Academic Statement

Thomas Dybdahl Ahle

September 2018

There is a gap between modern algorithmic theory and what gets used in practical machine learning. Contrast this with classical computing, where time proven algorithms are taught from 'Introduction to Algorithms", and implemented routinely. A large reason for this is the so called "curse of dimensionality", which happens when machine learning datasets are mapped into high-dimensional feature vector spaces. Classical algorithms for organizing data tend to fail in this setting due to an exponential time and/or space dependency on the dimension.

I want to make machine learning efficient both theoretically and practically. My research has focused on algorithms that are scalable, less error prone, and on finding the limits of what can be achieved computationally. The resulting algorithms have often been simple, even though the analysis may be mathematically complex.

During the last two years of my PhD I co-founded a startup called SupWiz, specialized in developing Natural Language Processing services. For more than a year, I was Chief Machine Learning Officer, primarily responsible for developing our chatbot program, which we developed and rolled out to three different customers during this time. Since rejoining academia, this has been a source of inspiration for problems to work on as well as a ton of valuable lessons in planning and teamwork, scaling a company from four to seventeen employees in a year.

1 Research Achievements

Much of my published work considers some form of approximate near neighbour search. The reason we study approximate near neighbours - rather than exact - is the mentioned "curse of dimensionality". Anything nontrivial for the exact version would imply a breakthrough in the k-satisfiability problem. Meanwhile the approximate version turns out to be feasible using so called Locality Sensitive Hashing (LSH) in which one designs a good partition of the vector space which is used for bucketing of the data points. In practice the resulting algorithms are often useful for exact nearest neighbour as well, since in many pratical data sets the nearest neighbour is nearest by a large margin.

Sketching. In TODO, together with Jakob Tejs Bæk Knudsen we show how to make Tensor Sketch much faster.

Impossibility. In [Ahle et al., 2016] I showed together with Rasmus Pagh, Francesco Silvestri, and Ilya Razenshteyn that even approximate nearest neighbour can be hard. (Meaning queries have to take time equal to brute force, given sub-exponential space for the data structure.) We showed this for the problem of "Maximum Inner Product" search, which is used in many practical applications such as multi-label linear classifiers. We showed this by reducing the approximate inner product search problem to the Strong Exponential Time Hypothesis (SETH), which had previously been used for reductions from other exact problems, but not from approximate ones. This work was later extended by the seminal work [Abboud and Rubinstein, 2017], which today lays the foundation for all approximative SETH lower bounds.

LSH Without false negatives. Since Indyk and Motwani introduced LSH, it was a major open problem if the randomized algorithms could be made deterministic. Other than a 3-approximation algorithm by Indyk in [Indyk, 2007] no sub-exponential algorithms are known. In [Ahle, 2017] I showed the partial result, that LSH can be made Las Vegas. That is, we guarantee that if any point in the data structure is within a given range, it is returned. This is in contrast to classical LSH (and all sub-exponential space data structures) which inherently suffers from a positive probability of false negatives. A prior algorithm of Rasmus Pagh [Pagh, 2016] achived a similar result, showing that Las Vegas approximate hamming distance near neighbour can be solved with query time roughly $n^{1.38/c}$. In my work I achived the optimal $n^{1/c}$ and with a general approach, which also gives the optimal query time for many other distances.

Output sensitivity. In [Ahle et al., 2017], together with Rasmus Pagh and Martin Aumüller we show how to achieve query time close to $n^\varepsilon + t$ where t is the number of returned points. Previous versions of LSH all require tn^ε which for many practical purposes is orders of magnitudes slower. Our algorithms have the side benefit of Parameter freeness, with which the optimal LSH parameters are chosen dynamically on a per query basis for no extra cost. This contrasts with classical hyper parameter tuning of LSH, which optimizes towards the average case query. Techniques inspired by our work was later used for so called "confirmation sampling".

The work on Las Vegas data structures helped me win one of two Travel Scholarships by the Stibo Foundation in 2016. This allowed me to travel to Austin for half a year to work with Eric Price at The University of Texas in Austin. With Eric Price I worked on using Semidefinite programming for solving robust learning problems, such as gaussian regression. A project that unfortunately

didn't end up fruitful, however, while there, I got to work with the other students on problems, such as expanders and the theory of boolean functions. Some of which work is still in the pipeline.

2 Research Agenda

I am interested in broadly in sketching, streaming, robust optimization, clustering, boolean functions, dimensionality reduction, fine grained lower bounds, random matrix theory, differential privacy and many other related areas. The common thread being a certain statistical and/or geometric intuition. Besides the published work, I am currently working on a unified approach to LSH for distances on set/boolean data. This involves new hypergeometric bounds in boolean functions. I am also working on explicit feature embeddings of polynomial kernels.

In the coming years, I would further like to pursue the following research directions:

The Medium dimensional regime. Recent results [Chan, 2017] have shown that classical data structures - here kd-trees - can be analyzed 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 funcitons than what is currently known.

Deterministic LSH and limited randomness. In most of randomized algorithms, we have a good understanding on the tradeoffs 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.

Unified theory of LSH metrics. 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 tradeoffs 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.

Nearest Neighbours beyond LSH. While modern LSH data structures have been improved using so called "data dependency" [Andoni and Razenshteyn, 2015, Andoni et al., 2018b], the ba-

sic 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.

3 Other work

During my undergrad at University of Oxford I was focused on programming languages and artificial intelligence. I started multiple open source chess engines, some projects of which I am still managing today. Eventually I got into programming competitions, which sparked my interest in algorithms. I since helped mentor young students wanting to take a similar path into theoretical computer science.

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 et al., 2017] Ahle, T. D., Aumüller, M., and Pagh, R. (2017). Parameter-free locality sensitive hashing for spherical range reporting. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 239–256. SIAM.

[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. P., and Waingarten, E. (2018b). 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.

[Indyk, 2007] Indyk, P. (2007). Uncertainty principles, extractors, and explicit embeddings of l2 into l1. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 615–620. ACM.

[Pagh, 2016] Pagh, R. (2016). Locality-sensitive hashing without false negatives. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 1–9. SIAM.