Seminar on Algorithms and Geometry 2014B
Lecture 2 – Doubling metrics and Nearest Neighbor Search (cont'd)*

Robert Krauthgamer

1 NNS algorithm

We continue in our proof of the following theorem. (We already saw the preprocessing algorithm.)

Theorem: One can preprocess a subset $S \subset M$ of size n, and build a data structure of size $2^{O(\mathrm{ddim}\, S)} \cdot n$, so as to answer $(1+\varepsilon)$-NNS queries (for every $\varepsilon < 1/2$) in time $(1/\varepsilon)^{O(\mathrm{ddim}\, S)} \cdot \log \Phi(S)$.

Query procedure (for $q \in M$):

Idea: maintain a set Z_i of level i net points "near" the query

1. Initialize $Z_m = Y_m$.

2. Iteratively for $i = m - 1$ going down to 0

3. Let Z_i contain all the points "pointed to" from Z_{i+1}, i.e., $\cup_{y \in Z_{i+1}} L_{y,i+1}$, that are within distance $(7/\varepsilon)2^i$ from q.

4. If $d(q, Z_i) \geq (3/\varepsilon) \cdot 2^i$ then return the point attaining this.

Query time: The number of iterations is $\log \Phi(S)$. In each iteration, we scan the pointers coming out of Z_{i+1}; We already bounded the size of any single list, and we can bound $|Z_i| \leq 2^{O(\mathrm{ddim}\, S)}$ by the packing lemma. Overall, the query time is $(1/\varepsilon)^{O(\mathrm{ddim}\, S)} \log \Phi(S)$.

To prove correctness of the query algorithm, we show it maintains a certain invariant.

Lemma (Invariant): Every set Z_i constructed by the query algorithm is exactly $Y_i \cap B(q, (7/\varepsilon) \cdot 2^i)$.

Proof: Was seen in class, by induction on i.

Correctness of the output: Let a^* be an optimal NNS, i.e., $d(q, a^*) = d(q, S) = OPT$. Consider the iteration i in which the stopping condition is met. (formally, it might never be met, as discussed below.) Then a^* is covered by some point $y \in Y_i$, and we get that

$$d(q, y) \leq (q, a^*) + d(a^*, y) \leq OPT + 2^i.$$

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

As seen in class, we can use the invariant to prove that $OPT > (1/\varepsilon)2^i$. We can also show that $y \in Z_i$, because the stopping condition was not met at the previous iteration $i+1$. Hence, the algorithm reports y or an even closer one, and then $d(q,y) \leq OPT + 2^i \leq (1+\varepsilon)OPT$.

Improved storage: The data structure can be stored using $2^{O(\operatorname{ddim} S)}n$ words.

Proof: As seen in class, we count how many lists $L_{y,i}$ are non-trivial in the sense that they contain at least other point than y itself. Bounding this number is done by charging each list to some point in the list, that is chosen carefully so that every point is charged $O(1)$ times in total.

Exer: Show that the following version of the query algorithm achieves $O(1)$-approximation: Instead of keeping a set Z_i, we keep only one point z_i, where each z_i is the closest point to q among $L_{z_{i+1},i+1}$.

Exer: Suppose we modify the preprocessing, as follows. At each level i, if some point z appears in more than one list $L_{y,i}$, then remove all but one of its occurrences. This means that for every i, the sets $L_{y,i}$ for $y \in Y_i$ are disjoint. Show that this reduces the preprocessing storage to $O(n)$, and that the query procedure still works. (Some changes in constants might be needed.)

2 Spanners of Doubling Spaces

We can think of a metric space (M,d) as a complete graph on vertex set M with edge weights.

Defn: A k-spanner of an (edge-weighted) graph $G=(V,E)$ is a subgraph $G'=(V,E')$ such that

$$\forall u,v \in V, \quad d_{G'}(u,v) \leq k \cdot d_G(u,v).$$

We also say that G' is a spanner of G with stretch k.

Applying this definition, a spanner of a metric space (M,d) is a collection of edges $E' \subset \binom{|M|}{2}$ (with weights according to d) such that $d_{G'}$ approximates d.

Theorem: For every $\varepsilon \in (0,1/2)$, every finite metric space (M,d) admits a $(1+\varepsilon)$-spanner with at most $(1/\varepsilon)^{O(\operatorname{ddim} M)}n$ edges.

Exer: Prove this theorem.

Conjecture: For every $\varepsilon \in (0,1/2)$, every finite metric space (M,d) admits a $(1+\varepsilon)$-spanner with total edge-weight $(1/\varepsilon)^{O(\operatorname{ddim} M)}\,\mathrm{MST}(M)$.

What is currently known: The known bound has an additional $O(\log n)$ term [Smid 2009]. In the special case of Euclidean k-dimensional metrics, the conjecture is true with $(1/\varepsilon)^{O(k)}\,\mathrm{MST}(M)$ [Das-Narasimhan-Salowe'95, Arya-Das-Mount-Salowe-Smid'95].