Data dependent LSH for ANN:

In pursuit of a simple yet near optimal algorithm

COMS 6998

Eliezer Zimble

Overview:

Survey of background for research goal and overview of approach:

1. **Introduction**: Nearest Neighbor search (**NNS**) & Approximate Near Neighbor search (**ANN**)

2. **Data-independent LSH:** Locality sensitive hashing quality & results.

3. **Data-dependent LSH**: Tradeoff of optimal quality vs. simplicity.

4. **Research goal**: Can we have both?

Nearest Neighbor Search:

Widespread interest, but challenging

- Goal: given a set *P* of *n* points in *d* dimensional space, a distance/similarity metric *D,* and query point *q,* we want to find the near or nearest neighbor $p \in P$.

- **Solutions**: We want to build a **data structure** to preprocess the dataset to allow us to **efficiently answer all queries** using as **little space as possible**.

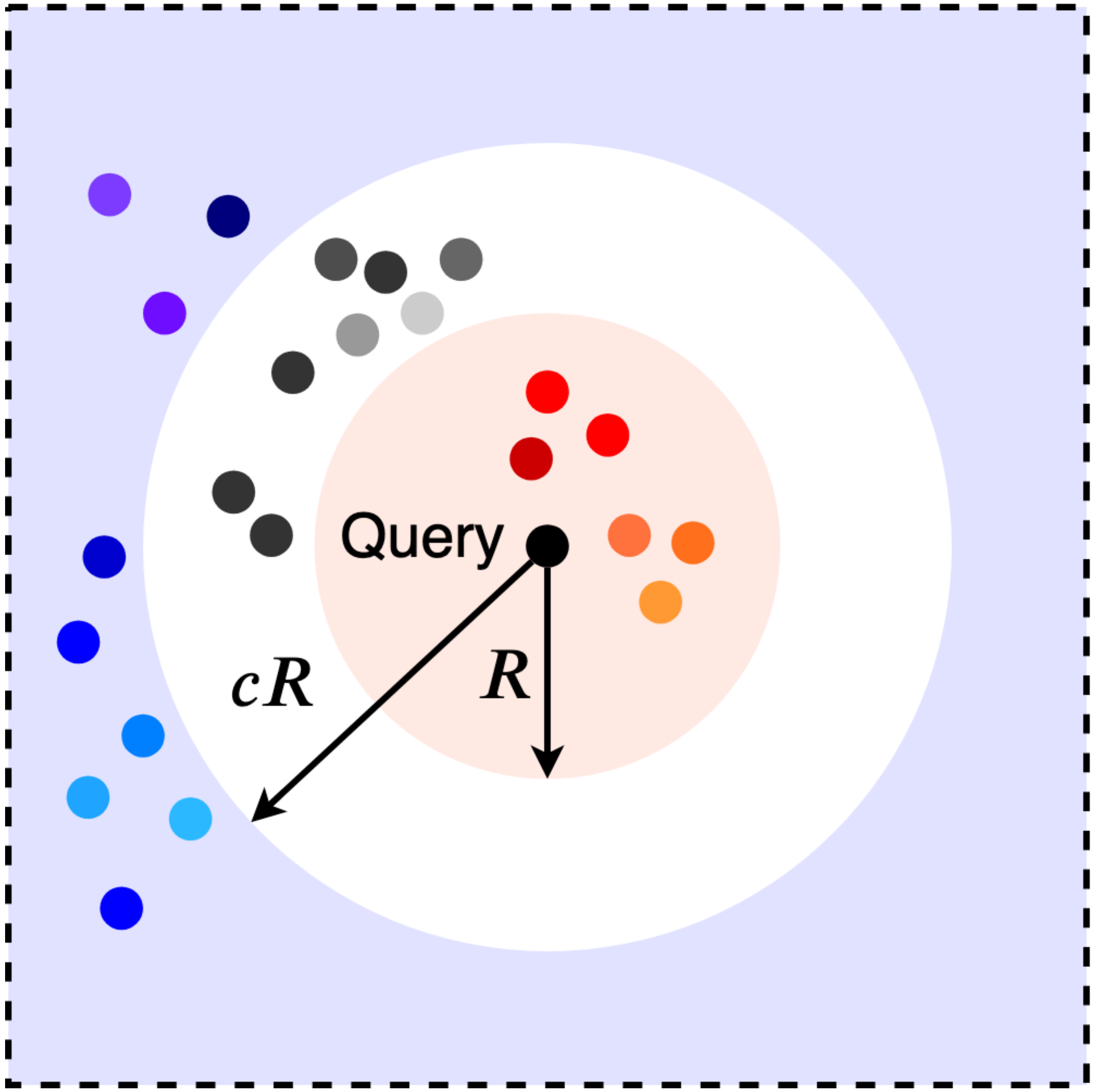
- Exact solution impractical for higher dimensional space as suffers from "curse of dimensionality": either space or query time exponential in *d.*

Approximate Near Neighbor

(c,r)-ANN: approximation factor *c*, distance *r*

- **New goal**: If a neighbor within "distance" *r* exists, then return any point within distance *cr*.

- As we use *randomized* algorithms, we aim for solutions that achieve the goal with high probability.



Credit: https://randorithms.com/2019/09/19/Visual-LSH.html

Locality Sensitive Hashing (LSH)

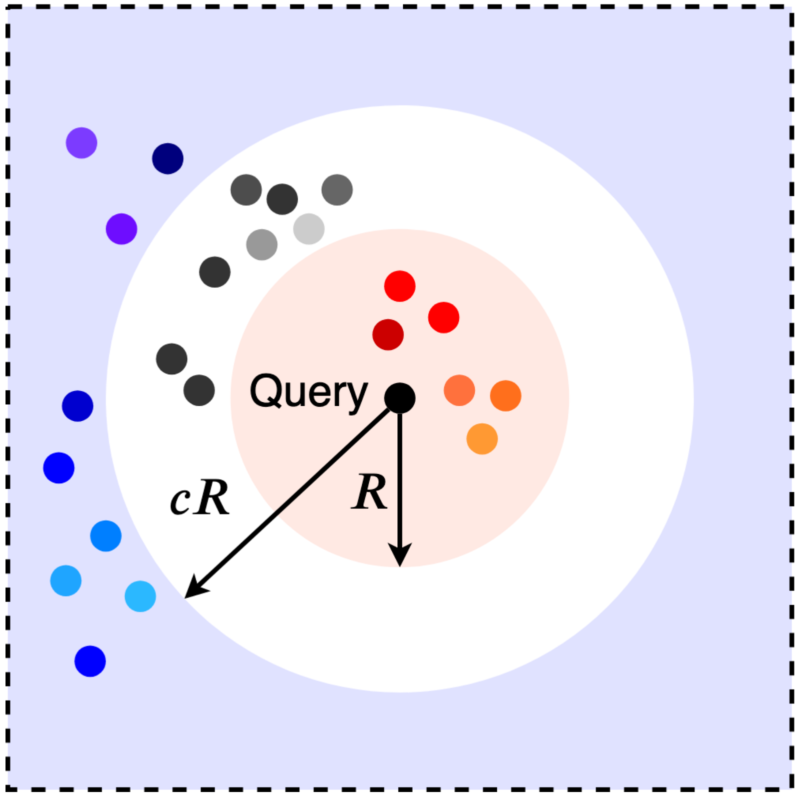
Equivalent to Space partitions

- **Intuition**: we want close points to have a high probability of hashing to the same bin, and far points to have a low probability.

- **Definition of LSH family**: A distribution $\mathcal{H}$ over hash functions *h* is (r, cr, p_1, p_2) -sensitive when for all points p, q:
  - $If\ D(p,q) \leq r,\ then\ Pr_h[h(p) = h(q)] \geq p_1$
  - $If\ D(p,q) > cr,\ then\ Pr_h[h(p) = h(q)] \leq p_2$

- **Quality of LSH family:**
  We define quality measure ρ:

$$\rho = \rho(\mathcal{H}) = \frac{log(1/p_1)}{log(1/p_2)}$$



LSH family theorem for ANN

Breakthrough result for (data independent) LSH

- **Theorem** [IM98,HIM12]**:** Given a (r, cr, p_1, p_2) **-**sensitive LSH family, there exists a (c,r)-ANN data structure with dn^ρ query time and using $n^{1+\rho}$ extra space.

- LSH families exist: partition on random coordinate (Hamming) or into Euclidean balls.

Bounds for data independent LSH:

- For Hamming space $\rho \geq 1/c - o(1)$
- For Euclidean space $\rho \geq 1/c^2 - o(1)$

Data dependent LSH offers better bounds!

Data dependent LSH

Definition:

- Data-dependent LSH uses the given dataset P to generate a randomized hash family.
- Our focus is on algorithms for datasets with **no assumed structure.**
- **Mind-blowingly**, even an arbitrary dataset has some structure that we can exploit to achieve better results!

LSH data dependent: optimal ρ

Less practical, optimal quality ρ

- **<u>LSH Quality</u>:** [AR15] achieve an optimal $\rho = 1/(2c^2 - 1)$ for Euclidean space for all $c > 1$.
  - (For $c = 2$ this amounts to improving the query time from $n^{1/4+o(1)}$ to $n^{1/7+o(1)}$)
- **Main idea**: reduce ANN on a generic dataset into ANN on a randomly distributed dataset:
  - The "good case" for data-independent LSH is when the points are randomly distributed on a unit sphere.
  - [AR15] perform the reduction by decomposing the dataset into "dense clusters" and "pseudo-random" remainders.
- **Drawback:** The algorithm is impractical as the decomposition is complex and hard to implement in practice.

LSH data dependent: simple

More practical algorithm, but suboptimal ρ

- **<u>Algorithm</u>**: [ARS17] offer a **simpler algorithm**:
  - Use LSH to build a decision tree and augment each node with a constant number of “pivot” points. Iterate down the tree until finding a near neighbor of the query.
  - To boost success probability, the process is repeated $O(n^\rho)$ number of times.

- **Drawback**: **Suboptimal <u>LSH Quality</u>** as $\rho \approx \dfrac{1}{ln(4) \cdot c}$

Research goal:

Can we design a simple algorithm with near optimal ρ?

- Overview of my approach: For “nice” dataset on Euclidean unit sphere, the algorithm modifies the “simple” approach of ARS17 using a Euclidean space LSH family with quality $\rho \to 1/c^2$

- Analysis: work in progress!

References

[AIR18] Alexandr Andoni, Piotr Indyk and Ilya Razenshteyn. Approximate Nearest Neighbor Search in High Dimensions. arXiv 1806.09823, 2018.

[AR15] Alexandr Andoni and Ilya Razenshteyn. Optimal data-dependent hashing for approximate near neighbors. In Proceedings of the Symposium on Theory of Computing (STOC), 2015.

[ARS17] Alexandr Andoni, Ilya Razenshteyn, and Negev Shekel Nosatzki. Lsh forest: Practical algorithms made theoretical. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), 2017.

[HIM12] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. Theory of computing, 8(1):321–350, 2012.

[IM98] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the thirtieth annual ACM symposium on Theory of computing, pages 604–613. ACM, 1998.