Research Statement

Akshay Balsubramani

My research has centered around machine learning, which has exploded in the past two decades. The field has addressed cornerstone problems that recur in myriad application areas, and solved them with general, useful algorithms that are ultimately state-of-the-art across many of these areas given enough data, using statistics and other mathematics. I have been motivated during my PhD by this remarkable data-driven combination of generality and practicality, working to devise new algorithms that exploit data structure in practical learning scenarios.

I have focused on two specific problem areas which share the above characteristics, answering the following questions over the course of my PhD:

1. **Semi-Supervised Learning with Ensembles of Predictors**: How can we devise an efficient, practical, and interpretable algorithm that uses large quantities of unlabeled data to best put together the predictions of a collection of classifiers of varying competence?

2. **Sequential Algorithms and Stopping**: How can we robustly adapt traditional statistical hypothesis tests to report accurate results when the sample size is unknown? How can we use as few samples as possible to detect an effect of unknown magnitude?

These areas have interested me because they recur frequently in common practical problems, and yet are still new enough to researchers to present many basic fresh challenges. I now discuss my work in each of them in the first two sections. In the third section, I discuss my future interests, which involve applications of such algorithms to interdisciplinary research.

**Combining Ensembles for Semi-Supervised Learning**

In the fundamental learning problem of (binary) classification, each datum has one of two labels; the learner is required to predict those of some unlabeled data, given a set of labeled data. Many widely-used methods exist for this problem, like linear predictors, decision trees, Bayesian classifiers, and deep neural nets. A typical workflow might involve using the labeled data to train several different such algorithms and estimate their error rates, and finally choosing the best of the algorithms with which to predict on the unlabeled data. Justifying this *empirical risk minimization* procedure is part of the bedrock of classical learning theory ([Vapnik, 1982]). However, practical cutting-edge usage very frequently involves aspects not addressed in this account, which I have focused on in my work.

a) First and foremost, reliably labeled data are expensive to obtain in many applications in medicine, NLP, and other areas – the phrase "big data" in classification often solely signifies abundant unlabeled data. Yet classification algorithms are typically designed to take advantage of only labeled data. This is the motivation for *semi-supervised learning*, which seeks scalable and practical methods that incorporate unlabeled data in a generic manner.

b) Second, practitioners often reasonably attempt to predict more accurately than the best single classifier by combining the algorithms together, often through some type of weighted majority vote; the results can be spectacular, and such *ensemble learning* methods routinely dominate open data mining competitions in practice, for instance winning the $1M Netflix Prize ([Lohr, 2009]) and the ImageNet challenge ([Russakovsky et al., 2015]) standard in deep learning.

My primary thesis work has been at the intersection of semi-supervised learning and ensemble learning. I have devised an algorithm ([Balsubramani and Freund, 2015b]) that uses unlabeled data scalably and generically to aggregate ensembles. The algorithm uses the ensemble classifiers’ estimated error rates, along with their predictions on the unlabeled data, to construct a predictor which is guaranteed to perform at
least as well as any single classifier. It adapts to the unlabeled data’s distribution without any assumptions at all on the structure or origin of the ensemble and its predictions; for instance, it can aggregate various heterogeneous methods like those mentioned above, and even handle *specialists* which do not attempt a prediction on some arbitrary subset of the data ([Balsubramani and Freund, 2015c](#)).

The new algorithm is as efficient as linear learning algorithms like logistic regression, and does not even have any parameters to tune. I prove that it produces the *best possible predictor*, given the ensemble and the unlabeled data – for instance, its guaranteed performance is as good as any weighted majority vote’s. Such a strong optimality guarantee has no precedent among semi-supervised learning algorithms.

The efficiency and simplicity of the method motivate practical usage. In [Balsubramani and Freund, 2015c](#), I directly aggregate an off-the-shelf random forest tree ensemble using unlabeled data, demonstrating an efficient and general algorithm that empirically performs significantly better than random forests on large datasets from recent data mining competitions. Performance is also superior to a range of other standard ensemble learning algorithms, particularly when there are relatively few labels and an abundance of unlabeled data.

These algorithmic advances are possible due to a framework and a set of algorithm design techniques I introduced in [Balsubramani and Freund, 2015b](#). I have continued to build upon this in recent work ([Balsubramani and Freund, 2015a](#)) which generalizes the algorithm to other loss functions representing different prediction tasks like predicting probabilities, while retaining all its aforementioned practical advantages. A large range of such situations can be treated with my techniques, leading in each case to efficient algorithms with strong optimality guarantees ([Balsubramani and Freund, 2015a](#)).

In the very near term, I will be completing ongoing projects extending the algorithms and theory, including to building classifiers which can abstain, and to ways of incrementally constructing the ensemble of classifiers. The goal of my theoretical and practical work is to realize the combined promise of semi-supervised ensemble learning:

1. **Semi-Supervised**: Adding unlabeled data should lead to better prediction.
2. **Ensemble**: The aggregated predictor should perform at least as well as any single classifier – the aggregation algorithm should be “safe” to use instead of the best classifier.

I am currently preparing the developed algorithms as an open-source code release (extending the code used in [Balsubramani and Freund, 2015c](#)), and will maintain the code indefinitely with further extensions to the core framework, which I have been devising as discussed above. In this task I expect to use the software engineering experience I have accrued during the PhD over summers at Google, Microsoft, and Outbrain, a content recommendation company.

**Sequential Algorithms and Stopping**

Sequential algorithms, which process data one at a time, are ubiquitous in learning and related fields, favored for their scalability, versatility, simplicity, and excellent theoretical properties ([Bousquet and Bottou, 2008](#)). Meanwhile, the sciences are rife with problems like *p-value peeking* (in which data are gathered until the results “look good”) and other stopping-related practices which make statistical tests unsound. I am interested in these topics, all related to variable sample size and stopping in learning.

I have explored the mathematics of “stopping times” to study such issues, in fresh results in submission to a prominent mathematics journal ([Balsubramani, 2015b](#)) that precisely quantify the natural amount of error unavoidable in these stopping problems, building on seminal 1920s work. I have since extended these ideas ([Balsubramani, 2015a](#)) in preparation) as part of an initiative to use them for practical learning problems.

I am also interested in *sequential hypothesis tests*, in which data are processed incrementally to calculate a test statistic until the algorithm decides to reject the null hypothesis. It has been widely known for decades that such tests use not many more samples than necessary, stopping earlier when the null hypothesis is easier to reject – but this had never been quantified theoretically. I have recently showed that basic well-known sequential tests have this desirable property in a very strong way, stopping essentially as soon as possible under any alternative to the null ([Balsubramani and Ramdas, 2015](#)). In the same work, my coauthor and I also present several new sequential tests designed to have particularly appealing statistical properties when
dealing with high-dimensional data, making them especially appropriate for “big data” settings.

**Future Research Directions**

I am interested in potential applications of and connections between efficient, principled algorithms like the ones I have devised to problems in the life sciences, where situations ripe for unlabeled data usage recur often. My overarching objective is to advance such scientific understanding, particularly by reducing supervision for fundamental algorithms and better exploiting the fine structure and largely untapped potential of unlabeled information. This manifests in several focused goals, the pursuit of which I would find exciting and fulfilling.

My interest in the life sciences stems from my background. Immediately before my PhD, I spent a couple of years at Strand Life Sciences, a computational genomics company, incorporating algorithms from recent research into statistical software tools used by thousands of researchers in industry and academia. I worked primarily on understanding gene expression through genome-wide association studies and on RNA sequencing methods during that period.

However, computational genomics has been revolutionized since then by much bigger data (an increased prevalence of technologies like ChIP-Seq) and better data (for instance, the ENCODE project data, or single-cell data and the microfluidics technologies behind it), and the scope for new methods to better understand the biology has increased apace. Such data can often be formulated as being unlabeled from a machine learning point of view, and I am extremely interested in designing new methods like the ones I have devised during my PhD to address these fresh new problems.

I have so far worked only within genomics among the computational life sciences, but I am excited about applying the quantitative techniques and expertise I have accrued to other fundamental disciplines of computational biology as well, such as population genetics and proteomics. I would like to stress that I am not solely interested in such areas as a vehicle for data and problems, even though I believe there is tremendous scope for applying algorithms like those I have designed. Rather, I am particularly enthusiastic about learning about the fundamental biology, as I did in my earlier genomics role, and view this versatility as imperative to my main goal of advancing basic scientific understanding, when allied to my existing algorithm design skills.

On a separate note, my sequential testing work also suggests simple, practical extensions that stoke my enthusiasm for future work in this area. It is clear that sequential methods are tremendously useful for their own sake; I again see enormous further potential to influence basic science. To give a methodological example, a significant application of sequential testing is in the design of clinical trials, which unavoidably suffer patient attrition and readmission, and I am interested in aiding the widespread incorporation of simple, sound sequential procedures to replace prevailing heuristic methods ([Peto et al., 1976](#)) in communities such as this. Devising robust and interpretable notions of confidence for such tests is also a problem that interests me for this reason, and I believe I have the tools to continue contributing to the area.
References


