Research statement – Naren Sarayu Manoj (nsm@ttic.edu)

1 Introduction

My research objective is to **develop a fundamental understanding of machine learning and algorithmic data science when the input is adversarial, unreliable, or drawn from an unpredictable source**. I am specifically interested in problems where the input arises from a *noisy data source* or when there are *resource constraints* that limit an algorithm's access to the input.

The algorithmic toolkit for data science problems has grown immensely over the last few years. However, it is not clear how brittle known algorithmic guarantees are to the assumptions upon which they are proven. In many cases, these conditions are chosen to make the algorithmic problems mathematically tractable. An unfortunate consequence is that these conditions may not accurately reflect real-world scenarios. For example, in practice, a machine learning algorithm's input dataset may be corrupted by malicious or unexpected data. Conceivably, the output of the algorithm run using this unclean data may be very different from its output when given a clean input (i.e., one that satisfies the assumptions under which its original guarantees were derived). On the other hand, the input data may be clean but too unwieldy to practically work with. Perhaps the dataset has too many features (high dimensionality), or there may be too many observations to process all at once (big data). It is therefore important to build algorithmic solutions for these more realistic scenarios.

As a step towards this goal, I have studied classical algorithmic data science questions under input assumptions that more closely reflect those seen in practice. Within this, my work is diverse, spanning robust and explainable machine learning [MB21; GM23], preference-based optimization [BGLMSY23], data summarization [MMO22; MMO23; MO23], and generalization from interpolating noisy data [MS23].

2 Prior and ongoing work

In this section, I describe my doctoral work in greater detail. My work can be split into two broad themes – one where the input arises from a semi-adversarial source and another where access to the input is restricted due to practical computational limitations.

2.1 Robustness of data science algorithms under input modifications

To understand how unclean data sources affect problems in machine learning algorithm design, I am interested in the questions – *What data models realistically capture real-world data science problem instances? When are algorithms robust to adversarial or unexpected data? If existing algorithms are not robust, can we design and analyze algorithms that are?*

As a step towards answering these questions, I studied the following problems.

Robustness against backdoor data poisoning attacks. To address security concerns with deep learning, a line of work studies the robustness of deep learning to training-time corruptions. One particularly insidious form of training-time corruption is the *backdoor data poisoning attack*. In a backdoor data poisoning attack, an adversary injects watermarked, mislabeled training examples into a training set. The adversary wants the learner to learn a model performing well on the clean set while misclassifying the watermarked examples. Hence, unlike in other malicious noise models, the attacker wants to impact the performance of the classifier only on watermarked examples while leaving the classifier unchanged on clean examples. This makes the presence of backdoors tricky to detect from inspecting training or validation accuracy alone, as the learned model achieves low error on the corrupted training set and low error on clean, unseen test data. Furthermore, it has been empirically shown that attacks of this flavor can be executed on modern machine learning models, even with random watermarks (e.g. [ABCPK18], [TJHAPJNT20], [CLLLS17], [WSRVASLP20], [SSP20], [TLM18]).

The existence of such adversarial modifications in practice begs the following questions – *Can we build an adversarial input model that captures the notion of backdoor adversarial attacks in machine learning classification settings? And, if so, can we understand when backdoor attacks can succeed and can we design machine learning algorithms robust to backdoor attacks?*.

In a joint work with Avrim Blum [MB21], I made significant progress towards positive answers to both

of these questions by building a formal framework within which it is possible to theoretically analyze backdoor data poisoning attacks. Within this framework, we gave a tight learning problem-dependent characterization for whether a backdoor attack is possible. This characterization then made it natural for us to analyze the robustness of several natural learning problems, including those of learning decision lists and overparameterized linear models. We also identified a few concrete connections between backdoor-robust learning algorithms and learning algorithms that output classifiers robust to test-time adversarial perturbations – namely, that adversarial training (a popular method to learn robust classifiers in practice) might be a promising approach to learn classifiers robust to backdoor attacks.

Dueling optimization with a monotone adversary. In a preference-based feedback model, a learning algorithm gets to observe qualitative feedback instead of quantitative feedback (i.e., it observes answers to the question "rank the recommended items in descending order of preference") [BV06; SJ11]. This model is heavily used to design recommendation systems, optimize search engines, and perform information retrieval. More recently, preference-based feedback has received a lot of attention as a mechanism to train large language models [OWJ+22].

One natural, theoretically appealing way to model this type of feedback is through the problem of *dueling convex optimization*. In this problem, the user's preferences are modeled by a convex function f. The recommender is allowed to suggest a pair of items $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, and the user chooses the item that has the lowest value according to f. The recommender's goal is to find a minimizer for f over several rounds of such interactions. However, in practice, the user could instead choose another item it prefers more than either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ – in other words, it could choose some item $\boldsymbol{x}_3$ that has lower f-value than both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This begs the question – *can we design a learner for dueling convex optimization that is robust to this monotone adversary (i.e., an adversary that can only make monotonically improving changes to the responses)?*

In a joint work with Avrim Blum, Meghal Gupta, Gene Li, Aadirupa Saha, and Chloe Yang [BGLMSY23], I gave the first algorithm to achieve both the optimal iteration complexity and total regret for dueling convex optimization with a monotone adversary under some natural conditions on the loss function f. Although the scope of possible responses from the user is much larger now that the user is allowed to give monotone feedback, we overcome this by exploiting a probabilistic observation about the correlation of uniform random vectors in high dimensions. The algorithm we derive from this observation is straightforward to implement, contributing to the arsenal of practically applicable data science algorithms.

Generalization of short program interpolators. The most central machine learning problem is that of *generalization* – given a sample of training data from some data distribution, recover a classifier that performs well on most unseen data from the distribution. On a technical level, one wants to supply a learning rule and a mathematical analysis that shows that the expected generalization error is not large. A common generalization issue faced in practice is that of overfitting, where a learning algorithm recovers a classifier that perfectly classifies a noisy training set but does not generalize any better than a trivial predictor that randomly guesses on each example. Conventional wisdom therefore suggests that models that perfectly fit their noisy training data will suffer from overfitting and will perform poorly at test time [HTFF09].

However, a recent line of work on benign overfitting suggests that under certain conditions, a learner that perfectly fits its noisy training set will still generalize well to unseen data. This reflects a commonly observed phenomenon concerning neural networks, i.e., that overparameterized neural networks generalize surprisingly well to test sets despite perfectly fitting their training sets. This cannot be explained by classical statistical learning theory. Therefore, building a theory for this phenomenon is one of the most pressing open questions in the theory of deep learning [MSAPBN22; KBK22]. Although directly analyzing neural networks appears exceedingly challenging, a compelling first step to study is the the shortest-program learning rule – i.e., "find the shortest computer program that correctly classifies the training set." Here, our question is – *can we analyze the overfitting incurred by running the shortest program learning rule on a noisy training set? Is the overfitting benign, catastrophic, or something else?*

In a joint work with Nathan Srebro [MS23], I obtained the first generalization bounds for the shortest-program learning rule in the presence of label noise, thereby answering the above question. Notably, we prove that the overfitting is neither catastrophic nor benign and the generalization error can be bounded as a function of the error of the Bayes-optimal classifier. Our result follows from explicitly constructing a short

program that can interpolate any training set, bounding its length, and proving an information-theoretic generalization bound in terms of the length of any short interpolating program.

2.2 Succinctly summarizing large datasets

In many machine learning and data science settings, a practitioner is faced with a dataset that is too large to store in memory or for downstream computation. We therefore desire fast algorithms that can construct a summary that is interpretable and preserves key properties of the original dataset. I am therefore interested in – *when and how we can construct small summaries of massive datasets?*

Ellipsoidal approximations and coresets. One natural geometric notion of a summary is that of an *ellipsoidal approximation*. Specifically, given a convex body $Z \subset \mathbb{R}^d$ we seek a center $\boldsymbol{c} \in \mathbb{R}^d$ and an ellipsoid $\mathcal{E} \subset \mathbb{R}^d$ for which $\boldsymbol{c} + \alpha \cdot \mathcal{E} \subseteq Z \subseteq \boldsymbol{c} + \mathcal{E}$, where $\alpha \in (0, 1)$ (observe that a larger α implies that $\mathcal{E}$ is a better approximation). A foundational result due to John [Joh48] states that for any convex body, $\alpha = 1/d$ is achievable and optimal, and when the body Z is origin-symmetric, $\alpha = 1/\sqrt{d}$ is achievable and optimal.

In a sequence of joint works with Yury Makarychev and Max Ovsiankin [MMO22; MMO23], I gave the first algorithms to construct ellipsoidal approximations of convex polytopes in a streaming and online model, where the points belonging to the dataset arrive one at a time and we are interested in preserving the convex hull of the points. Our result can be seen as an online variant of John's theorem for polytopes. Our algorithms are asymptotically nearly optimal in that they yield a dependence on d that nearly matches that given by John's theorem in both the symmetric and asymmetric settings. We also show how to use these algorithms to design the first algorithms to output *coresets* of the convex hull – that is, a subset of the vertices of the convex body whose convex hull approximates Z well.

Sparse approximations to convex objectives. In many data science applications, one wishes to optimize a function of the form $f(\boldsymbol{x}) = \|\mathbf{A}\boldsymbol{x}\|_p^p = \sum_{i=1}^n |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^p$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \geq 1$. If n is very large, then each evaluation of $f(\boldsymbol{x})$ could become expensive. It would therefore be convenient to find weights $\beta_1, \ldots, \beta_n$, most of which are 0, such that for all $\boldsymbol{x} \in \mathbb{R}^d$, we have $\|\mathbf{A}\boldsymbol{x}\|_p^p \approx \sum_{i=1}^n \beta_i |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^p$ (note that this is equivalent to constructing a weighted subset of the summands of f). A typical approach to this problem is importance sampling – one constructs scores $\rho_1, \ldots, \rho_n$ and includes summand $|\langle \boldsymbol{a}_i, \cdot\rangle|^p$ with probability proportional to ρ_i. The main technical challenge is to analyze the number of rounds of sampling required with given ρ_i such that the new objective formed by summing the (appropriately reweighted) subsampled elements is in fact a good approximation to f. The argument to do so typically involves viewing this sampling procedure as a random process and combining an understanding of the geometry of the resulting space of events with a careful high-dimensional probabilistic analysis.

In a joint, ongoing work with Max Ovsiankin [MO23], I defined a simple yet powerful geometric principle called the *ellipsoidal sampling framework*. We prove that the ellipsoidal sampling framework yields a unified method to analyze the sparsity arising from sampling using any set of given importances $\rho_1, \ldots, \rho_n$. This significantly generalizes earlier work (particularly [LT91] and [SZ01]) from the geometric functional analysis community that give similar arguments for a specific instantiation of the importances (known as Lewis weights). Our framework also allows one to go beyond the types of objectives described earlier and implies results for constructing sparse approximations for hypergraphs and mixed norm objectives.

Optimally certifying monotone functions. Another summarization problem that occurs in machine learning contexts is that of explainability. In this problem, we are given black-box access to some classifier f and a test point $\boldsymbol{x}^\star$. Our goal is to output a subset of $\boldsymbol{x}^\star$ that explains the prediction $f(\boldsymbol{x}^\star)$. A theoretically natural way to model this problem is from the lens of *certification*, a notion studied in complexity theory. In the certification problem, we are given query access to a Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ and a test point $\boldsymbol{x}^\star \in \{0,1\}^n$. Our task is to output a minimal subset of the bits of $\boldsymbol{x}^\star$ such that no matter how the remaining bits are varied, f returns the same value.

In a joint work with Meghal Gupta [GM23], I gave the first algorithm with optimal query complexity to solve the certification problem when f is *monotone*. This resolved an open problem posed at STOC 2022 [BKLT22]. We also showed that finding the shortest possible certificate may require up to $2^{\Omega(n)}$ queries to f, which yields a separation between finding a minimal certificate and finding the globally shortest certificate.

3 Future work

I am excited by modern challenges in algorithmic data science. I specifically plan to continue studying data science problems under unclean data sources and to focus more on the role that high-dimensional phenomena play in algorithm design. In particular, a common technical theme among many of my results is the idea that viewing problems from the lens of high-dimensional geometry or probability unlocks a vast set of tools that yield new, deep insights to the problem.

Adversarial data sources – malicious and monotone. In one direction, I will continue studying how we can build algorithms that are more robust to input scenarios seen in practice. For example, I will both investigate the sorts of malicious input perturbations that yield dangerous, undesirable behavior in machine learning models in practice (building off of my earlier work in backdoor poisoning) and build new algorithms for robust learning under training-time attacks.

I also intend to study data science problems in semi-random models. In a semi-random model, a problem instance with a planted solution is chosen from some underlying distribution. An adversary is then allowed to make monotone changes to the input – that is, changes that do not affect the correctness of the ground truth solution. For example, for a classification problem, an adversary may inject correctly labeled examples into the training set [BHQS21]. For a community detection problem, an underlying instance may be sampled from the stochastic block model and then an adversary may add several edges within the communities while deleting some of the crossing edges. Such changes clearly do not affect the correctness or optimality of the ground-truth solution, but perhaps surprisingly, many natural algorithms are not robust to these "helpful" changes. I therefore want to understand which commonly used algorithms are robust to monotone adversaries or build new algorithms for existing problems with a monotone adversary. In fact, I have already begun work in this direction; in an ongoing joint work with Agastya Jha, Michael Kapralov, Davide Mazzali, and Weronika Wrosz-Kaminska, I am attempting to analyze a widely used spectral algorithm for clustering in the semi-random stochastic block model.

Random matrix theory beyond the spectral norm. There is a rich theory studying the spectrum of random matrices under various structures [Wig55; LP91; Tro15]. However, spectral concentration guarantees cannot make sufficiently strong statements about other natural notions of distortion for random matrices. In fact, while we can view the aforementioned sparse approximation problem as a matrix concentration inequality for non-spectral properties, a further theory is lacking. Motivated by my prior work for the sparse approximation problem [MO23], I am deeply interested in building an understanding of matrix concentration with respect to non-Euclidean distortions.

Sampling from log-concave distributions. Finally, I would like to determine when we can obtain fast algorithms for the problem of log-concave sampling. In the log-concave sampling problem, the goal is to generate samples from a density proportional to $\exp\left(-f(\boldsymbol{x})\right)$ where f is convex. This is an overtly geometric problem (since the argument $\boldsymbol{x}$ is high-dimensional) whose applications to optimization and data science are endless. For example, a well-known result due to Bertsimas and Vempala [BV04] essentially reduces convex optimization to sampling uniformly from a convex set, which is a special case of a log-concave distribution. Faster log-concave sampling algorithms will also imply faster algorithms for a number of other optimization subroutines used in data science (e.g. the settings studied in [AHM15], [JLLS23]).

References

[ABCPK18] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: watermarking deep neural networks by backdooring, 2018. arXiv: 1802.04633 [cs.LG] (cited on page 1).

[AHM15] Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/38913e1d6a7b94cb0f55994f679f5956-Paper.pdf (cited on page 4).

[BV06] Eyal Beigman and Rakesh Vohra. Learning from revealed preference. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 36–42, 2006 (cited on page 2).

[BV04] Dimitris Bertsimas and Santosh Vempala. Solving convex programs by random walks. *Journal of the ACM (JACM)*, 51(4):540–556, 2004 (cited on page 4).

[BKLT22] Guy Blanc, Caleb Koch, Jane Lange, and Li-Yang Tan. The query complexity of certification. *arXiv preprint arXiv:2201.07736*, 2022 (cited on page 3).

[BGLMSY23] Avrim Blum, Meghal Gupta, Gene Li, Naren Sarayu Manoj, Aadirupa Saha, and Yuanyuan Yang. Dueling optimization with a monotone adversary. *under review*, September 2023 (cited on pages 1, 2).

[BHQS21] Avrim Blum, Steve Hanneke, Jian Qian, and Han Shao. Robust learning under clean-label attack, 2021. arXiv: 2103.00671 [cs.LG] (cited on page 4).

[CLLLS17] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning, 2017. arXiv: 1712.05526 [cs.CR] (cited on page 1).

[GM23] Meghal Gupta and Naren Sarayu Manoj. *An optimal algorithm for certifying monotone functions*. In *Symposium on Simplicity in Algorithms (SOSA)*. January 2023, pages 207–212 (cited on pages 1, 3).

[HTFF09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009 (cited on page 2).

[JLLS23] Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying sums of norms, 2023. arXiv: 2305.09049 [cs.DS] (cited on page 4).

[Joh48] Fritz John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays Presented to R. Courant on his 60th Birthday*, pages 187–204. Interscience Publishers, Inc, 1948 (cited on page 3).

[KBK22] K. Kawaguchi, Y. Bengio, and L. Kaelbling. Generalization in deep learning. In *Mathematical Aspects of Deep Learning*, pages 112–148. Cambridge University Press, December 2022. DOI: 10.1017/9781009025096.003. URL: https://doi.org/10.1017%2F9781009025096.003 (cited on page 2).

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991. ISBN: 9783540520139. URL: https://books.google.com/books?id=cyKYDfvxRjsC (cited on page 3).

[LP91] Françoise Lust-Piquard and Gilles Pisier. Non commutative khintchine and paley inequalities. *Arkiv för matematik*, 29:241–260, 1991 (cited on page 4).

[MMO22] Yury Makarychev, Naren Sarayu Manoj, and Max Ovsiankin. Streaming algorithms for ellipsoidal approximation of convex polytopes. In *Proceedings of Thirty Fifth Conference on Learning Theory (COLT)*, pages 3070–3093, July 2022 (cited on pages 1, 3).

[MMO23] Yury Makarychev, Naren Sarayu Manoj, and Max Ovsiankin. Near-optimal streaming ellipsoidal rounding for general convex polytopes. *working manuscript*, October 2023 (cited on pages 1, 3).

[MSAPBN22] Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: a taxonomy of overfitting, 2022. arXiv: 2207.06569 [cs.LG] (cited on page 2).

[MB21] Naren Sarayu Manoj and Avrim Blum. Excess capacity and backdoor poisoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021 (cited on page 1).

[MO23] Naren Sarayu Manoj and Max Ovsiankin. An ellipsoidal sampling condition for hypergraph sparsification and generalized matrix row sampling. *working manuscript*, October 2023 (cited on pages 1, 3, 4).

[MS23] Naren Sarayu Manoj and Nathan Srebro. Shortest program interpolation learning. In *Proceedings of Thirty Sixth Conference on Learning Theory (COLT)*, pages 4881–4901, July 2023 (cited on pages 1, 2).

[OWJ+22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022 (cited on page 2).

[SSP20] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. *AAAI 2020*, January 2020. arXiv: 1910.00033 [cs.CV] (cited on page 1).

[SZ01] Gideon Schechtman and Artem Zvavitch. Embedding subspaces of lp into np, 0< p< 1. *Mathematische Nachrichten*, 227(1):133–142, 2001 (cited on page 3).

[SJ11] Pannagadatta K Shivaswamy and Thorsten Joachims. Online learning with preference feedback. *arXiv preprint arXiv:1111.0712*, 2011 (cited on page 2).

[TLM18] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *NeurIPS 2018*, December 2018. arXiv: 1811.00636 [cs.LG] (cited on page 1).

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities, 2015. arXiv: 1501.01571 [math.PR] (cited on page 4).

[TJHAPJNT20] Loc Truong, Chace Jones, Brian Hutchinson, Andrew August, Brenda Praggastis, Robert Jasper, Nicole Nichols, and Aaron Tuor. Systematic evaluation of backdoor data poisoning attacks on image classifiers, April 2020. arXiv: 2004.11514 [cs.CV] (cited on page 1).

[WSRVASLP20] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: yes, you really can backdoor federated learning, 2020. arXiv: 2007.05084 [cs.LG] (cited on page 1).

[Wig55] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955. ISSN: 0003486X. URL: http://www.jstor.org/stable/1970079 (visited on 10/14/2023) (cited on page 4).