AA Lecture 8. 2/4/21.

Last time: Dimension Reduction

JL: distribution over $\varphi: \mathbb{R}^d \to \mathbb{R}^k$,
s.t. $\forall x, y \in \mathbb{R}^d$:

$$\Pr_{\varphi}\left[\|\varphi(x)-\varphi(y)\| = (1\pm\varepsilon)\cdot\|x-y\|\right] \geq 1 - e^{-\varepsilon^2 k/3}.$$

$\Pr[-(-)] \geq 1 - 1/n^3$ for $k = O\left(\frac{\lg n}{\varepsilon^2}\right)$.

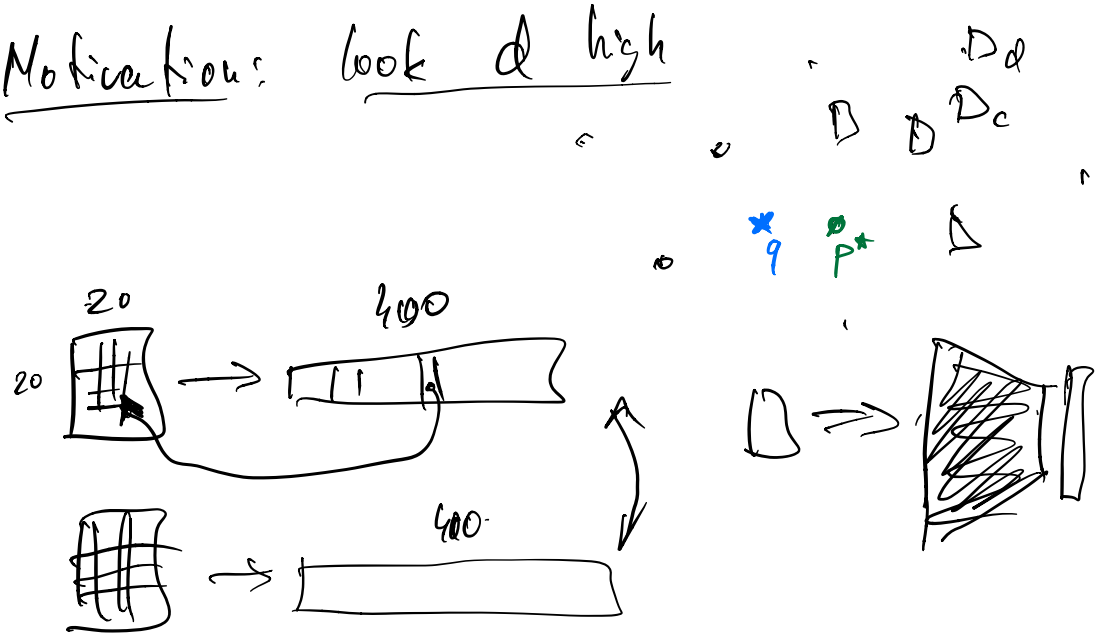
$\Rightarrow$ embed n pts in $\mathbb{R}^d$ into $\mathbb{R}^k$ preserving dist $1\pm\varepsilon$,
$k = O(\lg n/\varepsilon^2)$.

(metric embeddings).

Nearest Neighbor Search

Def: (NNS): given $D \subset \mathbb{R}^d$, $|D| = n$,
preprocess D s.t. can answer q's:
given $q \in \mathbb{R}^d$, output closest $p^* \in D$,
$\arg\min_{p \in D} \|q - p\|$.

Motivation: look at high



Performance measures of NNS:

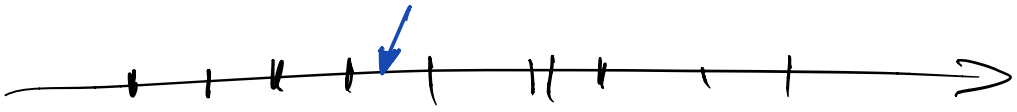
- space
- query time.

Solution #0: no preprocessing "lazy".

- space: $O(n \cdot d)$ ← good.
- q.t.: $O(n \cdot d)$. ← bad

Solution #0.1: $d = 1$.

Algo: sort all D. @query q, binary search.



space: $O(n)$.

q.t.: $O(\lg n)$. ← good.

Solution #0.2: $d = 2$

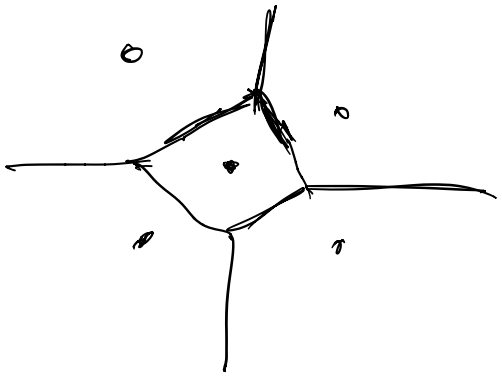
Voronoi diagram:

size = $O(n)$.

+ point location d.s.

- space: $O(n \cdot \lg n)$.

- q.t.: $O(\lg n)$.



$d >> 2$?

Voronoi d. size = $n^{\lfloor \frac{d}{2} \rfloor}$

curse of dimensionality.

Thm: if can solve NNS in dimension $d = (\lg^2 n)$ with $n^{O(1)}$ preproc. time & space, and query time $n^{0.99}$, then SETH is false.

↳ Strong Exponential Time Hypothesis.

"we can't solve CNF on n vars in time $\ll 2^n$".

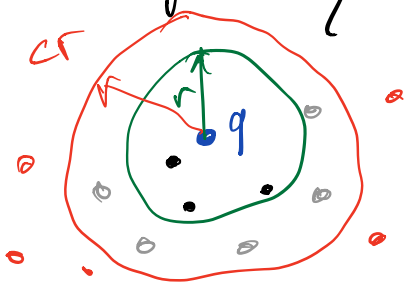
Approximate NNS in high-d.

Def: (r-near neighbor): given $r>0$, $D \subset \mathbb{R}^d$, build data str s.t. given query $q \in \mathbb{R}^d$, report *any* $p \in D$ s.t. $\|q-p\| \le r$.

"threshold".

Def: (c-approx r-near neighbor, c-ANN): given $c>1$, $r>0$, dataset $D \subset \mathbb{R}^d$, $n = |D|$, build d.s. on D s.t. can answer:

given $q \in \mathbb{R}^d$: - if $\exists p^* \in D$ s.t. $\|p^*-q\| \le r$, report $p \in D$ s.t. $\|p-q\| \le cr$. (w.p. $\ge 90\%$)



- if $\nexists p^* \in D$ s.t. $\|p^*-q\| \le r$,

may or not report it

Remark: 1) formal reduction from approx nearest neighbor to approx near n.

idea: guess r.

2) usually, algo for ANN can be modified to report LCD s.t:
- if $p \in D$ is $\|p-q\| \le r \Rightarrow p \in L$ with prob. $\ge 90\%$ per pt.
- if p is $\|p-q\| \ge cr \Rightarrow p \notin L$.

3) randomized algos,

Back to c-ANN:

Improving over:
space: $O(nd)$
q.t.: $O(nd)$

Solution #1:

Use dimensionality reduction (JL).

Algo: - pick φ at random $k = O\left(\frac{\log n}{\varepsilon^2}\right)$.
- store $\varphi(p)$ for all $p \in D$.
- @query q: compute $\varphi(q)$,

compute $\|\varphi(p) - \varphi(q)\|$ for $p \in D$.

Space: $O(nk) = O\left(n \frac{\lg n}{\varepsilon^2}\right)$

q. t.: $O(dk) + O(n \cdot k) = O\left(n \cdot \frac{\lg n}{\varepsilon^2}\right)$.

approx: $c = 1+\varepsilon$. (assuming $n \gg d$).

Correctness analysis:

Using Corollary of JL on points

$D \cup \{q\}$ $(N = n+1)$

$\Rightarrow$ distances $p \in D$, q are all preserved up to $(1+\varepsilon)$ factor, with prob. $\geq 1 - 1/N$.

$k = O\left(\frac{\lg N}{\varepsilon^2}\right) = O\left(\frac{\lg n}{\varepsilon^2}\right)$.

Crucially used φ is oblivious.

$\Rightarrow$ report an approx nearest neighbor with prob. $\geq 1 - 1/N \geq 1 - 1/n$.

---

what we used here:

JL.

exist φ : original space/metric $\downarrow$ "lower complexity" | $\mathbb{R}^d$ low-dim $\mathbb{R}^k$.

s.t. φ is defined on entire metric.

& $\forall p, q$, have that

distance b/w $\varphi(p)$ & $\varphi(q)$

$\propto$ distance b/w p & q.

with prob. $\geq 1 - 1/n^3$.

JL: $\varphi : (\mathbb{R}^d, \ell_2) \to (\mathbb{R}^k, \ell_2)$

Enough to use sketching map (instead of dim-reducing map).

Thm (sketching for ℓ_1): $\exists$ distrib over φ

s.t. $\varphi : \mathbb{R}^d \to \mathbb{R}^k$, & estimation algo

$E : \mathbb{R}^k \times \mathbb{R}^k \to \{0,1\}$:

$\forall x, y \in \mathbb{R}^d$, - if $\|x-y\|_1 \leq r \Rightarrow E(\varphi(x), \varphi(y)) = 0$

- if $\|x-y\|_1 \geq (1+\varepsilon)r \Rightarrow$

$$E(\varphi(x), \varphi(y)) = 1.$$

with probability $\geq 1-\delta$,

$$k = O\left(\frac{\lg 1/\delta}{\varepsilon^2}\right).$$

Thm: $\varphi: \ell_1 \to (\ell_2)^2$ preserving distances.

Consider $\{0,1\}^d$ with ℓ_1 distance

$\Rightarrow$ Hamming space

Thm: Fix d, $r \geq 1$, approx $1+\varepsilon$.

$\exists$ distrib. $\varphi: \{0,1\}^d \to \{0,1\}^k$,

$k = O\left(\frac{\lg n}{\varepsilon^2}\right)$ s.t.:

- if $\|x-y\|_1 \leq r \Rightarrow \|\varphi(x)-\varphi(y)\|_1 \leq \tau k$
- if $\|x-y\|_1 \geq (1+\varepsilon)r \Rightarrow$ — " — $\geq \left(\tau + \frac{\varepsilon}{9}\right)k$.

τ is a universal constant.

$\Rightarrow (1+\varepsilon)$ - ANN for Hamming space with same params as sol #1 for ℓ_2.