the value function V^* is the solution to the fixed-point equation

$$TV^* = V^*, \tag{1}$$

where $T : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ is the so-called *Bellman-operator*:

$$TV\,(x) = \int (r + \gamma V(x'))d\mathcal{P}(x', r|x), \quad V \in L^\infty(\mathcal{X}).$$

As it can be seen from the definition, T is an affine linear operator, so (1) is a linear fixed point equation. Also, T is a bounded and is in fact a γ-contraction.

The *projection method* for solving (1) consists of choosing two finite dimensional linear subspaces $\mathcal{F}, \mathcal{G}$ of a Banach space $\mathcal{B}$ of functions over $\mathcal{X}$, $\mathcal{F}, \mathcal{G}$ sharing a common dimension $d \in \natural$, and then solving the projected equation

$$\Pi_{\mathcal{G}} TV = \Pi_{\mathcal{G}} V \quad \text{for } V \in \mathcal{F}, \tag{2}$$

where $\Pi_{\mathcal{G}} : \mathcal{B} \to \mathcal{F}$ is a projection operator [Kirsch, 2011]. It can be easily seen that this leads to a $d \times d$ linear system of equations once one fixes some bases for $\mathcal{F}$ and $\mathcal{G}$. When $\mathcal{B}$ is a pre-Hilbert space, $\Pi_{\mathcal{G}}$ is the corresponding orthogonal projection and $T : \mathcal{F} \to \mathcal{B}$ is bounded, we arrive at a *Galerkin method.* In this case, (2) is equivalent to

$$\langle\, TV, g\,\rangle = \langle\, V, g\,\rangle \quad \text{for all } g \in \mathcal{G}. \tag{3}$$

Write $TV = \overline{r} + \gamma \mathcal{P} V$, where $\overline{r}(x) = \int r d\mathcal{P}(x', r|x)$ and $\mathcal{P} : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ is defined by $\mathcal{P}V = \int V(x') d\mathcal{P}(x', r|x)$. Then, assuming that $\mathcal{F} = \mathrm{span}(f_1, \ldots, f_d)$, $\mathcal{G} = \mathrm{span}(g_1, \ldots, g_d)$, writing $V = \sum_{i=1}^d \alpha_i f_i$, (3) leads to the linear system of equations

$$\sum_{i=1}^{d} \alpha_i \langle\, (I - \gamma\mathcal{P}) f_i, g_j\,\rangle = \langle\, \overline{r}, g_j\,\rangle\,, \quad 1 \le j \le d.$$

or

$$A\alpha = b,$$

where $A_{ij} = \langle\, (I - \gamma\mathcal{P}) f_i, g_j\,\rangle$, $b_j = \langle\, \overline{r}, g_j\,\rangle$. For the error analysis of Galerkin's method, see, e.g., Chapter 3 of [Kirsch, 2011]. Unsurprisingly, these results are identical to results derived, e.g., by [Yu and Bertsekas, 2010] or [Scherrer, 2010].

Now consider the case when we are given a sample D_n as described in the previous section. Further, for simplicity, assume that $(X_t)_{1 \le t \le n}$ is stationary and let μ be the common distribution underlying (X_t). Choose $\mathcal{B}$ to be the Hilbert space $L^2(\mu)$. Given the sample D_n, approximate A_{ij} by

$$\hat{A}_{ij}^{(n)} = \frac{1}{n} \sum_{t=1}^{n} (f_i(X_t) - \gamma f_i(X_t')) g_j(X_t)$$

and approximate b_j by

$$\hat{b}_j^{(n)} = \frac{1}{n} \sum_{t=1}^{n} R_t g_j(X_t)\,.$$

The method of [Bradtke and Barto, 1996] corresponds to the case when $\mathcal{F} = \mathcal{G}$ (the corresponding "exact" method is the so-called Bubnov-Galerkin method).

Under various additional assumptions (e.g., independence of the snippets), $\hat{A}_{ij}^{(n)}$ and $\hat{b}_{ij}^{(n)}$ converge with probability one to A_{ij} and b_j, respectively. A stability analysis of the equation $A\alpha = b$ can then be used to derive bounds on the quality of solutions obtained by solving $\hat{A}^{(n)}\alpha = \hat{b}^{(n)}$. One can further introduce appropriate regularization to appropriately stabilize the estimation process in the case of large dimensions (d) and small sample sizes (n). By regularizing using an ℓ^1-norm, an efficient procedure that shows only a mild dependence on d can be arrived at. Details are available in the forthcoming paper [Pires and Szepesvári, 2011].

4. Conclusions

We have discussed the connection between a popular method in reinforcement learning, the so-called least-squares temporal difference (LSTD) method, and Galerkin's method. Although this connection was recognized before, we found it useful to explain it as it is lesser known within the reinforcement learning community and because the connection leads to new insights into the issues related to statistical performance of this important algorithm. We hope that this short abstract will foster further research to explore this and other connections between reinforcement learning and statistical inverse problems.

References

[Antos et al., 2008] Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129. Published Online First: 14 Nov, 2007.

[Bradtke and Barto, 1996] Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57.

[Farahmand et al., 2009] Farahmand, A., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2009). Regularized policy iteration. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *NIPS-21*, pages 441–448. MIT Press.

[Kirsch, 2011] Kirsch, A. (2011). *An Introduction to the Mathematical Theory of Inverse Problems*. Springer, 2nd edition.

[Pires and Szepesvári, 2011] Pires, B. and Szepesvári, C. (2011). Statistical analysis of ℓ^1-penalized linear estimation with applications. under preparation.

[Scherrer, 2010] Scherrer, B. (2010). Should one compute the temporal difference fix point or minimize the Bellman residual? The unified oblique projection view. In *ICML 2010*, pages 959–966.

[Sutton, 1988] Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9–44.

[Yu and Bertsekas, 2010] Yu, H. and Bertsekas, D. (2010). Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 35(2):306–329.

Optimal rates of estimation of high-dimensional matrices

Alexandre Tsybakov

Assume that we observe n entries or linear combinations of entries of an unknown $m \times T$ matrix A corrupted by noise. We propose a new nuclear-norm penalized estimator of A called the linearized matrix Lasso, and establish a general sharp oracle inequality for this estimator for arbitrary values of n, m, T under the condition of isometry in expectation. Then this method is applied to the matrix completion problem. In this case, the estimator admits a simple explicit form and we prove that it satisfies oracle inequalities with faster rates of convergence than in the previous works. They are valid, in particular, in the high-dimensional setting $mT \gg n$. We show that the obtained rates are optimal up to logarithmic factors in a minimax sense and also derive, for any fixed matrix A, a non-minimax lower bound on the rate of convergence of our estimator, which coincides with the upper bound up to a constant factor. Finally, we show that our procedure provides an exact recovery of the rank of A with probability close to 1. We also discuss the statistical learning setting where there is no underlying model determined by A and the aim is to find the best trace regression model approximating the data. As a by-product, we show that, under the Restricted Eigenvalue condition, the usual vector Lasso estimator satisfies a sharp oracle inequality (i.e., an oracle inequality with leading constant 1). This is a joint work with Vladimir Koltchinskii and Karim Lounici.

References

[Koltchinskii et al.(2010)Koltchinskii, Lounici and Tsybakov] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2010). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. To appear in *Annals of Statist.*, `arXiv:1011.6256`.

Insuring against loss of evidence and capital

Vladimir Vovk

(joint work with A. Philip Dawid, Steven de Rooij, Peter Grünwald, Wouter M. Koolen, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin)

This talk is about worst-case results, à la prediction with expert advice. They admit two main interpretations:

Statistical:: Suppose you have a lot of evidence against a null hypothesis. How can you avoid losing it all?

Financial:: Suppose your current capital is large. Should you continue trading (risking losing all your money) or should you stop (preventing your capital from growing further)? Can we compromise? What trade-offs are open to us?

We are trading in one security X in a financial market. Normalize the initial price X_0 to 1 and the investor's initial capital $\mathbb{K}_0$ to 1. This is our trading protocol:

```
$X_0 := 1$ and $\mathbb{K}_0 := 1$
FOR $t = 1, 2, \ldots$:
  Investor announces $p_t \in \mathbb{R}$
  Market announces $X_t \in [0, \infty)$
  $\mathbb{K}_t := \mathbb{K}_{t-1} + p_t(X_t - X_{t-1})$
END FOR
```

$\mathbb{K}_t$: Investor's capital. A *trading strategy* is a strategy for Investor in this protocol.

Set $X_t^* := \max_{s \le t} X_s$. We would like to have a trading strategy that guarantees

$$\mathbb{K}_t \ge F(X_t^*)$$

for all t, where, as $y \to \infty$, $F(y) \to \infty$ almost as fast as y. If this inequality can be guaranteed, F is a *simple lookback adjuster* (*SLA*).

More generally, we can ask when Investor can guarantee

$$\mathbb{K}_t \geq F(X_t^*, X_t), \quad \forall t.$$

Such F will be called *lookback adjusters* (*LAs*). We are interested only in nonnegative SLAs and FAs F.

The set of SLAs and LAs is too big. More manageable subsets consist of admissible SLAs (*ASLA*s) and admissible LAs (*ALA*s), defined as follows. An SLA (or LA) G *dominates* an SLA (or LA) F if $G \geq F$. F is *admissible* if it is not dominated by any G different from F.

Simple lookback adjusters

Theorem 4. *Any SLA is dominated by an ASLA. A function $F : [1, \infty) \to [0, \infty)$ is an ASLA if and only if it is increasing, right-continuous, and satisfies*

$$\int_1^\infty \frac{F(y)}{y^2} dy = 1.$$

It is impossible to have $F(y) = y$ (it would mean guaranteeing $\mathbb{K}_t \geq X_t^*$ for all t), but we want to come as close to this as possible. Let $\alpha \in (0, 1)$. These are 2 simple examples:

- There exists a trading strategy guaranteeing
$$\mathbb{K}_t \geq \alpha \, (X_t^*)^{1-\alpha}$$
for all t.
- There exists a trading strategy guaranteeing
$$\mathbb{K}_t \geq \alpha(1+\alpha)^\alpha \frac{X_t^*}{\ln^{1+\alpha} X_t^*}$$
whenever $X_t^* \geq e^{1+\alpha}$.

The talk covered some simple statistical applications.

Lookback adjusters

Let f_r be the right derivative of f.

Theorem 5. *Every LA is dominated by an ALA. A positive function $F(X^*, X)$ with domain $X^* \in [1, \infty)$ and $X \in [0, X^*]$ is an ALA if and only if the following two conditions are satisfied:*

- *the function*
$$F^=(X^*) := F(X^*, X^*), \quad X^* \in [1, \infty),$$
is increasing, concave, and satisfies $F^=(1) = 1$ and $F_\mathrm{r}^=(1) \leq 1$;
- *for each $X^* \in [1, \infty)$, the function $F(X^*, X)$ is linear in X and its slope is equal to $F_\mathrm{r}^=(X^*)$.*

Option pricing

We would like to price the following exotic option O_G (a kind of perpetual American lookback): at the time t of her choice, the option's owner is entitled to $G(X_t^*)$, where G is a given nonnegative increasing function. The result about ASLAs can be restated as: the *upper price* of this option at time 0 (after learning X_0) is $\int_1^\infty G(y)y^{-2}dy$. Formally, $\int_1^\infty G(y)y^{-2}dy$ is the smallest initial capital c such that there exists a trading strategy starting with c and guaranteeing $\mathbb{K}_t \geq G(X_t^*)$ for all t (intuitively, the seller can always meet his obligation).

The formula $\int_1^\infty G(y)y^{-2}dy$ assumes $X_0 = 1$. Without this assumption, the upper price at time 0 is $\int_1^\infty G(X_0 y)y^{-2}dy$ (by re-scaling).

Notice that the upper price of O_G can be written as $X_0 \int_{X_0}^\infty G(y)y^{-2}dy$, which is the expected value of G with respect to the probability measure P on $[X_0, \infty)$ with density X_0/y^2. It plays the role of *risk-neutral probability* (but, unusually, emerges in a heavily incomplete market).

References

[1] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science* **26**, 84–101 (2011).

[2] A. Philip Dawid, Steven de Rooij, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Insuring against loss of evidence in game-theoretic probability. *Statistics and Probability Letters* **81**, 157–162 (2011).

[3] A. Philip Dawid, Steven de Rooij, Peter Grünwald, Wouter Koolen, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Probability-free pricing of adjusted American lookbacks. arXiv:1108.4113 [q-fin.PR], August 2011.

Weakly universally consistent static forecasting of stationary and ergodic time series via local averaging and least squares estimates

Harro Walk

(joint work with Tina Felber, Daniel Jones, Michael Kohler)

Given a stationary and ergodic time series the problem of estimating the conditional expectation of the dependent variable at time zero given the infinite past is considered. It is shown that the mean squared error of a combination of suitably defined local averaging or least squares estimates converges to zero for all distributions whenever the dependent variable is square integrable.

Let $((X_n, Y_n))_{n\in\mathbb{Z}}$ be a stationary and ergodic sequence of $\mathbb{R}^d \times \mathbb{R}$-valued random variables with $\mathbf{E}\left\{Y_0^2\right\} < \infty$. The abbreviations $X_k^l = (X_k, \ldots, X_l)$, $\quad Y_k^l = (Y_k, \ldots, Y_l)$ and $\mathcal{D}_k^l = \{(X_k, Y_k), \ldots, (X_l, Y_l)\}$, $k \leq l$, will be used, correspondingly for realizations x_k^l y_k^l, d_k^l. The following static forecasting problem is considered: On the basis of the data set $\mathcal{D}_{-n}^{-1}$ and X_0, construct simple estimates $m_n(X_0, \mathcal{D}_{-n}^{-1})$ of $\mathbf{E}\left\{Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1}\right\}$ which are weakly consistent in the sense that they satisfy

$$\mathbf{E}\left\{\left|m_n(X_0, \mathcal{D}_{-n}^{-1}) - \mathbf{E}\left\{Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1}\right\}\right|^2\right\} \to 0 \quad (n \to \infty).$$

One starts with defining a parameter set $\mathcal{P} = \{(k,r,N) : k,r,N \in \mathbb{N}\}$ and elementary estimates (experts) $\tilde{m}^{(i)}_{n,(k,r,N)}(x^0_{-k}, y^{-1}_{-k}; d^{-1}_{-n})$ and $\hat{m}^{(i)}_{n,(k,r,N)}(x_0, d^{-1}_{-n})$ ($i \in \{1,2,3,4\}$) of $\mathbf{E}\left\{Y_0 \mid X^0_{-k} = x^0_{-k}, Y^{-1}_{-k} = y^{-1}_{-k}\right\}$, where k indicates how far back the estimate will look, r determines the set of approximating functions and N indicates the level of truncation by the truncation operator T_N.

For the first estimate one uses bandwidths $h_r > 0$ satisfying $h_r \to 0 \quad (r \to \infty)$ and a usual continuous kernel function $K : \mathbb{R}^{(k+1)\cdot d+k} \to \mathbb{R}_+$. With $\frac{0}{0} := 0$ the kernel estimate $\tilde{m}^{(1)}_{n,(k,r,N)}$ is defined by

$$\begin{aligned}&\tilde{m}^{(1)}_{n,(k,r,N)}(u^0_{-k}, v^{-1}_{-k}; d^{-1}_{-n})\\ &:= \begin{cases} \dfrac{\sum_{i=-n+k-1}^{-2} T_N(y_{i+1}) K\left(\frac{(x^{i+1}_{i-k+1}, y^i_{i-k+1}) - (u^0_{-k}, v^{-1}_{-k})}{h_r}\right)}{\sum_{i=-n+k-1}^{-2} K\left(\frac{(x^{i+1}_{i-k+1}, y^i_{i-k+1}) - (u^0_{-k}, v^{-1}_{-k})}{h_r}\right)} & \text{if } n \geq k+1,\\ 0 & \text{else.}\end{cases}\end{aligned}$$

Correspondingly, estimates $\widetilde{m}^{(2)}_{n,(k,r,N)}$ and $\widetilde{m}^{(3)}_{n,(k,r,N)}$ of partitioning type and of nearest neighbors type (with suitable tie-breaking), respectively, are defined.

For the fourth estimate, let $B_1, \ldots, B_{K_r}$ be bounded and continuous functions $B_j : \left(\mathbb{R}^d\right)^{k+1} \times \mathbb{R}^k \to [-B, B]$ for some $B > 0$, and set

$$\mathcal{F}_{k,r} = \left\{ \sum_{j=1}^{K_r} a_j \cdot B_j \quad : \quad a_j \in [-L_r, L_r] \quad (j = 1, \ldots, K_r) \right\}$$

with $K_r \to \infty$, $L_r \to \infty$ $(r \to \infty)$, where $\cup_r \mathcal{F}_{k,r}$ satisfies an L_2-denseness assumption for each k and each probability measure (fulfilled for suitable spline function spaces).

Define the corresponding least squares estimate by

$$\tilde{m}^{(4)}_{n,(k,r,N)}\left(u^0_{-k}, v^{-1}_{-k}; d^{-1}_{-n}\right) := \begin{cases} \bar{m}^{(4)}_{n,(k,r,N)}\left(u^0_{-k}, v^{-1}_{-k}; d^{-1}_{-n}\right) & \text{if } n \geq k+1,\\ 0 & \text{else,}\end{cases}$$

where

$$\bar{m}^{(4)}_{n,(k,r,N)}(\cdot\,; d^{-1}_{-n}) := \arg\min_{f \in \mathcal{F}_{k,r}} \frac{1}{n-k} \sum_{i=-n+k}^{-1} \left| f(x^i_{i-k}, y^{i-1}_{i-k}) - T_N\left(y_i\right)\right|^2 .$$

For $i \in \{1,2,3,4\}$ choose $0 < s < \frac{1}{2}$ and set

$$\hat{m}^{(i)}_{n,(k,r,N)}(x_0, d^{-1}_{-n}) := T_{n^s}\left(\tilde{m}^{(i)}_{n,(k,r,N)}(x^0_{-k}, y^{-1}_{-k}; d^{-1}_{-n})\right).$$

The prediction strategy is defined as a convex combination of these experts using weights, which are the higher the better the expert performed in the past. After $n-1$ rounds of play the normalized cumulative squared prediction error of $\hat{m}^{(i)}_{n,(k,r,N)}$

on the string d_{-n}^{-1} defined by

$$L_n^{(i)}(k,r,N) := L_n^{(i)}(k,r,N)(d_{-n}^{-1})$$
$$:= \frac{1}{n-1}\sum_{j=-n}^{-2}(T_{n^s}(y_{j+1}) - \hat{m}_{j+n+1,(k,r,N)}^{(i)}(x_{j+1}, d_{-n}^{j}))^2$$

quantizes the performance of the expert in the past. Let $(p_{(k,r,N)})_{(k,r,N)\in\mathcal{P}}$ be a probability distribution such that $p_{(k,r,N)} > 0$ for all $(k,r,N) \in \mathcal{P}$. Put $c_n = 8n^{2s}$ and define weights, which depend on this cumulative loss, by

$$w_{n,(k,r,N)}^{(i)} := p_{(k,r,N)} \cdot \exp\left(\frac{-(n-1)L_n^{(i)}(k,r,N)}{c_n}\right).$$

Now with normalized weights

$$v_{n,(k,r,N)}^{(i)} := \frac{w_{n,(k,r,N)}^{(i)}}{\sum_{(k,r,N)\in\mathcal{P}} w_{n,(k,r,N)}^{(i)}},$$

$\hat{m}_n^{(i)}$ is defined by

$$\hat{m}_n^{(i)}(x_0, d_{-n}^{-1}) := \sum_{(k,r,N)\in\mathcal{P}} v_{n,(k,r,N)}^{(i)} \cdot \hat{m}_{n,(k,r,N)}^{(i)}(x_0, d_{-n}^{-1}).$$

In order to estimate $\mathbf{E}\left\{Y_0 \mid X_{-\infty}^0, Y_{-\infty}^{-1}\right\}$ the arithmetic mean is used:

$$m_n^{(i)}(X_0, R_{-n}^{-1}) := \frac{1}{n}\sum_{j=1}^{n}\hat{m}_j^{(i)}\left(X_0, R_{-j}^{-1}\right).$$

Then

$$\mathbf{E}\left\{\left|m_n^{(i)}(X_0, R_{-n}^{-1}) - \mathbf{E}\left\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\right\}\right|^2\right\} \to 0 \quad (n\to\infty),$$

$i \in \{1,\ldots,4\}$, for all stationary and ergodic sequences $((X_n, Y_n))_{n\in\mathbb{Z}}$ of $\mathbb{R}^d \times \mathbb{R}$-valued random variables with $\mathbf{E}\left\{Y_0^2\right\} < \infty$ (weak universal consistency).

For the proof one uses stationarity, an ergodic theorem for random variables in a separable Banach space and an inequality of Kivinen and Warmuth (1999) and Singer and Feder (1999) in the theory of estimation via experts.

In the talk erroneously $m_n^{(i)}(X_0, \mathcal{D}_{-n}^{-1}) \to \mathbf{E}\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\}$ a.s., $i \in \{1,2,3\}$, for all stationary and ergodic sequences $((X_n, Y_n))_{n\in\mathbb{Z}}$ with $\mathbf{E}\{|Y_0|\} < \infty$ was stated. The problem of strong universal consistency in static forecasting (under the assumption of mere integrability of Y_0) remains open.

References

[1] Felber, T., Jones, D., Kohler, M., and Walk, H. (2011). Weakly universally consistent static forecasting of stationary and ergodic time series via local averaging and least squares estimates. Universitt Stuttgart, Fachbereich Mathematik, Stuttgarter Mathematische Berichte 2011. Submitted for publication.

[2] Kivinen, J. and Warmuth, M. K. (1999). Averaging expert predictions. In *Computational Learning Theory: Proceedings of the Fourth European Conference, Eurocolt99,* Simon, H. U. and Fischer, P., eds., pp. 153-167. Springer, Berlin.

[3] Singer, A. and Feder, M. (1999). Universal linear prediction by model order weighting. *IEEE Transactions on Signal Processing 47*, pp. 2685-2699.

Reporter: Sebastien Bubeck

Participants

Prof. Dr. Peter L. Bartlett
Computer Science Division
University of California, Berkeley
Soda Hall
Berkeley , CA 94720-1776
USA

Prof. Dr. Misha Belkin
Department of Computer Science
and Engineering
Ohio State University
2015 Neil Ave.
Columbus , OH 43210-1277
USA

Prof. Dr. Sebastien Bubeck
Departament de Matematiques
Universitat Autonoma de Barcelona
Campus Universitari
E-08193 Bellaterra

Prof. Dr. Sara van de Geer
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
CH-8092 Zürich

Prof. Dr. Laszlo Györfi
Department of Computer Science and
Information Theory
Budapest University of Techn.& Economics
Stoczek u. 2
H-1521 Budapest

Prof. Dr. Elad Hazan
Department of Operations Research
TECHNION
Israel Institute of Technology
32000 Haifa
ISRAEL

Prof. Dr. Gabor Lugosi
Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
E-08005 Barcelona

Gergely Neu
Department of Computer Science and
Information Theory
Budapest University of Techn.& Economics
Stoczek u. 2
H-1521 Budapest

Dr. Guillaume Obozinski
INRIA/ENS
CS 81321
23, avenue d'Italie
F-75214 Paris Cedex 13

Prof. Dr. Massimiliano Pontil
Department of Computer Science
University College London
Gower Street
GB-London WC1E 6BT

Prof. Dr. Alexander Rakhlin
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia , PA 19104-6340
USA

Dr. Lorenzo Rosasco
MIT
Bldg. 46-5155
43 Vassar Street
Cambride , MA 02139
USA

Prof. Dr. Ingo Steinwart
Fachbereich Mathematik
Universität Stuttgart
Pfaffenwaldring 57
70569 Stuttgart

Prof. Dr. Wolfgang Stummer
Department Mathematik
Universität Erlangen-Nürnberg
Bismarckstr. 1 1/2
91054 Erlangen

Prof. Dr. Csaba Szepesvari
Department of Computing Science
University of Alberta
Edmonton T6G 2E8
CANADA

Prof. Dr. Alexandre B. Tsybakov
Laboratoire de Probabilites
Universite Paris 6
4 place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Vladimir Vovk
Department of Computer Science
Royal Holloway
University of London
Egham
Surrey TW20 0EX
UNITED KINGDOM

Prof. Dr. Harro Walk
Fachbereich Mathematik
Universität Stuttgart
Pfaffenwaldring 57
70569 Stuttgart