Problem Set 2

Due May 28th, 2021

Problem 1. (No-outlier ANN) Design the parameters of the LSH data structure for r-Near Neighbor, such that for any query point q, with constant probability, the following hold:

- For any point in the r-neighborhood of the query $p \in N(q, r)$, there exists one of the query buckets that has p in it.
- Any point p that is hashed to the same bucket as the query in at least one of the hash functions is not an outlier, i.e., $\bigcup_{i \leq L} g_i(q) \subseteq B(q, cr)$

What will be the query times and space usage of the algorithm for the Euclidean metric in this case?

Problem 2. Give an algorithm for Set Cover in the sub-linear query model, that has a constant factor approximation, and makes $\tilde{O}(mn/k)$ number of queries. Here k is the size of the minimum set-cover.

Problem 3.

- i) Show that to sample a vector v uniformly from a $(d-1)$-dimensional unit sphere in d dimensions, it is enough to let each of the d coordinates of v to be taken independently from a normal distribution $N(0, 1)$, and then normalize the resulting vector by its norm.
- ii) Show that if we take two vectors u and v uniformly at random from the surface of a $(d-1)$ dimensional sphere, then $Pr[|\langle u, v \rangle| > \epsilon] \leq \exp(-\Theta(\epsilon^2 m))$.