Senior Thesis Proposal

Ananya Kumar (ananyak)

September 6, 2016

I would like to work with Prof. Avrim Blum on algorithms for coresets and on the applications of coresets to machine learning.

1 Research Area

Given a set of points P, a coreset $Q \subseteq P$ approximates P with respect to some *extent measure*. The notion of a coreset depends on the definition of the extent measure. A simple type of coreset is an approximate convex hull, where every point in P is within distance ϵ from the convex hull of the coreset Q.

Coresets have numerous applications in computational geometry and machine learning. Blum et al explain how approximate convex hulls can be used for sparse non-negative matrix factorization, a useful technique in unsupervised learning, and give offline algorithms for finding approximate convex hulls [BHPR16]. Agarwal et al give a survey of computational geometry algorithms that can be approximately solved using coresets [AHPV05].

2 Research Questions

In my senior thesis I intend to work on the following questions:

1. [**Streaming Algorithm**] Devise a streaming algorithm for approximate convex hulls (or other types of coresets) that is asymptotically optimal in terms of the number of points stored. To simplify the problem, we could assume that the point set P is generated in a certain specified way.

2. [**Topic Recovery**] Assuming that the points in P are generated by taking convex combinations of k points in a point set T, recover the set T.

3. [**Parallel Reconstruction**] Given a point p in the convex hull of Q, we can efficiently find a sparse convex combination of points in Q that sum to p. However, existing algorithms that do this are inherently sequential - can we find an efficient parallel algorithm?

4. [**Supervised Learning Applications**] Coresets learn the structure of a data set - can we use this in meaningful ways in supervised learning tasks?

3 Research Plan

In a previous independent study I came up with basic results for the topic recovery problem, and examined a supervised learning application of coresets. I will build on this work and expect to have algorithms with proofs for some of the above research questions.

I will begin by reading [AHPV05] for a general survey of coresets, [HS08b], [HS08a], [Cla93] for previous work on streaming algorithms for coresets, and [TKC05], [BLK15], [FSS13] for applications of coresets to machine learning. By the end of October 2016 I will conjecture algorithms for some of the research questions, and by the end of December 2016 I will have proofs for some of these conjectures.

References

[AHPV05] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *COMBINATORIAL AND COMPUTATIONAL GEOMETRY, MSRI*, pages 1–30. University Press, 2005.

[BHPR16] Avrim Blum, Sariel Har-Peled, and Benjamin Raichel. Sparse approximation via generating point sets. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 548–557, 2016.

[BLK15] Olivier Bachem, Mario Lucic, and Andreas Krause. Coresets for nonparametric estimation - the case of dp-means. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 209–217. JMLR Workshop and Conference Proceedings, 2015.

[Cla93] Kenneth L. Clarkson. *Algorithms for polytope covering and approximation*, pages 246–252. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

[FSS13] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1434–1453, Philadelphia, PA, USA, 2013. Society for Industrial and Applied Mathematics.

[HS08a] John Hershberger and Subhash Suri. Adaptive sampling for geometric problems over data streams. *Comput. Geom. Theory Appl.*, 39(3):191–208, April 2008.

[HS08b] John Hershberger and Subhash Suri. *Simplified Planar Coresets for Data Streams*, pages 5–16. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[TKC05] Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast svm training on very large data sets. *J. Mach. Learn. Res.*, 6:363–392, December 2005.