CS 532 Homework 5

Qilin Ye

November 22, 2024

Solution to problem 1. We first define a hierarchical grid data structure that we use to store the points. Specifically, let $\mathcal{S}_1$ be a collection of $(1/\epsilon)^d$ d-dimensional cubes, such that they are placed in a lattice manner (adjacent, and aligned in all coordinate directions), and that their union cover all points in S. Say its side length is s_0. Iteratively, let $\mathcal{S}_j$ be the partition of the space into smaller cubes, arranged in a lattice order (adjacent, and aligned in all coordinate directions) of side length $s_0(1+\epsilon)^{-j}$. We repeat till $\mathcal{S}_L$ has side length no more than $d_{\min}\epsilon$, where $d_{\min}$ is the closest distance between all pairs of points. By the assumption that S has spread $n^{\mathcal{O}(1)}$, we know $L = \mathcal{O}(\log_{1+\epsilon} n) \leqslant \mathcal{O}(\log_{\exp(\epsilon)} n) = \mathcal{O}(\epsilon^{-1}\log n)$. We discard empty cubes.

Now, for query, we check the growing neighborhoods of query point q whose radii form a geometric sequence of ratio $1+\epsilon$. Define the neighborhood/ball $B(x,r) := \{y \in \mathbb{R}^d : \|x-y\| < r\}$.

We start with $s = d_{\min}$ and $B(q,s)$. In the first iteration, for each cube in $\mathcal{S}_L$ that intersects with $B(q,s)$, we check it contains a point $p_i \in S$ (it may or may not due to the stochastic nature of S), and if so, return any point found. For a cube C with side length ℓ that contains a point p_i, an easy necessary (but not sufficient) criterion for $C \cap B(q,s) \neq \varnothing$ is that $\|p_i - q\| - \ell < s$. We can use this to quickly determine the boxes that are completely disjoint from $B(q,s)$.

Analogously, in iteration i we check each of the boxes in $\mathcal{S}_{L-i}$ that intersect with the annulus $B(q,(1+\epsilon)s)\backslash B(q,s)$, and if nothing is found, update $s \leftarrow (1+\epsilon)s$. As long as there are points in S to begin with, this algorithm will terminate in no more than $L = \mathcal{O}(\epsilon^{-1}\log n)$ iterations, where in each iteration it checks up to $\mathcal{O}(\epsilon^{-d})$ boxes that lay on the annulus. Because we are inspecting cubes of side length $s(1+\epsilon)^i\epsilon$ in a d-dimensional annulus with inner radius $s(1+\epsilon)^i$, the number of candidate cubes is $\mathcal{O}(1/\epsilon^d)$. This gives the desired runtime $\mathcal{O}(\epsilon^{-(d+1)}\log(n))$.

Suppose our algorithm returns at iteration i. This means no point exists within distance $(1+\epsilon)^i d_{\min}$ of q, but there is some point p lying in a cube with radius $(1+\epsilon)^i d_{\min}\epsilon$ that intersects with the outer radius $(1+\epsilon)^{i+1}d_{\min}$. By triangle inequality,

$$\|p-q\| \leqslant (1+\epsilon)^{i+1}d_{\min} + \sqrt{2}(1+\epsilon)^i d_{\min}/\epsilon = (1+\epsilon)^i d_{\min}\cdot(1+\epsilon+\sqrt{2}\epsilon) \leqslant (1+3\epsilon)d^*_{\min}$$

where the true nearest neighbor must have distance $d^*_{\min} \geqslant (1+\epsilon)^i d_{\min}$.

Finally, we discretize the expected value and write

$$\Delta(q) = \mathbb{E}\|q - \mathrm{NN}(q,S)\| = \sum_{p_i\in S}\mathbb{P}(p_i = \mathrm{NN}(q,S))\cdot\|q-p_i\|.$$

Applying the bound to each individual conditionals, we obtain a $(1+3\epsilon)$-approximation as claimed. (This be modified into a $1+\epsilon$-approximation by choosing finer cubes and balls that increase slightly slower.)

Solution to problem 2. (1) For this problem, as in MAX-CUT we pick a random d-dimensional vector v so that it partitions $\mathcal{S}^d$ into two hemispheres $\{u \in \mathcal{S}^d : u^T v \leqslant 0\}$ and $\{u \in \mathcal{S}^d : u^T v > 0\}$. We simply define h_v to be the indicator function $h_v(u) = \mathbf{1}[u^T v > 0]$.

We have shown in the proof of MAX-CUT that given two vectors $u_1, u_2 \in \mathcal{S}^d$, the probability that $h_v(u_1) \neq h_v(u_2) = \theta_{u_1,u_2}/\pi$ where θ_{u_1,u_2} is the angle between them. Therefore, it follows directly from definition that h_v is $(r, (1+\epsilon)r, 1-r/\pi, 1-(1+\epsilon)r/\pi)$-sensitive:

$$\mathbb{P}(h_v(u_1) = h_v(u_2) \mid d_{\mathcal{S}}(u_1,u_2) \leqslant r) = 1 - \frac{d_{\mathcal{S}}(u_1,u_2)}{\pi} \geqslant 1 - \frac{r}{\pi}$$

and likewise for the other inequality.

(2) Consider an arrangement of d-dimensional grids with side length $\ell = 2r$. As suggested by the hint, we consider a hashing based on a randomly shifted grid. That is, let $\epsilon = (\epsilon_1, \cdots, \epsilon_d)$ where each ϵ_i is drawn uniformly from $(0, 2r)$, and let x, y be hashed to the same value if they belong to the same box, i.e., if $h_i(x_i) = h_i(y_i)$ for all i, where $h_i(z) = \lfloor (z + e_i)/\ell \rfloor$. For each individual dimension, $\mathbb{P}(h_i(x_i) = h_i(y_i)) = 1 - |x_i - y_i|/\ell$. Let $\ell = r/d$.

It follows that for each coordinate, $\mathbb{P}(h_i(x_i) = h_i(y_i)) = 1 - |x_i - y_i|/\ell$, so $\mathbb{P}(h(x) = h(y)) = \prod_{i=1}^{d}(1 - |x_i - y_i|/\ell)$.

If $\|x - y\|_1 \leqslant r$, to minimize the product above, we concentrate all values of $|x_i - y_i|$ onto one coordinate, yielding a product of $1 - r/d$.

Conversely, if $\|x - y\|_1 \geqslant (1 + \epsilon)r$ then certainly for some i, $|x_i - y_i| \geqslant (1 + \epsilon)r/d > r/d > \ell$, so there is no way to hash x, y into the same grid.

Combining the results above we conclude that such a function is $(r, (1 + \epsilon)r, r/d, 0)$-sensitive.

Solution to problem 3. (1) Recall that the determinant of a $n \times n$ matrix is simply given as a linear (with coefficients ± 1) combination of the product of n terms, where each term belongs to a different row and column. In other words, each term in this expansion corresponds to a perfect matching $M : [n] \to [n]$ by mapping the row indices to column indices. It then becomes clear that the determinant is nonzero if and only if there exists a nonzero term in this expansion, if and only if a perfect matching exists.

(2) To check if $Q \not\equiv 0$, we use the polynomial identity testing discussed in lecture (notes). We will choose a suitable interval I and sample each $x_{i,j}$ independently and randomly at uniform from integers in I. Repeat this k times, and conclude $Q \equiv 0$ if all of these trials result in 0. With appropriate choices we may conclude with probability at least $1/2$.

Solution to problem 4. Since ϵ is arbitrary, the JL-lemma also implies the existence of a linear mapping A preserving the squared norm up to a factor of ϵ:

$$(1 - \epsilon)\|x - y\|^2 \leqslant \|f(x) - f(y)\|^2 \leqslant (1 + \epsilon)\|x - y\|^2.$$

Likewise,

$$(1 - \epsilon)\|x + y\|^2 \leqslant \|f(x) + f(y)\|^2 \leqslant (1 + \epsilon)\|x + y\|^2.$$

In particular, comparing the difference between the red and the blue terms gives

$$\begin{aligned}(1 + \epsilon)\|x + y\|^2 - (1 - \epsilon)\|x - y\|^2 &= 2\epsilon\|x\|^2 + 2\epsilon\|y\|^2 + 4\epsilon\langle x, y\rangle \\ &\geqslant \|f(x) + f(y)\|^2 - \|f(x) - f(y)\|^2 = 4\langle f(x), f(y)\rangle.\end{aligned}$$

To show that JL-lemma preserves angles up to a factor of ϵ, it suffices to show that it $(1 \pm \epsilon)$-preserves dot products between unit vectors. To this end assume $\|x\| = \|y\| = 1$, so the above gives

$$4\epsilon + 4\epsilon\langle x, y\rangle \geqslant 4\langle f(x), f(y)\rangle \implies \langle x, y\rangle + \epsilon \geqslant \langle f(x), f(y)\rangle.$$

Likewise, by contrasting the difference of the blue terms against the two unused terms, we obtain

$$\langle f(x), x(y) \rangle \geqslant \langle x, y \rangle - \epsilon.$$

By a suitable choice of ϵ (likely different from the original ϵ provided in the formulation), we conclude that JL embedding also preserves angles arbitrarily well, as $\langle x, y \rangle - \epsilon \leqslant \langle f(x), f(y) \rangle \leqslant \langle x, y \rangle + \epsilon$.