Gregory Valiant—Research Statement (December, 2019)

My research explores how to extract as much information as possible from data, in various fundamental settings. Within this theme, I have focused on several lines of questions: What information can be accurately inferred given modest amounts of data, for example when the sample size is sublinear in the support size or dimensionality of the distribution in question? How do restrictions on the amount of available memory affect the time or amount of data required to learn certain concepts? Can we make learning and estimation algorithms robust—robust to significant fractions of "bad" or maliciously generated training data, and robust to deviations from the idealized case where the test and training sets are drawn from similar distributions?

My research into these questions has ranged from designing practically useful algorithms, to establishing impossibility results showing inherent limitations on the performance of any algorithm. Across this work, I have strived to develop theoretically deep algorithmic machinery and analysis techniques whose applicability extends beyond the specific problems at hand. Additionally, a significant fraction of my recent work has introduced new conjectures and problem formulations that distill aspects of the challenges emerging from the current practice of data science. Below, I describe some of the core themes of my research and concrete future directions.

Learning with Little Data

What can, and cannot be accomplished with little data? Given too little data to accurately learn the distribution underlying the data, what types of structure can be inferred and what hypotheses can be confidently tested? Can one accurately evaluate the value of collecting additional data? In contrast to my earlier work on these questions, at Stanford my goal shifted from developing worst-case (i.e. "minimax") optimal algorithms and estimators, to striving for *instance-optimal* algorithms: algorithms which perform as well as possible on *every* input. A crucial component of this line of work was defining reasonable notions of instance optimality for distributional learning and hypothesis testing settings [19, 18]. One key result [19], with Paul Valiant, is an algorithm for testing whether a set of samples was drawn from a specified distribution, p, versus having been drawn from a distribution with distance at least ϵ from p—a classic hypothesis testing problem studied for over a century. Our algorithm, which is a significantly modified version of Pearson's chi-squared test, is optimal for every distribution, p, and established the surprising result that the sample size necessary for performing such a test scales roughly as the $2/3$-norm of the vector of probabilities of distribution p. Beyond yielding conceptually pleasing theoretical results, this emphasis on instance optimality has yielded algorithms which perform very well in practice for a number of important problems. In a 2016 Nature Communications paper [21], we extended machinery developed for the theoretical paper on instance optimal learning [18], to reconstruct the "frequency spectrum" of rare genetic mutations, and provided concrete answers to questions such as "how many new mutations of a certain type will likely be discovered if we sequence an additional 500k, 1M, 10M genomes?"

A closely related branch of my effort to understand the power of little data focuses on high-dimensional real-valued settings. What sorts of *geometric* structure and other practically meaningful properties can be accurately inferred? In a 2017 Annals of Statistics paper [8] with my student, Weihao Kong, we provided the first algorithm for accurately recovering the eigenvalues of the covariance of a distribution that is accurate in the sublinear data regime where the sample size is asymptotically smaller than the dimension of the data—a significant open problem in the statistics community. These eigenvalues contain useful information about the distribution in question, including the effective dimensionality, and applicability of Principal Component Analysis and higher-level machine learning and multivariate statistical tools. Beyond the clean and practical algorithm, this work developed tools in random matrix theory, which we, and others, are fruitfully applying to other high-dimensional problems. In ongoing work with Weihao (beginning with our NeurIPS'18 paper [7]) we are repurposing some of these tools to estimate *learnability*. Given too little labeled data to have any hope of accurately learning a good predictor, we show that one can accurately estimate the quality of the best model, in certain fundamental settings including regression. This ability to accurately *estimate the optimal value*, in the regime in which there is insufficient data to *find where the optimum is attained*, raises a number of tantalizing open directions for more general optimization problems, which we are currently investigating.

The lay-of-the-land of optimal estimation has now been largely charted for many of the most basic estimation and testing problems; nevertheless, we are only beginning to scratch the surface for many important structured settings, including those with sequential structure, or those inspired by *federated learning* where data is grouped into batches corresponding to distinct heterogeneous data sources. My students and I are actively working on these increasingly relevant problems [17, 20]. In a slightly different direction, my students and I are also beginning to look at the problem of *sampling* from a distribution, given access to a limited number of independent draws. Curiously, sampling seems fundamentally distinct from learning: for example, in basic settings such as high-dimensional Gaussian distributions, or distributions with large discrete support, given n draws from an unknown distribution, one can return a set of $m > n$ datapoints that is indistinguishable from m i.i.d. draws, *despite the fact that learning D would require $\gg n$ samples* [1]. We are in the process of working to understand the differences between learning and sampling, and chart out potential connections to GANs. I am excited to apply some of the perspectives of sublinear sample testing and estimation to these new and natural sampling questions.

Memory-Bounded Learning

Beyond the amount of available data, there are a variety of other resource constraints which impact the ability to learn. One such resource that I am particularly interested in is *memory*. What are the fundamental tradeoffs between the amount of available memory, and the speed or data required to learn? Beyond the obvious practical importance of understanding these tradeoffs, there are several conceptual motivations for studying these questions. First, lower bounds on memory-bounded learning may offer new vantage points from which to understand the apparent computational challenge of classical problems, such as learning parity with noise (a problem with deep connections to learning theory, coding theory, and cryptography). Second, understanding the landscape of continuous optimization in terms of memory requirements may profoundly impact how we design optimization algorithms. Finally, I believe these questions *can* be answered. As opposed to some of the more traditional complexity theoretic open problems on memory/time tradeoffs in the cell-probe model whose difficulty seems akin to proving P vs NP, memory/data tradeoffs in learning are intrinsically information theoretic questions for which there do not seem to be fundamental obstacles. This optimism motivated the formulation of the following conjecture, which was the cornerstone of a COLT'16 paper with students Jacob Steinhardt and Stefan Wager [16]: *Learning a random parity function over length n inputs either requires a quadratic amount of memory (in which case one can apply Gaussian elimination), or an exponential number of examples.* In an impressive paper in 2017 [12], Ran Raz resolved this conjecture, leading a flurry of papers over the past two years further developing techniques for establishing sharp memory/example tradeoffs for broader classes of Boolean learning problems.

Despite the progress on memory/data tradeoffs for discrete (Boolean) learning problems, the landscape of continuous optimization is still largely unexplored. There is a huge body of work investigating gradient based "first-order" optimization methods, which require memory that is only linear in the number of parameters, and "second-order" schemes that require quadratic memory but converge in significantly fewer iterations (but are practically infeasible for large-scale learning due to their memory requirements). In work from STOC'19 [13] with Aaron Sidford and my student Vatsal Sharan, we provided the first rigorous explanation for the apparent difficulty of devising an algorithm with the converge rates of second-order methods yet the memory requirements of first-order methods: we showed that any algorithm that uses a subquadratic amount of memory, must have a convergence rate that is asymptotically slower than that achieved by the best quadratic-memory second-order method. We believe that this result, and the technical machinery developed in the proof, is the beginning of a line of work to chart out the landscape of memory requirements for convex optimization. One conjecture we are currently working on, posed in our paper [13], is that the apparent inability of first-order methods to efficiently handle ill-conditioned settings is inherent to any memory bounded algorithm: *Conjecture: Any algorithm for regression over d-dimensional datapoints either requires memory quadratic in d, or requires an amount of data that is polynomial in the condition-number of the datapoints,* *as opposed to the logarithmic dependence achieved by second-order methods.*

Robust Learning

A third direction of my recent research has been understanding robust learning and estimation. The majority of current approaches to learning and statistical estimation are predicated on the assumption that the training and test data are drawn from similar distributions. My research has examined a variety of fundamental deviations from this idealized setting. In a series of work with Moses Charikar and Jacob Steinhardt [15, 3, 14], and with my students Michela Meister [9] and with Mingda Qiao [11], we investigated the extent to which estimation, learning, and convex optimization over data, could be successfully accomplished even when a significant (but unknown) component of the available data was *untrusted*—encompassing arbitrarily outlying, biased, and maliciously constructed datapoints. Surprisingly, we showed that strong positive results were possible for robust learning, even with a *majority* of bad data. A component of this body of work was the introduction of novel frameworks in which to consider these questions, including the "semi-verified" model in which a learner has access to a tiny dataset of clean data in addition to a large dataset with a significant fraction of bad data, and the "list-decodable learning" model where the learner may return a small set of guesses, with the guarantee that at least one of them is correct (this model is analogous to list-decodable codes in the coding-theory community, and has also been considered in clustering settings by Balcan, Blum, and Vempala [2].)

A different side of my work on robust learning focuses on test-time attacks. In addition some theoretical work [5], I am beginning to consider the question of whether *humans* are susceptible to "adversarial examples" in auditory or vision settings. This is not a search for isolated instances (e.g optical illusions), but an attempt to understand whether there are settings where nearly every natural input can be turned into an "illusion". Preliminary results at CogSci'19 with my student Melody Guan [6], demonstrate that a significant fraction of natural language is susceptible to a certain type of subtle manipulation, confounding or altering the perceived words. I am optimistic that future work will yield further insights into human perception, which may inform the current discussion on the lack of robustness in trained computer models.

One final perspective on robust learning that I am considering is whether accurate predictions are possible without *any* assumptions on the data, and without switching to regret-based metrics as in on-line learning. One example is recent work from COLT'19 with Mingda Qiao on *selective prediction* [10], which builds on earlier work of Drucker [4]. These works consider the problem of predicting statistics of future datapoints, given access to a sequence of bounded observations. Provided the predictor has the power to select 1) when the prediction is made, and 2) the window-length to which the prediction pertains, accurate prediction is possible *even for worst-case data* without any assumptions about how future data is related to past data. Specifically, even for worst-case sequences of data $x_1,...,x_n$ with $x_i \in [0,1]$, there is an algorithm which selects a time point $t < n$, and window length w, and after observing $x_1,...,x_t$, outputs a prediction for the average value of $x_{t+1},...,x_{t+w}$, with the guarantee that the expected error of the prediction is bounded by $O(1/\log n)$, which is optimal to constant factors.

Given this surprising result on accurately predicting the future without assumptions on how the future is related to the past, we are in the process of considering significantly more general formulations where accurate inferences are possible for worst-case data. For example, given worst case data but an understanding of the process by which the data is partitioned into a test and training set, when is accurate prediction or learning possible? Clearly if each datapoint is independently assigned to either the test or training set, then strong positive results are trivial because basic statistics of the test and training set concentrate. But what about more interesting partitioning processes, such as "snowball sampling" that induce strong correlations between datapoints? We are optimistic that positive results are possible even in settings where statistics of the test and training set are *not* concentrated. I believe that this new line of work that my collaborators and I are initiating on worst-case data will serve as an enlightening counterpoint to the predominate theoretical frameworks for learning and statistics that either posits a distribution or generative process underlying the data, makes strong structural assumptions about the data, or measures performance with respect to a limited class of benchmarks.

References

[1] Brian Axelrod, Shivam Garg, Vatsal Sharan, and Gregory Valiant. Sample amplification: Increasing dataset size even when learning is impossible. *arXiv preprint arXiv:1904.12053*, 2019.

[2] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing (STOC)*, pages 671–680, 2008.

[3] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 47–60, 2017.

[4] Andrew Drucker. High-confidence predictions under adversarial uncertainty. In *3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

[5] Shivam Garg, Vatsal Sharan, Brian Zhang, and Gregory Valiant. A spectral view of adversarially robust features. In *Advances in Neural Information Processing Systems*, pages 10138–10148, 2018.

[6] Melody Guan and Gregory Valiant. A surprising density of illusionable natural speech. In *The 41st Annual Meeting of the Cognitive Science Society (CogSci)*, 2019.

[7] Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5455–5464, 2018.

[8] Weihao Kong, Gregory Valiant, et al. Spectrum estimation from samples. *The Annals of Statistics*, 45(5):2218–2247, 2017.

[9] Michela Meister and Gregory Valiant. A data prism: Semi-verified learning in the small-alpha regime. In *Conference on Learning Theory*, pages 1530–1546, 2016.

[10] Mingda Qiao and Gregory. A theory of selective prediction. In *Conference on Learning Theory (COLT)*, 2019.

[11] Mingda Qiao and Gregory Valiant. Learning discrete distributions from untrusted batches. In *9th Innovations in Theoretical Computer Science Conference (ITCS)*, 2018.

[12] Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. *Journal of the ACM (JACM)*, 66(1):3, 2018.

[13] Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Memory-sample tradeoffs for linear regression with small error. In *Proceedings of the fifty-first annual ACM symposium on Theory of Computing (STOC)*, 2019.

[14] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS)*, 2018.

[15] Jacob Steinhardt, Gregory Valiant, and Moses Charikar. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4439–4447, 2016.

[16] Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory (COLT)*, pages 1490–1516, 2016.

[17] Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5778–5787, 2017.

[18] Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing (STOC)*, pages 142–155, 2016.

[19] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[20] Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham M Kakade. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning (ICML)*, 2019.

[21] James Zou, Gregory Valiant, Paul Valiant, Konrad Karczewski, Siu On Chan, Kaitlin Samocha, Monkol Lek, Shamil Sunyaev, Mark Daly, and Daniel G MacArthur. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature communications*, 7:13293, 2016.