Information Theory for Trustworthy Machine Learning

Hao Wang, Harvard University

My research uses information theory to build the theoretical foundations of trustworthy machine learning (ML). I derive new generalization bounds for analyzing how the data distribution and the optimization method affect the generalization of complex ML models. I develop theory that delineates the fundamental limit of algorithmic fairness and privacy. The theory provides design guidelines to practitioners who deploy ML technology in applications of individual-level consequences. To date, my work has been published in top venues in information theory and machine learning, including IEEE Transactions on Information Theory, ISIT, ICML, and NeurIPS.

1 Generalization in Machine Learning

Deep neural networks (DNNs) have achieved enormous success in various applications. Despite being highly expressive, DNNs often generalize extremely well to unseen data. These empirical observations are not captured by the traditional wisdom, which attributes the generalization to the use of a hypothesis class with constrained complexity. Several recent efforts introduce new generalization bounds for complex models using the mutual information between the training set and the output from the learning algorithm. Alas, a major challenge that obstructs the deployment of these bounds is that computing the mutual information from data is intractable. Building on this information-theoretic framework, my research aims to provide generalization bounds that can be computed from data and help understand empirical observations of the generalization of DNNs.

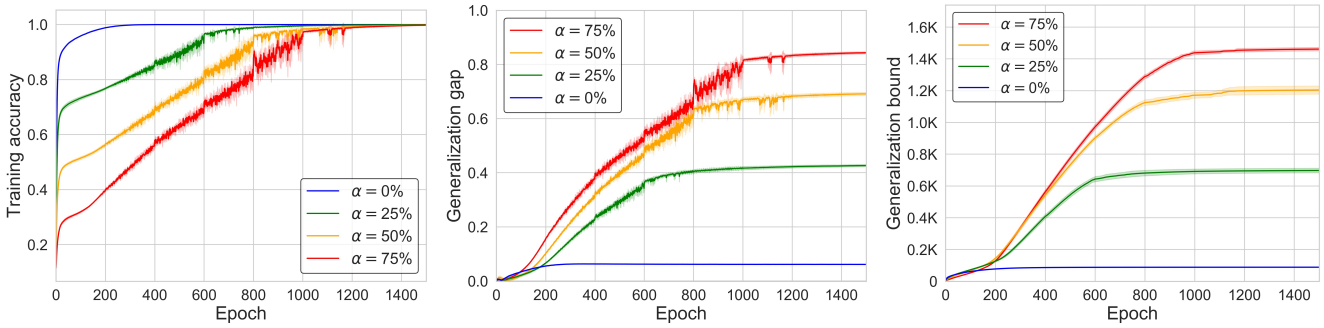

Fig. 1: Illustration of the generalization bound I derived for the stochastic gradient Langevin dynamics (SGLD) algorithm. I trained 3-layer neural networks on the MNIST dataset when the training data have different label corruption levels $\alpha \in \{0\%, 25\%, 50\%, 75\%\}$. The training process continues until the training accuracy is 1.0 (Left). As shown, my bound (Right) is highly correlated with the generalization gap (Middle). When α or the number of epochs is increasing, the generalization gap is increasing and my bound can capture this phenomenon.

Optimization-Based Generalization Bounds. Optimization is a key component for training DNNs. To understand how it influences the generalization of DNNs, my research derives generalization bounds that are tailored to specific optimization methods. My recent work [NeurIPS21] investigates the generalization of models trained by the stochastic gradient Langevin dynamics (SGLD) algorithm. The key contribution is to connect SGLD with *Gaussian channels* found in communication and information theory. This connection enables me to derive state-of-the-art generalization bounds that depend on the data distribution; can be computed from the training data; and incorporate the variance of gradients for quantifying the sharpness of the loss landscape. Numerical experiments suggest that my generalization bounds are *highly* correlated with the true generalization gap and can help understand some empirical observations of DNNs (e.g., label corruption).

My follow-up paper [WGC21] extends the analysis to a class of noisy iterative algorithms and improves existing approaches in two aspects. First, existing bounds are mostly monotonically increasing with time. In contrast, my bounds incorporate a time-decaying factor which enables the impact of early iterations to decrease with time. This decay factor is established by a fundamental tool from information theory, namely *strong data processing inequalities*, and prevents the bounds from blowing up when time increases. Second, I introduce a new framework for analyzing the generalization of noisy iterative algorithms. It uses the total variation distance for bounding the generalization and is built upon my previous work [ISIT19]. This framework not only leads to a significantly tighter bound but also deals with a much broader class of noisy iterative algorithms.

Future Direction: Benign Overfitting in the Overparameterized Regime. The uniform convergence results from statistical learning theory suggest that models that perfectly fit the training data will exhibit a poor generalization performance. Although these results are useful for understanding classical ML, they fail to explain the behavior of deep learning methods in the overparameterized regime where models can interpolate all training data and achieve perfect training accuracy. Surprisingly, in this regime, prediction models often exhibit "benign" overfitting—they generalize well to unseen data while overfitting the training data. In the future, I plan to develop theory to understand the generalization of deep learning methods in the overparameterized regime. In particular, I am interested in exploring the following directions, among others.

- Implicit Regularization. In the overparameterized regime, there exists a large number of solutions to the empirical risk minimization, each having distinct generalization properties. In this case, gradient-based optimization algorithms introduce a bias in selecting a minimizer with certain properties. For example, gradient descent for training a linear model with squared error loss converges to a solution with minimal L_2 norm. I intend to investigate the phenomenon of implicit regularization through the information-theoretic framework developed in my prior work and to derive new optimization-based generalization bounds beyond SGLD.

- Double Descent. The classical learning theory suggests that the test error exhibits a U-shaped curve with respect to model complexity as a result of the bias-variance tradeoff. However, this phenomenon only occurs in the underparameterized regime. When it comes to the overparameterized regime, the test error decreases again with the model complexity and highly overparameterized models often achieve a better test accuracy than the best underparameterized model. I intend to investigate this research direction and propose new model complexity measures for explaining the double descent phenomenon.

2 Fairness and Privacy in Machine Learning

ML models (e.g., DNNs) are increasingly complex and, consequently, increasingly opaque. As a result, when a model propagates bias or leaks private information, it becomes harder for the model designer to detect the bias and correct the model. Motivated by these practical concerns, my research in trustworthy ML has three intertwined goals: (i) develop a theoretical framework for investigating fairness and privacy in ML; (ii) apply this framework to probe ML models for discrimination and to analyze the fundamental limit of privacy-utility trade-off; and (iii) construct theoretically-grounded fair and private algorithms with performance guarantees. In what follows, I describe my recent work towards achieving these goals.

Ensuring the Fair Use of Group Attributes. In applications such as healthcare, ML models often include group attributes (e.g., age and sex). In this setting, it is crucial to ensure the fair use of group attributes—the designed model should tailor treatment to people's unique characteristics while preventing harm to any group of individuals. However, standard methods to encode group attributes may not improve performance for all groups. For example, if a data scientist trains a separate model for each group of individuals, the minority groups with little training data may suffer from overfitting.

Motivated by this practical issue, my recent work [ISIT21, T-IT21] introduces precise conditions for the fair use of group attributes—i.e. conditions under which training a separate model for each group produces the most performance improvement. These conditions are cast in terms of the difference in probability distributions between groups of individuals and are revealed by powerful two-point methods for studying min-max problems. Furthermore, I provide an algorithm that can quickly verify the conditions from data. Currently, I am extending this research to settings where more factors (e.g., sample diversity) are involved. My long-term ambition is to provide a theoretically-grounded framework that helps data scientists design algorithms under fair use of group attributes.

Identifying Proxies for Discrimination. If a ML model is deemed unfair, it is essential to understand the source of discrimination. Discrimination can occur if a ML model produces different decisions for individuals based on a protected attribute. More pervasive today is a phenomenon known as disparate impact, where the protected attribute is omitted from the model but affects the decision through correlations with "proxy" variables. For example, in the history of "redlining", lenders deliberately used zip code as a "proxy" for race, denying home mortgages in areas populated primarily by minorities.

Race	Age	Priors	Degree	Proxy score
0	26	2	Felony	**2.4**
0	32	3	Felony	**1.4**
1	45	0	Misd.	0.01
0	20	0	Misd.	0.8

Fig. 2: My framework creates a method to compute "proxy scores" that indicate which data features may explain bias for a given ML model. By applying this method to the COMPAS recidivism dataset, it flags young individuals with a large number of prior convictions as a potential "proxy".

My work [ISIT18a] develops an information-theoretic framework that can systematically detect "proxy" features for discrimination. This framework is based on a key observation: disparate impact can happen if the distribution of input features to a given ML model varies when conditioned on a protected attribute. Thus, by characterizing the variation of discrimination metrics under local perturbation of data distribution, my framework assigns a "proxy score" for each data point and feature.

Repairing without Retraining The performance of a ML model relies on the distribution of input features. Given a model which performs poorly on a target group, is there a hypothetical distribution of input features that minimizes this performance disparity? In my work [ICML19], I refer to this hypothetical distribution as a counterfactual distribution and design a (functional) gradient descent algorithm for learning this distribution from data. By leveraging the influence function from robust statistics, I derive a close-form expression of the (functional) gradient for a variety of group discrimination metrics. Finally, I construct a mapping from the original data distribution to a counterfactual distribution by solving an optimal transport problem. This mapping can be used as a data pre-processor for repairing the unfair model and has three practical benefits: (i) it minimizes intervention (as it only affects the target group); (ii) it improves the model performance for the target group (on average); (iii) it can be customized to satisfy real-world constraints. This original framework is further extended by my follow-up work [ISIT20] in collaboration with IBM Research.

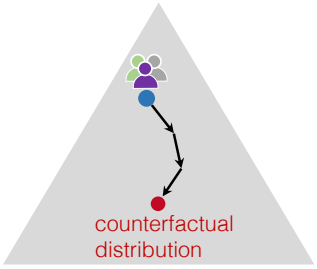

Fig. 3: A counterfactual distribution is a hypothetical distribution that minimizes a given group discrimination metric. I propose a (functional) gradient descent algorithm that learns a counterfactual distribution from data.

Fundamental Privacy-Utility Trade-off. Trustworthy ML hinges on the existence of mechanisms that protect data privacy. A popular method for accomplishing this goal is to perturb each data point so that the private information is masked before disclosing the whole dataset. However, ML models trained from noisy data suffer from an accuracy reduction. My research aims to characterize the fundamental privacy-utility trade-off.

My work [Allerton17, T-IT19] proposes an estimation-theoretic framework where an analyst is allowed to reconstruct certain useful functions from the disclosed data while other private functions should not be reconstructed (in a mean-squared error sense) with distortion under a threshold. Based on this framework, I develop theory to characterize the fundamental privacy-utility trade-off and propose a convex program for designing privacy mechanisms when the data distribution is known a priori. My follow-up papers [ISIT18b, T-IT20] consider when an empirical distribution is used for designing privacy mechanisms. I establish robustness guarantees by proving Lipschitz continuity properties for a variety of information leakage measures.

Future Direction: Auditing Discrimination in ML Models. Ensuring fairness requires being able to detect discrimination in the first place. However, a major challenge for testing group fairness is that the available data for conducting the test are limited but the number of groups can grow exponentially with the number of group attributes. If we conduct an independent hypothesis test for each group, then the probability that at least one null hypothesis is wrongly rejected can increase rapidly. I plan to design statistical tests to verify discrimination which will account for the multiplicity problem. These tests will be extremely useful to audit discrimination in ML models and are currently missing in the literature.

References

[Allerton17] **H. Wang** and F. P. Calmon. An Estimation-theoretic View of Privacy. In *Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[ICML19] **H. Wang**, B. Ustun, and F. P. Calmon. Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions. In *International Conference on Machine Learning (ICML)*, 2019.

[ISIT18a] **H. Wang**, B. Ustun, and F. P. Calmon. On the Direction of Discrimination: An Information-theoretic Analysis of Disparate Impact in Machine Learning. In *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[ISIT18b] **H. Wang**, M. Diaz, F. P. Calmon, and L. Sankar. The Utility Cost of Robust Privacy Guarantees. In *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[ISIT19] **H. Wang**, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon. An Information-theoretic View of Generalization via Wasserstein Distance. In *IEEE International Symposium on Information Theory (ISIT)*, 2019.

[ISIT20] W. Alghamdi, S. Asoodeh, **H. Wang**, F. P. Calmon, D. Wei, and K. N. Ramamurthy. Model Projection: Theory and Applications to Fair Machine Learning. In *IEEE International Symposium on Information Theory (ISIT)*, 2020.

[ISIT21] **H. Wang**, H. Hsu, M. Diaz, and F. P. Calmon. The Impact of Split Classifiers on Group Fairness. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[NeurIPS21] **H. Wang**, Y. Huang, R. Gao, and F. P. Calmon. Analyzing the Generalization Capability of SGLD Using Properties of Gaussian Channels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[T-IT19] **H. Wang**, L. Vo, F. P. Calmon, M. Médard, K. R. Duffy, and M. Varia. Privacy with Estimation Guarantees. *IEEE Transactions on Information Theory*, 2019.

[T-IT20] M. Diaz*, **H. Wang***, F. P. Calmon, and L. Sankar. On the Robustness of Information-theoretic Privacy Measures and Mechanisms. *IEEE Transactions on Information Theory*, 2020. ***Equal contribution**.

[T-IT21] **H. Wang**, H. Hsu, M. Diaz, and F. P. Calmon. To Split or Not to Split: The Impact of Disparate Treatment in Classification. *IEEE Transactions on Information Theory*, 2021.

[WGC21] **H. Wang**, R. Gao, and F. P. Calmon. Generalization Bounds for Noisy Iterative Algorithms Using Properties of Additive Noise Channels. *working paper*, 2021.