Hristo Paskov Research Statement

My research explores learning, data representations, and compression in order to provide large-scale statistical tools and a better understanding of intelligence. I take a multifaceted approach that incorporates computer science and engineering with my avid exploration of mathematics; I believe this is essential in order to design machine learning systems that can meet the challenges of science and industry's biggest problems. Such systems critically rely on an effective feature representation that highlights relevant information and a learning criterion that draws on prior knowledge. Scaling to massive datasets further requires careful algorithmic paradigms, such as compact data storage that natively provides feature representations, as well as a deep understanding of the mathematical structure of the learning criterion to find good approximate solutions and expedite convergence.

My work is an interdisciplinary endeavor that spans machine learning and statistics, continuous and discrete optimization, and algorithms. I have applied my work to security [8, 9], natural language processing [1-5], bioinformatics [7], and social network analysis [6]. I have investigated:

- a massively scalable framework for text storage and N-gram feature representations [3]
- an alternative paradigm to kernels for learning in very high (possibly infinite) dimensional spaces that is compatible with arbitrary regularization and admits fast learning algorithms [1]
- a new deep learning paradigm that uses compression to simultaneously reduce data size and provide useful feature representations [2, 4, 5].

Large-Scale Feature Representations and Learning

My earliest work in building large-scale machine learning systems was in security: creating algorithms that break audio captchas [9] or that de-anonymize authors at an internet-scale based on stylometric factors [8]. I was responsible for the machine learning behind collaborations with students from Tulane and Berkeley Universities. In both cases simple algorithms that scaled gracefully to large datasets were essential as they allowed for rapid experimentation to find the best feature representations and fine-tune prediction performance. The work on **audio captchas prompted companies like Microsoft, Yahoo, and eBay to switch to more secure schemes**, and the work on **de-anonymization is widely cited in the stylometrics community with over 100 citations**.

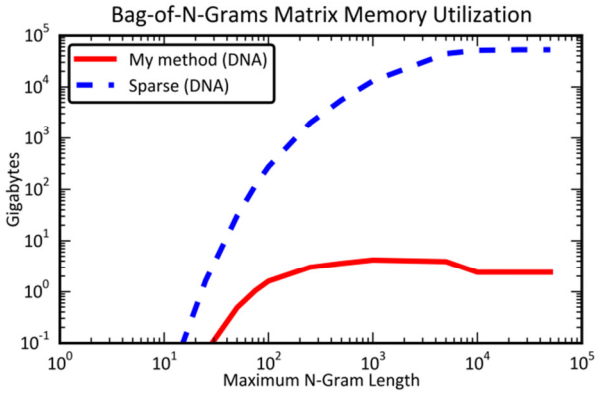

Figure 1. My suffix tree method requires substantially less memory than naïve sparse formats to store BoN matrices. Performance is shown on binary DNA SNP array data.

More recently my work has focused on efficient feature representations and learning algorithms for large text datasets. At their core, most learning algorithms rely on repeatedly multiplying a matrix (the feature representation) in order to learn, so improving matrix multiplication and storage can provide profound computational savings. For instance, bag-of-N-gram (BoN) features represent a document by counting the frequency of relevant strings in the document and are ubiquitous in natural language processing and bioinformatics. While longer N-grams capture more information, computational tractability limits most applications to using bi-grams and tri-grams since the number of possible N-grams increases with maximum N-gram length. In [3] I show how to leverage the topology of a document corpus' suffix tree to **store and multiply any BoN feature matrix derived from the corpus in memory and time that is, at worst, linear in the total corpus length**. This potentially quadratic computational improvement (over naive representations) is shown in Figure 1, which compares the memory requirements of my method with a standard sparse matrix format. My method is **17 times more efficient on natural language data and over 10,000 times more efficient on binary DNA SNP array**

data; multiplication times follow a similar pattern. These speed improvements allowed us to experimentally verify that **long N-grams improve prediction performance in sentiment analysis tasks involving millions of examples**. I also provide an efficient framework that stores text in a format more amenable to learning. Given an arbitrary learning problem, the framework customizes a BoN matrix by screening all N-grams for usefulness and it emits a data structure optimized for fast multiplication.

My most recent work is a large-scale learning algorithm that is inspired by applying the BoN framework to long binary text corpora such as DNA SNP array data. In [1] I introduce fine control kernels, a representation that focuses on the "$p \gg n$" learning scenario where there are so many features that it is impossible to store a vector of coefficients $\beta \in \mathbb{R}^p$ for each of them. The kernel trick, which represents β implicitly and never computes feature specific quantities, is traditionally used in such scenarios; however, it is incompatible with a number of regularizers that are essential for high-dimensional learning such as the Elastic-Net or Lasso. Fine control kernels represent β implicitly but are compatible with arbitrary regularizers because they assume that it is reasonable to compute feature specific quantities (without ever storing them); they lie in between kernels and explicit feature representations. I provide a **superlinearly convergent learning algorithm that is compatible with structured matrices** because it treats matrix multiplication as a black box. Moreover, this algorithm is fundamentally **based on the same duality theory that leads to fast feature screening rules for the Lasso (e.g. SAFE/ Strong rules), and it utilizes generalizations of these rules derived from monotone operator theory**.

Deep Learning: Features through Compression

Dictionary-based compression represents text via a dictionary of strings and a set of pointers, locations in which to paste copies of these strings so as to reconstruct the data. My work thus far has focused on a shallow scheme [5] that stores the dictionary strings in plaintext as well as a deep formulation termed Dracula [2] which can be exponentially more space efficient. Dracula recursively compresses its own dictionary, whereby dictionary strings construct longer dictionary strings, and its representation can be stratified into layers akin to those of a deep neural network. Figure 2 depicts a simple compression. Both compression schemes provide BoN features when dictionary strings are interpreted as N-grams, and they provide **state of the art performance in natural language processing, author de-anonymization, and unsupervised learning tasks**. Importantly, **both compression criteria exhibit considerable problem structure that allows them to scale to large datasets**. I provide an efficient large-scale homotopic algorithm for the shallow scheme in [4] and am currently working on a large-scale algorithm for Dracula based on concepts in the BoN multiplication and fine control kernel papers. Those papers were inspired by ideas (suffix trees and optimization in Hilbert spaces) I encountered in my investigation of compression, and they are representative of the research feedback cycle that jointly drives my work in large scale learning and compression.

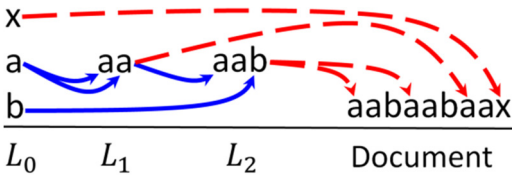

Figure 2. A 3-layer compressed representation from Dracula showing dictionary (blue solid arrows) and document (red striped arrows) pointers.

Dracula's problem structure also provides statistical insights into its representations. For instance, its compression criterion can be expressed as a binary linear program (BLP) the objective of which is parameterized by the cost of storing individual dictionary strings and pointers. Varying these costs provides a variety of representations. At one extreme *negative* costs lead to the traditional BoN representation (using every possible N-gram and pointer for maximal redundancy) whereas cost models inspired by traditional linear memories (e.g. disk or RAM) provide sparse representations that are superior for learning in a number of problem domains [2]. Finding the *best* representation for a specific learning problem requires tuning the costs to the problem. Here the perspective offered by viewing Dracula's BLP as an equivalent linear program over a sufficiently constrained polyhedron is invaluable. Varying the costs continuously corresponds to walking around the surface of this polyhedron, so Dracula's representations

will not jump unexpectedly. It is therefore reasonable to fine tune the cost model, and a simple parameterized cost model – inspired by linear memory – directly modulates the depth and diversity of the resulting dictionary.

Future Research Directions

My research explores data representation schemes, inference procedures, and optimization by uncovering key problem structure and exploiting it to provide new computational and statistical methods. Going forward there are three main areas that I would like to pursue:

1. **Feature representations and large scale solvers**: My work has largely focused on text because its underlying structure is simple and well characterized by suffix trees. I next plan to focus on continuous data (e.g. audio and images) where error tolerance and invariance are necessary to draw connections between data samples. I look forward to exploring these powerful paradigms and the challenges they bring to the underlying problem structure and data representations used for learning. I am also excited to continue exploring new algorithms for convex and non-convex optimization and to uncover new structured matrix multiplication routines.
2. **Compression**: I similarly plan to extend my compression work to continuous data where I hope to establish it as a new form of deep learning that has strong algorithmic and statistical guarantees because of its rich polyhedral structure. This work will have a fruitful symbiosis with my aforementioned research into features for continuous data, and it holds promising insights for error tolerant text compression.
3. **Applications**: Applying my algorithms to new problem domains will be a focal point of my research. I am particularly excited about biological data such as DNA sequence reads or neuronal spike trains because of its potential to benefit public health, its underlying structure, and the scale of modern datasets. For instance, the multiple hypothesis testing problem is limiting in whole genome association studies because of the large number of genetic loci. I believe that compression may decrease the number of independent locations that need to be tested and improve interaction detection. I also plan to apply my work to neuronal signals in hopes of advancing neuroprosthetics. More broadly, I have been fortunate to work with experts in a variety of domains, and I am actively looking to collaborate with domain experts to help expand the boundaries of all disciplines of study.

References

[1] **Hristo Paskov**, Julian McAuley, John Mitchell, and Trevor Hastie. Fine Control Kernels for Massive Scale Feature Selection. Submitted for Publication.

[2] **Hristo Paskov**, John Mitchell, and Trevor Hastie. Data Representation and Compression Using Linear-Programming Approximations. To appear in the *International Conference on Learning Representations*, 2016.

[3] **Hristo Paskov**, John Mitchell, and Trevor Hastie. Fast Algorithms for Learning with Long N-Grams via Suffix Tree Based Matrix Multiplication. *Uncertainty in Artificial Intelligence*, pages 672-681, 2015.

[4] **Hristo Paskov**, John Mitchell, and Trevor Hastie. An Efficient Algorithm for Large Scale Compressive Feature Learning. *AISTATS*, pages 760-768, 2014.

[5] **Hristo Paskov**, Robert West, John Mitchell, and Trevor Hastie. Compressive Feature Learning. *Neural Information Processing Systems*, pages 2931-2939, 2013.

[6] Robert West, **Hristo Paskov**, Jure Leskovec, and Christopher Potts. Exploiting Social Network Structure for Person-to-Person Sentiment Analysis. *Transactions of the Association for Computational Linguistics,* 2 (Oct), pp 297-310, 2014.

[7] Ivan Paskov, Han Yuan, **Hristo Paskov**, Alvaro Gonzalez, and Christina Leslie. Joint Learning Over Drugs Improves Prediction of Cancer Drug Response. *RECOMB/ISCB Conference on Regulatory and Systems Genomics,* Abstract and Oral Presentation, 2014.

[8] Arvind Narayanan, **Hristo Paskov**, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the Feasibility of Internet-Scale Author Identification. *IEEE Symposium on Security and Privacy*, pages 300-314, IEEE Computer Society, 2012.

[9] Elie Bursztein, Romain Beauxis, **Hristo Paskov**, Daniele Perito, Celine Fabry, and John Mitchell. The Failure of Noise-Based Non-Continuous Audio Captchas. *IEEE Symposium on Security and Privacy*, pages 19-31, IEEE Computer Society, 2011.