Rajiv A. Khanna — Research Statement

Well-grounded theoretical understanding can aid a practitioner in making crucial design decisions in the modeling process. This is especially relevant for demystifying and building upon the remarkable empirical successes of neural networks. These theoretical insights include identifying key properties of the data, the model and the algorithm which ensure good statistical performance or faster learning. The modeling process is also often iterative, which makes it imperative for the practitioner to refine the model using human-interpretable markers of the predictive process. *My research vision is to develop theoretical tools to elucidate the mechanisms of machine learning models and algorithms beyond just pure predictive performance, and for developing a human-interpretable understanding of model predictions.* My research body uses tools from several sub-fields of machine learning including discrete/continuous optimization, Bayesian modeling, high dimensional statistics and interpretability.

Consider the problem of interpretable dimensionality reduction. Prior theoretical understanding postulated that interpretability and quantitative performance are at odds with each other. In our recent paper, using beyond worst-case theoretical analysis, we identified key properties of the data to show that it is often possible to have good quantitative performance while retaining interpretability, as is also observed in practice. Our paper won *the best paper award at NeurIPS 2020*. Similarly, in another work, we identify key traits that enable robust generalization in adversarially trained neural networks. This also leads to algorithmic developments and to the discovery that robust generalization can imply cross-dataset generalization. In another line of work on model interpretation, I have worked on identifying key training data points which were vital to making a prediction.

In the following sections, I present a representative sample of my recent works and future interests. Having worked in industry for four years before joining my PhD, I understand the importance and challenges of real-world applications. As such, my research draws its motivation from practical problems, with the objective of bridging the gaps in our understanding of these problems and their respective solutions. Towards this goal, my research makes novel theoretical contributions to the fields of optimization and learning theory as well as to algorithmic development guided by these analyses. My work also uncovers new connections across different subfields of machine learning. For example, my work on interpretability using Fisher kernels develops a probabilistic analog of the frequentist approach of influence functions. Apart from the works presented here, I have also worked on optimization theory, sparse approximate inference problems in Bayesian settings, click prediction and recommendation systems.

Bridging theory and practice in feature selection: In this line of research, my focus is to study feature selection as a means of interpretable dimensionality reduction. While in practice simple variants of the greedy forward selection algorithm often work well, the known theoretical guarantees were unable to explain these empirical successes. To bridge this gap, we show that feature selection satisfies a weaker form of submodularity. Submodular functions are a special class of set variate functions that have several interesting properties. For example, a simple greedy algorithm to maximize a monotone, normalized submodular function guarantees a solution that is within $(1 - {}^{1}\!/_{e})$ of the best possible solution of the same size, i.e. greedy gives a constant multiplicative factor approximation guarantee. But in fact submodularity is not necessary to obtain constant factor guarantees. It is known that greedy forward support selection for linear regression also guarantees a constant factor approximation if the design matrix satisfies a technical condition called the Restricted Isometry Property (RIP). These bounds are slightly weaker than the standard submodular bounds. We codify this concept as *weak submodularity*, and extend these results to greedy support selection for general functions based on Restricted Strong Convexity (RSC) and Smoothness (RSM) properties of the function [2]. Submodularity is generally considered for discrete optimization problems. By associating it with sparse continuous optimization, this thread of my research establishes new relationships at the intersection of discrete and continuous optimization. We also obtain statistical generalization guarantees to connect weak submodularity to high dimensional statistics. Furthermore, we extend the applicability of a distributed greedy algorithm typically used for submodular functions to obtain novel bounds for weak submodular functions [4]. We further show that greedy low rank approximation can also be considered within this framework's umbrella [3]. This improves the existing approximation guarantees for greedy low rank approximation by an exponential factor.

A fundamental question in this area relates to the cost of interpretability: how well can a subset of columns of a matrix of size k compete with the best rank k approximation? An established paradigm to quantify the hardness of the problem is the use of constructive lower bounds for the worst-case. For the problem of subset selection, prior worst-case analysis shows that there exist matrices for which the cost of interpretability scales linearly with the size of the summary. However, this is not generally observed in practice. In a recent work that was awarded *the best paper award at NeurIPS 2020* [1], we develop techniques which exploit spectral properties of the data matrix to obtain improved approximation guarantees which go beyond the standard worst-case analysis and uncover a non-linear relationship between the approximation ratio and the summary size, which is much closer to the practical performance of column selection. Our theoretical analysis also reveals an interesting phenomenon: the approximation factor as a function of the number of chosen columns may exhibit multiple peaks and valleys, which we call a multiple-descent curve. A lower bound we establish shows that this behavior is not an artifact of our analysis, but rather it is an inherent property of the problem. As an ongoing work, we are working on expanding these results to more general cost functions.

Bridging adversarial training and generalization: My goal in this thread of research is to illuminate the underlying factors for empirical successes of adversarial training in neural networks. Neural networks are adversarially trained to make them more resistant to perturbation based attacks. While traditionally a larger margin has been associated with better generalization, in a recently presented poster at NeurIPS 2020, we design experiments to show that margin can fall short of explaining the robust generalization gap [8]. To remedy this, we generalize margin to the new concept of *boundary thickness*. Intuitively, boundary thickness measures the expected distance to travel along line segments between different classes across a decision boundary. We empirically show that boundary thickness correlates highly with robust generalization across many training schemes including regularization, different batch-sizes and several data augmentation schemes. On the theoretical side, we establish that maximizing boundary thickness during training is akin to the so-called mixup training. In mixup, training data is augmented using pseudo data points generated by convex combination of original data points. Using these observations, we introduce a new algorithm that uses noise-augmentation on mixup training to further increase boundary thickness, thereby combating vulnerability to various forms of adversarial attacks and out of distribution transforms.

It has been also recently observed that the features in the data can be partitioned into robust and non-robust ones. Adversarial training retains robust features, and removes non-robust features. Non-robust features are known to be responsible for standard generalization while the robust features help against perturbation-based attacks. Hence a trade-off between robust generalization and standard generalization is established. But can the robust features also aid in any other form of generalization? In a recent work [7], we discover some interesting side-effects of adversarial training. We demonstrate that robust models trained for defense against adversarial attacks on the ImageNet dataset generalize better on other image datasets for the task of classification, especially if only limited data is available for the new domain task. We also observe that adversarial training biases the learnt representations to retaining shapes, as opposed to textures, which impacts the transferability of the source models. Finally, through the lens of influence functions, we discover that transferred adversarially-trained models contain more human-identifiable semantic information, which explains – at least partly – why adversarially-trained models transfer better. As ongoing work, we are working towards a theoretical explanation of transferability of robust models and on using self-training to improve transferability.

Bridging interpretability in machine learning and human learning: My goal in this thread of research is to develop principled methodologies for understanding of model predictions independent or agnostic of the model used. This is especially important for critical applications (such as self-driving cars and medical diagnosis), as well as for judging societal impact of machine learning by unearthing unintentional biases in the learnt model. Interpretability can also help bring a human in the learning loop to refine or debug a model prediction. Studies of human reasoning have shown that the use of examples or prototypes is fundamental to the development of effective strategies for tactical decision-making. Example-based explanations are widely used in the effort to improve interpretability. However, in a paper accepted as an oral presentation to NeurIPS 2016, we posit that examples are not enough [6]. Relying only on examples to explain the models' behavior can lead over-generalization and misunderstanding. Thus, to maintain interpretability, it is important, along with prototypical examples, to deliver insights signifying the parts of the input space where prototypical examples do not provide good explanations. We call the data points that do not fit the model criticism samples. We use the Maximum Mean Discrepancy (MMD) statistic as a measure of similarity between points and potential prototypes, and efficiently select prototypes that maximize the statistic. The scalability follows from our theoretical analysis, where we show that under certain conditions, the MMD for prototype selection allows constant-factor approximation. While we are primarily concerned with prototype selection and criticism, we quantitatively evaluate the performance of MMD-critic as a nearest prototype classifier, and show that it achieves comparable performance to existing methods. We also present results from a human subject pilot study which shows that including the criticism together with prototypes is helpful for an end-task that requires the data-distributions to be well-explained.

In a follow up work, we present a novel way to select prototypes by generalizing influence functions [5]. Recently, influence functions were used to address the question "which training data point is most responsible for a given test prediction". We extend the use case to a *set* of training data points responsible for making a *set* of test predictions. For example, the classical influence functions can be used to address "which training data point is most important for classifying a given picture as that of a horse" but can not address "which set of training data points are responsible for distinguishing cats from horses in the test set". We make use of a greedy forward selection algorithm called Sequential Bayesian Quadrature on Fisher Kernels to address the second question, and show that we recover the influence function method as a special case of our framework. We present strong empirical performance of our method for real world use-cases of data compression, coreset construction and data cleaning. We also establish theoretical approximation guarantees for kernel herding on discrete spaces. More specifically, we show that under certain mild conditions, the cost function satisfies weak submodularity, and hence guarantees constant factor approximation using the greedy forward selection procedure. This connection also establishes a probabilistic interpretation of influence functions. As ongoing work, we are investigating the relationship of influence scores with leverage scores.

Ongoing and Future Works For future directions, I plan to expand my investigations into uncovering the interplays among data, model and the algorithm for explaining performance and interpretability. Motivated by my earlier works, my research agenda is to pursue the following threads of research.

Algorithmic generalization bounds: Traditional generalization bounds for neural networks are derived from stability or compression based arguments for the model using the parameters obtained after the training is complete, thereby ignoring the exploratory phase of the stochastic optimization. This is especially interesting in context of neural networks, where understanding generalization [11] and robust generalization [12] are still active research areas. My work on robust generalization provides empirical markers that correlate with robust generalization, but a well-rounded theoretical understanding is still lacking. An important conclusion we derived from my work was that these markers were heavily impacted by algorithmic choices, at a finer level than the visible impact on the overall error on any given neural network. Motivated by this insight, an interesting direction I am pursuing is to quantify the algorithmic generalization bounds for machine learning models. Recently, [15] have provided probabilistic bounds under the assumption that the path of the stochastic optimization can be modeled as a Feller process. This was accomplished by examining the fractal dimensions of sample paths of the continuous time stochastic process. My objective in this thread is to first extend the analysis to the case when the process is discrete time, which is non-trivial since a direct discrete analog of the continuous fractal dimension does not exist. The next steps would be to derive similar bounds for stochastic optimization algorithms using their tail behaviors using tools from optimization theory, and combine them with existing approaches to provide a unified view of generalization in machine learning.

Approximate influence functions: My vision for my research on interpretability is to encourage more mainstream adoption by developing faster and reliable tools. Influence functions require solving a linear system with the Hessian of the system for every query. This can get prohibitively expensive on scale. It is also not clear how the definition will change if the Hessian is not invertible. An alternative way would be to make use of variational calculus [13] to approximate the influence without making use of the Hessian. This would allow us to identify a set within which the solution lies, as opposed to the exact solution. For many practical problems, this may be enough and can be practically very useful since it will be much faster. This line of research will also have a broader impact on Hessian-free approximate stochastic second order optimization in the longer term. We have some preliminary results that establish a relationship between leverage scores and influence. Based on these, my goal is to explore the use of influence-based importance sampling algorithms and to characterize the information content of a model by differentiating learning and memorization [10].

Beyond worst-case analysis: It is common to characterize hardness of a problem using the lower bounds obtained from the hardest instance of the problem. While this is convenient, it can leave a big gap in characterizing the practical hardness of the problem especially if the most common instances are easier than the hardest instance. This is exacerbated when there is a reduction involved from one problem to another to make a statement about the hardness of the target domain. At the same time, it is often non-trivial to delineate the more commonly occurring instances without making distributional assumptions. There have been limited number of studies in machine learning that undertake finer-grained analyses to address this gap of worst-case vs practical performance [14]. In my work on beyond worst case analysis for column subset selection, we identified key properties of the data that characterize the hardness of the problem at a finer level, showing that the worst case instances are only corner cases. Motivated by this work, I would like to further investigate related problems in sparse optimization, to bridge the gap between theory and practice. The class of submodular and weakly submodular functions should benefit from such studies, because the reported bounds (e.g., the guarantee of $(1 - {}^{1}\!/_{e})$) for these functions are indeed worst-case bounds, and are not indicative of practical performance. The goal is to do finer and beyond worst case analysis for specific problem instances to identify the precise conditions under which the worst-case bounds hold as well as the conditions under which better theoretical guarantees that are more representative of practical performance can be obtained.

Selected Publications

[1] Michal Derezinski, Rajiv Khanna, and Michael W. Mahoney. Improved guarantees and a multiple-descent curve for column subset selection and the nystrom method. In *NeurIPS*, 2020.

[2] Ethan Elenberg, Rajiv Khanna, Alexandros Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *Annals of Statistics*, 2018.

[3] Rajiv Khanna, Ethan Elenberg, Alexandros Dimakis, Joydeep Ghosh, and Sahand Neghaban. On approximation guarantees for greedy low rank optimization. *ICML*, 2017.

[4] Rajiv Khanna, Ethan Elenberg, Alexandros Dimakis, Sahand Neghaban, and Joydeep Ghosh. Scalable greedy support selection via weak submodularity. *AISTATS*, 2017.

[5] Rajiv Khanna, Been Kim, Joydeep Ghosh, and Oluwasanmi Koyejo. Interpreting black box predictions using fisher kernels. In *AISTATS*. 2019.

[6] Rajiv Khanna*, Been Kim*, and Sanmi Koyejo*. Examples are not enough, learn to criticize! criticism for interpretability. In *NeurIPS*, 2016.

[7] Francisco Utrera, Evan Kravitz, N. Benjamin Erichson, Rajiv Khanna, and Michael W. Mahoney. Adversarially-trained deep nets transfer better. *CoRR*, abs/2007.0586, 2020.

[8] Yaoqing Yang, Rajiv Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph E. Gonzalez, Kannan Ramchandran, and Michael W. Mahoney. Boundary thickness and robustness in learning models. In *NeurIPS*, 2020.

Other References

[9] Ethan Elenberg, Alexandros G Dimakis, Moran Feldman, and Amin Karbasi. Streaming weak submodularity: Interpreting neural networks on the fly. In *NeurIPS*. 2017.

[10] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation, 2020.

[11] Yiding Jiang*, Behnam Neyshabur*, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *ICLR*, 2020.

[12] Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *CoRR*, abs/2002.11080, 2020.

[13] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, 1998.

[14] Tim Roughgarden. Beyond worst-case analysis. *Communications of the ACM*, 62(3):88–96, 2019.

[15] Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A. Erdogdu. Hausdorff dimension, stochastic differential equations, and generalization in neural networks. In *NeurIPS 2020*, 2020.