Research Statement

Sudeep Salgia

I have been working on sequential learning problems arising in stochastic optimization, distributed learning, and active learning. My research focuses on establishing fundamental limits on feasible performance and developing machine learning algorithms that achieve or approach the performance limits under practical constraints in terms of computation complexity and communication costs.

Learning efficiency (i.e., how fast—in terms of the number of data points used in learning—the decisions converge to the optimal point) is the key performance metric for learning algorithms. Many prevailing algorithms achieve the optimal learning efficiency by assuming certain prior knowledge on the learning task at hand (e.g., the smoothness of the underlying unknown loss function or the probabilistic model of the noise). However, acquiring such auxiliary knowledge can be difficult in practice. It is highly desirable to have learning algorithms that *adapt* to unknown parameters: to automatically offer the optimal learning efficiency dictated by the underlying unknown problem parameters without assuming any prior knowledge.

In addition to learning efficiency, computation and communication costs are equally important performance measures, especially in distributed settings involving a large number of local devices with limited computation and communication power. There lack systematic studies on the fundamental tradeoff between learning efficiency and computation-communication costs as well as learning algorithms that operate at the optimal tradeoff.

My research has focused on addressing adaptive learning and learning under practical constraints on computation and communication costs. I present below a summary of my research contributions.

1 Stochastic Optimization

Stochastic Optimization aims at minimizing the expectation of a random loss function for which probabilistic model is unknown. The archetypal statistical learning problem of classification based on random instances is a stochastic optimization problem, where the decision variable is the classifier and the randomness corresponds to the unknown probabilistic model relating the feature vector and the hidden label of the chosen instance. With the objective function unknown, the decision maker can only take a trial-and-error learning approach by choosing, sequentially in time, a sequence of query points with the hope that the decisions improve over time. Without any structural assumptions on the unknown function, any hope to solve the problem is completely forgone. There are two commonly adopted structural assumptions in the literature. The first class of studies considers convex objective functions. The second class allows for more general, non-convex functions and assumes that the objective function belongs to a Reproducing Kernel Hilbert Space (RKHS) associated with a known kernel. Together with my collaborators, I have worked on developing computationally efficient and adaptive algorithms for optimizing functions satisfying the different structural assumptions described above.

Stochastic Convex Optimization: Stochastic convex optimization (SCO) was pioneered by Robbins and Monro in 1951 [6], who studied the problem of approximating the root of a monotone function based on successive observations of noisy function values at chosen query points. Its equivalence to the minimization of

an unknown convex function is immediate when monotone function is viewed as the gradient of the objective function. The Stochastic Gradient Descent (SGD) approach developed by Robbins and Monro [6] has long become a classic and continues to be widely used in ML applications till date. SCO continues to be an active area of research and numerous variants of the classical SGD have been proposed since.

Despite their popularity, SGD and its variants have their own share of drawbacks. They require a manual control of the learning rate that requires prior knowledge of the underlying objective function such as strong convexity parameters and smoothness. Additionally, they also incur a high cost while computing full gradients in large-scale high-dimensional problems and resist natural extensions to parallel implementation. These shortcomings have prompted the search for alternative approaches that enjoy better adaptivity, scalability and parallelizability. This has been the driving force of my work in SCO.

I am a key contributor to a stochastic optimization algorithm that is robust, has no tuning parameters and self-adapts to the unknown function characteristics while matching the same optimal learning efficiency of SGD.

In addition to adaptivity, I also addressed the issue of parallelizability through our novel framework called Progressive Coordinate Minimization (PCM) [11]. PCM is rooted in the methodology of decomposing a high-dimensional optimization problem into a sequence of low-dimensional problems which can be implemented in parallel. PCM, when used in conjunction with an arbitrary low-dimensional optimization routine, provides a general, routine-agnostic framework for a scalable, parallelizable and computationally-efficient high-dimensional extension of the low-dimensional routine, that retains its desirable characteristics. The extension of SGD within the PCM framework leads to a marriage between the efficiency of SGD with the scalability and parallelizability offered by PCM.

Kernel-based Optimization: In contrast to convexity, the RKHS-based modelling endows a structural regularity on the unknown function through a known kernel. The high expressive power of kernel-based modelling makes it a useful tool to model blackbox functions, especially those which are expensive to evaluate and can be accessed only through their values at the queried points. Consequently, kernel-based optimization is often studied in the context of maximizing a blackbox function with applications in tuning hyperparameters in experiment design, optimising control strategies in complex systems, neural networks, and scientific simulation based studies [5].

The higher flexibility and expressivity of RKHS-based modelling, however, makes kernel-based optimization significantly more challenging than its convex counterpart. Popular existing approaches either include heuristics with little to no theoretical guarantees or are based on the popular Upper Confidence Bound strategy [1, 12] from bandit literature and suffer from sub-optimal performance guarantees. Even in the simpler noiseless scenario, achievable learning efficiency continues to be elusive and was posted as a COLT 2022 open problem [13]. Additionally, all existing approaches involve computationally intensive subroutines making them practically infeasible even in low-dimensional problems.

My work in kernel-based optimization has been towards systematically addressing these large gaps in our fundamental understanding in this challenging setup. For the noisy setup, I proposed the *first* algorithm that achieves the optimal performance guarantees in the kernel-based setting, closing the gap between the lower and the upper bounds [7]. The proposed algorithm was also significantly more *computationally efficient* than existing algorithms, thereby improving upon the state of the art simultaneously on both the theoretical as well as the practical fronts.

2 Distributed Learning

Distributed learning refers to the strategy of employing several networked entities, called clients, to collaborate together to achieve a common learning goal. Recent years have witnessed distributed learning become the de facto approach in ML applications. One of the major factors that has led to this transformation is the ever-increasing amount of data that is being collected that necessitates the storage across several devices for easier retrieval and management. Another factor that has contributed towards this change is the emergence of applications where the data needs to be physically separated for privacy and security reasons, e.g., Federated Learning. As data gets more decentralized, learning algorithms need to incorporate a collaborative effort from the various devices, leading to the adoption of distributed learning strategies.

Distributed Learning brings forth a new set of challenges that are not encountered in the traditional centralized learning setups. Arguably the major challenge in design of distributed learning algorithms is the control of the communication overhead. Since multiple devices participate in the learning process, communication is essential for collaboration. This leads to an accuracy-communication trade-off as better accuracy is facilitated by more information exchange resulting is greater communication overhead. While there has been a considerable effort towards designing communication-efficient algorithms [14], this accuracy-communication trade-off is not well-understood at the fundamental level. Another challenge, which is more typical in Federated Learning (FL) [4], is the heterogeneity of data distribution across devices. In FL, typically the participating devices are single-user devices, e.g. cellphones and hence have private data drawn from user-dependent distribution. While the availability of more data helps in the learning task at a macro-level, the challenge arises at the micro-level where a single global model may not be able to cater to a variety of user-dependent target distributions.

Considering linear bandits as a representative problem, I undertook the *first* study [10] to rigorously analyze the accuracy-communication trade-off frontier at an information-theoretic level. The result establishes the minimum number of bits that need to be communicated to achieve the optimal performance and proposes a new algorithm based on progressive learning and sharing that achieves the same performance both in terms of *accuracy and communication*. This tight characterization of the trade-off frontier provides a fundamental insight for developing communication-efficient distributed algorithms and the necessary machinery to extend these ideas to other sequential decision problems. Building upon this work, I also proposed a novel algorithm for distributed stochastic convex optimization that achieves optimal performance, incurs the same *low communication cost*, and is *adaptive*. To understand the impact of heterogeneity, my collaborators and I analysed a distributed kernel-based optimization problem with each client having a personalized objective function. Our proposed framework [8] is based on the fundamental insight that designing algorithms for statistically heterogeneous clients requires striking a balance between data sharing and targeted local performance. It involves each client taking altruistic actions to help others learn from the information they have (collaborative exploration) while balancing it with maximizing their own reward (personalized exploitation). The proposed algorithm incurs a *low communication cost* while simultaneously achieving near-optimal performance, as corroborated by our matching lower bounds.

3 Active Learning

Active learning, referred to as the optimal design of experiments in the statistics community, is a learning strategy that involves the decision maker *actively* choosing which data points to use or which experiments to carry out in order to maximize the relevant information for the learning task while minimizing the number of data points or experiments. Active learning approaches are particularly important in this era of Big Data when the over abundance of data may well become an obstacle to gathering meaningful information.

There has been a considerable effort towards developing statistical theories in the offline setting, where all the data is available prior to the deployment of the learning algorithm. However, in practice, the problem of active learning is often sequential in nature. As more data is collected, more informative points become available that aid in improving the accuracy as well as in designing query strategies for future points. Despite its pervasiveness, the online/sequential setting has received lesser attention than its offline counterpart. To better understand this less explored setting, I worked on designing active learning strategies for online classification along with my collaborators. We proposed a novel active learning algorithm that makes no more than a fixed number of mistakes over the best classifier, independent of the learning horizon [3]. Moreover, this classifier is learnt with a diminishing fraction of the data demonstrating the efficacy of active learning. Such strategies are particularly useful for applications like spam detection and event detection in real-time surveillance that are characterized by high-volume streaming of instances and nuanced definition of labels.

While classification tasks represent the most general use case of active learning, group testing is a widely-studied special case of active learning. The objective in group testing is to quickly and reliably identify the anomalous subset from a given set of items by judiciously choosing which subsets to test at each time. Group testing finds applications in cyber security, biology and medicine and was widely used during the recent COVID-19 pandemic to improve the efficiency of testing [2]. My work in group testing has been towards designing practically efficient strategies by relaxing the assumption of requiring an exact probabilistic model of noise and the restriction to very specific observation models, both of which limit the practical feasibility of existing results. To this effect, I proposed a novel group testing algorithm [9] that works for general observation models, is agnostic to the noise distribution, *adapts* to the signal to noise ratio and achieves order-optimal detection delay with respect to the noise level and detection accuracy, thus making it a theoretically optimal as well as a practically viable alternative to existing strategies.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, may 2002.

[2] J. N. Eberhardt, N. P. Breuckmann, and C. S. Eberhardt. Multi-Stage Group Testing Improves Efficiency of Large-Scale COVID-19 Screening. *Journal of Clinical Virology*, 128:104382, Jul 2020.

[3] B. Huang, S. Salgia, and Q. Zhao. Disagreement-Based Active Learning in Online Settings. *IEEE Transactions on Signal Processing*, 70:1947–1958, 2022.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[5] J. Mockus, V. Tiesis, and A. Zilinskas. *The application of Bayesian methods for seeking the extremum*, volume 2, pages 117–129. 09 2014.

[6] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Statistics*, 22(3):400–407, 1951.

[7] S. Salgia, S. Vakili, and Q. Zhao. A Domain-Shrinking based Bayesian Optimization Algorithm with Order-Optimal Regret Performance. Oct 2020.

[8] S. Salgia, S. Vakili, and Q. Zhao. Kernel-based federated learning with personalization. *arXiv preprint arXiv:2207.07948*, 2022.

[9] S. Salgia and Q. Zhao. An order-optimal adaptive test plan for noisy group testing under unknown noise models. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4035–4039, 2021.

[10] S. Salgia and Q. Zhao. Distributed linear bandits under communication constraints. *arXiv preprint arXiv:2211.02212*, 2022.

[11] S. Salgia, Q. Zhao, and S. Vakili. Stochastic coordinate minimization with progressive precision for stochastic convex optimization. In *37th International Conference on Machine Learning, ICML 2020*, pages 8396–8406, Mar 2020.

[12] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pages 1015–1022, 2010.

[13] S. Vakili, V. Picheny, and N. Durrande. Regret bounds for noise-free bayesian optimization. *arXiv preprint arXiv:2002.05096*, 2020.

[14] Z. Zhao, Y. Mao, Y. Liu, L. Song, Y. Ouyang, X. Chen, and W. Ding. Towards efficient communications in federated learning: A contemporary survey. *arXiv preprint arXiv:2208.01200*, 2022.