

## Research statement

Lee-Ad Gottlieb

The central theme unifying my research is the study of a broad spectrum of proximity problems – problems that address the relationship between points in space. My goal in pursuing this research is to understand how the underlying space hosting data points affects or limits the complexity of the relationship between these points. For example, it is known that many important geometric problems on high-dimensional Euclidean space suffer from the “curse of dimensionality,” wherein some feature of the solution (storage space, algorithm runtime, etc.) grows exponentially in the dimension. For these problems, a more profound understanding of the underlying space can potentially lead to tractable solutions.

My research initially focused on the areas of string matching [CGL04], and subsequently branched out to consider proximity algorithms and data structures for metric spaces [CG06, GR08a, GR08b, GK10, BGK<sup>+</sup>11, GK11] – including a recent breakthrough result concerning the well-studied **traveling salesman problem**. A special emphasis of my recent work relates to problems and applications for machine learning [GN10, GKK10, GKK11, GKM11]).

**String matching.** In [CGL04], we considered a family of classical problems in string matching, most notably the  $k$ -error **dictionary matching problem**, where an input dictionary of strings is preprocessed so that all  $k$ -error matches of a query string to the dictionary (under edit distance and wildcards) can be efficiently returned. We gave a data structure of roughly  $O(n \log^k n)$  space and query time (where  $n$  is the size of the dictionary). This paper is considered a breakthrough in the field, has been cited more than 100 times.

**Proximity algorithm for metric space.** A major motif of my research considers proximity problems in metric space of low *doubling dimension*. This program asks whether algorithm tailored for low-dimensional Euclidean space can be adapted for metric spaces of low intrinsic dimension. Following the framework introduced in [GKL03], given a metric space  $M$ , the *doubling constant*  $c = c(M)$  be the smallest value such any ball in  $M$  can be covered by  $c$  balls of half its radius. The doubling dimension of  $M$  is  $\dim(M) = \log_2 c$ , and this parallels the dimension of Euclidean space. We have adapted algorithms for Euclidean spaces to metric spaces with low doubling dimension:

- We first considered the classic **approximate nearest neighbor search problem**. In this problem, a point set must be preprocessed to return a near-neighbor of a query point. We presented a dynamic data structure that supports searches in time  $2^{O(\dim(M))} \log n$  using linear space (where  $n$  is the number of points) [CG06]. This structure generalizes the best known results for Euclidean spaces to metric spaces, improving upon [KL04, HM06]. I am pursuing an implementation of this work in conjunction with Ben Kimia (Brown University).
- We then considered the problem of **dynamic spanners**. A spanner for a point set is a graph on these points containing a subset of edges of the full graph, a construction commonly utilized in network simulations. The *stretch* of the spanner compares the spanner distance between two points to their true distance. We constructed a low-stretch and low-degree dynamic spanner that can be updated in time  $2^{O(\dim(M))} \log n$  per point [GR08a, GR08b]. This again matches for metric spaces the best known results for Euclidean space.

- We extended the techniques developed for spanner construction towards the creation of an **approximate distance oracle**. A distance oracle is a data structure that returns the interpoint distance between any point pair without explicitly storing all  $\Theta(n^2)$  distances. We gave the first approximate distance oracle that answers queries with arbitrary precision in (universal) constant time while using only near-linear space [BGK<sup>+</sup>11]. The oracle can further be made fully dynamic, and improves for metric space what was previously known for the more restrictive Euclidean space.
- In a very recent breakthrough, we applied the techniques in the above works to the famous **traveling salesman problem (TSP)**. The celebrated results of Arora [Aro98] and Mitchell [Mit99] gave polynomial time approximation schemes (PTAS) for Euclidean TSP. We show that a similar PTAS can be obtained for metric space with low doubling dimension, meaning that the dependence on Euclidean geometry in [Aro98, Mit99] is not necessary.

**Dimension reduction.** The most celebrated result in this field, the Johnson-Lindenstrauss Lemma, states that any (high dimensional)  $n$ -point Euclidean set can be projected into an  $O(\log n)$ -dimensional Euclidean space with arbitrarily small interpoint distortion. In [GK11], we considered Euclidean points residing in low doubling dimension, and used a spectrum of embedding techniques to reduce their Euclidean dimension to the doubling dimension. This may be viewed as an improvement over the Johnson-Lindrauss Lemma (albeit with the important caveat that our distortion features a “snowflake” property).

In recent work [BG11], we showed that  $l_p$  ( $1 < p < \infty$ ) sets with small aspect ratio admit dimension reduction into  $O(\log n)$  dimensions with arbitrarily low distortion (with dimension dependent on the aspect ratio). Point sets with small aspect ratio occur naturally in data sets associated with image recognition, and also as subproblems in nearest neighbor search and snowflake embeddings. An immediate consequence of our result is an improvement over the best known run times for nearest neighbor search for  $l_p$ .

**Machine learning.** A recent focus of my research has been in the area of machine learning, where the techniques developed in the above papers can be readily applied.

In [GN10], we considered the **matrix sparsification problem**, which seeks to minimize the number of non-zero entries in a matrix by using only elementary row reductions. This problem is of broad interest, having applications in machine learning, structural analysis, mathematical programming, and detecting time-series correlations. We are the first to show conditions under which this problem is tractable, and gave rigorous hardness bounds for the general case. One heuristic in this paper (an  $\ell_1$  solver) has already been implemented by Xuan Vinh Doan (University of Waterloo) for use in operations research, and separately by Daniel Cohen (Weizmann Institute of Science) for use in mathematical programming.

In [GKM11] we tackled the problem of determining the VC-dimension of nearly-orthogonal vectors. This problem is of interest in learning by statistical queries (where characterizations and algorithms consider nearly orthogonal or decorrelated function classes) and fits squarely within our purview of interest. We used random projections and a tight bound on the sum of binomial coefficients to improve upon a 15 year old result of Haussler [Hau95].

Finally, we turned our attention to **metric learning**, specifically towards efficiently learning classifiers for binary-labeled points residing in metric space, and evaluating a classifier for new data. By harnessing the proximity techniques developed in the above works, we gave the first metric classifier that can be efficiently determined and evaluated [GKK10] and extended this approach to address metric regression [GKK11].

## References

- [Aro98] Sanjeev Arora. Polynomial time approximation schemes for euclidean traveling salesman and other geometric problems. *J. ACM*, 45(5):753–782, 1998.
- [BG11] Y. Bartal and L. Gottlieb. Dimension reduction techniques for  $l_p$ ,  $1 < p < \infty$ . 2011. Manuscript.
- [BGK<sup>+</sup>11] Y. Bartal, L. Gottlieb, T. Kopelowitz, M. Lewenstein, and L. Roditty. Fast and precise distance queries. In *Proc. of Symposium on Discrete Algorithms (SODA)*, pages 840–853, 2011.
- [CG06] R. Cole and L. Gottlieb. Searching dynamic point sets in spaces with bounded doubling dimension. In *Proc. of ACM Symposium on Theory of Computing (STOC)*, pages 574–583, 2006.
- [CGL04] R. Cole, L. Gottlieb, and M. Lewenstein. Dictionary matching and indexing with errors and don’t cares. In *Proc. of ACM Symposium on Theory of Computing (STOC)*, pages 91–100, 2004.
- [GK10] L. Gottlieb and R. Krauthgamer. Proximity algorithms for nearly-doubling spaces. In *Proc. of APPROX*, pages 192–204, 2010.
- [GK11] L. Gottlieb and R. Krauthgamer. A nonlinear approach to dimension reduction. In *Proc. of Symposium on Discrete Algorithms (SODA)*, pages 888–899, 2011.
- [GKK10] L. Gottlieb, A. Kontorovich, and R. Krauthgamer. Efficient classification for metric data. In *Proc. of Conference on Learning Theory (COLT)*, pages 433–440, 2010.
- [GKK11] L. Gottlieb, A. Kontorovich, and R. Krauthgamer. Efficient regression in metric space via approximate lipschitz extension. 2011. Submitted to SIAM Journal on Computing (SICOMP).
- [GKL03] Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pages 534–543, 2003.
- [GKM11] L. Gottlieb, A. Kontorovich, and E. Mossel. VC bounds on the cardinality of nearly orthogonal function classes. *Discrete Mathematics*, 2011. pending minor revision.
- [GN10] L. Gottlieb and T. Neylon. Matrix sparsification and the sparse null space problem. In *Proc. of APPROX*, pages 205–218, 2010.
- [GR08a] L. Gottlieb and L. Roditty. Improved algorithms for fully dynamic geometric spanners and geometric routing. In *Proc. of Symposium on Discrete Algorithms (SODA)*, pages 591–600, 2008.
- [GR08b] L. Gottlieb and L. Roditty. An optimal dynamic spanner for doubling metric spaces. In *Proc. of the Annual European Symposium on Algorithms (ESA)*, pages 478–489, 2008.
- [Hau95] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vavnik-chervonenkis dimension. *J. Comb. Theory Ser. A*, 69(2), 1995.

- [HM06] S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics, and their applications. *SIAM Journal on Computing*, 35(5):1148–1184, 2006.
- [KL04] R. Krauthgamer and J. Lee. Navigating nets: Simple algorithms for proximity search. In *Proc. of Symposium on Discrete Algorithms (SODA)*, pages 798–807, 2004.
- [Mit99] Joseph S. B. Mitchell. Guillotine subdivisions approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric tsp, k-mst, and related problems. *SIAM Journal on Computing*, 28(4):1298–1309, 1999.