Academic Statement

Thomas Dybdahl Ahle

September 2018

Images, sound or genomic sequences are all objects that modern computers can handle because of one key idea: the idea of high dimensional geometry. Even something as diverse as a tweet can be written as a (long) vector where the ith entry specifies how many copies the tweet contains of the ith word in the English language. Looking at data as points in a geometry allows us to use the concept of distance to represent similarity and of isoperimetry to understand inherent structure in data domains. Different applications lead to different geometries, and only recently did we fully understand the most basic – the Euclidean geometry.

In my thesis I worked on multiple problems related to search and data structures. In [Ahle et al., 2016] we showed that there are certain limits of how fast you can search in any geometry. We did this by showing that the existence of a very fast algorithm would allow other too fast algorithms in the fundamental computational area of boolean satisfiability, breaking the so called Strong Exponential Time Hypothesis. The work was extended in [Abboud and Rubinstein, 2017] to become an entire field of study. Another key idea in high dimensional algorithms is randomization. It turns out that even simple high dimensional problems like packing spheres tightly is very hard mathematically. Meanwhile simply placing the spheres at random locations works very well, this has problems related to the predictability and fairness of the computation. Today all the most efficient algorithms in the field has some chance of failing without any chance of verifying their result! In [Ahle, 2017] I found a way to solve this problem for the two most common geometries, and in [Wei, 2018] the results were extended to cover the Euclidean case as well. My most recent work includes a unified approach to LSH for distances on set/boolean data and explicit feature embeddings of polynomial kernels, giving new state of the art results in linear methods in machine learning.

Research Plan At Columbia I will tackle the most important problem in search: Edit Distance. The edit distance between two strings like SIMON and SALOON is the minimum number of insertions, deletions or substitutions required to turn one into the other. In the particular case, the answer is 3: delete an O and change AL into SI. Given a database of strings and a query, we would like to quickly retrieve the most similar string in the database. This problem is deeply linked to natural language processing [Sidorov et al., 2015] and computational biology [McGrane and Charleston, 2016], and as of yet a very big computational mystery.

Practical solutions include transforming strings into sets (using so called k-mers) in which case my recent work [Ahle, 2019] gives the state of the art algorithm. Another approach embeds the strings into a geometry called L1 [Jowhari, 2012, Ostrovsky and Rabani, 2005] which is related to the Euclidian geometry and allows fast search. Unfortunately both transformations distort the edit distance a great deal, which makes the approach less than ideal. In fact for L1 [Krauthgamer and Rabani, 2009] showed that there is no way around this.

I believe that a more direct approach will yield much better algorithms, leading to many break through from machine learning to biology. We know from the very recent paper [Andoni et al., 2018] that all geometric search can be viewed through so-called non-linear spectral cuts. I will study these quantities from both an upper and lower bound perspective. For upper bounds, new analytical methods developed in my recent papers allow studying candidate algorithms that this far has been a mystery. For lower bounds, new hypercontractive inequalities I have discovered give a new view at spectral cuts. Omri Weinstein is the top most expert on lower bounds for data structures, and working with him, and other experts and Columbia University, will make this project much more likely to succeed.

,

References

[Abboud and Rubinstein, 2017] Abboud, A. and Rubinstein, A. (2017). Distributed PCP theorems for hardness of approximation in P. *CoRR*, abs/1706.06407.

[Ahle, 2017] Ahle, T. D. (2017). Optimal las vegas locality sensitive data structures. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 938–949. IEEE.

[Ahle, 2019] Ahle, T. D. (2019). Subsets and supermajorities: Unifying hashing-based set similarity search. *arXiv preprint arXiv:1904.04045*.

[Ahle et al., 2016] Ahle, T. D., Pagh, R., Razenshteyn, I., and Silvestri, F. (2016). On the complexity of inner product similarity join. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 151–164. ACM.

[Andoni et al., 2018] Andoni, A., Naor, A., Nikolov, A., Razenshteyn, I., and Waingarten, E. (2018). Data-dependent hashing via nonlinear spectral gaps. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 787–800. ACM.

[Jowhari, 2012] Jowhari, H. (2012). Efficient communication protocols for deciding edit distance. In *European Symposium on Algorithms*, pages 648–658. Springer.

[Krauthgamer and Rabani, 2009] Krauthgamer, R. and Rabani, Y. (2009). Improved lower bounds for embeddings into l_1. *SIAM Journal on Computing*, 38(6):2487–2498.

[McGrane and Charleston, 2016] McGrane, M. and Charleston, M. A. (2016). Biological network edit distance. *Journal of Computational Biology*, 23(9):776–788.

[Ostrovsky and Rabani, 2005] Ostrovsky, R. and Rabani, Y. (2005). Low distortion embeddings for edit distance. In *STOC*, pages 218–224.

[Sidorov et al., 2015] Sidorov, G., Gómez-Adorno, H., Markov, I., Pinto, D., and Loya, N. (2015). Computing text similarity using tree edit distance. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*, pages 1–4. IEEE.

[Wei, 2018] Wei, A. (2018). Optimal las vegas approximate near neighbors in ℓ_p. *arXiv preprint arXiv:1807.07527*.