# New Directions in Machine Learning: Research Statement

Liu Yang[*]

Machine learning has changed the way we approach many applications in computer science, and has enabled us to approach many applications of computer science that were not previously possible: for instance, webpage classification, image retrieval, and natural language question answering. However, there are some applications that escape the traditional approach to machine learning and require new insights to extend the benefits of machine learning to these applications. One of my interests is in advancing the set of possible applications by extending the current formalism for machine learning: for example, settings where the concept to be learned changes over time, settings where it is possible to gain further benefits by interacting with an expert, and settings where we have access to a sequence of related learning problems and wish to get improvements from that relatedness. In particular, one common thread throughout much of my work is the use of *interaction* to gain improvements in performance compared to standard non-interactive protocols. For instance, I have a series of papers on the *active learning* setting, in which a learning algorithm is able to request the target label of selected instances sequentially, and the goal is to learn an accurate classifier using a number of label requests smaller than the number of random samples that would be required to achieve the same accuracy. I also have work on the problem of *property testing* based on this protocol, which finds that the sample complexity of testing with this active testing approach is often superior to that of passive testing, and is a more realistic model of interaction for practice than the membership query model. Another of my interests is in applying the techniques of theoretical machine learning to other areas such as algorithmic economics.

## 1 Machine Learning over Time (NIPS 2011, JMLR submission)

One setting in which active learning can be quite useful is when data are presented to the learner in a stream, and for each example the algorithm is required to make a label prediction, and then may optionally request the true label of the example. We are then interested in both the number of prediction mistakes and the number of labels the algorithm requests. Most existing analyses of active learning are based on an i.i.d. assumption on the data; but in many stream-based learning scenarios, either the distribution of the data or the target concept drifts over time.

In a paper published at NIPS 2011, I studied a variant of stream-based learning in which the examples are independent, but the distribution from which the data are drawn can change over time (while the target function and noise conditions remain fixed), as long as it remains in a (possibly unknown) totally bounded family of distributions (e.g., smooth densities). Surprisingly, even with this drifting distribution, both the number of extra mistakes compared to the best function in hindsight (i.e., regret), and the number of label requests, can often be *sublinear* in the number of examples observed. This means that it is *possible* to learn the target function in this scenario, and furthermore that active learning provides a significant advantage in terms of the number of label requests, compared to passive learning (which requests all of the labels). I further characterized the rates of growth of the number of mistakes and the number of label requests, for a particular active learning algorithm designed for this setting, as a function of the complexities of the concept space, noise conditions, and class of possible distributions. Interestingly, I also obtained minimax lower bounds on these quantities that match these upper bounds in certain cases, indicating a sense of optimality for this method.

The above work left open the question of a drifting target concept. To bridge this gap, my recent work (joint with Steve Hanneke and Varun Kanade) studies the problem of active learning (and passive learning) with a drifting target concept. As a concrete model, consider a statistical learning setting, in which data arrive i.i.d. in a stream, and for each data point the learner is required to predict a label for the data point at that time, and then optionally request the true (target) label of that point. We are then interested in making a small number of queries and mistakes (including mistakes on unqueried labels) as a function of the number of points processed so far at any given time. The target labels are generated from a function known to reside in a given concept space, and at each time the target function is allowed to change by a distance $\epsilon$ (that is, the probability the new target function disagrees with the old target

[*] Carnegie Mellon University, Computer Science Department. Email: liuy@cs.cmu.edu.

function on a random sample is at most $\epsilon$). The recent work of (Crammer, et al) studies this problem in the context of passive learning of linear separators. In this theoretical study, we have broadened the scope of that work, to other concept spaces and distributions, improving the guarantees on performance, establishing lower bounds on achievable performance, and extending the framework to study the number of labels requested by an active learning algorithm while maintaining the performance guarantees established for passive learning. In particular, we proved bounds on the number of queries and mistakes made by a particular algorithm, as a function of $\epsilon$, the VC dimension of the concept space, and the number of time steps so far. We also considered variants of this in which $\epsilon$ is also allowed to change over time, and then the bounds on the number of mistakes and queries should depend on the sequence of $\epsilon$ values.

There are many other possible extensions of this model that I plan to explore in the near future; for instance, I am particularly interested in considering noise in the concept drift setting, considering the problem of simultaneously drifting target concept and data distribution, considering a stronger model in which the target concept is drifting according to an unobservable but potentially learnable Markov process, and considering a setting where the interpretation of the features is changing over time.

## 2 Transfer Learning (COLT 2011, Machine Learning Journal)

Some of my work (joint with Steve Hanneke) showed that, in a Bayesian learning setting, knowledge of the prior distribution of the target concept can provide strong benefits in the context of active learning. However, direct knowledge of the target's distribution may be too strong a requirement for many realistic scenarios. Fortunately, we can remove this assumption if we are tasked with a sequence of learning problems. Specifically, we explore a "transfer learning" setting, in which a sequence of target concepts are sampled independently with an unknown distribution from a known family. We then study the total number of label requests required to learn all targets to an arbitrary specified expected accuracy (by self-verifying algorithms), focusing on the asymptotics in the number of tasks and the desired accuracy.

The main result of this work is that, as the number of tasks grows large, we can obtain an average number of label requests per task equal to the expected number of label requests for learning with direct access to the target's distribution. Thus, we effectively replace the direct access to the distribution mentioned above with indirect access via a small number of labeled examples from each of a sequence of learning problems. In particular, when combined with the result mentioned above for Bayesian active learning, we find that there are quantifiable benefits from applying this method in the context of self-verifying active learning.

Our technique involves estimating the target's distribution, which poses a challenge since we have only indirect access to the sequence of target functions via a small number of labeled examples from each learning problem. The key insight driving our approach is that the distribution of the target concept is identifiable from the joint distribution over a number of random labeled data points equal the VC dimension of the concept space. This is not necessarily the case for the joint distribution over any smaller number of points. This observation, and the study of estimating the target's distribution from labeled examples in general, may also be of independent interest.

In recent follow-up work, we have further studied this technique, and can now bound the *rate* of convergence of this estimate of the prior, as a function of the number of tasks observed so far, and the number of labeled samples per task. This rate has implications for quantifying the precise benefits of transfer learning, compared to learning each task independently.

## 3 Efficient Active Learning with a Surrogate Loss (AISTATS 2010, Submission to the Annals of Statistics)

Much of the recent progress in studying the sample complexity of active learning with noise has made use of algorithms which have excessively high running times. This is because they perform optimizations of empirical error rates, measured in terms of the 0-1 loss, which for many hypothesis classes are known to be NP-Hard. In passive learning, practical learning algorithms circumvent these computational barriers by replacing the 0-1 loss with a convex *surrogate* loss function. One can then show that, under certain conditions, having small risk under the surrogate loss implies small error rate under the 0-1 loss as well. This has become the dominant paradigm in modern approaches to machine learning, such as AdaBoost, Support Vector Machines, Logistic Regression, and many others.

Given this fact, it only makes sense to make use of surrogate losses in active learning as well, to circumvent the computational barriers in that case as well. In recent work, we develop an approach to active learning with surrogate losses, which is both computationally efficient and provides provable reductions in sample complexity compared to passive learning.

Interestingly, we find that the naïve approach of designing an active learning algorithm that directly optimizes the surrogate loss does not lead to improved sample complexity guarantees compared to passive learning. However, we then develop more subtle methods, which make use of the surrogate loss in internally-constructed optimization problems, but do not necessarily optimize the surrogate risk overall, and yet provably converge in error rate (measured under 0-1 loss) at a rate often faster than possible by known passive methods that use the given surrogate loss.

## 4 Active Testing (FOCS 2012)

One of the motivations for property testing of boolean functions is the idea that testing can serve as a preprocessing step before learning. However, while most of the work in property testing has focused on the Membership Query model (i.e., the ability to query functions on arbitrary points), there have been several convincing experimental studies showing that membership queries are not realistic for most machine learning problems: for instance, image recognition, medical diagnosis, handwritten digit recognition, etc. The crux of the problem is that algorithms based on membership queries tend to query highly ambiguous points, which appear unnatural or bizarre to the human oracle.

As a result, the machine learning community has largely abandoned the Membership Query model, returning to the classic model of learning from random samples. But since random samples are often highly redundant in their information content, there has more recently been renewed interest in adding interaction into the learning process, in the form of “Active Learning”. The idea is that we get a pool of random unlabeled examples, and the algorithm can request the label of any example in the pool. The hope is that, by carefully choosing only the informative examples, we can reduce the number of labels necessary for learning. So active learning does not suffer the “strange examples” issues faced by learning with membership queries. In this work, we bring this well-studied model in learning to the domain of testing. In particular, we assume that as in active learning, our algorithm can make a polynomial number of draws of unlabeled examples from the underlying distribution D, and then can make a small number of label queries but only over the unlabeled examples drawn. The unlabeled examples are viewed as cheap, whereas label queries are viewed as expensive.

We show that for a number of important properties, testing can still yield substantial benefits in this setting. This includes testing unions of intervals, testing linear separators, and testing various assumptions used in semisupervised learning. For example, we show that testing unions of $d$ intervals can be done with $O(1)$ label requests in our setting, whereas it is known to require $\Omega(\sqrt{d})$ labeled examples for passive testing (where the algorithm must pay for labels on all examples drawn from D) and $\Omega(d)$ for learning. In fact, our results for testing unions of intervals also yield improvements on prior work in both the membership query model (where any point in the domain can be queried) and the passive testing model as well. In the case of testing linear separators in $R^n$, we show that both active and passive testing can be done with $O(\sqrt{n})$ queries, substantially less than the $\Omega(n)$ needed for learning. We also show a general combination result that any disjoint union of testable properties remains testable in the active testing model, a feature that does not hold for passive testing.

In addition to these specific results, we also develop a general notion of the testing dimension of a given property with respect to a given distribution. We show this dimension characterizes (up to constant factors) the intrinsic number of label requests needed to test that property; we do this for both the active and passive testing models. We then use this dimension to prove a number of lower bounds. For instance, interestingly, one case where we show active testing does not help is for dictator functions, where we give $\Omega(log(n))$ lower bounds that match the upper bounds for learning this class. In particular, this implies that any class that contains dictator functions and is not so large as to contain almost all functions, requires at least $\Omega(log(n))$ queries to test in the active and passive models, including decision trees, functions of low Fourier degree, juntas, DNFs, etc.

Our results show that testing can be a powerful tool in realistic models for learning, and further that active testing exhibits an interesting and rich structure. Our work in addition develops new characterizations of common function classes that may be of independent interest.

# 5 Online Allocation and Pricing with Economies of Scale (STOC 2014 submission)

Allocating multiple goods to customers in a way that maximizes some desired objective is a fundamental part of Algorithmic Mechanism Design. In recent joint work with Avrim Blum and Yishay Mansour, I consider the problem of offline and online allocation of goods that have economies of scale, or decreasing marginal cost per item for the seller. In particular, we analyze the case where customers have unit-demand and arrive one at a time with valuations on items, sampled iid from some unknown underlying distribution over valuations. Our strategy operates by using an initial sample to learn enough about the distribution to determine how best to allocate to future customers, together with an analysis of structural properties of optimal solutions that allow for uniform convergence analysis. We show, for instance, if customers have $\{0, 1\}$ valuations over items, and the goal of the allocator is to give each customer an item he or she values, we can efficiently produce such an allocation with cost at most a constant factor greater than the minimum over such allocations in hindsight, so long as the marginal costs do not decrease too rapidly. We also give a bicriteria approximation to social welfare for the case of more general valuation functions when the allocator is budget constrained. The techniques involved in proving that optimization on the small initial sample of customers have implications for near-optimal performance on a larger population of customers are heavily rooted in the theory of machine learning, including results on uniform concentration based on the VC dimension and related quantities.

# 6 Concluding Remarks

The above topics give some of the overall flavor of my work. I am currently working to advance the state-of-the-art in machine learning in several areas. In active learning, I am working to extend the current approaches to make them more robust to noise, computationally efficient, and provably optimal. In property testing, I am pursuing the largely-unexplored topic of testing properties of real-valued functions; I am also working to extend the active and passive testing frameworks and results to the tolerant testing framework, and to establish formal connections between learning and testing. As mentioned, although many machine learning applications involve change over time, the vast majority of the techniques in the literature require a static learning environment; to bridge this gap, I am working on learning with a changing target concept, a drifting data distribution, and changing interpretations of the features over time. In the area of transfer learning, I am working to extend my existing results to the setting of real-valued functions, to further characterize the rates of convergence, and to allow dependence among the learning tasks. In the area of algorithmic economics, I am currently studying the general problem of mechanism design with customer valuations that have an unknown distribution, which is estimable, and achieving optimal performance in an online allocation setting; I am also interested in extending this setting to allow the distribution of customer valuations to drift over time. In the future, I plan to continue pushing the frontiers of theoretical machine learning, identifying new directions that require significant advancement to become practically useful, and advancing those areas via deeper theoretical investigations; ultimately, the aim is to enable new applications of machine learning that are not approachable with the current techniques.