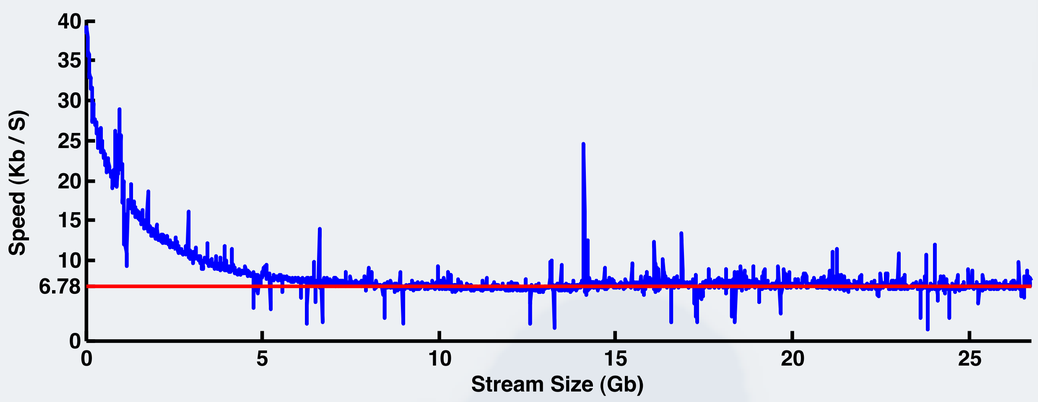
Modeling Streaming Data In the Absence of Sufficiency

Frank Wood, Department of Statistics, Columbia University

Abstract

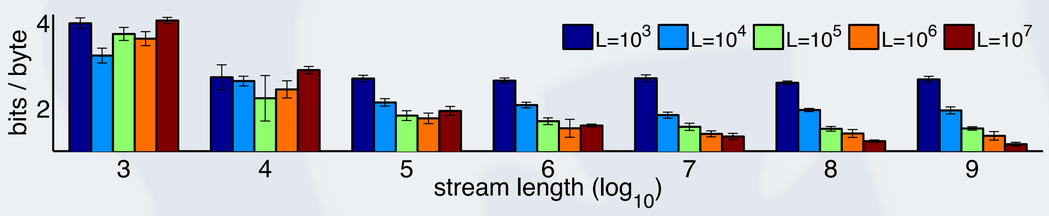
We interpret results from a paper [2] in which data were modeled using constant space approximations to the sequence memoizer. The sequence memoizer (SM) is a non-constant-space, Bayesian nonparametric model in which the data are the sufficient statistic in the streaming setup. We review approximations to the probabilistic model underpinning the SM that yield the computational asymptotic complexities necessary for modeling very large (streaming) datasets with fixed computational resource. Results from modeling a benchmark corpus are shown for both the effectively parametric, approximate models and the fully nonparametric SM. We find that the approximations perform nearly as well in terms of predictive likelihood. We argue from this single example that, due to the lack of sufficiency, Bayesian nonparametric models may, in general, not be suitable as models of streaming data, and suggest that parametric models and estimators for the same inspired by Bayesian nonparametric models may be worth investigating more fully.

Constant Space, Linear Time Sequence Memoizer [2]



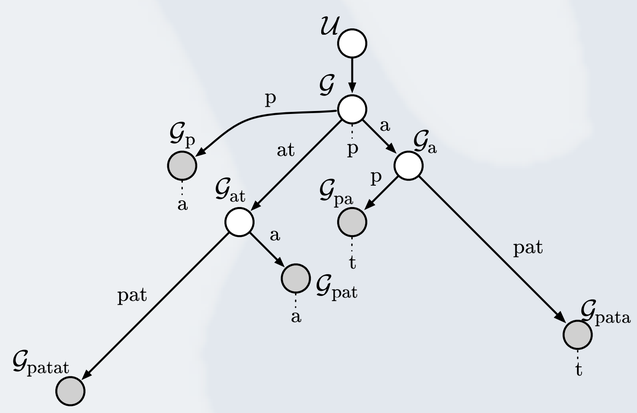
Speed of a lossless compressor based on the parametric sequence memoizer applied to the task of compressing the complete Wikipedia .xml corpus [5]. The speed of the compressor is plotted against the cumulative length of the input stream. After an initial warm-up period, compressor speed remains constant as the stream length increases (evidence of attained asymptotics). The compressor reduces the full 26.8Gb corpus to 4.0Gb, compared to 7.8Gb with gzip and 3.8Gb with paq9a.

More Data is Good



The performance of streaming deplump as a function of stream length was evaluated using the complete Wikipedia .xml dump [5]. For this result stream lengths ranging from 10^3 to 10^9 bytes were compressed using models of varying size (with customer count constrained to 8,196 and depth limited to 16).

Sequence Memoizer [6] Review

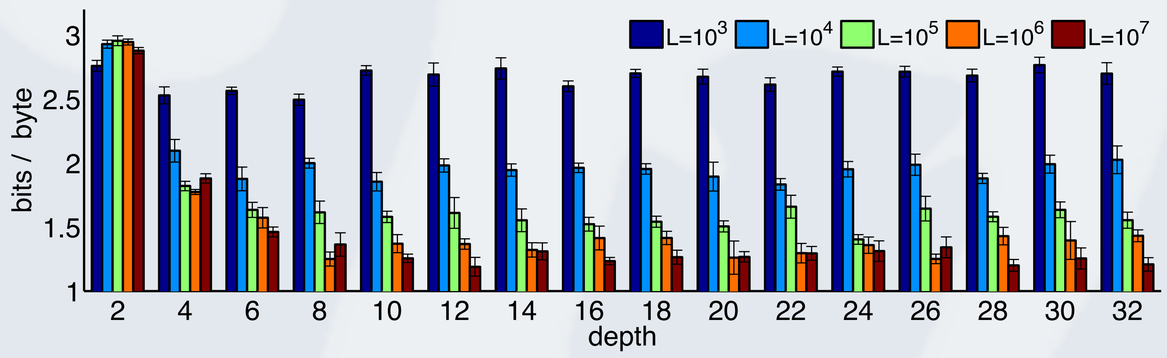


Main Point

Whac-a-Mole game play consists of using a hammer to bash the head of a mole that randomly pops up out of an array of holes. Modifying the sequence memoizer to work for streaming data was like playing a game of whac-a-mole with the nonparametric nature of the SM playing the role of the mole – it kept popping up, sometimes in unexpected ways, despite all our bashing. Ultimately we succeeded in producing an asymptotically scalable version of the SM for streaming data, however, to do so we had to abandon essentially all of its nonparametric nature. Intuitively the result can be thought of as an estimator for a complex, parametric model. The lessons learned from playing this game are obvious in retrospect and extend to all fundamentally nonparametric models where the data is the sufficient statistic: asymptotic scalability is a problem. For testing and statistical discovery of patterns in finite datasets, perhaps Bayesian nonparametric models have a role to play; however, our experience suggests a no-free-lunch result: it seems that one can not abandon sufficiency and have scalability too.

Mole	Hammer
Space complexity of original sequence memoizer was of the order of the number of nodes in the suffix-tree representation of the input sequence which grows as a function of stream length.	Dependent hierarchical Pitman-Yor process [1]: suffix-tree graphical model approximated by graphical model with node count of asymptotically constant order
Storage requirement at each node was an uncharacterized but not-constant function of the input sequence length	Collapsed restaurant franchise representation [4]: restaurant storage constrained be of constant order per node
Computational cost of inference grew as a super-linear function of the length of the training sequence	[2] introduced approximations that rendered the cost of inference asymptotically linear in the observation sequence length and cost of storage asymptotically constant in the same.

Bigger Models are Good



Here 10, 100MB chunks of Wikipedia were sampled with replacement and compressed using models limited to depths varying from 2 to 32. Each group contains results for models with a different upper limit L on the node count of the data structure. As expected, larger models generally perform better. Using a larger depth appears advantageous up to depth ≈ 16.

References

[1] Bartlett, N.; Pfau, D. & Wood, F. Forgetting Counts: Constant Memory Inference for a Dependent Hierarchical Pitman-Yor Process. Proceedings of the 27th International Conference on Machine Learning, 2010, 63-70.

[2] Bartlett, N.; Wood, F. Deplump for Streaming Data. Data Compression Conference, 2011, 363-372.

[3] Gasthaus, J.; Wood, F. and Teh, Y. W. Lossless compression based on the Sequence Memoizer. Data Compression Conference, 2010, 337-345.

[4] Gasthaus, J. and Teh, Y. W. Improvements to the Sequence Memoizer. Proceedings of Neural Information Processing Systems, 2011, 685-693.

[5] Wikipedia, 2010. URL: http://download.wikimedia.org/enwiki/.

[6] Wood, F., Archambeau, C., Gasthaus, J., James, L., and Teh, Y. W. A stochastic memoizer for sequence data. In Proceedings of the 26th International Conference on Machine Learning, 2009, 1129-1136.