

Research Statement

Lev Reyzin

Overview

Whether answering billions of search queries or detecting credit fraud, machine learning algorithms have revolutionized virtually every domain where data is abundant. This is no coincidence – the key insight of machine learning is that generic algorithms that learn and evolve as they see data often outperform specialized methods painstakingly devised by experts. Training a machine learning algorithm, however, is often expensive because determining labels for training data is quite costly; in medicine, it sometimes requires experimenting with drugs and can hurt patients in the process.

Fortunately, many settings allow learning algorithms to decide which data they want to have labeled or to otherwise engage with their data, and this can greatly reduce the initial costs. The study of such algorithms falls into the category of active learning, or more broadly **interactive learning**. Designing interactive algorithms can be difficult – a learner whose feedback depends on its own choices can inadvertently subject itself to systematic error and bias. Nonetheless, the ability to interact with data presents both great opportunities and exciting research challenges, and a recurring goal of my research is to understand and characterize the power of such interaction.

To this end, a large part of my research focuses on **designing practical and principled learning algorithms**, especially for interactive learning. In my Ph.D. work at Yale, I developed new models and techniques for actively learning interaction networks: hidden graphs (with applications to genome sequencing), circuits (for identifying gene regulatory networks), social networks (for analyzing viral marketing), and automata (for language learning). At Yahoo! Research, I tackled a variety of interactive learning problems in computational advertising, focusing on the theory and practice of learning under limited feedback in internet-scale applications. During my time at Georgia Tech, I've developed algorithms that examine only a few features before classifying, extending active learning to a feature scale.

Another major direction in my research has been to tackle **foundational questions** in learning theory. **Learning theory** is a fascinating field, with deep connections to many areas in theoretical computer science, and moreover, advances in learning theory have had tremendous impact on the practice of machine learning, from the advent of boosting to support vector machines.

One of my recent interests has been to understand the power of algorithms that work by looking at statistical properties of their input data; I showed unconditional lower bounds for statistical algorithms for a variety of optimization problems. I also tackled a related problem in statistical learning theory, giving the first improvement for the classical problem of learning sparse parities in the presence of noise. Other threads in my research involved studying finite automata in the context of learning, looking closely at data-dependent assumptions in clustering, making faster algorithms for online learning and bandit problems, and contributing to the fundamental understanding of the success of ensemble predictors. Some of this research has been incorporated into undergraduate and graduate machine learning courses at many universities.

1 Previous Work

1.1 Active and Interactive Learning

Algorithms for Computational Advertising. An important problem for many search engines is what advertisements to display alongside its search results. Search engines receive context about their users (queries, IP addresses, browser settings, etc.), and after displaying advertisements, they get feedback by seeing which ads the users clicked on. They get no explicit information, however, about how specific users would react if they were shown a different set of ads – partial feedback problems of this nature are called “bandit” problems. A good bandit algorithm must balance exploring new options and exploiting what has already been learned. In a series of papers, my colleagues at Yahoo! Research and I have developed the first efficient high probability optimal algorithm for the contextual bandit problem [9, 14], generalized the problem for the case when multiple ads can be displayed to users at once [19], and proved the first theoretical guarantees for a natural algorithm that works under a restricted “linear payoffs” setting [13].

Learning Circuits by Injecting Values. In a different line of research, my coauthors and I greatly extended the understanding of Value Injection Queries (VIQs) – a query model for learning circuits that was motivated by problems in discovering gene regulatory networks. Traditional circuit-learning models focus on the complexity of learning circuits by manipulating their inputs, but few classes of circuits are learnable in those models. VIQs give the learner the power to “inject values” into the gates of the hidden circuit, but only observe the value on the output gate. While the complexity of learning boolean circuits with VIQs was known, we showed that the results become surprisingly different when larger alphabet sizes are considered [2]. In a separate work, we also considered the case of learning probabilistic circuits (or Bayesian networks) in this model [1].

Discovering Social Networks. Dana Angluin, James Aspnes, and I adapted the Value Injection Query model to learning independent cascade social networks [4]. VIQs on social networks correspond to the natural operations of activating and suppressing agents in the network. We devised an optimal algorithm for learning social networks with VIQs. My more recent work on social networks focused on the ability of a passive learner to infer a social network by making observations. In [3] we tackled the problem of a learner inferring the most likely structure connecting a population after it observes how diseases, or other outbreaks, spread through that population. There, we gave algorithms for reconstructing social networks from such data.

Learning Hidden Graphs and Evolutionary Trees. An important graph learning problem is evolutionary tree reconstruction. A costly experiment can uncover the genetic distance between any two species, and these distances are tree-realizable – the goal is to reconstruct the species’ evolutionary tree using as few experiments as possible. Nikhil Srivastava and I discovered that a widely-cited analysis of a natural “longest path” algorithm for this task is incorrect and showed that the method is less efficient than was thought [26]. In another work on graph learning, Nikhil Srivastava and I analyzed the theoretical properties of a large variety of queries for the problems of graph learning and verification [25]. There, we proved new results for “edge-counting” queries (also called the additive-model). In addition, we showed how our results in graph verification lead to new techniques for matrix fingerprinting for a large class of matrices.

1.2 Other Foundational Questions in Learning Theory

Statistical Algorithms and Statistical Queries Some of my recent work has involved characterizing the power of statistical algorithms. Inspired by the statistical query model in learning theory, my colleagues and I characterized a wide class of optimization algorithms that use statistical properties of their inputs. This characterization essentially captures many well-studied heuristics, including local search, MCMC, and simulated annealing. We showed that for a wide variety of optimization problems over distributions, statistical algorithms unconditionally require time exponential in their input parameters [16]. A common barrier to devising efficient algorithms for statistical optimization algorithms and for statistical query algorithms in learning are parity functions, which are known not to be learnable in the presence of noise for statistical query algorithms; this is known as the classical noisy parity problem. In a different work, Elena Grigorescu, Santosh Vempala, and I improved on the state-of-the-art for learning sparse noisy parity functions, and this improvement also implied better bounds for other learning classes, such as juntas and DNF [18].

Learning Automata. Finite state automata are one of the most fundamental objects in all of computer science, but their learnability has not been fully explored. While it is known that automata cannot be efficiently learned from membership queries (very basic queries) alone, my colleagues and I showed that access to random bits “sprinkled” on the states a hidden automaton allows for efficient learning [5]. In that work, we also devised methods that rely on other types of help the learner can receive from a teacher. In a different line of research, I am currently working on other problems related to learning finite automata. My coauthors and I recently proved lower bounds for PAC learning random automata, as well as random DNF expressions and random decision trees, with statistical queries [6], and preliminary results give me hope that it will be possible to devise an algorithm for learning random automata under the uniform distribution on inputs.

Margins in Boosting. Robert Schapire and I tackled the problem of reconciling theoretical predictions from the margins theory of boosting with experimental data [24]. Experiments by Leo Breiman put the margins theory for the effectiveness boosting into serious doubt, but we showed how the margins theory explains his results and also discovered new phenomena. Our examination revealed a complex interplay between the choice of boosting algorithm, the margins it achieves on training examples, and the complexity of the weak learners it produces. This work has influenced subsequent margin bounds and the design of new boosting algorithms that aim to optimize the margins distribution. Then, in a later work, I explored the relationship between the ability of an algorithm to achieve a good margins distribution on training data and its ability to make use of only a few features while making predictions on test data [22].

Clustering. One current trend in clustering is to get around the NP-hardness of optimizing various objectives by examining which properties of the data, if satisfied, allow for polynomial time algorithms. One such assumption is perturbation resilience – that the optimum doesn’t change even when the distances among the data points are perturbed by certain factors [10]. In a recent work [23], I carefully analyzed this data resilience/stability assumption and showed that the choice of resilience parameter can drastically affect certain clustering problems, allowing only for a narrow range of parameter values between where the problems are NP-hard and where they are trivial.

Directions for Future Research

Here I list some research directions that I find interesting and I feel I have the background to tackle. The research I propose both explores fundamental problems in machine learning and at the same time addresses real-world challenges. My proposed work ranges from the theoretical to practical and impacts broad areas, from biological sciences to serving content on the internet. I am also interested in working in many other diverse areas of learning theory and machine learning.

Characterizing Statistical Approaches for Optimization. Building on the success of the statistical query model in learning theory [20], some of my recent work [16] (see previous Section) has dealt with trying to understand the power of statistical algorithms for optimization problems. While we have made progress in understanding the power of statistical approaches, there is still quite a lot we do not understand. Current techniques for analyzing the power of statistical algorithms rely on the underlying optimization problems being over distributions. However, when a statistical algorithm is employed on a fixed input, as often happens in practice, we have little theoretical understanding of its power. Yet we have strong experimental evidence that can guide us in trying to characterize the situations where such algorithms will be successful, and also where their power will be limited. My goal is to develop a theoretical understanding of statistical algorithms, which should be useful for designing better algorithms in practice.

Fast Prediction with Strong Guarantees. In most applications of machine learning, a learning algorithm has access to all of the features of its data before making predictions. This assumption, however, is not always true, and is becoming less so. In internet applications, users increasingly expect near-instantaneous results in response to their queries, and there is a tradeoff between producing good results and producing them quickly. In medical applications, the same tradeoff lies between doing costly (and sometimes dangerous) tests and getting a correct diagnosis. While related topics have been studied [17], and there have been some recent developments [12], a fundamental understanding of the problem is still missing. I have begun designing and analyzing feature-efficient algorithms [22], and I hope to develop a fundamental understanding as well as practical algorithms for this important problem.

Learning, from Biological Systems to Computer Networks. In the course of my research, I have considered various models for learning hidden graphs. A typical setup has the learner trying to discover the edges (or some property) of a hidden graph with some form of query access into the graph. This is related to property testing in graphs, in which the learner is constrained to a small number of queries. Much of this field is inspired by bioinformatics, and this application continues to interest me. There is, however, also potential to apply machine learning techniques to many problems of this character in systems research. One such problem is mapping the topology of the Internet [11]. We have natural queries into its structure, e.g. traceroute, and while there has been some theoretical work on network discovery [8], it can benefit from a machine learning perspective. Another problem I would like to pursue is ISP privacy. Recent proposals [27] for P2P systems allow clients to have some access to their ISP's preferences regarding which other clients they should connect to. There is some debate about which system can better prevent the clients from learning their ISP's underlying network, and I believe there is opportunity to design more principled protocols in this domain.

Discovering Social Interactions. Social networks are used to model interactions within populations of individuals. These interactions can include distributing information, spreading a disease, or passing trends among friends. Social networks present quite a few interesting learning challenges, especially with the abundant availability of real-world data. One research direction I have already pursued and plan to pursue further is on inferring social connections from data. For example, while social networks have been extensively used to model the spread of diseases (e.g. [15, 21]), we can ask the opposite question: by examining data from outbreaks of diseases, what can we learn about the underlying social connections? Another question is how can we take our knowledge of real-world social networks and incorporate it into learning algorithms? Often, in learning theory, one analyzes algorithms by how they perform on worst-case distributions of data, but this is generally not a realistic assumption. We have good models of the structure of real-world social networks [7] (e.g. the node degrees follow a power law distribution, the underlying graphs have certain expansion properties, etc.), and we should take advantage of this information.

References

- [1] ANGLUIN, D., ASPNES, J., CHEN, J., EISENSTAT, D., AND REYZIN, L. Learning acyclic probabilistic circuits using test paths. *Journal of Machine Learning Research* 10 (2009), 1881–1911. Previous version in *COLT* (2008).
- [2] ANGLUIN, D., ASPNES, J., CHEN, J., AND REYZIN, L. Learning large-alphabet and analog circuits with value injection queries. *Machine Learning* 72, 1-2 (2008), 113–138. Previous version in *COLT* (2007).
- [3] ANGLUIN, D., ASPNES, J., AND REYZIN, L. Inferring social networks from outbreaks. In *ALT* (2010), pp. 104–118.
- [4] ANGLUIN, D., ASPNES, J., AND REYZIN, L. Optimally learning social networks with activations and suppressions. *Theor. Comput. Sci.* 411, 29-30 (2010), 2729–2740. Previous version in *ALT* (2008).
- [5] ANGLUIN, D., BECERRA-BONACHE, L., DEDIU, A. H., AND REYZIN, L. Learning finite automata using label queries. In *ALT* (2009), pp. 171–185.
- [6] ANGLUIN, D., EISENSTAT, D., KONTOROVICH, L., AND REYZIN, L. Lower bounds on learning random structures with statistical queries. In *ALT* (2010), pp. 194–208.
- [7] BARABASI, A. L., AND ALBERT, R. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.
- [8] BEERLIOVA, Z., EBERHARD, F., ERLEBACH, T., HALL, A., HOFFMANN, M., MIHALÁK, M., AND RAM, L. S. Network discovery and verification. *IEEE Journal on Selected Areas in Communications* 24, 12 (2006), 2168–2181.
- [9] BEYGELZIMER, A., LANGFORD, J., LI, L., REYZIN, L., AND SCHAPIRE, R. E. Contextual bandit algorithms with supervised learning guarantees. *Journal of Machine Learning Research - Proceedings Track* 15 (2011), 19–26.
- [10] BILU, Y., AND LINIAL, N. Are stable instances easy? In *ICS* (2010), pp. 332 – 341.

- [11] BURCH, H., AND CHESWICK, B. Mapping the internet. *Computer* 32, 4 (1999), 97–98,102.
- [12] CESA-BIANCHI, N., SHALEV-SHWARTZ, S., AND SHAMIR, O. Efficient learning with partially observed attributes. In *ICML* (2010), pp. 183–190.
- [13] CHU, W., LI, L., REYZIN, L., AND SCHAPIRE, R. E. Contextual bandits with linear payoff functions. *Journal of Machine Learning Research - Proceedings Track 15* (2011), 208–214.
- [14] DUDIK, M., HSU, D., KALE, S., KARAMPATZIAKIS, N., LANGFORD, J., REYZIN, L., AND ZHANG, T. Efficient optimal learning for contextual bandits. In *UAI* (2011), pp. 169–178.
- [15] EUBANK, S., GUCLU, H., ANIL, MARATHE, M. V., SRINIVASAN, A., TOROCZKAI, Z., AND WANG, N. Modeling disease outbreaks in realistic urban social networks. *Nature* 429, 6988 (2004), 180–184.
- [16] FELDMAN, V., GRIGORESCU, E., REYZIN, L., AND VEMPALA, S. The Complexity of Statistical Algorithms. *ArXiv e-prints* (Jan. 2012).
- [17] GLOBERSON, A., AND ROWEIS, S. T. Nightmare at test time: robust learning by feature deletion. In *ICML* (2006), pp. 353–360.
- [18] GRIGORESCU, E., REYZIN, L., AND VEMPALA, S. On noise-tolerant learning of sparse parities and related problems. In *ALT* (2011).
- [19] KALE, S., REYZIN, L., AND SCHAPIRE, R. Non-stochastic bandit slate problems. In *NIPS* (2010), pp. 1045–1053.
- [20] KEARNS, M. J. Efficient noise-tolerant learning from statistical queries. In *STOC* (1993), pp. 392–401.
- [21] MOORE, C., AND NEWMAN, M. E. J. Epidemics and percolation in small-world networks. *Physical Review E* 61 (2000), 5678.
- [22] REYZIN, L. Boosting on a budget: Sampling for feature-efficient prediction. In *ICML* (2011), ACM, pp. 529–536.
- [23] REYZIN, L. Data Stability in Clustering: A Closer Look. *ArXiv e-prints* (July 2011).
- [24] REYZIN, L., AND SCHAPIRE, R. E. How boosting the margin can also boost classifier complexity. In *ICML* (2006), pp. 753–760.
- [25] REYZIN, L., AND SRIVASTAVA, N. Learning and verifying graphs using queries with a focus on edge counting. In *ALT* (2007), pp. 285–297.
- [26] REYZIN, L., AND SRIVASTAVA, N. On the longest path algorithm for reconstructing trees from distance matrices. *Inf. Process. Lett.* 101, 3 (2007), 98–100.
- [27] XIE, H., YANG, Y. R., KRISHNAMURTHY, A., LIU, Y., AND SILBERSCHATZ, A. P4P: Provider portal for applications. In *ACM SIGCOMM* (2008).