Academic Statement

Thomas Dybdahl Ahle

September 2018

The key idea that has let today's computers efficiently work with complicated objects like images, sound, or genomic sequences, is the idea of high dimensional geometry. Even something like 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 concepts of distance to represent similarity and isoperimetry to prove optimal ways of organizing data. 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 a very fast algorithm for any of a number of problems would allow other too fast algorithms in the fundamental computational area of boolean satisfiability, breaking the so called Strong Exponential Time Hypothesis. Another key idea in high dimensional algorithms is randomization. It turns out that even simple high dimensional problems like packing sphere tightly is very hard mathematically. Meanwhile simply placing the spheres at random locations gives a near optimal tightness. This however has problems related to the predictability and fairness of the computation. 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.

Besides the published work, I am currently working on a unified approach to LSH for distances on set/boolean data. This involves new hyper-geometric bounds in boolean functions. I am also working on explicit feature embeddings of polynomial kernels.

Research Plan A unified theory of geometric search. Machine learning has brought a

From the beautiful paper [Andoni et al., 2018b] we now have a general approach to data structures for so called normed spaces. It also handles other things via those spectral cut things. Just need to upper and lower bound it. I really need to read it.

We now have really promising results in LSH for symmetric norms [Andoni et al., 2018a]. It is however not clear if their approximation factors are optimal, and lots of work still has to be done before this important work becomes near practical - or the trade-offs are properly understood. Going beyond normed spacers, there are a large number of metrics we don't have any good data structures for. Important examples are edit distance and earth mover distance, which are prevailing in both computational biology and natural language processing. Other metrics are implicitly induced by neural networks, and similar modern machine learning architectures.

Simple things like sparse data still needs lots of work.

Better analysis of algorithms we already know. Exciting? ... What is Rasmus vision? Making computation fair? Mikkel? Making randomness real? Unification Information theory?

The Medium dimensional regime. Recent results [Chan, 2017] have shown that classical data structures - here kd-trees - can be analysed more tightly when the dimension is close to $\log n$. At the same time LSH algorithms have been shown in [Becker et al., 2016] to perform somewhat better in this range. Unifying the these two approaches is a major open problem with big implications for how data is processes

in practice. From a theoretical side, proving optimality in this range requires new, sharper bounds on the noise stability of boolean functions than what is currently known.

Deterministic LSH and limited randomness. In most of randomized algorithms, we have a good understanding on the trade-offs between randomized and deterministic variants, and the importance of high quality random bits, k-independence, tabulation hashing etc. A natural continuation of my work in [Ahle, 2017] is to make a completely deterministic LSH data structure with little or no loss in the various performance parameters.

Nearest Neighbours beyond LSH. While modern LSH data structures have been improved using so called "data dependency" [Andoni and Razenshteyn, 2015, Andoni et al., 2018c], the basic algorithm hasn't changed since Indyk and Motwani. Using LSH for Approximate Closest Pair yields a $n^{1-\Omega(\epsilon)}$ algorithm, but we know that algebraic algorithms allow an $n^{1-\Omega(\epsilon^{1/3})}$ algorithm [Alman et al., 2016]. It is a very interesting open problem whether these techniques generalizes to data structures, or conversely, if lower bounds can be shown separating Closest Pair from Nearest Neighbour.

At Columbia I plan to work with Omri Weinstein as well as a number of world class researchers in the fields of data structures and high dimensional geometry.

What do I want to do

How do I want to do it with Omri? Omri Weinstein is an expert in proving complexity results about data structures.

We might be able to show new results using information theory.

References

[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 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.

[Alman et al., 2016] Alman, J., Chan, T. M., and Williams, R. (2016). Polynomial representations of threshold functions and algorithmic applications. *CoRR*, abs/1608.04355.

[Andoni et al., 2018a] Andoni, A., Krauthgamer, R., and Razenshteyn, I. P. (2018a). Sketching and embedding are equivalent for norms. *SIAM J. Comput.*, 47(3):890–916.

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

[Andoni et al., 2018c] Andoni, A., Naor, A., Nikolov, A., Razenshteyn, I. P., and Waingarten, E. (2018c). Data-dependent hashing via nonlinear spectral gaps. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 787–800.

[Andoni and Razenshteyn, 2015] Andoni, A. and Razenshteyn, I. P. (2015). Optimal data-dependent hashing for approximate near neighbors. *CoRR*, abs/1501.01062.

[Becker et al., 2016] Becker, A., Ducas, L., Gama, N., and Laarhoven, T. (2016). New directions in nearest neighbor searching with applications to lattice sieving. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 10–24. Society for Industrial and Applied Mathematics.

[Chan, 2017] Chan, T. M. (2017). Orthogonal range searching in moderate dimensions: k-d trees and range trees strike back. In *33rd International Symposium on Computational Geometry, SoCG 2017, July 4-7, 2017, Brisbane, Australia*, pages 27:1–27:15.

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