Research Statement

Chen Cheng*

Datasets are central to the development of statistical learning theory, and the evolution of models. The burgeoning success of modern machine learning in sophisticated tasks crucially relies on the vast growth of massive datasets (cf. Donoho [12]), such as ImageNet [11], SuperGLUE [16] and Laion-5b [15]. However, such evolution breaks standard statistical learning assumptions and tools. My research revolves around understanding *the complexity of modern datasets* by developing new theoretical tools and studying unconventional models.

The study of traditional datasets consisting of $\{(X_i, Y_i)\}_{i=1}^n$ produced the rich and mature theory of text book statistical learning theory such as empirical risk minimization, witnessed the development of renowned models (e.g. EM algorithms and random forests), and they keep playing vital roles in casual inference and conformal prediction.

The truly phenomenal, in recent years, is the explosive emergence of modern datasets. They drive the development of statistical machine learning models and make unimaginably accurate predictions in complex tasks such as computer vision and artificial intelligence, harnessing the power of revolutionary models including deep neural networks, reinforcement learning from human feedbacks (RLHF) and large language models (LLM). The modern datasets despite their great powers, often break the textbook assumptions for classical theory and models—the data may not have single labels such as $\{(X_i, Y_i)\}_{i=1}^n$, features X_i may be high-dimensional or missing elements, and the underlying distribution $\mathbb{P}_X$ may shift from interaction with learning, etc.—such complexities challenge traditional tools of statistical machine learning. How should we think about fitting those models beyond conventional wisdom? What mathematical foundations we can lay that we can leverage to do more?

Therefore, my research goal is to study the success of modern ML and ground-breaking models by unraveling the mystery of *the complexity of modern datasets*—more precisely, I aim to develop statistical theory to explain unusual behavior of modern ML, make and test hypotheses in both datasets and methodologies, and predict model behaviors that we can leverage to streamline learning methods. I underwent a PhD journey of understanding modern datasets by addressing a few outstanding aspects unfamiliar to traditional models and theory—I gained both unique intuition and developed novel theoretical tools that contribute to unraveling the power of modern datasets. It also comes to my realization that those angles merely scratch the surface of modern datasets, and there are other problems that substantially interest me such as online learning and diffusion models. I aim to delve deeper into my current areas of focus and expand into these fields in my future career. My current research encompasses the following topics, highlighted by selected representative works.

High dimensional data and overparametrized models. In classical statistics, the feature dimension or the number of model parameters d is much smaller than the number of data points n, i.e. $d/n \to 0$. As modern datasets grow increasingly high dimensional and models become overly parameterized, it is crucial to understand the behavior of statistical machine learning under such settings, i.e. $d/n \to \gamma \in (0, \infty)$ or even $d/n \to \infty$. We approach various aspects of this problem in a series of work [4, 2, 7, 5, 9]. Notably, those works tackle the property of datasets with the design matrix $X = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n \times d}$ from a parametric model $y = f(x; \theta, \epsilon), y, \epsilon \in \mathbb{R}, \theta \in \mathbb{R}^{d \times k}$, where the dimension d is comparable or much larger than the sample size n.

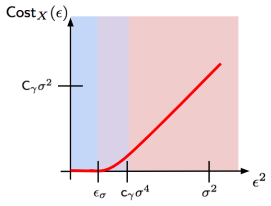

Figure 1: Cost of not fitting vs. training error.

[*] Department of Statistics, Stanford University; Email: chencheng@stanford.edu.

- *Memorize to generalize: on the necessity of interpolation in high dimensional linear regression* [9]. Recent work has sought to develop an understanding of "implicit regularization" in overparameterized interpolating models: whereas most minimizers of the empirical risk may generalize poorly, standard learning algorithms used in practice such as (stochastic) gradient descent tend to converge to solutions that do generalize well in many "benign-overfitting" scenarios (cf. [13, 14]). In those cases, it is sufficient to generalize well by interpolation.

  There nonetheless remains at least some reason to be skeptical of the notion that interpolation is necessarily "benign". In this work, we examine the necessity of interpolation in overparameterized models, that is, when achieving optimal predictive risk in machine learning problems requires (nearly) interpolating the training data. In particular, we consider simple overparameterized linear regression $y = X\theta + w$ with random design $X \in \mathbb{R}^{n\times d}$ under the proportional asymptotics $d/n \to \gamma \in (1, \infty)$. We precisely characterize how prediction (test) error necessarily scales with training error in this setting.

  Intriguingly, it implies as the label noise variance $\sigma^2 \to 0$, any estimator that incurs at least $\mathsf{c}\sigma^4$ training error for some constant c is necessarily suboptimal and will suffer growth in excess prediction error at least linear in the training error (see Fig. 1). Thus, optimal performance requires fitting training data to substantially higher accuracy than the inherent noise floor of the problem.

- *The high-dimensional asymptotics of first order methods with random data* [2]. First order methods are key to modern machine learning framework, which requires understanding the behavior of gradient descent dynamics in high-dimensional and non-convex landscapes. The main challenge is the data reuse at each step and the correlation it induces, as we cannot use new batch of data when n is comparable to d. In this work, we study a class of deterministic flows $\{\theta_t\}_{t=0}^{T} \in C(\mathbb{R}^{d\times k}, [0, T])$, parametrized by a random design matrix $X \in \mathbb{R}^{n\times d}$ with i.i.d. centered entries. We characterize the asymptotic behavior of these flows over bounded time horizons, in the high-dimensional limit in which $n, d \to \infty$ with k fixed and converging aspect ratios $d/n \to \gamma \in (0, \infty)$. The asymptotic characterization we prove is in terms of a system of a nonlinear stochastic process in k dimensions, whose parameters are determined by a fixed point condition.

  This type of characterization is known in physics as dynamical mean field theory (DMFT). Our proof is based on time discretization and a reduction to certain iterative schemes known as approximate message passing (AMP) algorithms, as opposed to earlier work that was based on large deviations theory and stochastic processes theory. As the asymptotic characterizations by DMFT are processes in low dimension, we can cheaply make predictions to first order methods behaviors in high-dimensional and non-convex landscapes, which may greatly reduce the cost of hyper-parameter tuning via retraining in practice.

- *Dimension free ridge regression* [5]. In previous two highlighted works, random matrix theory plays crucial roles in our theoretical analysis. However, random matrix theory is largely focused on the proportional asymptotics in which the number of columns grows proportionally to the number of rows of the data matrix. This is not always the most natural setting in statistics where columns correspond to covariates and rows to samples. With the objective to move beyond the proportional asymptotics, we revisit ridge regression (ℓ_2-penalized least squares) on i.i.d. data $X \in \mathbb{R}^{n\times d}, y \in \mathbb{R}^n$. We allow the feature vector to be infinite-dimensional ($d = \infty$), in which case it belongs to a separable Hilbert space. Within this setting, we establish non-asymptotic bounds that approximate the bias and variance of ridge regression in terms of the bias and variance of an 'equivalent' sequence model (a regression model with diagonal design matrix). The approximation is up to multiplicative factors bounded by $(1\pm\Delta)$ for some explicitly small Δ. Previously, such an approximation result was known only in the proportional regime and only up to additive errors: in particular, it did not allow to characterize the behavior of the excess risk when this converges to 0.

  Our general theory recovers earlier results in the proportional regime (with better error rates). As a new application, we obtain a completely explicit and sharp characterization of ridge regression for

Hilbert covariates with regularly varying spectrum. Finally, we analyze the overparametrized near-interpolation setting and obtain sharp 'benign overfitting' guarantees.

Multiple labels and data aggregation. Multilabeling is another curious aspect of modern human-labeled datasets that is often missing in statistical machine learning literature. This could arise from aggregating non-expert labels, or combining different models trained on their own data. We address this problem in [8, 10].

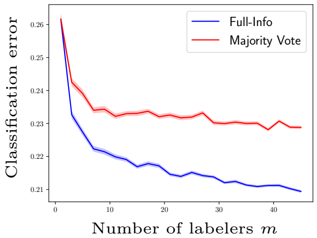

Figure 2: Experiments on CIFAR-10H dataset.

- *How many labelers do you have? A closer look at gold-standard labels* [8]. In earlier years, experts could provide reliable labels for reasonably sized datasets, the cost and size of modern datasets often precludes this expert annotation, motivating a growing literature on crowdsourcing and other sophisticated dataset generation strategies that aggregate expert and non-expert feedback or collect internet-based loosely supervised and multilabeled data. By aggregating multiple labels, one typically hopes to obtain clean, true, "gold-standard" data. Yet most statistical machine learning development—theoretical or methodological—does not investigate this full data generating process, assuming only that data comes in the form of (X, Y) pairs of covariates X and targets (labels) Y. To that end, we develop a stylized theoretical model to capture uncertainties in the labeling process, allowing us to understand the contrasts, limitations and possible improvements of using aggregated or non-aggregated data in a statistical learning pipeline. We model each example as a pair $(X_i, (Y_{i1}, \ldots, Y_{im}))$ where X_i is a data point and Y_{ij} are noisy labels. In these stylized models, we show how access to non-aggregated label information can make training well-calibrated models more feasible than it is with gold-standard labels. The theory makes several predictions for real-world datasets, including when non-aggregate labels should improve learning performance, which we test to corroborate the validity of our predictions (cf. Fig. 2).

  Concretely, the contrasts between aggregated and fuller label information depend on the particulars of the problem, where estimators that use aggregated information exhibit robust but slower rates of convergence, while estimators that can effectively leverage all labels converge more quickly if they have fidelity to (or can learn) the true labeling process.

Robustness quantification and fundamental limits for structured datasets. Apart from high-dimensionality and multilabeling, we also study datasets with specific structures inspired from modern machine learning, distributionally and geometrically. In particular, we develop theory that quantifies robustness of the problem and provides information-theoretic limits under these structural assumptions in [10, 1, 6].

- *Geometry, computation, and optimality in stochastic optimization* [6]. In stochastic convex optimization, linear updates are desirable as they are computationally cheap and easy to scale, which also motivates the study for general first order methods [2] in the last section. However, it generally remains unclear whether cheap computation is sufficient for optimality.

  We ask the question of what geometric constraints are required for linear methods to be minimax optimal, drawing the parallel between stochastic optimization and the classical Gaussian sequence models. By focusing on constraint set and gradient geometry, we characterize the problem families for which stochastic- and adaptive-gradient methods are (minimax) optimal and, conversely, when nonlinear updates—such as those mirror descent employs—are necessary for optimal convergence.

  Specifically, when the constraint set is quadratically convex, diagonally pre-conditioned stochastic gradient methods are minimax optimal. We provide quantitative converses showing that the "distance" of the underlying constraints from quadratic convexity determines the sub-optimality of subgradient methods.

- *Collaboratively learning linear models with structured missing data* [10]. As models keep evolving in real world, it is nature to ask how do we leverage data from different models and jointly combine them to make better predictions—for example, photo qualities improve as satellites' sensors upgrade. Instead of just throwing away previous models and data, we wish to leverage shared structure to achieve better accuracy. This work also aligns with my previous topic of data aggregation, and we study the problem of collaboratively learning least squares estimates for m agents. Each agent observes a different subset of the features—e.g., containing data collected from sensors of varying resolution.

  Our goal is to determine how to coordinate the agents in order to produce the best estimator for each agent. We propose a distributed, semi-supervised algorithm, consisting of three steps: local training, aggregation, and distribution. Our procedure does not require communicating the labeled data, making it communication efficient and useful in settings where the labeled data is inaccessible. Despite this handicap, our procedure is nearly asymptotically, local-minimax optimal—even among estimators allowed to communicate the labeled data such as imputation methods.

Other adventures. Besides the above main topics, I enjoy a few side-quests during my PhD career. My industrial research intern experience at LinkedIn involved studying backwards martingale and differential privacy. I also worked on natural policy gradient methods in reinforcement learning with entropy regularization [3].

Conclusion and research plan. My research, spanning across fields of statistical machine learning, focus on demystifying the complex aspects of modern datasets that are central to the success of modern machine learning.

Moving forward, I plan to further my current research focus, harnessing my expertise in random matrix theory, high dimensional statistics and information theory. I also plan to explore broader areas around the theme of understanding modern datasets, such as high dimensional inverse problems and diffusion models, as well as establishing the connection between backwards martingales and differential privacy thoroughly.

References

[1] F. Areces, C. Cheng, J. Duchi, and K. Rohith. Two fundamental limits for uncertainty quantification in predictive inference. In *Conference on Learning Theory*, pages 186–218. PMLR, 2024.

[2] M. Celentano, C. Cheng, and A. Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[3] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

[4] Y. Chen, C. Cheng, and J. Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *Annals of Statistics*, 49(1):435, 2021.

[5] C. Cheng and A. Montanari. Dimension free ridge regression. *Annals of Statistics, to appear*, 2024.

[6] C. Cheng, J. C. Duchi, and D. Levy. Geometry, Computation, and Optimality in Stochastic Optimization.

[7] C. Cheng, Y. Wei, and Y. Chen. Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Transactions on Information Theory*, 67(11):7380–7419, 2021.

[8] C. Cheng, H. Asi, and J. Duchi. How many labelers do you have? A closer look at gold-standard labels. *arXiv preprint arXiv:2206.12041*, 2022.

[9] C. Cheng, J. Duchi, and R. Kuditipudi. Memorize to generalize: on the necessity of interpolation in high dimensional linear regression. In *Conference on Learning Theory*, pages 5528–5560. PMLR, 2022.

[10] C. Cheng, G. Cheng, and J. Duchi. Collaboratively learning linear models with structured missing data. *Advances in Neural Information Processing Systems*, 36, 2024.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[12] D. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[13] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

[14] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2): 949, 2022.

[15] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[16] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in Neural Information Processing Systems*, 32, 2019.