Statistical Query Learning (1993; Kearns)

Vitaly Feldman, IBM Research - Almaden
researcher.ibm.com/view.php?person=us-vitaly

entry editor: Rocco A. Servedio

INDEX TERMS: Statistical query, PAC learning, classification noise, noise-tolerant learning, SQ dimension.

1 PROBLEM DEFINITION

The problem deals with learning to classify from random labeled examples in Valiant's PAC model [Val84]. In the *random classification noise* model of Angluin and Laird [AL88] the label of each example given to the learning algorithm is flipped randomly and independently with some fixed probability η called the *noise rate*. Robustness to such benign form of noise is an important goal in the design of learning algorithms. Kearns defined a powerful and convenient framework for constructing noise-tolerant algorithms based on *statistical queries*. Statistical query (SQ) learning is a natural restriction of PAC learning that models algorithms that use statistical properties of a data set rather than individual examples. Kearns demonstrated that any learning algorithm that is based on statistical queries can be automatically converted to a learning algorithm in the presence of random classification noise of arbitrary rate smaller than the information-theoretic barrier of $1/2$. This result was used to give the first noise-tolerant algorithm for a number of important learning problems. In fact, virtually all known noise-tolerant PAC algorithms were either obtained from SQ algorithms or can be easily cast into the SQ model.

In subsequent work the model of Kearns has been extended to other settings and found a number of additional applications in machine learning and theoretical computer science.

1.1 Definitions and Notation

Let $\mathcal{C}$ be a class of $\{-1, +1\}$-valued functions (also called *concepts*) over an input space X. In the basic PAC model a learning algorithm is given examples of an unknown function f from $\mathcal{C}$ on points randomly chosen from some unknown distribution $\mathcal{D}$ over X and should produce a hypothesis h that approximates f. More formally, an *example oracle* $\mathrm{EX}(f, \mathcal{D})$ is an oracle that upon being invoked returns an example $\langle x, f(x)\rangle$, where x is chosen randomly with respect to $\mathcal{D}$, independently of any previous examples. A learning algorithm for $\mathcal{C}$ is an algorithm that for every $\epsilon > 0$, $\delta > 0$, $f \in \mathcal{C}$, and distribution $\mathcal{D}$ over X, given ϵ, δ, and access to $\mathrm{EX}(f, \mathcal{D})$ outputs, with probability at least $1 - \delta$, a hypothesis h that ϵ-approximates f with respect to $\mathcal{D}$ (i.e. $\mathbf{Pr}_{\mathcal{D}}[f(x) \neq h(x)] \leq \epsilon$). We denote this distribution over labeled examples by $\mathcal{D}_f$. Efficient learning algorithms are algorithms that run in time polynomial in $1/\epsilon$, $1/\delta$, and the size of the learning problem s. The size of a learning problem is determined by the description length of f under some fixed representation scheme for functions in $\mathcal{C}$ and the description length of an element in X (often proportional to the dimension n of the input space).

A number of variants of this basic framework are commonly considered. The basic PAC model is also referred to as *distribution-independent* learning to distinguish it from *distribution-specific* PAC learning in which the learning algorithm is required to learn with respect to a single distribution $\mathcal{D}$ known in advance. A *weak* learning algorithm is a learning algorithm that can produce a hypothesis whose error on the target concept is noticeably less than $1/2$ (and not necessarily any $\epsilon > 0$). More precisely, a weak learning algorithm produces a hypothesis h such that $\mathbf{Pr}_{\mathcal{D}}[f(x) \neq h(x)] \leq 1/2 - 1/p(s)$ for some fixed polynomial p. The basic PAC model is often referred to as *strong* learning in this context.

In the random classification noise model $\mathrm{EX}(f,\mathcal{D})$ is replaced by a faulty oracle $\mathrm{EX}^{\eta}(f,\mathcal{D})$, where η is the noise rate. When queried, this oracle returns a noisy example $\langle x, b\rangle$ where $b = f(x)$ with probability $1-\eta$ and $\neg f(x)$ with probability η independently of previous examples. When η approaches $1/2$ the label of the corrupted example approaches the result of a random coin flip, and therefore the running time of learning algorithms in this model is allowed to depend on $\frac{1}{1-2\eta}$ (the dependence must be polynomial for the algorithm to be considered efficient). For simplicity one usually assumes that η is known to the learning algorithm. This assumption can be removed using a simple technique of Laird [Lai88].

To formalize the idea of learning from statistical properties of a large number of examples, Kearns introduced a new oracle $\mathrm{STAT}(\mathcal{D}_f,\tau)$ that replaces $\mathrm{EX}(f,\mathcal{D})$, where $\tau \in [0,1]$ is the *tolerance* parameter. The oracle $\mathrm{STAT}(\mathcal{D}_f,\tau)$ takes as input a *statistical query* (SQ) defined by a real-valued function $\phi : X \times \{-1,+1\} \to [-1,1]$ on labeled examples. Given such a query the oracle responds with an estimate v of $\mathbf{E}_{(x,y)\sim\mathcal{D}_f}[\phi(x,y)]$ that is accurate to within an additive $\pm\tau$.

Note that the oracle does not guarantee anything else on the value v beyond $|v - \mathbf{E}_{\mathcal{D}_f}[\phi(x,y)]| \leq \tau$ and an SQ learning algorithm needs to work with any possible implementation of the oracle. However, SQ oracle is known to be equivalent (up to polynomial factors) to a 1-bit- per-sample oracle which to a call with Boolean function ϕ returns the value of $\phi(x, f(x))$, where x is chosen randomly and independently according to $\mathcal{D}$ [BD98]. Such oracle can be used to get an estimate of $\mathbf{E}_{\mathcal{D}_f}[\phi(x,y)]$ that is distributed in in the same way as an estimate based on fresh samples.

Chernoff-Hoeffding bounds easily imply that for a single query, $\mathrm{STAT}(\mathcal{D}_f,\tau)$ can, with high probability, be simulated using $\mathrm{EX}(f,\mathcal{D})$ by estimating $\mathbf{E}_{\mathcal{D}_f}[\phi(x,y)]$ on $O(\tau^{-2})$ examples. Therefore the SQ model is a restriction of the PAC model. Efficient SQ algorithms allow only efficiently evaluatable ϕ's and impose an inverse polynomial lower bound on the tolerance parameter over all oracle calls. Kearns also observes that in order to simulate all the statistical queries used by an algorithm one does not necessarily need new examples for each estimation. Instead, assuming that the set of possible queries of the algorithm has Vapnik-Chervonenkis dimension d, all its statistical queries can be simulated using $\tilde{O}(d\tau^{-2}(1-2\eta)^{-2}\log(1/\delta))$ examples [Kea98]. Without such assumptions, the best upper bounds on the number of samples sufficient to answer multiple statistical queries were recently given in [DFH$^+$14, BNS$^+$16].

Further, to make the correspondence between the number of samples n and the accuracy of the estimate more precise, Feldman *et al.* [FGR$^+$12] introduced a strengthening of the SQ oracle that incorporates the variance of the random variable $\phi(x,y)$ into the estimate. More formally, given as input any function $\phi : X \times \{-1,+1\} \to [0,1]$, $\mathrm{VSTAT}(\mathcal{D}_f, n)$ returns a value v such that $|v - p| \leq \max\left\{\frac{1}{n}, \sqrt{\frac{p(1-p)}{n}}\right\}$, where $p = \mathbf{E}_{\mathcal{D}_f}[\phi]$. Note that $\frac{p(1-p)}{n}$ is the variance of the empirical mean when ϕ is Boolean. More generally, the oracle can be used to estimate of the expectation $\mathbf{E}_{\mathcal{D}_f}[\phi]$ for any real-valued function ϕ within $\tilde{O}(\sigma/\sqrt{n})$, where σ is the standard deviation of $\phi(x)$ [Fel16a].

A natural generalization of the SQ oracle that provides an estimate of the expectation of any function of k examples: $\phi : (X \times \{-1,+1\})^k \to [-1,1]$ was defined and studied by Blum *et al.* [BKW03]. Additional upper and lower bound on the power of this oracle are given in [FG17].

2 KEY RESULTS

2.1 Statistical Queries and Noise-tolerance

The main result given by Kearns is a way to simulate statistical queries using noisy examples.

Lemma 1 ([Kea98])**.** *Let ϕ be a statistical query such that ϕ can be evaluated on any input in time T, $\tau > 0$ and let $EX^\eta(f,\mathcal{D})$ be a noisy oracle. The value $\mathbf{E}_{\mathcal{D}}[\phi(x,f(x))]$ can, with probability at least $1-\delta$, be estimated within τ using $O(\tau^{-2}(1-2\eta)^{-2}\log(1/\delta))$ examples from $EX^\eta(f,\mathcal{D})$ and time $O(\tau^{-2}(1-2\eta)^{-2}\log(1/\delta)\cdot T)$.*

This simulation is based on estimating several probabilities using examples from the noisy oracle and then offsetting the effect of noise. The lemma implies that any efficient SQ algorithm for a concept class $\mathcal{C}$ can be converted to an efficient learning algorithm for $\mathcal{C}$ tolerating random classification noise of any rate $\eta < 1/2$.

Theorem 2 ([Kea98])**.** *Let $\mathcal{C}$ be a concept class efficiently PAC learnable from statistical queries. Then $\mathcal{C}$ is efficiently PAC learnable in the presence of random classification noise of rate η for any $\eta < 1/2$.*

Balcan and Feldman describe more general conditions on noise under which a specific SQ algorithm can be simulated in the presence of noise [BF13].

2.2 Statistical Query Algorithms

Kearns showed that, despite the major restriction on the way an SQ algorithm accesses the examples, many PAC learning algorithms known at the time can be modified to use statistical queries instead of random examples [Kea98]. Examples of learning algorithms for which he described an SQ analogue and thereby obtained a noise-tolerant learning algorithm include:

- Learning decision trees of constant rank.
- *Attribute-efficient* algorithms for learning conjunctions.
- Learning axis-aligned rectangles over $\mathbb{R}^n$.
- Learning AC^0 (constant-depth unbounded fan-in) Boolean circuits over $\{0,1\}^n$ with respect to the uniform distribution in quasipolynomial time.

Subsequent works have provided numerous additional examples of algorithms used in theory and practice of machine learning that can either be implemented using statistical queries or can be replaced by an alternative SQ-based algorithm of similar complexity, for example, the Perceptron algorithm and learning of linear threshold functions [BFKV97, DV04, BF15, FGV15], boosting [AD98], attribute-efficient learning via the Winnow algorithm (*cf.* [Fel14]), k-means clustering [BDMN05] and stochastic convex optimization [FGV15]. We note that many learning algorithms rely only on evaluations of functions on random examples and therefore can be seen as using access to the honest statistical query oracle. In such cases the SQ implementation follows immediately from the equivalence of the Kearns' SQ oracle and the honest one [FGR$^+$12].

The only known example of a technique for which there is no SQ analogue is Gaussian elimination for solving linear equations over a finite field. This technique can be used to learn parity functions that are not learnable using SQs (as we discuss below) and, more generally, affine subspaces over finite fields. As a result, with the exception of affine subspace learning problem (e.g. [Fel16b]), known bounds on the complexity of learning from random examples are, up to polynomial factors, the same as known bound for learning with statistical queries.

2.3 Statistical Query Dimension

The restricted way in which SQ algorithms use examples makes it simpler to understand the limitations of efficient learning in this model. A long-standing open problem in learning theory is learning of the concept class of all parity functions over $\{0,1\}^n$ with noise (a parity function is a XOR of some subset of n Boolean inputs). Kearns has demonstrated that parities cannot be efficiently learned using statistical queries even under the uniform distribution over $\{0,1\}^n$ [Kea98]. This hardness result is unconditional in the sense that it does not rely on any unproven complexity assumptions.

The technique of Kearns was generalized by Blum *et al.* who proved that efficient SQ learnability of a concept class $\mathcal{C}$ is characterized by a relatively simple combinatorial parameter of $\mathcal{C}$ called the *statistical query dimension* [BFJ$^+$94]. The quantity they defined measures the maximum number of "nearly uncorrelated" functions in a concept class. (The definition and the results were simplified and strengthened in subsequent works [Szo09, FGR$^+$12], and we use the improved statements here.) More formally,

Definition 3. *For a concept class $\mathcal{C}$ and distribution $\mathcal{D}$, the* statistical query dimension *of $\mathcal{C}$ with respect to $\mathcal{D}$, denoted SQ-DIM($\mathcal{C},\mathcal{D}$), is the largest number d such that $\mathcal{C}$ contains d functions $f_1, f_2, \ldots, f_d$ such that for all $i \neq j$, $|\mathbf{E}_{\mathcal{D}}[f_i \cdot f_j]| \leq \frac{1}{d}$.*

Blum *et al.* relate the SQ dimension to learning in the SQ model as follows.

Theorem 4 ([BFJ$^+$94, FGR$^+$12])**.** *Let $\mathcal{C}$ be a concept class and $\mathcal{D}$ be a distribution such that SQ-DIM($\mathcal{C},\mathcal{D}$) $= d$.*

- *If all queries are made with tolerance of at least $1/d^{1/3}$, then at least $d^{1/3} - 2$ queries are required to learn $\mathcal{C}$ with error $1/2 - 1/(2d^3)$ in the SQ model.*

- *There exists an algorithm for learning $\mathcal{C}$ with respect to $\mathcal{D}$ that makes d fixed queries, each of tolerance $1/(4d)$, and finds a hypothesis with error at most $1/2 - 1/(2d)$.*

Thus SQ-DIM characterizes weak SQ learnability relative to a fixed distribution $\mathcal{D}$ up to a polynomial factor. Parity functions are uncorrelated with respect to the uniform distribution, and therefore any concept class that contains a superpolynomial number of parity functions cannot be learned by statistical queries with respect to the uniform distribution. This, for example, includes such important concept classes as *k-juntas* over $\{0,1\}^n$ (or functions that depend on at most k input variables) for $k = \omega(1)$ and *decision trees* of superconstant size.

Simon showed that (strong) PAC learning relative to a fixed distribution $\mathcal{D}$ using SQs can also be characterized by a more general and involved dimension [Sim07]. Simpler and tighter characterizations of distribution-specific PAC learning using SQs have been demonstrated by Feldman [Fel12] and Szörényi [Szo09]. Feldman also extended the characterization to the agnostic learning model [Fel12]. More recently a notion of SQ dimension that characterizes distribution-independent learning and other problems was given in [Fel16b].

Despite characterizing the number of queries of certain tolerance, the SQ-DIM and its generalizations capture surprisingly well the computational complexity of SQ learning of most concept classes. One reason for this is that if a concept class has polynomial SQ-DIM then it can be learned by a polynomial-time algorithm with advice also referred to as a "non-uniform" algorithm (*cf.* [FK12]). However it was shown by Feldman and Kanade that for strong PAC learning there exist artificial problems whose computational complexity is larger than their statistical query complexity [FK12].

Applications of these characterizations to proving lower bounds on SQ algorithms can be found in [KS07, Fel12, FLS11, DSFT$^+$14]. Relationships of SQ-DIM to other notions of complexity of concept classes were investigated in [She08, KS11].

3 APPLICATIONS

The restricted way in which an SQ algorithm uses data implies that it can be used to obtain learning algorithms with additional useful properties. For example, SQ learning algorithms can be easily converted to algorithms for learning from positive and unlabeled examples [Den98, DGL05] and learning algorithms from multiple-instance examples [BK98]. Blum *et al.* [BDMN05] show that an SQ algorithm can be used to obtain a *differentially private* [DMNS06] algorithm for the problem. In fact, SQ algorithms are equivalent to *local* (or *randomized-response*) differentially private algorithms [KLN$^+$11]. Chu *et al.* [CKL$^+$06] show that SQ algorithms can be automatically parallelized on multicore architectures and give many examples of popular machine learning algorithms that can be sped up using this approach. Steinhardt *et al.* show how to obtain learning algorithms in the streaming setting with limited memory from SQ algorithms and derive new algorithms for sparse regression from this reduction [SVW16, Fel16b].

The SQ model of learning was generalized to active learning (or learning where labels are requested only for some of the points) and used to obtain new efficient noise-tolerant active learning algorithms [BF13].

The SQ learning model has also been instrumental in understanding Valiant's model of evolution as learning [Val09]. Feldman showed that the model is equivalent to learning with a restricted form of SQs referred to as correlational SQs [Fel08]. A correlational SQ is a query of the form $\phi(x, y) = g(x) \cdot y$ for some $g : X \to [-1, 1]$. Such queries were first studied by Ben-David *et al.* [BDIK90] (remarkably, before the introduction of the SQ model itself) and distribution-specific learning with such queries is equivalent to learning with (unrestricted) SQs. The correspondence between evolvability and SQ algorithms has been used in a number of subsequent works [Fel09, Fel12, KVV10, Kan11, Val12] on evolvability.

Statistical query-based access can naturally be defined for any problem where the input is a set of i.i.d. samples from some distribution over the domain of the problem (and not just labeled examples). Feldman *et al.* show that lower bounds based on SQ-DIM can be extended to this more general setting and give examples of applications [FGR$^+$12, FPV13]. A general statistical dimension that characterizes the statistical query complexity of an arbitrary problem over distributions is described in [Fel16b].

4 OPEN PROBLEMS

The main questions related to learning with random classification noise are still open. Is every concept class efficiently learnable in the PAC model also learnable in the presence of random classification noise? Is every concept class efficiently learnable in the presence of random classification noise of arbitrarily high rate (less than 1/2) also efficiently learnable using statistical queries? A partial answer to this question was provided by Blum *et al.* who show that Gaussian elimination can be used in low dimension to obtain a class learnable with random classification noise of constant rate $\eta < 1/2$ but not learnable using SQs [BKW03]. For both questions a central issue seems to be obtaining a better understanding of the complexity of learning parities with noise.

The complexity of learning from statistical queries remains an active area of research with many open problems. For example, there is currently an exponential gap between known lower and upper bounds on the complexity of distribution-independent SQ learning of polynomial-size DNF formulae and AC^0 circuits (*cf.* [She08]). Several additional open problems on complexity of SQ learning can be found in [FLS11, KS11, Fel14].

References

[AD98] J. Aslam and S. Decatur. General bounds on statistical query learning and pac learning with noise via hypothesis boosting. *Information and Computation*, 141(2):85–118, 1998.

[AL88] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

[BD98] Shai Ben-David and Eli Dichterman. Learning with restricted focus of attention. *J. Comput. Syst. Sci.*, 56(3):277–298, 1998.

[BDIK90] S. Ben-David, A. Itai, and E. Kushilevitz. Learning by distances. In *COLT*, pages 232–245, 1990.

[BDMN05] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *PODS*, pages 128–138, 2005.

[BF13] Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms. In *NIPS*, pages 1295–1303, 2013.

[BF15] Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms for noise tolerance and differential privacy. *Algorithmica*, 72(1):282–315, 2015.

[BFJ$^+$94] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *STOC*, pages 253–262, 1994.

[BFKV97] A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1997.

[BK98] Avrim Blum and Adam Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998.

[BKW03] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003.

[BNS$^+$16] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016.

[CKL$^+$06] C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *NIPS*, pages 281–288, 2006.

[Den98] FranÇois Denis. *PAC Learning from Positive Statistical Queries*, pages 112–126. 1998.

[DFH$^+$14] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

[DGL05] Franois Denis, Rmi Gilleron, and Fabien Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70 – 83, 2005.

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[DSFT$^+$14] Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. *arXiv, CoRR*, abs/1405.5268, 2014.

[DV04] J. Dunagan and S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *STOC*, pages 315–320, 2004.

[Fel08] V. Feldman. Evolvability from learning algorithms. In *STOC*, pages 619–628, 2008.

[Fel09] V. Feldman. Robustness of evolvability. In *COLT*, pages 277–292, 2009.

[Fel12] V. Feldman. A complete characterization of statistical query learning with applications to evolvability. *Journal of Computer System Sciences*, 78(5):1444–1459, 2012.

[Fel14] Vitaly Feldman. Open problem: The statistical query complexity of learning sparse halfspaces. In *COLT*, pages 1283–1289, 2014.

[Fel16a] Vitaly Feldman. Dealing with range anxiety in mean estimation via statistical queries. *arXiv*, abs/1611.06475, 2016.

[Fel16b] Vitaly Feldman. A general characterization of the statistical query complexity. *CoRR*, abs/1608.02198, 2016.

[FG17] Vitaly Feldman and Badih Ghazi. On the power of learning from k-wise queries. Innovations in Theoretical Computer Science (ITCS), 2017.

[FGR$^+$12] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *arXiv, CoRR*, abs/1201.1214, 2012. Extended abstract in STOC 2013.

[FGR$^+$13] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for planted clique. In *STOC*, pages 655–664. ACM, 2013.

[FGV15] Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. *CoRR*, abs/1512.09170, 2015. Extended abstract in SODA 2017.

[FK12] Vitaly Feldman and Varun Kanade. Computational bounds on statistical query learning. In *COLT*, pages 16.1–16.22, 2012.

[FLS11] V. Feldman, H. Lee, and R. Servedio. Lower bounds and hardness amplification for learning shallow monotone formulas. In *COLT*, volume 19, pages 273–292, 2011.

[FPV13] Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. *CoRR*, abs/1311.4821, 2013. Extended abstract in STOC 2015.

[Kan11] Varun Kanade. Evolution with recombination. In *FOCS*, pages 837–846, 2011.

[Kea98] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[KLN$^+$11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, June 2011.

[KS07] A. Klivans and A. Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2-3):97–114, 2007.

[KS11] M. Kallweit and H. Simon. A close look to margin complexity and related parameters. In *COLT*, pages 437–456, 2011.

[KVV10] V. Kanade, L. G. Valiant, and J. Wortman Vaughan. Evolution with drifting targets. In *COLT*, pages 155–167, 2010.

[Lai88] P. Laird. *Learning from good and bad data.* Kluwer Academic Publishers, 1988.

[She08] Alexander A. Sherstov. Halfspace matrices. *Computational Complexity*, 17(2):149–178, 2008.

[Sim07] H. Simon. A characterization of strong learnability in the statistical query model. In *Symposium on Theoretical Aspects of Computer Science*, pages 393–404, 2007.

[SVW16] J. Steinhardt, G. Valiant, and S. Wager. Memory, communication, and statistical queries. In *COLT*, pages 1490–1516, 2016.

[Szo09] B. Szorenyi. Characterizing statistical query learning:simplified notions and proofs. In *ALT*, pages 186–200, 2009.

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

[Val09] L. G. Valiant. Evolvability. *Journal of the ACM*, 56(1):3.1–3.21, 2009. Earlier version in ECCC, 2006.

[Val12] Paul Valiant. Distribution free evolvability of polynomial functions over all convex loss functions. In *ITCS*, pages 142–148, 2012.