Research Statement

Anup Bhattacharya

The primary motivation in my research is to design polynomial time approximation algorithms for computationally *hard* problems. I also want to design efficient algorithms for problems in various models of interest like online, streaming etc. Showing lower bounds is also an area of interest.

Problems studied and Results

Algorithms based on D^2-sampling Given dataset $X \subseteq \mathbb{R}^d$ and integer k, the goal in k-means problem is to find a set C of k centers that minimizes the Euclidean sum of squares cost. k-means++ uses D^2-sampling (points sampled with probability proportional to squared Euclidean distance) to choose centers and gives $O(\log k)$-approximation in expectation for k-means [3]. Our results on D^2-sampling based algorithms for k-means are summarized as follows.

- k-means++ seeding was conjectured to yield $O(\log d)$-approximation with high probability on d-dimensional instances [11]. We refuted the conjecture by giving construction of *bad* instances [8].
- Ding and Xu [14] gave a polynomial time approximation scheme (PTAS) for constrained k-means problem. We designed a simpler D^2-sampling based PTAS with much better running time [9].
- (Under submission) Given $\epsilon > 0$, a number l is computed such that Euclidean sum of squares cost with l centers sampled using D^2-sampling is at most ϵ times the optimal k-means cost. These l centers form a (k, ϵ)-*coreset*. We also give a D^2-sampling based heuristic to estimate *intrinsic dimension* of data.

Streaming algorithms for Sampling We designed a uniform sampling algorithm in the streaming setting [7]. Our algorithm uses $O(\log n)$-random bits, matching the randomness used by any offline algorithm.

Clustering with Oracle Ashtiani et al. [4] gave an efficient algorithm for k-means on well-separated datasets in a semi-supervised setting given a same-cluster oracle. Our results in this setting are as follows.

- (Under submission) We designed a $(1 + \epsilon)$-approximation for k-means without any separation assumption, and provided almost matching upper and lower bounds on the query complexity [1].
- (Under submission) Similar upper and lower bounds on query complexity were obtained for $(1 + \epsilon)$-approximation for correlation clustering where the number of optimal clusters is given.
- We are working on upper bounds where the query oracle is allowed to err with some probability $q < 1/2$.

Future Work

I am interested in working on problems of both theoretical and practical importance, and willing to learn new things. Some problems of particular interest are mentioned below.

Cluster Recovery This is inspired by the clustering with faulty oracle problem. Exact recovery algorithms are known for noisy correlation clustering [15] and stochastic block model [12] but they require all clusters to have size at least $\Omega(\sqrt{n})$ where n is the number of vertices. We don't know of any lower bound on cluster sizes for exact recovery. It is interesting to show lower bounds or improve the recovery guarantees.

Clustering with Stability assumptions If the dataset satisfies *stability* conditions [10], then clustering as well as some *hard* combinatorial optimization problems become easy [2]. Awasthi et al. [5] recovered *ground truth* clustering by showing integrality of convex relaxations given that the dataset is generated in a specific manner. It is interesting to explore whether convex relaxations of clustering problems become integral when the instances are stable. It is also interesting to study the same under weak-stability conditions [6].

Testing Clusterability of Instances There are efficient algorithms [2] for stable instances of clustering problems. But how do we know whether an instance is stable or not? Czumaj et al. [13] studied cluster structure of graphs from property testing perspective. It would be interesting to know more on this.

References

[1] Nir Ailon, Anup Bhattacharya, Ragesh Jaiswal, and Amit Kumar. Approximate clustering with same-cluster queries. *arXiv preprint arXiv:1704.01862*, 2017.

[2] Haris Angelidakis, Konstantin Makarychev, and Yury Makarychev. Algorithms for stable and perturbation-resilient problems. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 438–451. ACM, 2017.

[3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[4] Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. In *NIPS*, pages 3216–3224, 2016.

[5] Pranjal Awasthi, Afonso S Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200. ACM, 2015.

[6] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102–155, 2016.

[7] Anup Bhattacharya, Davis Issac, Ragesh Jaiswal, and Amit Kumar. Sampling in space restricted settings. *Algorithmica*, 2017.

[8] Anup Bhattacharya, Ragesh Jaiswal, and Nir Ailon. Tight lower bound instances for k-means++ in two dimensions. *Theoretical Computer Science*, 634:55–66, 2016.

[9] Anup Bhattacharya, Ragesh Jaiswal, and Amit Kumar. Faster Algorithms for the Constrained k-Means Problem. In *33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016)*, volume 47, pages 16:1–16:13, 2016.

[10] Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(5):643–660, 2012.

[11] Tobias Brunsch and Heiko Röglin. A bad instance for k-means++. *Theoretical Computer Science*, 505:19–26, 2013.

[12] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research*, 15(1):2213–2238, 2014.

[13] Artur Czumaj, Pan Peng, and Christian Sohler. Testing cluster structure of graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 723–732. ACM, 2015.

[14] Hu Ding and Jinhui Xu. A unified framework for clustering constrained data without locality property. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1471–1490. Society for Industrial and Applied Mathematics, 2015.

[15] Claire Mathieu and Warren Schudy. Correlation clustering with noisy input. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 712–728. SIAM, 2010.