Research Statement

Ben London
University of Maryland
blondon@cs.umd.edu

The momentum in modern machine intelligence and data analytics is towards learning from massive amounts of data. Many of these data sources—such as images, text, proteins and social networks—have a natural underlying structure. My research focuses on developing core **machine learning** approaches to model these data, and **statistical learning theory** to analyze these approaches. My goal is to understand under what circumstances leveraging structure can enhance representational power and improve efficiency, and how it affects learning-theoretic guarantees. I believe this is how we will address today's challenging problems in artificial intelligence, computer vision and natural language understanding.

Generalization Guarantees for Structured Prediction

A central question in statistical learning theory is how well a trained model generalizes to new data; that is, what is the expected error over draws from a distribution over instances? Often, this is shown by bounding the difference of the expected and empirical error rates (referred to as the *generalization error*) by some function of the size of the training set and the complexity of the model. Structured prediction problems are characterized by multiple interdependent prediction tasks, whose dependence relationships are represented by a network topology. Instances of a structured prediction problem (e.g., a document, social network, or high-resolution image) typically have large internal structure. Therefore, training sets for these types of problems usually contain very few examples—sometimes only one. According to traditional analyses, generalization should not be possible in this scenario. But in practice, we often find that the training error is a good approximation of the test error. This prompts the question: why is generalization possible in this setting?

Structured models often employ a technique known as *templating* (or, *parameter-tying*), in which a finite, typically small, set of parameters are applied to "similar" substructures throughout each structured example. For instance, Markov logic networks are defined using a weighted combination of first-order logic rules; the number of weights is proportional to the number of rules, which is independent of, and smaller than, the number of rule *groundings*. As such, each structured (i.e., grounded) example in the training set actually contains many "micro examples" (i.e., groundings). One should be able to estimate the rule weights from a single grounded example. However, the micro examples are interdependent; this means that traditional analyses, which assume independently and identically distributed examples, do not apply to this setting.

In a series of papers [2, 5, 6, 10, 11], I developed generalization bounds based on the above intuition, leveraging and extending recent results in measure concentration and learning theory. I showed that, under certain assumptions on the data distribution and hypothesis class, the generalization error decreases at a rate of $\mathrm{O}(1/\sqrt{mn})$, where m is the number of structured examples and n is the number of output variables per example. This proved that

one can indeed generalize from a few, large examples—*even just one*. A key takeaway is that the dependence in the data distribution, and the distribution induced by the model, must decay quickly as a function of graph distance. The latter condition is equivalent to a *stability* property for structured inference. Stability is related to robustness to noise, in that it measures a predictor's sensitivity to small perturbations in the input. I showed that stability follows from regularization, in both the learning and inference objectives.

When Does Strong Convexity Help?

Interestingly, the relationship between generalization and stability motivates the use of inference with strongly convex free energies. The modulus (i.e., strength) of convexity is determined by the entropy term, which acts as a regularizer on the output space. In recent work [9], I showed that several forms of (variational) inference are strongly convex, where the modulus is independent of the number of variables in the model. I am currently investigating the implications of this work for learning graphical models with approximate inference. My new theoretical analysis suggests that learning with a strongly convex free energy, with a provably constant modulus, can substantially improve the quality of learned marginal probabilities. This claim is supported by my empirical results with a new variational method, which exhibits dramatically reduced error.

Other Forms of Structured Learning

Structure can be found or imposed on a number of problems, in ways that don't necessarily fit into the typical structured learning paradigm. An example of this is learning a binary relation from a sample of pairwise combinations. This is a common setting in entity resolution. Pairwise sampling induces a dependency structure on the training data, which complicates analysis. I developed generalization bounds for this setting [4], using ideas from graph theory. A related problem is learning multiple binary relations. In [3, 8], I proposed a latent-factor tensor decomposition to jointly model multi-relational data. Structure is also found in computer vision, particularly in scene understanding. In [7], I applied collective reasoning to detect human activities in video data.

Future Research

Stable inference can improve generalization, but it can also improve efficiency in scenarios where predictions must be updated regularly. An example of this is in knowledge base construction, in which inference may be continually updated as new evidence is collected. At the scale of a knowledge base, repeated calls to inference may be prohibitively expensive. If prediction is stable, then small updates to the evidence should not induce much change in the predictions. Accordingly, inference may be deferred until a time when it is convenient, or until the stability guarantees suggest the predictions will change significantly. A related problem is how to update inference *efficiently*, based on knowledge of which predictions are stable. I am actively examining these questions, and plan to continue this research in the future. I believe this research will be crucial to tackling today's large-scale structured prediction problems.

I am also interested in the connections between structured prediction and deep learning, whether ideas from one can inform the other. For example, in graphical models, structure is typically imposed *a priori* by domain knowledge; it is possible that representation learning, which has proven so effective in deep learning, could be used to automatically learn the structure of a graphical model. Similarly, one of the central questions in deep learning is why *drop-out* training improves accuracy; if we view a deep circuit as a structured predictor, it is possible that the drop-out promotes stability, which thereby improves generalizability.

My overall agenda is to develop core machine learning algorithms and analysis for important problems. I would like to conduct research that has great impact, to theoreticians and practitioners alike. My work thus far has been informed by a belief that machine learning techniques should be based in a solid theoretical foundation. I also believe that theory cannot exist in a vacuum, and that it must have relevance to real situations. In my future work, I hope to tackle challenging new problems, apply my work to interesting data sources, and foster new interdisciplinary collaborations.

Selected Publications

[1] B. London and L. Getoor. Collective classification of network data. In Charu C. Aggarwal, editor, *Data Classification: Algorithms and Applications*. CRC Press, 2013.

[2] B. London, B. Huang, and L. Getoor. Improved generalization bounds for large-scale structured prediction. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, 2012.

[3] B. London, T. Rekatsinas, B. Huang, and L. Getoor. Multi-relational weighted tensor decomposition. In *NIPS Workshop on Spectral Learning*, 2012.

[4] B. London, B. Huang, and L. Getoor. Graph-based generalization bounds for learning binary relations. *CoRR*, abs/1302.5348, 2013.

[5] B. London, B. Huang, B. Taskar, and L. Getoor. Collective stability in structured prediction: Generalization from one example. In *International Conference on Machine Learning*, 2013.

[6] B. London, B. Huang, B. Taskar, and L. Getoor. PAC-Bayes generalization bounds for randomized structured prediction. In *NIP Workshop on Perturbation, Optimization and Statistics*, 2013.

[7] B. London, S. Khamis, S. Bach, B. Huang, L. Getoor, and L. Davis. Collective activity detection using hinge-loss Markov random fields. In *CVPR Workshop on Structured Prediction: Tractability, Learning and Inference*, 2013.

[8] B. London, T. Rekatsinas, B. Huang, and L. Getoor. Multi-relational learning using weighted tensor decomposition with modular loss. *CoRR*, abs/1303.1733, 2013.

[9] B. London, B. Huang, and L. Getoor. On the strong convexity of variational inference. In *NIPS Workshop on Advances in Variational Inference*, 2014.

[10] B. London, B. Huang, B. Taskar, and L. Getoor. PAC-Bayesian collective stability. In *Artificial Intelligence and Statistics*, 2014.

[11] B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. In preparation, 2015.