My research interests lie broadly in learning theory, especially developing new paradigms to understand how complex Machine Learning (ML) models extract information effectively from data. I begin by describing my research in *Four Strands* of efforts to advance the frontier of learning theory and mathematical statistics and then lay out the future research agenda. All numbered references below refer to items on my vitae.

Strand I: Classic Empirical Risk Minimization (ERM) Principle. Uniform Law of Large Numbers (ULLN) laid the probabilistic foundation for classical learning theory and mathematical statistics. When the model class is convex with controlled complexity, the ERM principle ensures that learning from empirical data is possible and often optimal. The complexity control is typically achieved by explicit regularization. In this research strand, I studied the extent to which analysis based on classic ULLN and the ERM principle can provide rigorous mathematical guarantees for complex ML models, including deep neural networks. [16, *Econometrica*] answered how well standard deep ReLU neural networks estimate classic nonparametric functions and how to perform downstream semiparametric inference robustly. The analysis leverages careful localization arguments for the non-convex class (to improve ULLN) and recent approximation theory developments. [3, *COLT*] discovered a new localization approach by proposing "offset Rademacher complexity" to derive rates for general non-convex and unbounded function spaces. For learning probability distributions, [18, *J. Mach. Learn. Res.*] formulated the first statistical framework to understand generative adversarial networks. The framework identified the curious regularization effect on the tradeoffs of the generator-discriminator-pair.

Strand II: New Minimum-Norm Interpolation (MNI) Principle. According to conventional wisdom, explicit regularization should be added to complex models to prevent "interpolating" the training data. Curiously, abundant empirical evidence suggests that modern ML methods perform well statistically even at interpolation, without "explicit regularization." ULLN and the ERM principle fail to explain learning in this new interpolation regime. One likely hypothesis is that methods or algorithms favor a certain "minimal" way of interpolating the data, typically measured by certain norms induced by algorithms. In this research strand, I investigated the MNI principle for practical ML models in the interpolation regime and discovered new phenomena in high dimensions. [13, *Ann. Stat.*] studied interpolation with kernel ridgeless regression and identified that, surprisingly, high dimensionality and non-linear inner-product kernel generate an "implicit regularization" effect. Hence, the MNI principle generalizes well with favorable geometric properties on the design. [15, *COLT*] extended the above linear growth regime (dimension proportional to sample-size) to a broader polynomial growth regime and discovered a curious "multiple-descent shape" of the risk curve as the dimension grows. [20, *Ann. Stat., R&R*] examined the MNI principle for classification problems and established a precise high dimensional asymptotic theory for boosting. At the heart of the analysis lies an exact characterization of the max-min margin and the generalization error of MNI induced by AdaBoost. My contributions are among the first to the rapidly growing literature on understanding the statistical properties of the MNI principle.

Strand III: Representation Puzzles in Neural Networks. How to quantify the adaptive representation aspects of modern learning such as deep neural networks and random forests still remains a puzzle in theory. [14, *J. Am. Stat. Assoc.*] considered gradient flow training of two-layer neural networks and established provable benefits of the adaptive representation in neural networks compared to the prespecified fixed-basis representation in classical nonparametrics. [17, *J. Am. Stat. Assoc.*] studied the mathematical role of non-linear activation in deep neural networks and how it affects interpolation and memorization capacity.

Strand IV: Algorithmic Component of Learning Theory. One distinctive advantage of modern ML models over classic nonparametric models is the flexible yet tractable algorithmic component. In this research strand, I examined the algorithmic component of learning theory and unveiled the behavior of stochastic-gradient-type algorithms for learning non-convex and overparametrized ML models. [8, *COLT*] and [12, *J. Royal Stat. Soc. B*] considered Langevin dynamics as the stochastic optimization procedure for training generic learning models with possibly non-convex landscapes. We proposed and analyzed new stochastic-gradient-type algorithms for local inference using Langevin dynamics.

Future Work: Overparametrization and Infinite-Dimensional Models. Historically, nonparametric statistics concerns the extreme form of overparametrization: infinite-dimensional models. These models leverage the Nadaraya-Watson idea, which enjoys mathematical flexibility but has modest empirical performance. Recently, practical optimization of overparametrized or infinite-dimensional models attracts a renewed interest due to the empirical success of modern machine learning. What can insights from infinite-dimensional linear or semi-definite programs bring to mathematical statistics and learning theory? My recent papers [23, 22] aim to devise the next generation nonparametric/infinite-dimensional statistical models with computation in mind and address the question mentioned above.