My primary research interests are in machine learning theory and algorithms. The common thread in my research is the study of modern paradigms of machine learning, involving both the development of mathematical models and the design of efficient algorithms.

Overview

I am particularly excited about gaining an understanding of learning paradigms that do not yet enjoy a solid theoretical foundation and designing efficient algorithms based on the insight gained. In recent years, Machine Learning has evolved into a broad discipline with enormous practical success in a wide range of fields, and has produced fundamental theories of learning processes. The primary theoretical advances have been for supervised learning problems, where a target function is estimated using only labeled data. However, there is often cheap and plentiful unlabeled data available in contemporary applications. There are opportunities to explore the utilization of unlabeled data, and how to incorporate them with labeled data in learning.

A significant part of my work aims to provide rigorous theoretical understanding of unsupervised learning, where one hopes to find hidden structures in unlabeled data. Clustering is the prototypical example of unsupervised learning. For this task, I provided a beyond worst-case analysis [4, 5], explaining why the clustering problem seems very hard from the perspective of computational complexity theory while clustering instances in practice can still be solved efficiently. I also developed a robust hierarchical clustering algorithm [7], which is proven to cluster accurately in many cases where the traditional hierarchical algorithms fail. For community detection, the recent extension of clustering onto network data, I proposed a theoretical model and presented an efficient algorithm that provably detects all the communities in the model [6].

Another component of my work is to design efficient algorithms that exploit unlabeled data to save labeled data and improve learning. While in principle incorporating unlabeled data can provide benefits over fully supervised learning, the more challenging algorithmic problems remain largely open. In [2], I studied learning disjunctions in this setting, and provided efficient algorithms with nearly optimal label complexity.

I am also enthusiastic about studying new learning paradigms that are not captured by existing theoretical frameworks. For example, in many applications the data is collected from different locations, which leads to a recent interest in Distributed Learning. I studied the problems of clustering and Principal Component Analysis (PCA) in this setting, and designed efficient algorithms with provable guarantees of the solution quality and the communication needed [3, 13].

Finally, I am interested in applying these techniques to problems in related areas. One approach we have successfully applied is influence maximization (or viral marketing) over social networks, where the goal is to recommend items (e.g., news and products) to a selected set of users such that the overall spread of the items is maximized. In [9], I provided a novel formulation of practical influence maximization as a submodular maximization over the intersection of a matroid and multiple knapsack constraints, and exploited the structure of our formulation to design an efficient algorithm which achieves approximation better than the known results in the literature.

Selected Research Directions

Clustering. Clustering is one of the most widely-used techniques in machine learning and data analysis. Generally, the goal of clustering is to partition the data points into several meaningful subsets (called clusters) given a distance function on the points. A common approach for clustering is to optimize a chosen objective function on pairwise distances between the points. However, most of the natural clustering objectives, including k-median, k-means and min-sum, are NP-hard to optimize. It is, therefore, unsurprising that many

of the clustering algorithms used in practice come with few guarantees. In joint work with Maria-Florina Balcan [4, 5], we seek to overcome the NP-hardness of the problem by exploiting some properties of practical clustering instances, thus providing justification for why such instances can be dealt with efficiently even if the general problem is NP-hard. We considered the model recently introduced by [8], which assumes that the instances have a perturbation resilience property that the optimal clustering is preserved under small multiplicative perturbations to the distances between the points. Our work pointed out that this property leads to efficient algorithms for k-median, k-means and min-sum, and thus can be incorporated into the clustering framework to better describe practical applications. Furthermore, we considered a more challenging and more general notion of perturbation resilience, where we allow the optimal clustering after perturbation to be close to the original. This relaxed assumption fits better into the practical scenarios, and our results showed that the instances under this assumption can indeed be dealt with efficiently. Additionally, we studied the inductive setting, where the algorithm is only run on a sample of the entire dataset. This setting naturally arises from increasingly critical large scale applications, such as Internet scale image clustering, topic discovery in web documents, and so on. The analysis provided supportive evidence for the approach of first clustering a random sample, then inserting the remaining data into the produced clusters based on some simple computations.

Another widely used approach for clustering is hierarchical clustering, which has long been used across many different fields as a basic tool. Unfortunately, it is well known that many classic algorithms are not robust to noise. In joint work with Maria-Florina Balcan and Pramod Gupta [7], we proposed a new robust hierarchical clustering algorithm, which can be used to cluster accurately in cases where the data satisfies a number of natural properties and where the traditional hierarchical algorithms perform poorly. Such properties allow the presence of a substantial degree of noise, including corrupted data and points close to the boundaries between clusters.

In the context of social networks, clustering is generalized to the problem of identifying a collection of communities based on the affinity between the members. In joint work with Maria-Florina Balcan [6], we proposed a theoretical model that explicitly formalizes the active interaction and the hierarchical nature of the communities observed in practice. Given this formalization, we proposed an efficient algorithm that detects all the communities in this model, and proved that all the communities form a tree hierarchy.

Efficient Semi-supervised and Active Learning. In many modern applications, unlabeled data is abundant but labeling it is expensive. As a consequence, Semi-Supervised Learning (using large amounts of unlabeled data to augment limited labeled data) and Active Learning (where the algorithm itself asks for labels of carefully chosen examples) have become important areas of machine learning. Much of the theoretical work has focused either on sample complexity or on providing polynomial time algorithms with error bounds for surrogate losses only. In joint work with Maria-Florina Balcan, Christopher Berlind, and Steven Ehrlich [2], we provided efficient algorithms with nearly optimal label complexity for semi-supervised and active learning of disjunctions under a natural regularity assumption. This work provided a concrete example of a polynomial time algorithm with theoretical guarantees on the sample complexity and the classification error.

Distributed Learning. Most classic learning algorithms are designed for the centralized setting, but in recent years data is often distributed over different locations, and thus it has become crucial to develop effective algorithms in the distributed setting. In joint work with Maria-Florina Balcan and Steven Ehrlich [3], we studied the problem of distributed clustering and provided algorithms with small communication cost and provable guarantees on the clustering quality. Our technique is based on the construction of a small set of points called coreset [11], which acts as a proxy for the entire dataset. We showed that each node can construct a local portion of a global coreset and then share these local portions to compute the final

solution, resulting in communication cost that is independent of the number of points in the global data.
Most recently, in joint work with David Woodruff, Maria-Florina Balcan, and Vandana Kanchanapally [13], we proposed a distributed algorithm for Principal Component Analysis (PCA). It has state of the art communication cost for computing low rank approximations. It can also be used as a pre-processing step for k-means clustering, non-negative matrix factorization, Latent Dirichlet Allocation and many related problems, such that any good approximation solution on the projected data is also a good approximation on the original data, while the projected dimension required is independent of the original dimension. This approach provides a general way to reduce communication requirements for all these problems. For example, when combined with our distributed clustering algorithm, this approach leads to an algorithm with communication cost independent of the number of data points and linear in their dimension.

Learning and Maximizing Influence over Networks. The rising role of social networks in the spread of information leads to a substantial interest in viral marketing. In this problem, a set of source nodes are selected to initiate the information diffusion process for an item such that the influence, the expected number of follow-ups, is maximized. The first difficulty is to learn the influence directly from the information diffusion traces (or cascades). In joint work with Nan Du, Maria-Florina Balcan, and Le Song [10], we exploited the insight that the influence functions in many diffusion models are coverage functions, and proposed to learn them by convex combinations of random basis functions. We designed an efficient maximum likelihood based algorithm which can provably learn the influence function with low sample complexity.
Based on the learned influence function, we study the influence maximization problem for multiple items from the perspective of the owner of an online social platform. Previous studies have not sufficiently addressed the following important practical aspects: entities can only be selected as source nodes for a small number of times, and influencing any entity has a cost while advertising is limited to given budgets. In [9], we provided a novel formulation as a submodular maximization under the intersection of a matroid and multiple knapsack constraints. The special structure of our formulation allows us to design an algorithm with approximation better than known guarantees in the combinatorial optimization literature. Extensive synthetic and real world experiments demonstrated that our approach achieves the-state-of-the-art in terms of both effectiveness and scalability, often beating the next best by significant margins.

Future Directions

Communication Complexity of Distributed Learning. Due to the explosion of distributed data, different Distributed Learning algorithms and platforms have been developed and numerous successful experimental results have been reported. However, our understanding of the communication complexity is still limited; essentially the existing theoretical results use very strong models of communication which are not particularly true to applications or are loose compared to the known upper bounds. It is an exciting challenge to develop a better understanding of the communication complexity in a more general setting.

Deep Learning. The success of Deep Learning raises fascinating and challenging questions for theoretical research on learning [12]. However, a main criticism of deep learning concerns the lack of theory surrounding many of the algorithms. Recent work [1, 14] provided theoretical guarantees for specific models of deep neural nets, although the assumptions in the models are generally not satisfied in those nets with impressive practical success. The existing work opens the door to systematic understanding of deep neural nets and the derivation of principled deep learning algorithms, with many possible avenues for future work.

Final Thoughts

I believe that Machine Learning will continue to be one of the most important aspects in a computational world. More challenges will emerge as machine learning applications are used in a dynamic, interactive environment and face rapidly growing large-scale heterogeneous datasets. The insights from Machine Learning theory are thus crucial for designing effective and efficient algorithms, and understanding the inherent limitations of learning. I am excited to contribute to the advancement of this fascinating field.

References

[1] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. *arXiv preprint arXiv:1310.6343*, 2013.

[2] Maria-Florina Balcan, Christopher Berlind, Steven Ehrlich, and Yingyu Liang. Efficient semi-supervised and active learning of disjunctions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[3] Maria-Florina Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k-means and k-median clustering on general communication topologies. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[4] Maria-Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. In *Proceedings of the International Colloquium on Automata, Languages, and Programming (ICALP)*, 2012.

[5] Maria-Florina Balcan and Yingyu Liang. Clustering perturbation resilient k-median instances. In *the Learning Faster from Easy Data Workshop in Advances in Neural Information Processing Systems*, 2013.

[6] Maria-Florina Balcan and Yingyu Liang. Modeling and detecting community hierarchies. In *Proceedings of the International Workshop on Similarity-Based Pattern Analysis and Recognition (SIMBAD)*, 2013.

[7] Maria-Florina Balcan, Yingyu Liang, and Pramod Gupta. Robust hierarchical clustering. *arXiv preprint arXiv:1401.0247*, 2013.

[8] Yonatan Bilu and Nathan Linial. Are stable instances easy? In *Proceedings of Innovations in Computer Science (ICS)*, 2010.

[9] Nan Du, Yingyu Liang, Maria-Florina Balcan, and Le Song. Budgeted influence maximization for multiple products. *Submitted for publication.*

[10] Nan Du, Yingyu Liang, Maria-Florina Balcan, and Le Song. Influence function learning in information diffusion networks. *Submitted for publication.*

[11] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, 2004.

[12] Nicola Jones. Computer science: The learning machines. *Nature*, 505, 2014.

[13] Yingyu Liang, David Woodruff, Maria-Florina Balcan, and Vandana Kanchanapally. Fast and communication efficient algorithms for distributed pca. *Submitted for publication.*

[14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. A provably efficient algorithm for training deep networks. *arXiv preprint arXiv:1304.7045*, 2013.