Research Statement

Baharan Mirzasoleiman
Stanford University

My research focuses on developing new methods that enable efficient learning from massive datasets. More specifically, I am interested in designing techniques that can gain insights from the underlying data structure by utilizing complex and higher-order interactions between data points. The extracted information can be used to efficiently explore and robustly learn from datasets that are too large to be dealt with by traditional approaches. My methods have immediate application to high-impact problems where massive data volumes prohibit efficient learning and inference, such as huge image collections, recommender systems, Web and social services, video and other large data streams.

While datasets are steadily growing, their information content is much smaller than their actual volume, due to the built-in redundancy. Existing techniques are not effective in identifying and extracting the information volume—the non-redundant information content—from sheer data volumes. Hence, machine learning models are trained on massive data volumes. This is contingent on exceptionally large and expensive computational resources, and incurs a substantial environmental cost due to the significant energy consumption. A grand challenge of machine learning today is to develop methods that can extract the information volume, and accurately and robustly learn from the extracted information content.

My research approach to uniquely address this challenge consists of: (1) extracting the information volume by summarizing the most representative subsets; and (2) developing data-efficient and robust optimization methods to learn from summaries (Figure 1). At the core of my research lie theoretically rigorous techniques that provide strong guarantees for the quality of the extracted summaries, and the learned models’ accuracy and robustness against noise. My techniques open up new avenues for learning from massive datasets.

My research has made significant contributions to machine learning, and has been published in premier machine learning and data mining venues (NeurIPS’13, KDD’14, NeurIPS’15, AAAI’15, NeurIPS’16, ICML’16a, ICML’16b, JMLR’16, ICML’17, AISTATS’17, AAAI’18, NeurIPS’18). I received an outstanding doctoral thesis award from ETH Zurich, and was selected as a Rising Star in EECS by MIT.

My research broke new ground by developing truly scalable methods to extract the information volume for machine learning. For example, my techniques enabled me to tackle the largest exemplar clustering problem considered in the literature on 80 million images, and to extract real-time video summaries 1,700× faster than existing methods. My research has had a substantial real-world impact, which has garnered interests of academic, and industry researchers. For example, my large-scale summarization algorithms have been utilized in Google and YouTube, and are now a core data science platform of several startups (e.g., Summary Analytics).

(1) Extracting the Information Volume from Big Data

Extracting the information volume by summarizing the most representative subsets is generally an NP-hard problem. However, many natural notions of representativeness satisfy submodularity, an intuitive diminishing returns property: Selecting new data points help less if similar data points have already been selected. Thus, summarization problems can often be reduced to maximizing a submodular set function. The classical greedy algorithm produces solutions competitive with the optimal (intractable) solution for submodular maximization. However, the greedy algorithm is inherently sequential in nature, and sequentially selecting elements on a single machine
is heavily constrained in terms of speed and memory. As a result, the greedy algorithm is highly impractical for massive datasets. My research pioneered submodular maximization at scale by developing practical distributed and streaming methods with strong theoretical guarantees.

Identifying Representative Elements in Massive Datasets

Data volumes are growing faster than the ability of individual computers to process them. Distributed and parallel processing is therefore necessary to keep up with modern massive datasets. The classical algorithms that work well for centralized submodular maximization are poorly suited for parallel architectures, due to their sequential nature. This mismatch makes it inefficient to apply classical algorithms directly to distributed setups.

To address this challenge, I developed the first distributed framework for submodular maximization that scales to truly massive datasets [NeurIPS'13]. My distributed framework runs in a constant number of rounds. It hence requires minimal communication and can be easily implemented in MapReduce-style parallel computation models. I showed that under some natural conditions, for large datasets the quality of the obtained solution is provably competitive with the best centralized solution. I further extended this framework to obtain approximation algorithms for (not-necessarily monotone) submodular maximization subject to more general types of constraints [JMLR'16]. Constraints such as matroid and knapsack allow for customizing the summaries. Continuing this line of research, I made several further pivotal algorithmic contributions to large-scale submodular maximization. These include the first randomized algorithm that reduces the cost of selecting $k$ elements out of $n$ to $O(n)$ [AAAI'15], compared to $O(nk)$ for classical algorithms; a fast algorithm for handling constraints [ICML'16a], compared to prior work that were prohibitive even for medium-sized datasets; and the first algorithm for distributed submodular coverage [NeurIPS'15, NeurIPS'16].

My distributed algorithms implemented on a Spark compute cluster allowed me to tackle the largest submodular maximization problems considered in the literature. These include exemplar based clustering on 80 million images, active-set selection on more than 45 million user visits from Yahoo! Front Page, and finding influential nodes in Friendster social network with more than 65 million nodes and 1.8 billion edges. My large-scale summarization algorithms have been utilized in Google and YouTube, and several startups (e.g. Summary Analytics). I have given invited talks about this research at MIT, Harvard, and UIUC seminar series on distributed algorithms. My research has also been discussed in tutorials in ISIT, CVPR, and in lectures at the Simons Institute’s Foundations of Machine Learning program.

Massive Data Summarization “on the Fly”

The need for real-time analysis of rapidly-producing data streams (e.g., video and image streams) motivated the design of streaming algorithms that can efficiently extract and summarize the information volume from massive data on the fly. Streaming methods enable predictions in a timely manner based on the data seen so far, facilitating real-time analytics. Existing methods for streaming submodular maximization either made strong assumptions about the way the data stream is generated, or required multiple passes over the data and had a prohibitive update time (computational cost to process one data point). They were, therefore, impractical for online analysis.

To address this problem, I proposed the first single-pass streaming algorithm for monotone submodular maximization [KDD’14]. My method guarantees a constant factor $1/2 - \epsilon$ approximation to the optimum solution, requires memory independent of data size, and has a small update time. Hence, it permits real-time analysis of fast data streams. I further extended my approach to maximize a (not-necessarily monotone) submodular function under various important constraint to consider privacy or personalization [AAAI'18]. In addition, motivated by privacy concerns of analyzing personal information in online services, I proposed a novel formulation for privacy preserving data summarization. My developed deletion-robust streaming framework [ICML’17] efficiently maintains an accurate summary, yet allows individuals to have their data deleted at any time, thus preserving their “right to be forgotten”.

Through extensive experiments, I demonstrated that my streaming methods run more than $100 \times$ faster for training large-scale kernel methods for click-through prediction on 45 million feature vectors, and more than $1,700 \times$ faster for video summarization, compared to previous work. For click-through prediction, my deletion-robust algorithm can almost recover the performance of the classifier trained on the full dataset, even after 99% deletion. I have given invited talks about this research at several industry labs such as Google Research, and Xerox Research. My research has also been discussed in tutorials in CVPR, and in lectures at the Harvard Data Science Initiative.
(2) Data-efficient and Robust Learning from the Information Volume

Stochastic gradient methods are widely used to learn from massive data volumes. These methods, however, are slow to converge and tend to overfit to noisy labels prevalent in big datasets. There have been sustained efforts to develop data-efficient methods for training machine learning models. Nevertheless, it had remained a challenging open question how to develop a practical and theoretically sound method to find a subset that can generalize on par with the full data when trained on. My research addressed this challenge by extracting summaries that are representative for training machine learning models, and developing theoretically rigorous optimization techniques to effectively and robustly learn from the extracted summaries.

Data-efficient Learning from Summaries

Training machine learning models often reduces to the problem of minimizing a regularized empirical risk function. Stochastic gradient descent can find the minimizer of this problem, but has slow convergence asymptotically due to the inherent variance of the random gradient estimates. The majority of the work on speeding up stochastic gradient descent has primarily focused on reducing this variance, or carefully selecting stepsizes. Nevertheless, modern machine learning models require days to weeks to train on big datasets.

My research addressed this challenge by developing the first data-efficient method with theoretical guarantee for training machine learning models. I proposed a novel formulation for extracting representative subsets of training data that closely approximate the full gradient. I proved that the optimal subset is simply medoids of the dataset in the gradient space weighted by their cluster size. Medoids can be efficiently extracted from massive datasets using my fast and scalable submodular maximization methods [NeurIPS'15, NeurIPS'16]. I showed that training on the set of medoids with per-element stepsizes equal to the weight of every medoid is guaranteed to (approximately) converge to the globally optimal solution (i.e., the model parameters that would be obtained if training/optimizing on the full dataset). At the same time, training on the subset of $k$ medoids directly leads to a speedup that is inversely proportional to $k$. I further showed that processing medoids in the order provided by submodular maximization algorithms can further speed up training [ArXiv'19a].

In an extensive set of experiments, I demonstrated that my method speeds up various stochastic gradient methods, including SGD, SAGA, SVRG, Adam, Adagrad, and NAG by up to $10 \times$ for convex and $3 \times$ for non-convex (training deep networks) loss functions, while achieving better generalization performance. I have given invited talks about this recent result at University of Michigan Data Science Consortium, and Microsoft Research AI Breakthrough workshop. My research has also been discussed in invited talks in ICML.

Robust Deep Learning from Summaries

Over-parametrized neural networks—where the parameters of the model far exceed the size of the data volume—trained with first order methods have the capacity to (over)fit any set of labels including pure noise. To prevent overfitting to noisy labels, existing methods work by either reweighting training examples, designing robust loss functions, or using explicit regularization techniques. However, these techniques are either challenging to implement or cannot achieve optimal performance. In addition, existing results are very limited in providing theoretical guarantee for the performance of the trained neural networks with noisy labels.

In an ongoing work, I developed a robust and data-efficient framework for training deep networks with noisy labels that provides rigorous theoretical guarantees while enjoying a superior empirical performance. Specifically, I proved that stochastic gradient methods applied to summaries obtained in [ArXiv'19a] are robust to noisy labels for over-parametrized neural networks. Summaries provide a low-rank approximation of the neural network Jacobian matrix, and thus stochastic gradient methods applied to summaries cannot overfit to noise despite over-parameterization. Training on summaries prevents the parameters to get too far away from initialization and eliminates the need for early stopping. Extensive numerical experiments corroborated on my theory and verified that deep networks trained on the summaries achieve state-of-the-art accuracy in presence of up to 50% label noise.

My research on data-efficient and robust machine learning proves the great potential for developing novel optimization methods that can benefit from the underlying data structure and higher order interactions between data points. Such methods can potentially revolutionize the way machine learning models are trained on massive data, and enable learning accurate and robust general-purpose models on truly massive datasets.
Future Research Plans

My research vision is to develop the next generation of resource-efficient and reliable techniques to learn from massive datasets. To this end, I plan to focus on the following challenges: designing efficient methods to learn from multimodal data streams, where information comes as different modalities such as images, text, sensory data, etc.; developing resource-efficient methods to reduce substantial costs of learning from massive datasets; designing reliable and safe learning algorithms with rigorous guarantees for safety-critical systems; and providing generalization guarantee for the performance of deep neural networks trained on big datasets.

I am in particular excited about applications of my techniques in domains such as medical and health-care, Internet of Things (IoT), self-driving cars, financial services, and urban planning, where a tremendous amount of data is generated every second, and demands fast and accurate analysis.

Multimodal Learning from Massive Streams. In many domains, such as Internet of Things (IoT), stock exchange, computational social science and health-care, rapid streams of data are generated from various sources. For example, a typical self driving car equipped with radar, cameras, lidar, and ultrasonic sensors, produce more than 4TB of data per day. Similarly, various sensors and smart devices in IoT applications generate a huge amount of data at a high velocity. The challenge in making use of such data is to design algorithms that can efficiently extract and fuse data from various sources to learn and make inference on the fly. My current research has already addressed the challenge of extracting online summaries from a rapidly-producing data stream. Moving forward, I intend to focus on efficient learning from summaries of multimodal large data streams.

Resource-efficient Learning from Big Data. Training modern machine learning models on massive datasets incurs a substantial financial and environmental cost. My recent research focused on extracting representative subsets of data to reduce these costs and accelerate training of machine learning models. I believe this direction poses great opportunities for further scholarly study. For example, in a collaboration with Stanford DAWN project, we showed that by removing hidden layers from deep target networks, we can create proxies that are an order of magnitude faster to train, and can identify which points to select nearly as well as the larger target model but up to 50× faster [ArXiv'19b]. I would like to continue this line of research to develop a new generation of resource-efficient and fast techniques for learning from massive datasets.

Reliable and Safe Machine Learning. The functional safety of many intelligent systems, such as autonomous robots and self-driving cars, remains largely dependent on the robustness of the underlying machine learning model. These systems are expected to operate flawlessly, and hence need to be trained on examples of all possible situational conditions. My current research demonstrated that training models on representative summaries highly improves the robustness of the learned models against noisy labels, both in theory and practice. One concrete direction I plan to pursue along these lines is to build scalable robust frameworks for safety critical systems with very high accuracy requirements. In particular, I am interested in developing optimization methods that can provide guarantees for the quality of inference under noisy data, noisy labels, and adversarial attacks.

Deep Learning with Generalization Guarantee. Despite the great empirical success of deep models, we have very little understanding of the way such models work. The neural network architectures often contain more parameters than size of the training data, and can interpolate any set of (random) labels. Nevertheless, over-parametrized neural networks generalize well to real datasets. My research on robust deep learning against noisy labels made initial steps towards better understanding of generalization in over-parametrized neural networks. In particular, I showed that training on summaries provides a better generalization performance by avoiding overfitting. For the future, I plan to continue this line of research and provide generalization guarantee on the true data distribution for deep neural networks.

I strongly believe that the above research directions can advance the current state of machine learning research, and will have tremendous real-world influence. The high impact of these problems and my grant writing experience will enable me to obtain funding for my research through a variety of sources including NSF, DARPA, IARPA, and industrial organizations. I am confident that my research and collaborations with experts in various fields, including data science, machine learning, statistics, mathematics, and theoretical computer science have equipped me with the necessary background to approach the above challenging and impactful research directions.
References


*Authors appear in alphabetical order.