Streams, Sketching and Big Data – References

Graham Cormode
G.Cormode@warwick.ac.uk

July 11, 2014

1 Lecture 1: Sketches and Frequency Moments

1.1 General Background on streams and concentration bounds

- Survey of streaming algorithms: S. Muthukrishnan. *Data Streams: Algorithms and Applications*. Now Publishers, 2005
- Survey of sketching algorithms: G. Cormode. Sketch techniques for massive data. In G. Cormode, M. Garofalakis, P. Haas, and C. Jermaine, editors, *Synposes for Massive Data: Samples, Histograms, Wavelets and Sketches*, Foundations and Trends in Databases. *NOW* publishers, 2012
- Textbook on randomized algorithms: R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995
- Textbook on randomized algorithms: M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005

1.2 Sketches for frequencies

- Count-Min sketch: G. Cormode and S. Muthukrishnan. An improved data stream summary: The Count-Min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005
- Count sketch: M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Procedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, 2002
- AMS sketch: N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *ACM Symposium on Theory of Computing*, pages 20–29, 1996
- "Fast" AMS sketch: M. Thorup and Y. Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *ACM-SIAM Symposium on Discrete Algorithms*, 2004
- Lower bound for Johnson-Lindenstrauss: J. Nelson and H. L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *ACM Symposium on Theory of Computing*, pages 101–110, 2013
- "Hash kernels" in machine learning: K. Q. Weinberger, A. Dasgupta, J. Langford, A. J. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *International Conference on Machine Learning (ICML)*, 2009

1.3 Sketches for F_0

- Flajolet-Martin algorithm: P. Flajolet and G. N. Martin. Probabilistic counting. In *IEEE Conference on Foundations of Computer Science*, pages 76–82, 1983. Journal version in *Journal of Computer and System Sciences*, 31:182–209, 1985

- K-Minimum values for F_0: Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisian. Counting distinct elements in a data stream. In *Proceedings of RANDOM 2002*, pages 1–10, 2002

- Hyperloglog algorithm: P. Flajolet, E. Fusy, O. Gandouet, and F. Meunier. Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. In *International Conference on Analysis of Algorithms*, 2007

- Subset size estimation via KMV: M. Thorup. Bottom-k and priority sampling, set similarity and subset sums with minimal independence. In *ACM Symposium on Theory of Computing*, pages 371–380, 2013

1.4 Extensions

- Combined frequency moments: G. Cormode and S. Muthukrishnan. Space efficient mining of multigraph streams. In *ACM Principles of Database Systems*, 2005

- Range efficient F_0: A. Pavan and S. Tirthapura. Range-efficient counting of distinct elements in a massive data stream. *SIAM Journal on Computing*, 37(2):359–379, 2007

- Range efficient F_2: F. Rusu and A. Dobra. Fast range-summable random variables for efficient aggregate estimation. In *ACM SIGMOD International Conference on Management of Data*, 2006

- Rectangle efficient F_0: S. Tirthapura and D. P. Woodruff. Rectangle-efficient aggregation in spatial data streams. In *ACM Principles of Database Systems*, pages 283–294, 2012

2 Lecture 2: Advanced Topics

2.1 L_p Sampling

- Initial L_p sampling constructions: M. Monemizadeh and D. P. Woodruff. 1-pass relative-error $\mathrm{l_p}$-sampling with applications. In *ACM-SIAM Symposium on Discrete Algorithms*, 2010

- Improved L_p sampling constructions: H. Jowhari, M. Saglam, and G. Tardos. Tight bounds for lp samplers, finding duplicates in streams, and related problems. In *ACM Principles of Database Systems*, 2011

- Graph sketching: K. J. Ahn, S. Guha, and A. McGregor. Analyzing graph structure via linear measurements. In *ACM-SIAM Symposium on Discrete Algorithms*, 2012

- L_0 sampling (1): G. Frahling, P. Indyk, and C. Sohler. Sampling in dynamic data streams and applications. In *Symposium on Computational Geometry*, June 2005

- L_0 sampling (2): G. Cormode, S. Muthukrishnan, and I. Rozenbaum. Summarizing and mining inverse distributions on data streams via dynamic inverse sampling. In *International Conference on Very Large Data Bases*, 2005

- L_0 sampling evaluation: G. Cormode and D. Firmani. On unifying the space of ℓ_0-sampling algorithms. In *Algorithm Engineering and Experiments*, 2013

- Precision sampling: A. Andoni, R. Krauthgamer, and K. Onak. Streaming algorithms via precision sampling. In *IEEE Conference on Foundations of Computer Science*, 2011

2.2 Matrix Sketching

- Matrix sketching (1): T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *IEEE Conference on Foundations of Computer Science*, pages 143–152, 2006

- Matrix sketching (2): K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *ACM Symposium on Theory of Computing*, pages 205–214, 2009

- Matrix sketching (3): K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *ACM Symposium on Theory of Computing*, pages 81–90, 2013

- Compressed Matrix multiplication: R. Pagh. Compressed matrix multiplication. In *ITCS*, pages 442–451, 2012

2.3 Verification of Computation

- Freivald's algorithm: R. Freivalds. Probabilistic machines can use less running time. In *IFIP Congress*, pages 839–842, 1977

- Inner-product protocol: A. Chakrabarti, G. Cormode, and A. McGregor. Annotations in data streams. In *International Colloquium on Automata, Languages and Programming (ICALP)*, 2009

- Interactive proofs for muggles: S. Goldwasser, Y. T. Kalai, and G. N. Rothblum. Delegating computation: interactive proofs for muggles. In *ACM Symposium on Theory of Computing*, 2008

- Implementation and improvements on 'muggles': J. Thaler. Time-optimal interactive proofs for circuit evaluation. In *CRYPTO (2)*, pages 71–89, 2013

2.4 Lower bounds

- Communication complexity textbook: E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997

- Hardness of Matrix Multiplication: K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *ACM Symposium on Theory of Computing*, pages 205–214, 2009

- Hardness of Entropy: A. Chakrabarti, G. Cormode, and A. McGregor. A near-optimal algorithm for computing the entropy of a stream. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007