

# RESEARCH STATEMENT—SUMMARY

Modern research attempts to further complex systems like the Internet, computationally intensive projects like the genome mapping, and harness the huge quantities of data generated better. While a lot of statistical work has been answering questions on sequences of symbols much longer than the set of possible symbols, the *alphabet size*—a number of current research problems require solutions for the opposite scenario, the *large alphabet* setting.

The theoretical focus of the research program will be the intersection of statistical learning, information theory, and signal processing. In particular, on probability estimation, inference, classification and clustering in large alphabet settings—in each case, older approaches simply do not extend easily to the large alphabet scenario. Practical applications include text classification and compression, **language modeling**, image segmentation, analysis of gene expression data, and **risk analysis for network insurance policies**.

Many principles behind the approaches we take for the above problems are not just new and theoretically interesting, but have proved to be general in nature. In the compression problem, as in [1] which won the 2006 Information Theory Society Award, we separate the description of data strings, say “abracadabra” into two parts: (i) the symbols appearing in the sequence, and (ii) of their *pattern*, 12314151231, the order in which the symbols appear. We subsequently develop two provably asymptotically optimal estimators using similar principles. These results appeared in Science [2].

Perhaps, the power of these optimality results can be best demonstrated by their application in real life contexts. Text classifiers based on these principles are comparable to the state of the art results (and faster) on several benchmarks [3] using nothing but simple minded implementations of our algorithms, with no tuning whatsoever. We are currently combining the approach above with new **graphical model inference** techniques to yield a richer class of **classifiers**.

The constrained distribution estimation problem—for example, the distribution has to be a mixture of gaussians, a mixture of binomials, etc. has similar implications for **clustering** problems—the theoretical challenge here is to incorporate this constraint into our large alphabet approaches. Practically, large alphabet clustering problems crop up in data mining, **image segmentation**, biology (MRI image segmentation), and bioinformatics.

The problem of *distribution modeling*, where we estimate the values of the probabilities in the distribution, but not their mapping to the support, combines the above approach with the Expectation-Maximization algorithm and Markov chain Monte Carlo sampling heuristics [4], see [5] for consistency results. Theoretically, the Markov chain sampling approximations involved are related to the estimation of the *permanent* of a matrix, a problem that is #P-complete [6]. **Computation of the permanent** is a well studied problem as well, and it arises in the contexts of estimating the number of bipartite matchings, computation of network reliability, etc. While polynomial time algorithms to compute the permanent have been obtained [7], they are still not practical, and we are working on improving existing results for certain subsets of matrices.

Distribution modeling described above serve to **estimate support independent statistics**—such as entropy and support size. Pattern based estimators seem to approximate the distribution well even when the sample size is comparable to the alphabet. Support size questions are relevant for chip testing and software testing. A preliminary set of results was published in [8]. While estimation of entropy or mutual information is not identical to estimation of the underlying distribution, a good approach to the distribution estimation problem does help, and is currently the best approach known under some conditions [9]. Results from [10] show promise in using a plug-in estimator based on distribution modeling from patterns.

The host of problems identified above attempt to integrate novel information theory into large alphabet (high dimensional) problems already encountered in many modern problems. The proof of any applied science is in the results, and in this regard, our optimism is justified by the ease with which the principles above perform well in practical problems.

## References

- [1] A. Orlitsky, N.P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469–1481, July 2004.
- [2] A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, October 17 2003.
- [3] N.P. Santhanam, A. Orlitsky, and K. Viswanathan. New tricks for old dogs: Large alphabet probability estimation. In *Information Theory Workshop (invited)*, Lake Tahoe, CA, 2007.
- [4] A. Orlitsky, Sajama, N.P. Santhanam, K. Viswanathan, and J. Zhang. Practical algorithms for modeling sparse data. Proceedings of the 2004 Proceedings of IEEE Symposium on Information Theory.
- [5] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Convergence of profile based estimators. In *Proceedings of the IEEE Symposium on Information Theory*, 2005.
- [6] L.G.Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8:189–201, 1979.
- [7] M. Jerrum, A.Sinclair, and Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries. *Theory of computing*, pages 712–721, 2001.
- [8] A. Orlitsky, N.P. Santhanam, and K. Viswanathan. Population estimation with performance guarantees. In *Proceedings of IEEE Symposium on Information Theory*, 2007.
- [9] A. Wyner and D. Foster. On the lower limits of entropy estimation. Submitted to IEEE Transactions on Information Theory.
- [10] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J.Zhang. On modeling profiles instead of values. In *Uncertainty in Artificial Intelligence*, 2004.