My research is on theoretical foundations of machine learning. During graduate school, I primarily analyzed standard algorithms under practical assumptions, particularly in areas not covered by existing theory. While this was satisfying and the understanding it granted will color all my future work, during my postdoctoral studies I have been more interested in design of new algorithms.

In this research statement I will exemplify these two phases with the following two groups of works.

Theoretical analyses under practical assumptions. Consider clustering, the task of partitioning data into k groups according to some notion of similarity. Clustering is one of the oldest statistical techniques, and continues to be a tool of choice in modern data analysis. Consequently, it is applied to data of arbitrary shapes and sizes; by contrast, the theory often dictates specific algorithms, specific properties of the data, and even a specific model generating the data. To address this, I investigated standard clustering methods under only the assumption of bounded moments, and showed the procedures are still statistically well-behaved, meaning good performance on a data sample will translate to future data points. This result will be discussed in Section 1, along with some of my work on standard linear prediction algorithms.

Representation learning. Consider boosting, an algorithm which takes a collection of elementary classifiers, forms a linear combination of them, and predicts with its sign. In my initial studies of boosting, I would fix a standard set of elementary classifiers (say, small decision trees), and focus on the combining phase of the algorithm. While classical results indicate this approach can adapt to arbitrarily complex prediction problems, there is evidence that the learned representation — the set of linear combination weights and corresponding elementary classifiers — can fail to be succinct, thus in my current work I am studying non-linear combination rules. In Section 2 I will discuss both my prior work, and also recent studies of polynomial and neural networks, in the latter case obtaining exponential savings.

To close, I will summarize a few other older works, as well as ongoing work in non-convex optimization.

1. Theoretical analyses under practical assumptions

This section will discuss three results where the focus was on practical settings: statistical properties of clustering (Telgarsky and Dasgupta, 2013), margin properties of boosting (Telgarsky, 2013b), and convergence properties of linear classifiers (Telgarsky et al., 2015).

1.1. Clustering. Two standard methods in data analysis are k-means — which fits spherical clusters to the data, thereby inducing a partition — and Gaussian mixture modeling — which fits Gaussians to the data, thereby producing a probability vector for each point specifying the relative likelihood of generating it from each Gaussian mean. These methods are ubiquitous, and while much of their use and analysis is in situations where some underlying model may be recovered (Dasgupta 1999, Balcan et al. 2009), it is often the case that one simply wishes to fit the data, perhaps to merely compress or summarize it (Lugosi and Zeger 1994).

Consequently, together with my advisor Sanjoy Dasgupta, I analyzed these two methods under the following setting (Telgarsky and Dasgupta, 2013): the algorithm need only satisfy an approximation guarantee (thus avoiding NP-hardness woes (Vattani 2009)), and the data need not be drawn from any model or even be bounded, it must merely have some bounded moments. (The issue of boundedness is the most difficult, and will reappear in subsequent discussions.) In this setting, it is still possible to show the standard statistical sanity check: a good fit on a data set translates into a good fit on future data points.
1.2. Margins. The standard boosting algorithms from above can be formulated as a nearly standard convex optimization problem, where the optimization variables are the weights in the linear combination of elementary classifiers. The only non-standard aspect is that these formulations generally shun a basic tenant of statistics and optimization methodology: they do not require boundedness of the optimization variables. In consequence of this, a new theory arose to describe the statistical stability of these methods: they find classifiers which are not only correct, but moreover correct by some margin [Schapire et al., 1997].

This discovery spawned a great deal of research into specialized methods achieving a better margin (Rätsch and Warmuth, 2005; Shalev-Shwartz and Singer, 2008; Rudin et al., 2007). Unfortunately, these methods did not enter practical use. Consequently, years after the boom, I studied the original method, and showed that merely scaling the steps taken by the algorithm suffices to achieve arbitrarily good margins (Telgarsky, 2013b). Moreover, in the spirit of the other results of this section, I also dropped some impractical assumptions, for instance giving a characterization when only some of the data may be separated by some margin.

1.3. Linear classifiers. More generally, boosting algorithms are members of a broader methodology, the very popular approach of linear prediction. Here as well, the standard optimization and statistical theory requires strong constraints and/or regularization (Bubeck, 2014), whereas the folklore methodology eschews them (Telgarsky et al., 2015, Appendix D). Together with my colleagues Miroslav Dudík and Robert Schapire, I established convergence properties of linear prediction schemes (including boosting) without these assumptions, thus justifying the practical methodology (Telgarsky et al., 2015).

1.4. Future work. The goal of studying practical settings and algorithms will influence all of my work; more concretely, my colleagues and I are already continuing from the immediately previous project, both providing a tighter characterization, and performing more extensive experimentation to confirm which regimes are indeed observed in practice.

2. Representation learning

In the preceding discussions of general linear learning (and boosting), the algorithm is given the data in some processed form (or, in the case of boosting, with some fixed set of elementary classifiers). This choice of representation can make or break a method in practice, and has contributed to the success of neural networks, where the lower layers of the network may be interpreted as producing a rich data representation.

2.1. Flat representations. Classical arguments indicate that even linear algorithms can do well on arbitrary prediction problems so long as their feature representation is sufficiently rich (Barron, 1993). Along these lines, when I established the consistency of boosting with the logistic loss, I also proved that the popular choice of small decision trees as elementary classifiers suffices to fit arbitrary classification problems (Telgarsky, 2013a, Proposition 3).

Unfortunately, this representation is far from ideal; namely, it can easily fail to be succinct, adversely affecting both optimization and statistical aspects of the algorithm. In one of my clustering works, I tried a standard technique of augmenting data representation with distances to learned clusters (Telgarsky and Dasgupta, 2012, Section 7); this purely empirical investigation showed major gains, and while no theoretical treatment was included, it is direct that a well-separated clustering also induces a well-separated classification problem. Before investigating this line further, however, I became interested in layered representations, where new features can be built on top of others.

2.2. Layered representations. A natural layered representation is a polynomial, in the sense that higher degree terms can be written as summations and compositions of lower degree terms. While there exist a variety of algorithms to learn polynomials under specialized assumptions, my colleagues and I focused on designing an algorithm which is efficient in practice (Agarwal et al., 2014). Unfortunately, our theoretical results were incomplete, effectively requiring an oracle to suggest how to extend a representation, and the heuristic from our experiments seemed to only provide very mild guarantees. This shortcoming was on my mind as I turned next to study neural networks.
With neural networks, classical results also show they can adapt to arbitrary prediction problems \cite{Barron1993}; unfortunately, these representation results used only two layers: one layer mimicking the elementary classifiers from boosting, and another summing them up. Consequently, my first investigation here was to show that multiple layers can help; specifically, in unpublished, current work, I have established that there exist prediction problems where a deep network achieves zero error, whereas any shallow network with low error requires exponentially as many nodes \cite{Telgarsky2015}. Due to the form of shallow neural networks, the same result also gives exponential savings by deep networks over boosting of small decision trees.

2.3. Future work. I have become entranced by the question of the exact representation power of neural networks; the preceding result only identifies a potentially pathological set of problems with improved representation, thus I am currently developing a more general characterization of the prediction problems amenable to multi-layered representations.

The goal of these works is partly mathematical, but also algorithmic. In another preliminary work — appearing in a NIPS 2015 workshop — myself and two colleagues have used the preceding structural results to design a layer-wise training method which is guaranteed to find those earlier difficult representations, along with some more general guarantees and promising empirical results \cite{Abernethy2015}.

3. Other work

In my earlier work, there were two main threads: clustering, and boosting. I will mention two results here not mentioned before: first, together with my colleague Andrea Vattani, I showed that a simple classical algorithm can escape certain local optima of the standard $k$-means algorithm \cite{Telgarsky2010}; second, in boosting, my work not only showed how to circumvent boundedness issues in statistical analyses, but also in optimization analyses \cite{Telgarsky2012}.

More recently, I have become interested in non-convex optimization in general; here, both neural networks and clustering serve as concrete motivating problems. During a summer internship at Microsoft Research New England, I was fortunate to take part in a project showing that the structure of many latent variable models can be discerned by running a power method on a tensor constructed from observed data, a problem not amenable to standard (convex) techniques \cite{Anandkumar2014}. Lastly, I am currently involved in extending certain recent advances in simulated annealing to handle generic non-convex functions \cite{Vempala2006}, the goal being good performance on natural problems.

References


Matus Telgarsky. Boosting with the logistic loss is consistent. In *COLT*, 2013a.


Andrea Vattani. k-means requires exponentially many iterations even in the plane. In *SOCG*, 2009.