Approximate Nearest Neighbor using Locality Sensitive Hashing

IM (1998)

Based on Lecture Notes by Sariel Har-Peled

Presented by

K.V.N. Sreenivas, PhD Student, IISc Bengaluru

The Nearest Neighbor Problem

→ **Input:** Set of points P in a metric space $(\mathcal{X}, \mathrm{dist})$

→ **Goal:** Preprocess them such that the following queries can be efficiently solved.

Query: Given a point $q \in \mathcal{X}$, output the point $p \in P$ such that $\mathrm{dist}(p, q)$ is minimized.

The Nearest Neighbor Problem

Trade off between Preprocessing Space and Query time.

One extreme:-

Don't preprocess (constant preprocessing space)

$|P| =: n$ query time

Other extreme:

Compute Voronoi Diagram ($n^{d/2}$ space required)

Constant Query time.

Approximate Nearest Neighbor (ANN)

Input : A set of points P in a metric space $(\mathcal{X}, dist)$

Goal : Preprocess set P s.t. for query $q \in \mathcal{X}$, return $p \in P$

$$dist(q, p) \leq (1+\epsilon)\, dist(q, p^*)$$

where

p^* = nearest neighbor of q in P $=: nn(q)$.

Approximate Nearest Neighbor (ANN)

We will assume $X \subseteq \mathbb{R}^d$.

Naive method: nd query time

Goal: $\text{Sub}(n) \cdot d$ query time with reasonable space

This Presentation:

Query time: $\tilde{O}\left(d\, n^{\frac{1}{1+\epsilon}}\right)$ whp

Preprocessing space: $\tilde{O}\left(nd + n^{\frac{1}{1+\epsilon}}\right)$

A Simpler Problem: Proximity Neighbor (PN)

$PN(q, r)$:

For a given query point q:

1) If there is $p \in Ball(q, r)$, return it.
2) If no point in $Ball(q, r)$, say so.

Approximate Proximity Neighbor (APN)

Fix accuracy parameter ϵ.

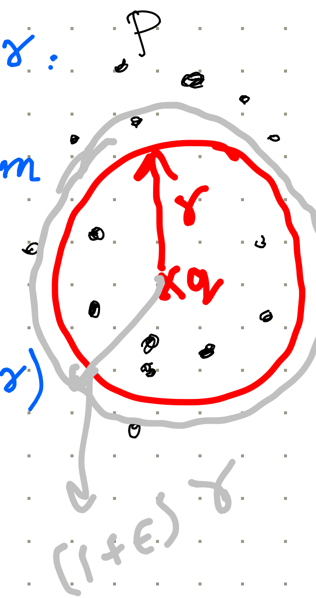
APN(q, r):

For a query point q, and a radius r:

1) If there is a point in Ball(q, r), return some point in Ball$(q, (1+\epsilon) r)$.

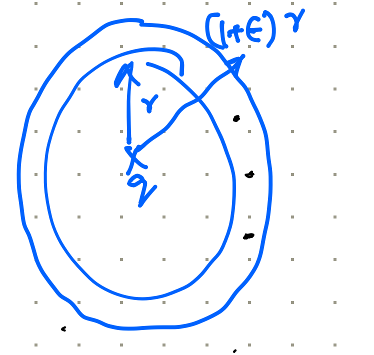
2) If all points are outside Ball$(q, (1+\epsilon) r)$ then report "No point in Ball(q, r)."

3) The intermediate case is immaterial.



ANN $\cong$ APN

APN $\leq$ ANN :



Input to APN : query q, radius r

Directly solve the ANN problem with query q.

Returns p s.t. $dist(p, q) \leq (1+\epsilon)\, dist(nn(q), q)$

Two Cases : Case 1 : $dist(p, q) \leq (1+\epsilon) r$. Then we return q

Case 2 : $dist(p, q) > (1+\epsilon) r \implies dist(nn(q), q) > r$

Return "No point inside the ball of radius r".

ANN $\cong$ APN

$[1, \text{diam}(P)]$

$\text{ANN} \leq \left(\log\left(\log_{1+\epsilon} \text{diam}(P)\right)\right) \text{APN}$

$\downarrow$ $\log(k)$

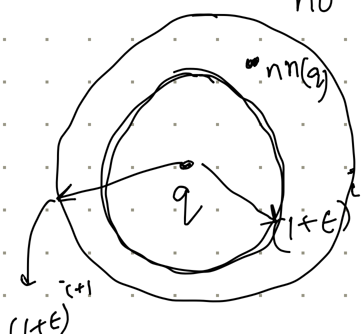
$\text{dist}(q, nn(q))$

2^k

r: 1, $(1+\epsilon)$, $(1+\epsilon)^2$, ..., $(1+\epsilon)^i$, $(1+\epsilon)^{i+1}$, $(1+\epsilon)^{i+2}$, ..., $(1+\epsilon)^{\log_{1+\epsilon} \text{diam}(P)}$

no, no, no, yes/no, yes, ..., yes

$$\boxed{\log\left(\log_{1+\epsilon} \text{diam}(P)\right)}$$



$$\text{dist}(p, q) \leq (1+\epsilon)^{i+2} = (1+\epsilon)^2 (1+\epsilon)^i$$

$$\leq (1+3\epsilon)\, \text{dist}(nn(q), q)$$

SUMMARY SO FAR

Nearest Neighbor (Hard) → Proximity Query

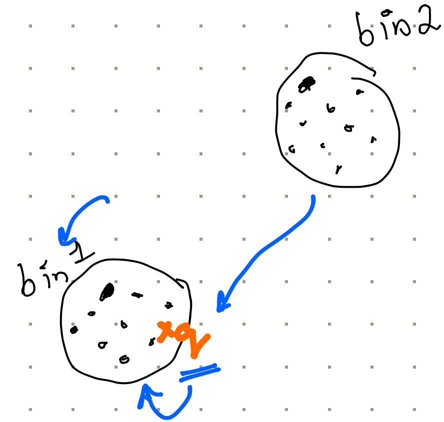
↓ ↓

Approximate Nearest Neighbor - ANN ≈ Approximate Proximity Query - APQ

IDEAL SCENARIO

Store points in P in a hash table such that

- Close points → Same bin
- Far points → different bins.



LOCALITY SENSITIVE HASHING (LSH)

A family $\mathcal{F}$ of hash functions is said to be (r, R, α, β)-sensitive if for any $p, q \in \mathcal{U}$, we have the following:

A) If $q \in \text{Ball}(p, r)$, then $\mathbb{P}_{f \sim \mathcal{F}}[f(p) = f(q)] \geq \alpha$

B) If $q \notin \text{Ball}(p, R)$, then $\mathbb{P}[f(p) = f(q)] \leq \beta$ $\;1-\beta$

We also require $\boxed{\alpha > \beta.}$

AN LSH FAMILY FOR $(\{0,1\}^d, \|\cdot\|_1)$

The universe is the corners of d-dim unit hypercube.

$$f_i(p) = i^{th} \text{ coord of } p \text{ for all } i \in [d]$$

Claim: $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ is (r, R, α, β)-sensitive where

$$\alpha = 1 - \frac{r}{d} \qquad \beta = 1 - \frac{R}{d} \qquad (\text{for } r < R,\ \alpha > \beta).$$

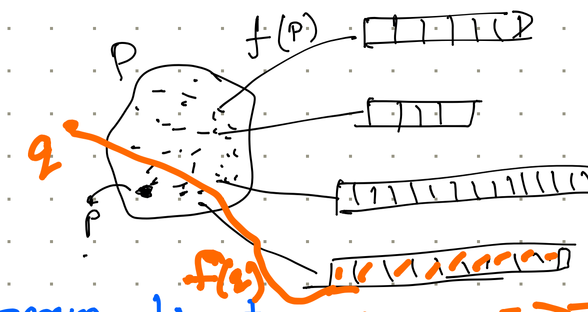
Proof: (A) $\|p-q\|_1 \leq r$. We chose $f_i \sim \mathcal{F}$. $\mathbb{P}[f_i(p) = f_i(q)] \geq 1 - \frac{r}{d}$

(B) $\underbrace{\|p-q\|_1 \geq R}_{d-R}$. $\mathbb{P}[f_i(p) = f_i(q)] \leq 1 - \frac{R}{d}$

NAIVE APG$^{(r)}_{=}$ ($\exists\ \mathcal{F}$ which is $(r,(1+\epsilon)r,\alpha,\beta)$-LSH)

Preprocessing (The point set is P)

For every $p \in P$, we compute $f(p)$ where $f \sim \mathcal{F}$.



For a query q.

We compute $f(q)$

We search through the list/bucket corresponding to $f(q)$ for a point p s.t. $\text{dist}(p,q) \le r(1+\epsilon)$.

If such point p exists, return p.

O/w, say "no point in $\text{Ball}(q,r)$".

NAIVE APQ (Analysis)

Suppose query q.

Expec Time to answer query q :- in bucket corresp. to $f(q)$

Once we encounter a point p, s.t. $dist(p,q) \leq (1+\epsilon)r$.

# of points p s.t. $dist(p,q) > (1+\epsilon)r$:

the prob that p, q match to same bucket

(i.e, $f(p) = f(q)$) is at most β. (β is a constant).

So, # $\leq$ $\beta(n)$ [(constant) n]

NAIVE APQ

If there is p s.t dist $(p,q) \le (1+\epsilon)r$,

then we have a probability of at least α [α : constant]

$(f_1, f_2, \ldots, f_J) \sim \mathcal{F}$

We deem that two points p, q collide iff

& $f(p) = f(q)$

p s.t dist $(p,q) > (1+\epsilon)r$,

$\le \beta n$ (linear)

$f \sim \mathcal{F}$

$(f_1, f_2, \ldots, f_k) \sim \mathcal{F} \rightarrow \Lambda(\mathcal{F}, k)$

We deem that two points p & q "collide" iff

$f_1(p) = f_1(q)$ & $f_2(p) = f_2(q) \ldots$

---

p s.t dist $(p,q) > (1+\epsilon)r$ &

$\Lambda(\mathcal{F},k)(p) = \Lambda(\mathcal{F},k)(q) \rightarrow \boxed{\beta^k n}$

$f_1(p) = f_1(q) \vee f_2(p) = f_2(q) \vee \ldots$

Then,

Case 1: (p, q) dist is $\leq (1+\epsilon) r$

$$\mathbb{P}[V(\mathcal{F}, T)(p) = V(\mathcal{F}, T)(q)]$$

$$= 1 - (1-\alpha)^T$$

It can be shown that

$\#\, p$ s.t $V(\mathcal{F}, T)(p) = V(\mathcal{F}, T)(q)$

but dist $(p, q) \geq (1+\epsilon) r$.

$$\leq \left(1 - (1-\beta)^T\right) n = (1 - o(1)) n$$

Suppose dist $(p, q) \leq (1+\epsilon) r$,

then

$f_1(p) = f_1(q)$ & $f_2(p) = f_2(q)$ & $\ldots$

$$\downarrow$$

$$\mathbb{P} \longrightarrow \alpha^k$$

COMBINED APQ.

$\mathcal{H} = V(\Lambda(\mathcal{F}, k), J)$ $\longrightarrow$ k, J are unknowns. will be fixed later.

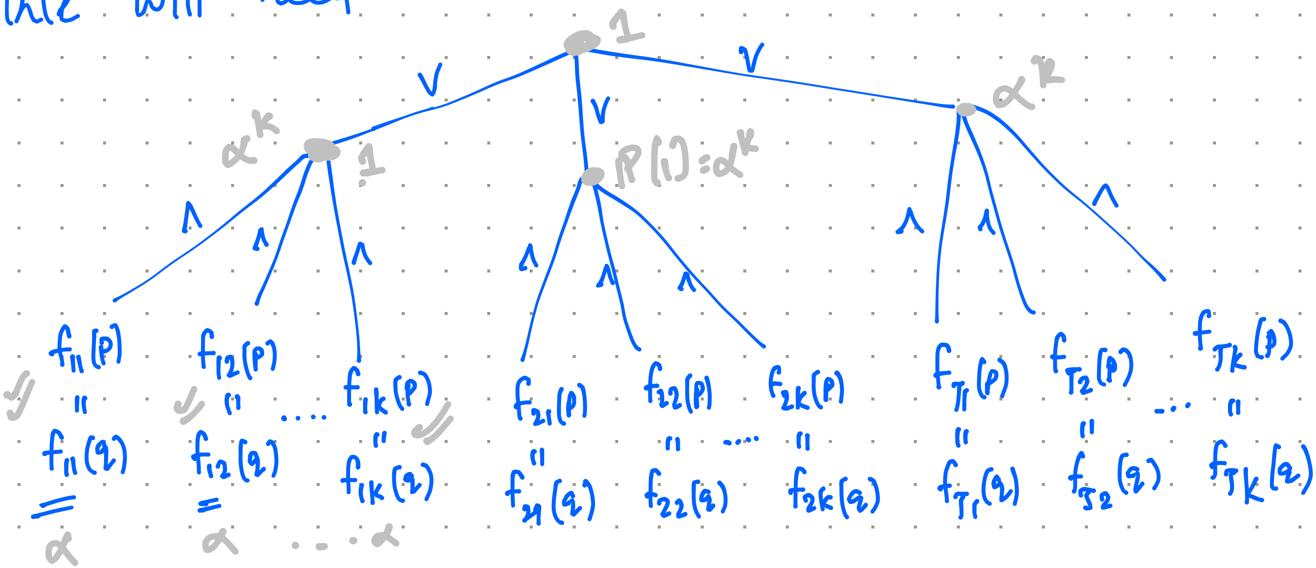
A hash function $h \in \mathcal{H}$, given as input a point p, computes the following matrix.

$$h(p) = \begin{bmatrix} f_{11}(p) & f_{12}(p) & \cdots & f_{1k}(p) \\ f_{21}(p) & f_{22}(p) & \cdots & f_{2k}(p) \\ & \vdots & & \\ f_{J1}(p) & f_{J2}(p) & \cdots & f_{Jk}(p) \end{bmatrix}$$

each f_{ij} ($i \in [J]$ and $j \in [k]$) is unif. rand. $\sim \mathcal{F}$.

COMBINED APQ:

We will need a collision criterion to decide when $\underline{\underline{h(p)}} = \underline{\underline{h(q)}}$



Lemma:- Given an (r, R, α, β)-sensitive LSH $\mathcal{F}$, the family $\mathcal{H} = \vee(\wedge(\mathcal{F}, k), J)$ is a $(r, R, 1-(1-\alpha^k)^J, 1-(1-\beta^k)^J)$-sensitive LSH.

Proof: Suppose $dist(p,q) \leq r$. We know that $\mathbb{P}_{f \sim \mathcal{F}}[f(p) = f(q)] > \alpha$

So, $\mathbb{P}_{h \sim \mathcal{H}}\left[h(p) = h(q)\right] \geq 1-(1-\alpha^k)^J$

Similarly, if $dist(p,q) > R$, we can show that

$$\mathbb{P}_{h \sim \mathcal{H}}\left[h(q) = h(p)\right] \leq 1-(1-\beta^k)^J$$

$\blacksquare$

Choosing k, J

$\text{dist}(p,q) \leq r \implies \mathbb{P}[h(p) = h(q)] \geq 1 - (1-\alpha^k)^J$ // $\alpha > \beta$

$\text{dist}(p,q) > (1+\epsilon)r \implies \mathbb{P}[h(p) = h(q)] \leq 1 - (1-\beta^k)^J$

Impose the cond^n that $1 - (1-\alpha^k)^J = 3/4$

$$(1-\alpha^k)^J = 1/4$$

$$\frac{1}{4} = (1-\alpha^k)^J \leq \exp(-\alpha^k J)$$

$$e^{-4} > 1/4$$

It sufficies to take $J = 4/\alpha^k$

$$\boxed{J = 4/\alpha^k}$$

Prob that a point in $P \setminus \text{Ball}(q, \underline{(1+\epsilon)\gamma})$ ~~that~~ collides with a given point q is

$$\leq 1 - (1-\beta^k)^J = \left(1-(1-\beta^k)\right)\left(1 + \overbrace{(1-\beta^k)}^{\leq 1} + \overbrace{(1-\beta^k)^2}^{\leq 1} + \dots\right)$$

$$\leq \beta^k J = 4\left(\frac{\beta}{\alpha}\right)^k$$

Query time :- $h(q)$ and then go thru the Bucket B which contains all points p s.t $h(q) = h(p)$.

If we find p s.t $\text{dist}(p,q) \leq (1+\epsilon)\gamma$,

If the num of points p s.t $\text{dist}(p,q) > (1+\epsilon)\gamma$

$\mathbb{E}$xpec num of P s.t $dist(p,q) > (1+\epsilon)r$ and $h(p) = h(q)$

is given by $= \underline{4n\left(\beta/\alpha\right)^k}$.

and to compute $h(q)$ itself, we need $\underline{Jk}$ $\underline{\text{time}}$.

Query time $\approx \left(\underline{Jk + 4n\left(\frac{\beta}{\alpha}\right)^k}\right)d$

~~By~~ Take $\boxed{JK = 4n\left(\beta/\alpha\right)^k}$

$= n\beta^k J$

$\Rightarrow \boxed{K = n\beta^k}$

$\left(\frac{1}{p}\right)^k = n$

$\boxed{n\beta^k = 1}$

If we choose $\boxed{k = \log_{(1/\beta)} n}$

Query time $= d\left(Tk + 4n\left(\frac{P}{\alpha}\right)^k\right)$

$$\leq d\left(Tk + 4\frac{(k)}{\alpha^k}\right)$$

$$= \underline{\underline{2dTk}} \simeq \underline{\underline{O(dTk)}}$$

$$k = \frac{\ln n}{\ln(1/\beta)} \qquad T = 4/\alpha^k$$

$$= \frac{4}{\alpha^{\frac{\ln n}{\ln(1/\beta)}}} = \frac{4}{\left(\alpha^{\ln n}\right)^{1/\ln(1/\beta)}}$$

$$= 4\left(n^{\ln(1/\alpha)}\right)^{1/\ln(1/\beta)}$$

$$= 4\, n^{\frac{\ln(\frac{1}{\alpha})}{\ln(1/\beta)}}$$

$$\underline{\underline{\alpha = 1 - \frac{r}{d}}}, \quad \underline{\underline{\beta = 1 - \frac{(1+\epsilon) r}{d}}} \quad = \underline{\underline{T}}$$

$$\text{Query time} \simeq 2d \left(4\, n^{\frac{\ln(1/\alpha)}{\ln(1/\beta)}} \cdot \frac{\ln(n)}{\ln(1/\beta)} \right)$$

$$\approx O\left(n^{\underline{\frac{\ln(1/\alpha)_\epsilon}{\ln(1/\beta)_\epsilon}}} \ln(n)\, d \right)$$

Prop: For $x \in (0,1)$ & $t \geq 1$ s.t $1-tx > 0$, we have

$$\frac{\ln\left(\frac{1}{1-x}\right)}{\ln\left(\frac{1}{1-tx}\right)} \leq \frac{1}{t}$$

A) $$g(x) = (1-tx) - (1-x)^t$$

$$g'(x) = -t + t(1-x)^{t-1}$$

$$= t\left((1-x)^{t-1} - 1\right)$$

$$\leq 0 .$$

$$g(0) = (1-0) - (1-0)^t = 0.$$

$g(x) \leq 0$ for all $x \in (0,1)$.

$$(1-tx) \leq (1-x)^t$$

$$\ln(1-tx) \leq t\ln(1-x)$$

$$\ln\left(\frac{1}{1-tx}\right) \geq t\ln\left(\frac{1}{1-x}\right)$$

$$\Longrightarrow \boxed{\frac{\ln\left(\frac{1}{1-x}\right)}{\ln\left(\frac{1}{1-tx}\right)} \leq \frac{1}{t}.}$$

Query time =

$$O\left(n^{\frac{\ln\left(\frac{1}{1-r/d}\right)}{\ln\left(\frac{1}{1-(1+\epsilon)r/d}\right)}} \ln n\, d \right)$$

$$\le O\left(\underbrace{n^{\frac{1}{1+\epsilon}}}_{n^{1/3}} (\ln n)\, d \right)$$

Preprocessing space :-

$$(n)(d)(\underline{\underline{Jk}}) = O\left(\underbrace{n^{1+\frac{1}{1+\epsilon}}}_{n^{4/3}} (\ln n)\, d \right)$$

Main Thm: (Assuming $(X, dist)$ has (r, R, α, β)-LSH family)

(APQ) It is possible to preprocess a set of n points in $\mathbb{R}^d$ such that

Preprocessing space is $O\left(n^{1+\frac{\ln(1/\alpha)}{\ln(1/\beta)}} (\ln n)\, d\right)$

Query time is $O\left(n^{\frac{\ln(1/\alpha)}{\ln(1/\beta)}} (\ln n)\, d\right)$

Each query is successful with prob at least $3/4$.

LSH on Euclidean Metric $(\mathbb{R}^d, dist)$

Lemma:- Let $X_1, X_2, \ldots, X_d$ be normally distributed independent random variables. Let $v_1, \ldots, v_d \in \mathbb{R}$. Then the variable $Y = v_1X_1 + v_2X_2 + \ldots + v_dX_d$ will be distributed as $\mathcal{N}(0, \sigma^2 = \|v\|^2)$ where $\vec{v} = (v_1, v_2, \ldots, v_d)$

Proof:- $\phi_X(t) = \mathbb{E}\left[e^{itX}\right]$

It can be shown that for $\mathcal{N}(0, \sigma^2)$, the

$$\phi_{N(0,\sigma^2)}(t) = \exp\left(-\frac{\sigma^2 t^2}{2}\right)$$

$$Y = v_1 X_1 + \dots + v_d X_d$$

$$\phi_Y(t) = \mathbb{E}\left[e^{it(v_1 X_1 + v_2 X_2 + \dots + v_d X_d)}\right]$$

$$= \prod_{j=1}^{d} \mathbb{E}\left[e^{it(v_j X_j)}\right]$$

$$= \prod_{j=1}^{d} \exp\left(-\frac{v_j^2 t^2}{2}\right)$$

$$\phi_Y(t) = \exp\left(-(v_1^2 + v_2^2 + \dots + v_d^2)\frac{t^2}{2}\right)$$

$$Y \sim \mathcal{N}(0, \sigma^2 = \|v\|^2).$$ $\blacksquare$

LSH for Euclidean Metric:

Fix a vector $\vec{v}$. Then, let's define for a point $p \in \mathbb{R}^d$

$$h(p) = \left\lfloor \frac{\langle p, \vec{v} \rangle + t}{x} \right\rfloor$$ where $t \sim \text{Unif}[0, x]$ and

x is a parameter which we will choose shortly.

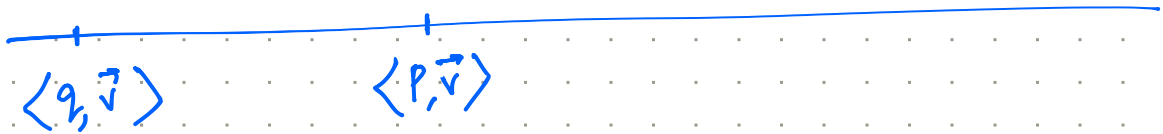
$$\mathcal{H} = \{ h : \vec{v} \text{ is a vector} \sim \mathcal{N}(0,1)^d \}$$

We need to choose x s.t $\mathcal{H}$ is LSH.

Fix $p, q, \vec{v}, x$. Let $\|p-q\| =: r$ and

$$|\langle p, \vec{v}\rangle - \langle q, \vec{v}\rangle| =: \beta.$$

$\langle p, \vec{v}\rangle - \langle q, \vec{v}\rangle = \beta$ (Assume)



Fix some $p, q, \vec{v}, x$.

$$\mathbb{P}\left[h(p) = h(q)\right]$$

$$h(p), h(q) \iff \left\lfloor \frac{\langle p, \vec{v}\rangle + t}{x} \right\rfloor = \left\lfloor \frac{\langle q, \vec{v}\rangle + t}{x} \right\rfloor$$

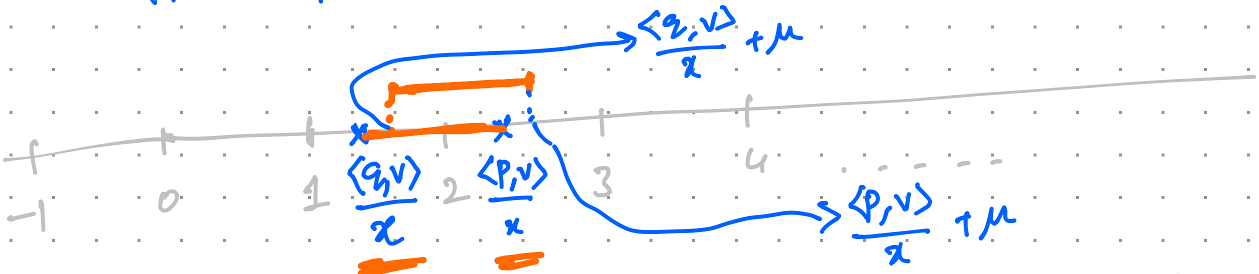
<u>Claim</u>: If $\beta = \langle p, v\rangle - \langle q, v\rangle > x$, then $h(p) \neq h(q)$.

<u>Pf</u>: $\underline{h(p)} = \left\lfloor \frac{\langle q, v\rangle + \beta + t}{x} \right\rfloor = \left\lfloor \frac{\beta}{x} + \frac{\langle q, v\rangle + t}{x} \right\rfloor$

$$\underline{>} \left\lfloor \frac{\langle q, v\rangle + t}{x} \right\rfloor = \underline{h(q)}$$

$\blacksquare$

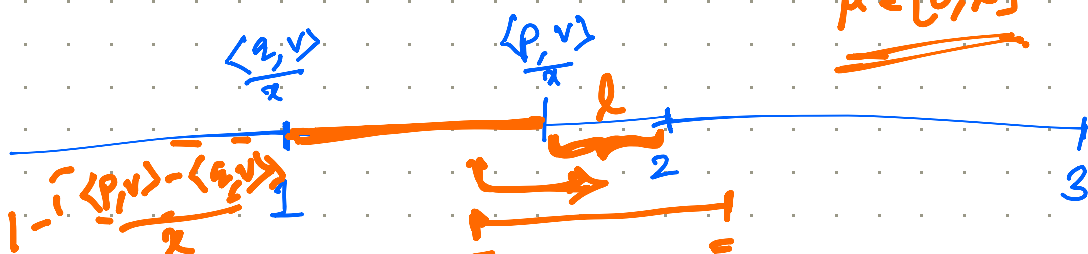
Suppose $\beta = \langle p,v\rangle - \langle q,v\rangle \le x.$



$h(p) = \left\lfloor \frac{\langle p,v\rangle}{x} + \mu \right\rfloor$ where $\mu \sim \mathrm{Unif}[0,1]$

$\mathbb{P}[h(p) = h(q)] = 1 - \frac{\beta}{x}$ $\quad (\beta = \langle p,v\rangle - \langle q,v\rangle)$

$\mu \in [0, l]$



$l = 1 - \frac{\langle p,v\rangle - \langle q,v\rangle}{x}$

$$\mathbb{P}\left[h(p) = h(q)\right] = \int_0^{\beta} \mathbb{P}\left[\,|\langle p,v\rangle - \langle q,v\rangle| = \beta\right]\left(1-\frac{\beta}{\alpha}\right) d\beta$$

$$\langle p,v\rangle - \langle q,v\rangle = \langle p-q, v\rangle \sim \mathcal{N}\left(0, \|p-q\|^2\right)$$

Say $\eta = \|p-q\|$.

$$\mathbb{P}\left[\,|\langle p,v\rangle - \langle q,v\rangle| = \beta\right] = \mathbb{P}\left[\langle p-q,v\rangle = \beta \text{ or } \langle q-p,v\rangle = \beta\right]$$

$$= 2\,\mathbb{P}\left[\langle p-q, v\rangle = \beta\right]$$

$$\mathbb{P}\left[\langle p-q, v\rangle = \beta\right]_{x} \to \text{the pdf of } \mathcal{N}(0,\eta^2) \text{ at } \beta$$

$$2\left[\begin{array}{c}\text{pdf of } \mathcal{N}(0,\eta^2) \\ \text{at } \beta\end{array}\right] = \frac{2}{\sqrt{2\pi}\,\eta}\exp\left(-\frac{\beta^2}{2\eta^2}\right)$$

$$\mathbb{P}\left[h(p) = h(q)\right] = \boxed{\int_0^x \frac{2}{\sqrt{2\pi}\,\eta}\exp\left(-\frac{\beta^2}{2\eta^2}\right)\left(1-\frac{\beta}{x}\right)d\beta}$$

$\eta \;\; = \;\; \|p-q\|$.

$p(\eta, x)$

$$\tilde{O}\left(n^{\frac{\ln(1/\alpha)}{\ln(1/\beta)}}\right) \quad \text{where} \quad \alpha \leq \mathbb{P}[h(p) = h(q)] \text{ given } \|p-q\| \leq r$$

$$\beta \geq \mathbb{P}[h(p) = h(q)] \text{ given } \|p-q\| \geq r(1+\epsilon)$$

Choose x s.t

$$\frac{\ln\left(\frac{1}{p(r,x)}\right)}{\ln\left(\frac{1}{p(r(1+\epsilon),x)}\right)} \quad \text{is minimized}$$

Datar, Immorlica, Indyk, Mirrocni (2004)

Using computations, they show that it is possible to choose param α s.t

$$\frac{\ln\left(\frac{1}{p(\eta,\alpha)}\right)}{\ln\left(\frac{1}{p(\eta(1+\epsilon),\alpha)}\right)} \leq \underline{\underline{\frac{1}{1+\epsilon}}} \quad \text{for all } \eta > 0.$$

→ <u>APQ</u>!

Query time → $\tilde{O}\left(d\, n^{\frac{1}{1+\epsilon}}\right)$

Preprocessing space → $\tilde{O}\left(d\left(n^{1+\frac{1}{1+\epsilon}}\right)\right)$.