Editors' Introduction

Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto

Philosophers have pondered the phenomenon of learning for millennia; scientists and psychologists have studied learning for more than a century. But the analysis of learning as a *computational* and *algorithmic* phenomenon is much more recent, going back only a few decades. Learning theory is now an active research area that incorporates ideas, problems, and techniques from a wide range of disciplines including statistics, artificial intelligence, information theory, pattern recognition, and theoretical computer science. Learning theory has many robust connections with more applied research in machine learning and has made significant contributions to the development of applied systems and to fields such as electronic commerce and computational biology.

Since learning is a complex and multi-faceted phenomenon, it should come as no surprise that a wide range of different theoretical models of learning have been developed and analyzed. This diversity in the field is well reflected in the topics addressed by the invited speakers to ALT 2007 and DS 2007, and by the range of different research topics that have been covered by the contributors to this volume in their papers. The research reported here ranges over areas such as unsupervised learning, inductive inference, complexity and learning, boosting and reinforcement learning, query learning models, grammatical inference, online learning and defensive forecasting, and kernel methods. In this introduction we give an overview first of the five invited talks of ALT 2007 and DS 2007 and then of the regular contributions in this volume. We have grouped the papers under different headings to highlight certain similarities in subject matter or approach, but many papers span more than one area and other alternative groupings are certainly possible; the taxonomy we offer is by no means absolute.

Avrim Blum works on learning theory, online algorithms, approximation algorithms, and algorithmic game theory. His interests within learning theory include similarity functions and clustering, semi-supervised learning and co-training, online learning algorithms, kernels, preference elicitation and query learning, noise-tolerant learning, and attribute-efficient learning. In his invited talk for ALT 2007, Avrim spoke about developing a theory of similarity functions for learning and clustering problems. Some of the aims of this work are to provide new insights into what makes kernel functions useful for learning, and to understand what are the minimal conditions on a similarity function that allow it to be useful for clustering.

Alexander Smola works on nonparametric methods for estimation, in particular kernel methods and exponential families. He studies estimation techniques including Support Vector Machines, Gaussian Processes and Conditional Random Fields, and uses these techniques on problems in bioinformatics, pattern recognition, text analysis, computer vision, network security, and optimization for parallel processing. In his invited lecture for ALT 2007, co-authored with Arthur Gretton, Le Song, and Bernhard Schölkopf, Alexander spoke about a technique for comparing distributions without the need for density estimation as an intermediate step. The approach relies on mapping the distributions into a reproducing kernel Hilbert space, and has a range of applications that were presented in the talk.

Masaru Kitsuregawa works on data mining, high performance data warehousing, high performance disk and tape arrays, parallel database processing, data storage and the Web, and related topics. His invited lecture for DS 2007 was about "Challenges for Info-plosion."

Thomas G. Dietterich studies topics in machine learning including sequential and spatial supervised learning, transfer learning, and combining knowledge and data to learn in knowledge-rich/data-poor application problems. He works on applying machine learning to a range of problems such as ecosystem informatics, intelligent desktop assistants, and applying AI to computer games. His invited lecture for DS 2007 discussed the role that machine learning can play in ecosystem informatics; this is a field that brings together mathematical and computational tools to address fundamental scientific and application problems in the ecosystem sciences. He described two on-going research efforts in ecosystem informatics at Oregon State University: (a) the application of machine learning and computer vision for automated arthropod population counting, and (b) the application of linear Gaussian dynamic Bayesian networks for automated cleaning of data from environmental sensor networks.

Jürgen Schmidhuber has worked on a range of topics related to learning, including artificial evolution, learning agents, reinforcement learning, metalearning, universal learning algorithms, Kolmogorov complexity and algorithmic probability. This work has led to applications in areas such as finance, robotics, and optimization. In his invited lecture (joint for ALT 2007 and DS 2007), Jürgen spoke about the algorithmic nature of discovery, perceived beauty, and curiosity. Jürgen has been thinking about this topic since 1994, when he postulated that among several patterns classified as "comparable" by some subjective observer, the subjectively most beautiful is the one with the simplest (shortest) description, given the observer's particular method for encoding and memorizing it. As one example of this phenomenon, mathematicians find beauty in a simple proof with a short description in the formal language they are using.

We now turn our attention to the regular contributions contained in this volume.

Inductive Inference. Research in inductive inference follows the pioneering work of Gold, who introduced a recursion-theoretic model of "learning in the limit." In the basic inductive inference setting, a learning machine is given a sequence of (arbitrarily ordered) examples drawn from a (recursive or recursively enumerable) language L, which belongs to a known class C of possible languages. The learning machine maintains a hypothesis which may be updated after each successive element of the sequence is received; very roughly speaking, the goal is for the learning machine's hypothesis to converge to the target language after finitely many steps. Many variants of this basic scenario have been studied in inductive inference during the decades since Gold's original work.

John Case, Timo Kötzing and Todd Paddock study a setting of learning in the limit in which the time to produce the final hypothesis is derived from some ordinal which is updated step by step downwards until it reaches zero, via some "feasible" functional. Their work first proposes a definition of feasible iteration of feasible learning functionals, and then studies learning hierarchies defined in terms of these notions; both collapse results and strict hierarchies are established under suitable conditions. The paper also gives upper and lower runtime bounds for learning hierarchies related to these definitions, expressed in terms of exponential polynomials.

John Case and Samuel Moelius III study *iterative learning.* This is a variant of the Gold-style learning model described above in which each of a learner's output conjectures may depend only on the learner's current conjecture and on the current input element. Case and Moelius analyze two extensions of this iterative model which incorporate parallelism in different ways. Roughly speaking, one of their results shows that running several distinct instantiations of a single learner in parallel can actually increase the power of iterative learners. This provides an interesting contrast with many standard settings where allowing parallelism only provides an efficiency improvement. Another result deals with a "collective" learner which is composed of a collection of communicating individual learners that run in parallel.

Sanjay Jain, Frank Stephan and Nan Ye study some basic questions about how hypothesis spaces connect to the class of languages being learned in Gold-style models. Building on work by Angluin, Lange and Zeugmann, their paper introduces a comprehensive unified approach to studying learning languages in the limit relative to different hypothesis spaces. Their work distinguishes between four different types of learning as they relate to hypothesis spaces, and gives results for vacillatory and behaviorally correct learning. They further show that every behaviorally correct learnable class has a *prudent* learner, i.e., a learner using a hypothesis space such that it learns every set in the hypothesis space.

Sanjay Jain and Frank Stephan study Gold-style learning of languages in some special numberings such as Friedberg numberings, in which each set has exactly one number. They show that while explanatorily learnable classes can all be learned in some Friedberg numberings, this is not the case for either behaviorally correct learning or finite learning. They also give results on how other properties of learners, such as consistency, conservativeness, prudence, iterativeness, and non U-shaped learning, relate to Friedberg numberings and other numberings.

Complexity aspects of learning. Connections between complexity and learning have been studied from a range of different angles. Work along these lines has been done in an effort to understand the computational complexity of various learning tasks; to measure the complexity of classes of functions using parameters such as the Vapnik-Chervonenkis dimension; to study functions of interest in learning theory from a complexity-theoretic perspective; and to understand connections between Kolmogorov-style complexity and learning. All four of these aspects were explored in research presented at ALT 2007.

Vitaly Feldman, Shrenek Shah, and Neal Wadhwa analyze two previously studied variants of Angluin's exact learning model that make learning more challenging: learning from equivalence and incomplete membership queries, and learning with random persistent classification noise in membership queries. They show that under cryptographic assumptions about the computational complexity of solving various problems the former oracle is strictly stronger than the latter, by demonstrating a concept class that is polynomial-time learnable from the former oracle but is not polynomial-time learnable from the latter oracle. They also resolve an open question of Bshouty and Eiron by showing that the incomplete membership query oracle is strictly weaker than a standard perfect membership query oracle under cryptographic assumptions.

César Alonso and José Montaña study the Vapnik-Chervonenkis dimension of concept classes that are defined in terms of arithmetic operations over real numbers. Such bounds are of interest in learning theory because of the fundamental role the Vapnik-Chervonenkis dimension plays in characterizing the sample complexity required to learn concept classes. Strengthening previous results of Goldberg and Jerrum, Alonso and Montaña give upper bounds on the VC dimension of concept classes in which the membership test for whether an input belongs to a concept in the class can be performed by an arithmetic circuit of bounded depth. These new bounds are polynomial both in the depth of the circuit and in the number of parameters needed to codify the concept.

Vikraman Arvind, Johannes Köbler, and Wolfgang Lindner study the problem of properly learning k-juntas and variants of k-juntas. Their work is done from the vantage point of parameterized complexity, which is a natural setting in which to consider the junta learning problem. Among other results, they show that the consistency problem for k-juntas is $W[2]$-complete, that the class of k-juntas is fixed parameter PAC learnable given access to a $W[2]$ oracle, and that k-juntas can be fixed parameter improperly learned with equivalence queries given access to a $W[2]$ oracle. These results give considerable insight on the junta learning problem.

The goal in transfer learning is to solve new learning problems more efficiently by leveraging information that was gained in solving previous related learning problems. One challenge in this area is to clearly define the notion of "relatedness" between tasks in a rigorous yet useful way. M. M. Hassan Mahmud analyzes transfer learning from the perspective of Kolmogorov complexity. Roughly speaking, he shows that if tasks are related in a particular precise sense, then joint learning is indeed faster than separate learning. This work strengthens previous work by Bennett, Gács, Li, Vitányi and Zurek.

Online Learning. Online learning proceeds in a sequence of rounds, where in each round the learning algorithm is presented with an input x and must generate a prediction y (a bit, a real number, or something else) for the label of x. Then the learner discovers the true value of the label, and incurs some loss which depends on the prediction and the true label. The usual overall goal is to keep the total loss small, often measured relative to the optimal loss over functions from some fixed class of predictors.

Jean-Yves Audibert, Rémi Munos and Csaba Szepesvári deal with the stochastic multi-armed bandit setting. They study an Upper Confidence Bound algorithm that takes into account the empirical variance of the different arms. They give an upper bound on the expected regret of the algorithm, and also analyze the concentration of the regret; this risk analysis is of interest since it is clearly useful to know how likely the algorithm is to have regret much higher than its expected value. The risk analysis reveals some unexpected tradeoffs between logarithmic expected regret and concentration of regret.

Jussi Kujala and Tapio Elomaa also consider a multi-armed bandit setting. They show that the "Follow the Perturbed Leader" technique can be used to obtain strong regret bounds (which hold against the best choice of a fixed lever in hindsight) against adaptive adversaries in this setting. This extends previous results for FPL's performance against non-adaptive adversaries in this setting.

Vovk's Aggregating Algorithm is a method of combining hypothesis predictors from a pool of candidates. Steven Busuttil and Yuri Kalnishkan show how Vovk's Aggregating Algorithm (AA) can be applied to online linear regression in a setting where the target predictor may change with time. Previous work had only used the Aggregating Algorithm in a static setting; the paper thus sheds new light on the methods that can be used to effectively perform regression with a changing target. Busuttil and Kalnishkan also analyze a kernel version of the algorithm and prove bounds on its square loss.

Unsupervised Learning. Many of the standard problems and frameworks in learning theory fall under the category of "supervised learning" in that learning is done from labeled data. In contrast, in unsupervised learning there are no labels provided for data points; the goal, roughly speaking, is to infer some underlying structure from the unlabeled data points that are received. Typically this means clustering the unlabeled data points or learning something about a probability distribution from which the points were obtained.

Markus Maier, Matthias Hein, and Ulrike von Luxburg study a scenario in which a learning algorithm receives a sample of points from an unknown distribution which contains a number of distinct clusters. The goal in this setting is to construct a "neighborhood graph" from the sample, such that the connected component structure of the graph mirrors the cluster ancestry of the sample points. They prove bounds on the performance of the k-nearest neighbor algorithm for this problem and also give some supporting experimental results. Markus received the E. M. Gold Award for this paper, as the program committee felt that it was the most outstanding contribution to ALT 2007 which was co-authored by a student.

Kevin Chang considers an unsupervised learning scenario in which a learner is given access to a sequence of samples drawn from a mixture of uniform distributions over rectangles in d-dimensional Euclidean space. He gives a streaming algorithm which makes only a small number of passes over such a sequence, uses a small amount of memory, and constructs a high-accuracy (in terms of statistical distance) hypothesis density function for the mixture. A notable feature of the algorithm is that it can handle samples from the mixture that are presented in any arbitrary order. This result extends earlier work of Chang and Kannan which dealt with mixtures of uniform distributions over rectangles in one or two dimensions.

Language Learning. The papers in this group deal with various notions of learning languages in the limit from positive data. Ryo Yoshinaka's paper addresses the question of what precisely is meant by the notion of efficient language learning in the limit; despite the clear intuitive importance of such a notion, there is no single accepted definition. The discussion focuses particularly on learning very simple grammars and minimal simple grammars from positive data, giving both positive and negative results on efficient learnability under various notions.

François Denis and Amaury Habrard study the problem of learning stochastic tree languages, based on a sample of trees independently drawn according to an unknown stochastic language. They extend the notion of rational stochastic languages over strings to the domain of trees. Their paper introduces a canonical representation for rational stochastic languages over trees, and uses this representation to give an efficient inference algorithm that identifies the class of rational stochastic tree languages in the limit with probability 1.

Query Learning. In query learning the learning algorithm works by making queries of various types to an oracle or teacher; this is in contrast with "passive" statistical models where the learner typically only has access to random examples and cannot ask questions. The most commonly studied types of queries are membership queries (requests for the value of the target function at specified points) and equivalence queries (requests for counterexamples to a given hypothesis). Other types of queries, such as subset queries (in which the learner asks whether the current hypothesis is a subset of the target hypothesis, and if not, receives a negative counterexample) and superset queries, are studied as well.

Sanjay Jain and Efim Kinber study a query learning framework in which the queries used are variants of the standard queries described above. In their model the learner receives the *least* negative counterexample to subset queries, and is also given a "correction" in the form of a positive example which is nearest to the negative example; they also consider similarly modified membership queries. These variants are motivated in part by considerations of human language learning, in which corrected versions of incorrect utterances are often provided as part of the learning process. Their results show that "correcting" positive examples can sometimes give significant additional power to learners.

Cristina Tîrnăucă and Timo Knuutila study query learning under a different notion of correction queries, in which the prefix of a string (the query) is "corrected" by the teacher responding with the lexicographically first suffix that yields a string in the language. They give polynomial time algorithms for pattern languages and k-reversible languages using correction queries of this sort. These results go beyond what is possible for polynomial-time algorithms using membership queries alone, and thus demonstrate the power of learning from these types of correction queries.

Lev Reyzin and Nikhil Srivastava study various problems of learning and verifying properties of hidden graphs given query access to the graphs. This setting lends itself naturally to a range of query types that are somewhat different from those described above; these include edge detection, edge counting, and shortest path queries. Reyzin and Srivastava give bounds on learning and verifying general graphs, degree-bounded graphs, and trees with these types of queries. These results extend our understanding of what these types of queries can accomplish.

Rika Okada, Satoshi Matsumoto, Tomoyuki Uchida, Yusuke Suzuki and Takayoshi Shoudai study learnability of finite unions of linear graph patterns from equivalence queries and subset queries. These types of graph patterns are useful for data mining from semi-structured data. The authors show that positive results can be achieved for learning from equivalence and subset queries (with counterexamples), and give negative results for learning from restricted subset queries (in which no counterexamples are given).

Kernel-Based Learning. A kernel function is a mapping which, given two inputs, implicitly represents each input as a vector in some (possibly high-dimensional or infinite dimensional) feature space and outputs the inner product between these two vectors. Kernel methods have received much attention in recent years in part because it is often possible to compute the value of the kernel function much more efficiently than would be possible by performing an explicit representation of the input as a vector in feature space. Kernel functions play a crucial role in Support Vector Machines and have a rich theory as well as many uses in practical systems.

Developing new kernel functions, and selecting the most appropriate kernels for particular learning tasks, is an active area of research. One difficulty in constructing kernel functions is in ensuring that they obey the condition of positive semidefiniteness. Kilho Shin and Tetsuji Kuboyama give a sufficient condition under which it is ensured that new candidate kernels constructed in a particular way from known positive semidefinite kernels will themselves be positive semidefinite and hence will indeed be legitimate kernel functions. Their work gives new insights into several kernel functions that have been studied recently such as principal-angle kernels, determinant kernels, and codon-improved kernels.

Guillaume Stempfel and Liva Ralaivola study how kernels can be used to learn data separable in the feature space except for the presence of random classification noise. They describe an algorithm which combines kernel methods, random projections, and known noise tolerant approaches for learning linear separators over finite dimensional feature spaces, and give a PAC style analysis of the algorithm. Given noisy data which is such that the noise-free version would be linearly separable with a suitable margin in the implicit feature space, their approach yields an efficient algorithm for learning even if the implicit feature space has infinitely many dimensions.

Adam Kowalczyk's paper deals with analyzing hypothesis classes that consist of linear functionals superimposed with "smooth" feature maps; these are the types of hypotheses generated by many kernel methods. The paper studies continuity of two important performance metrics, namely the error rate and the area under the ROC (receiver operating characteristic curve), for hypotheses of this sort. Using tools from real analysis, specifically transversality theory, he shows that pointwise convergence of hypotheses implies convergence of these measures with probability 1 over the selection of the test sample from a suitable probability density.

Other Directions. Several papers presented at ALT do not fit neatly into the above categories, but as described below each of these deals with an active and interesting area of research in learning theory.

Hypothesis boosting is an approach to combining many weak classifiers, or "rules of thumb," each of which performs only slightly better than random guessing, to obtain a high-accuracy final hypothesis. Boosting algorithms have been intensively studied and play an important role in many practical applications. In his paper, Takafumi Kanamori studies how boosting can be applied to estimate conditional probabilities of output labels in a multiclass classification setting. He proposes loss functions for boosting algorithms that generalize the known margin-based loss function and shows how regularization can be introduced with an appropriate instantiation of the loss function.

Reinforcement learning is a widely studied approach to sequential decision problems that has achieved considerable success in practice. Dealing with the "curse of dimensionality," which arises from large state spaces in Markov decision processes, is a major challenge. One approach to dealing with this challenge is *state aggregation*, which is based on the idea that similar states can be grouped together into meta-states. In his paper Ronald Ortner studies pseudometrics for measuring similarity in state aggregation. He proves an upper bound on the loss incurred by working with aggregated states rather than original states and analyzes how online aggregation can be performed when the MDP is not known to the learner in advance.

In defensive forecasting, the problem studied is that of online prediction of the binary label associated with each instance in a sequence of instances. In this line of work no assumption is made that there exists a hidden function dictating the labels, and in contrast with other work in online learning there is no comparison class or "best expert" that is compared with. One well-studied parameter of algorithms in this setting is the calibration error, which roughly speaking measures the extent to which the forecasts are accurate on average. In his paper Vladimir V. V'yugin establishes a tradeoff between the calibration error and the "coarseness" of any prediction strategy by showing that if the coarseness is small then the calibration error can also not be too small. This negative result comes close to matching the bounds given in previous work by Kakade and Foster on a particular forecasting system.

July 2007

Marcus Hutter
Rocco A. Servedio
Eiji Takimoto