Proximity problems in high dimensions

Ioannis Psarros

National & Kapodistrian University of Athens

March 31, 2017

Problem definition

Definition (Proximity problems)

Problems in computational geometry which involve estimation of distances between geometric objects.

Examples:

- Approximate Nearest Neighbor Search,
- Closest Pair of points,
- Minimum Spanning Tree,
- etc.

We consider n points in $\mathbb{R}^d$.

Problem definition

Proximity Problems

Low dimensional

High dimensional

Problem definition

“Low dimensional”

Time/space complexity: $exp(d)$, but “good” dependence on n.

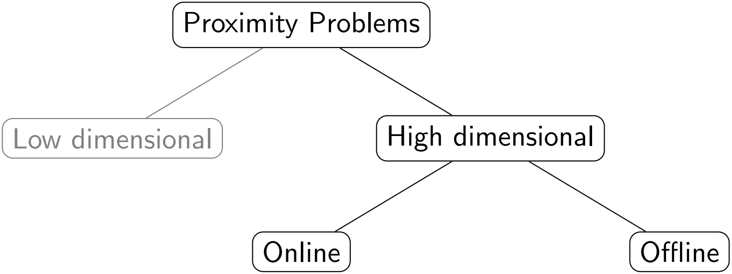
Example: $(1+\epsilon)$-ANN in space $\tilde{O}(dn)$ and query time $O(\frac{1}{\epsilon})^d$.

“High dimensional”

Time/space complexity: $poly(d)$, but “worse” dependence on n.

Example: $(1+\epsilon)$-ANN in space $\tilde{O}(dn^{1+\rho})$ and query time $\tilde{O}(dn^{\rho})$, where $\rho = \rho(\epsilon) < 1$.

Problem definition



Problem definition

"Online"

Not all points are given in advance. Query points are allowed.

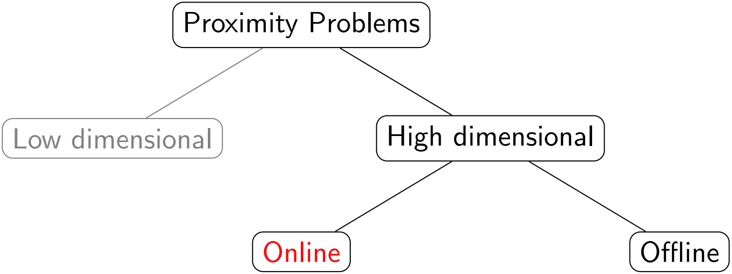
Example: ANN problem.

"Offline"

All points are given as input.

Example: ANNs with red/blue points, closest pair, r-nets.

Problem definition



Problem definition

If not stated otherwise, $\|\cdot\|$ is $\|\cdot\|_2$.

Definition (Approximate Nearest Neighbor)

Given set $X \subset \mathbb{R}^d$, error parameter $\epsilon > 0$, an ANN of some query point q is a point $p^* \in X$ s.t.:

$$\forall p \in X,\ \|p^* - q\| \leq (1+\epsilon)\|p - q\|.$$

Definition (Approximate Nearest Neighbor Problem)

Consider set $X \subset \mathbb{R}^d$. Build a data structure on X which given a query point $q \in \mathbb{R}^d$ reports an ANN of q.

Aim for (near) linear space.

Random Projections

Johnson-Lindenstrauss lemma

Let $X \subset \mathbb{R}^d$ and $|X| = n$. There exists a distribution over linear maps $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ with $d' = O(\epsilon^{-2} \log n)$ s.t., for any $p, q \in X$:

$$\|f(p) - f(q)\| \in (1 \pm \epsilon)\|p - q\|.$$

Low dimension + JL

- Space: $O(dn)$.
- Query time: $(\frac{1}{\epsilon})^{\Theta(\epsilon^{-2} \log n)} = \omega(n)$.

Random projections with slack

Observation

k distances are arbitrarily distorted $\implies d' = \Theta(\epsilon^{-2}\log(\frac{n}{k}))$ *is sufficient.*

Theorem (Anagnostopoulos, Emiris, P '15)

Consider $X \subset \mathbb{R}^d$, query $q \in \mathbb{R}^d$ and approximation error $\epsilon > 0$. Sample linear map $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ from a JL distribution, with $d' = \Theta(\epsilon^{-2}\log(\frac{n}{k}))$. Then, w.c.p. the following hold:

- *if p^* is the NN of q, then $\|f(p^*) - f(q)\| \in (1 \pm \epsilon)\|p^* - q\|$,*
- $|\{p \in X \setminus \{p^*\} : \|f(p) - f(q)\| \notin (1 \pm \epsilon)\|p - q\|\}| \leq k$

Low dimension + JL with slack

- Space: $O(dn)$.
- Query time: $(\frac{1}{\epsilon})^{\Theta(\epsilon^{-2}\log(n/k))} + k = dn^{1-\Theta(\epsilon^2/\log(1/\epsilon))}$.

LSH + Random projections with slack

Definition (Datar et al.)

Let $w > 0$ *be a parameter, and let* t *be a number distributed uniformly in* $[0, w]$*. Define:*

$$h(p) = \left\lfloor \frac{\langle p, v \rangle + t}{w} \right\rfloor, \quad p \in \mathbb{R}^d, v \in N(0,1)^d$$

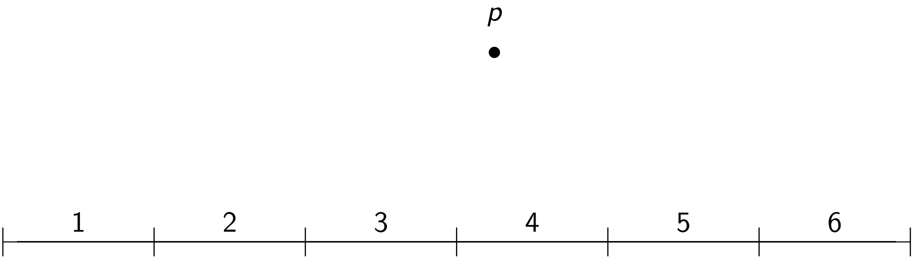
Definition (P, Avarikioti, Samaras, Emiris '17)

Define:

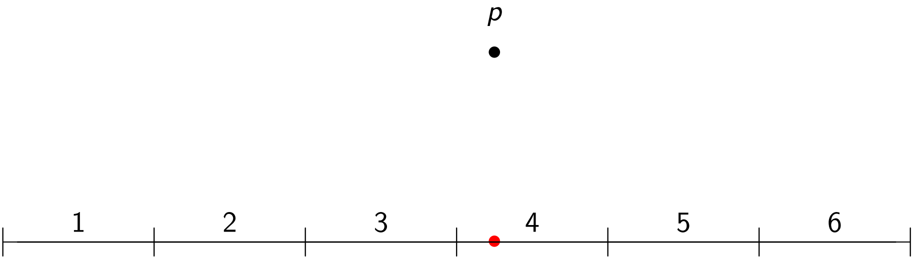
$$f(p) = (f_1(h(p)), \ldots, f_{d'}(h(p))), \quad p \in \mathbb{R}^d,$$

where $h : \mathbb{R}^d \to \mathbb{N}$ *is chosen uniformly at random as above and* $f_i : \mathbb{N}^d \to \{0, 1\}$ *random function.*

LSH + Random projections with slack

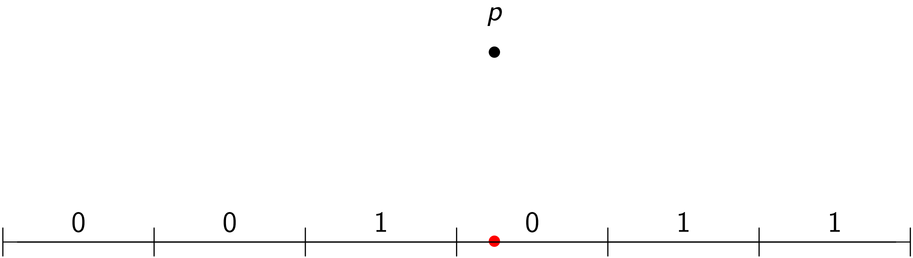


LSH + Random projections with slack



$h(p) = 4$

LSH + Random projections with slack



$f_1(p) = 0$. Repeat d' times: $f(p) = (0, \ldots) \in \{0, 1\}^{d'}$.

Random projections with slack

Theorem

Consider $X \subset \mathbb{R}^d$, query $q \in \mathbb{R}^d$ and radius $r > 0$, approximation error $\epsilon > 0$. Sample mapping $f : \mathbb{R}^d \to \{0,1\}^{d'}$ from a distribution as in the previous Definition, with $d' = \Theta(\epsilon^{-2}\log(\frac{n}{k}))$. Then, w.c.p. the following hold:

- $\|p - q\| \leq r$ *implies* $\|f(p) - f(q)\|_1 \leq r'$,
- $|\{p \in X : \|p - q\| \geq (1+\epsilon)r$ *and* $\|f(p) - f(q)\|_1 \leq r'\}| \leq k$,

Low dimension Hamming + LSH projection with slack

- Space: $O(dn)$.
- Query time: $2^{\Theta(\epsilon^{-2}\log(n/k))} + k = dn^{1-\Theta(\epsilon^2)}$.

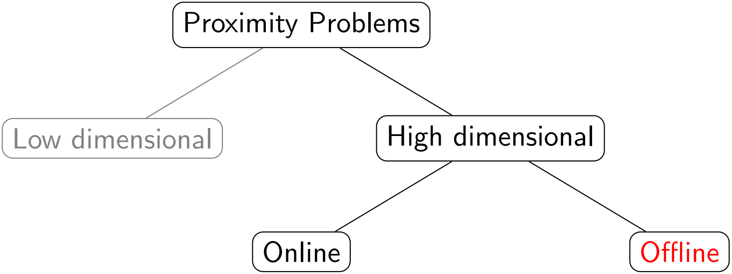
For any LSHable metric, we obtain linear space and sublinear query.

Summary

Near-linear space regime.

	Space	Query
Entropy-based LSH [Panigrahy '06]	$\tilde{O}(dn)$	$dn^{O((1+\epsilon)^{-1})}$
Entropy-based LSH [Andoni '08]	$\tilde{O}(dn)$	$dn^{O((1+\epsilon)^{-2})}$
JL with slack	$\tilde{O}(dn)$	$dn^{1-\Theta(\epsilon^2/\log(1/\epsilon))}$
LSH tradeoffs [Andoni et al. '17]	$\tilde{O}(dn)$	$O(dn^{(2(1+\epsilon)^2-1)/(1+\epsilon)^4})$
LSH-projection with slack	$\tilde{O}(dn)$	$dn^{1-\Theta(\epsilon^2)}$

Problem definition



Problem definition

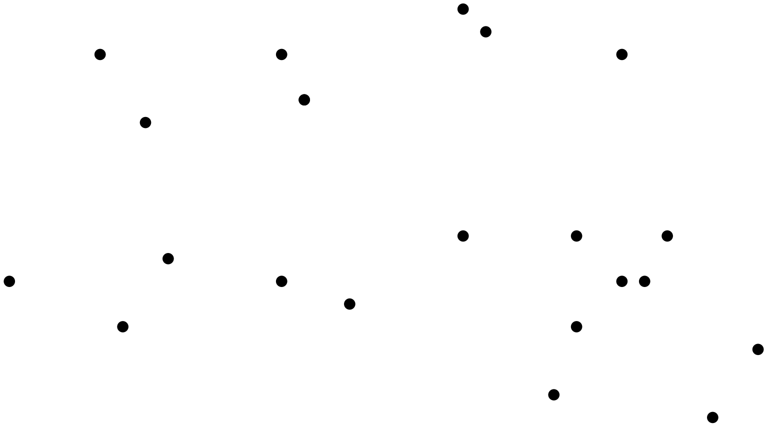
Definition

Given a pointset $X \subseteq \mathbb{R}^d$, a parameter $r > 0$, an r-net of X is a subset $N \subseteq X$ s.t. the following properties hold:

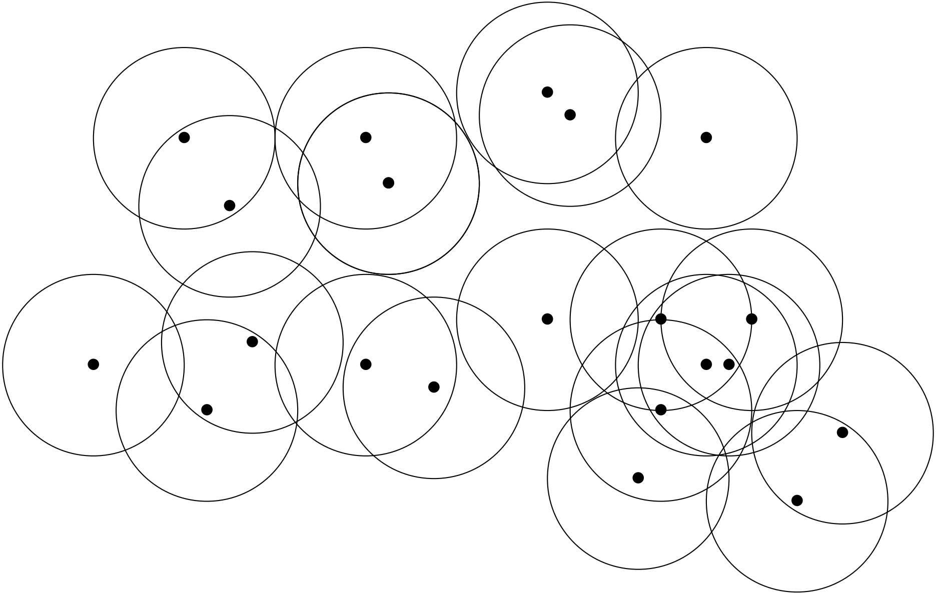
- *(packing) For every $p \neq q \in N$, we have that $\|p - q\|_2 > r$.*
- *(covering) For every $p \in X$, there exists $q \in N$ s.t. $\|p - q\|_2 \leq r$.*

Equivalently, an r-net is a maximal r-packing subset of X, or a minimal r-covering subset of X.

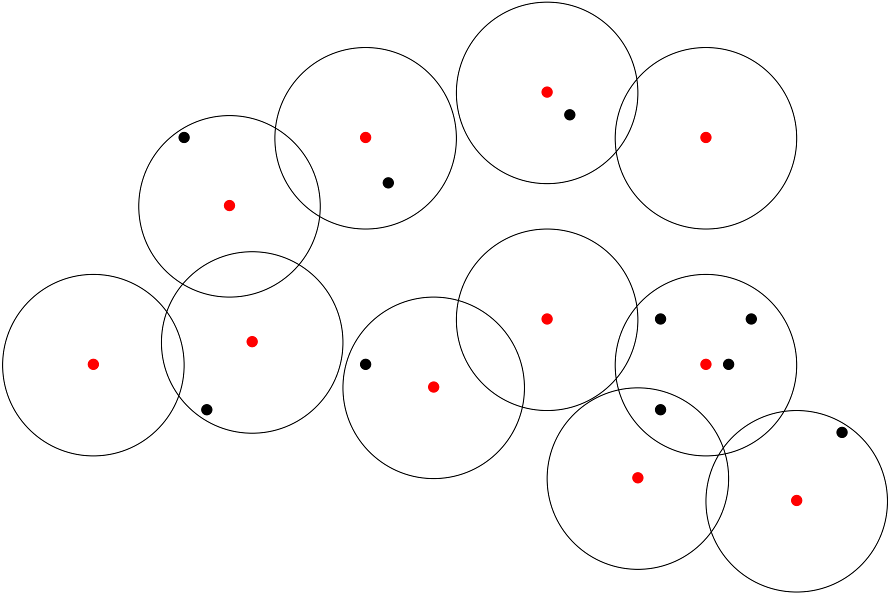
Problem definition



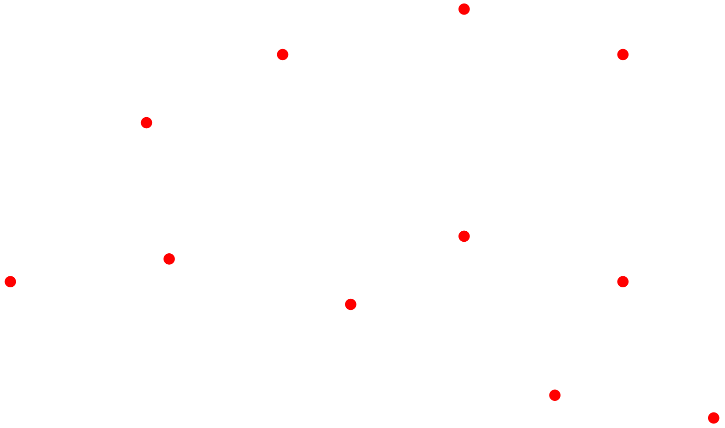
Problem definition



Problem definition



Problem definition



Problem definition

Definition (Approximate r-nets)

Given a pointset $X \subseteq \mathbb{R}^d$, a parameter $r > 0$ and an approximation parameter $\epsilon > 0$, a $(1+\epsilon)r$-net of X is a subset $N \subseteq X$ s.t. the following properties hold:

1. *(packing) For every $p \neq q \in N$, we have that $\|p - q\|_2 \geq r$.*
2. *(covering) For every $p \in X$, there exists $q \in N$ s.t. $\|p - q\|_2 \leq (1+\epsilon)r$.*

Why r-nets?

Computing r-nets is a fundamental primitive in Computational Geometry.

Recent improvements in high dimensional "offline" problems:

- LSH: Approximate closest pair in time $\tilde{O}(dn^{2-\Theta(\epsilon)})$.
- [Valiant '12]: Approximate closest pair in time $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$.

Can we extend this improvement for the problem of computing r-nets?

Previous work

Approach	Time	Output
Grid [Har-Peled '04]	$O(d^{d/2}n)$	r-net
Grid (Folklore)	$O(dn) \times O(\frac{1}{\epsilon})^d$	$(1+\epsilon)r$-net
LSH [Eppstein et al. '15]	$\tilde{O}(dn^{2-\Theta(\epsilon)})$	$(1+\epsilon)r$-net whp
This work	$\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$	$(1+\epsilon)r$-net whp

1 Previous Work

2 High dimensional approximate r-nets

- Random instance
- Nets under inner product
- Nets under Euclidean distance

1. Previous Work

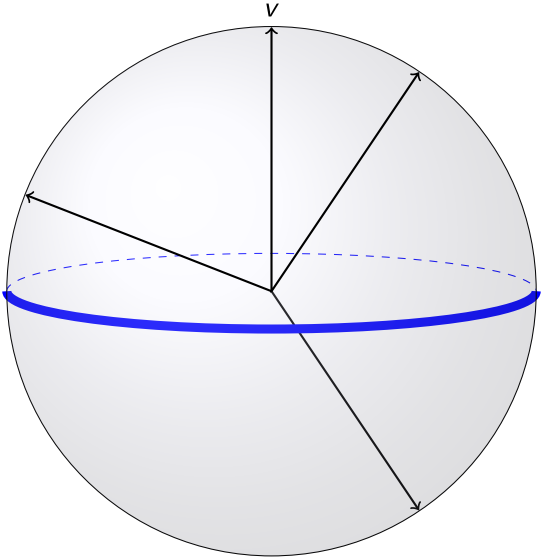
2. High dimensional approximate r-nets
   - Random instance
   - Nets under inner product
   - Nets under Euclidean distance

Random instance

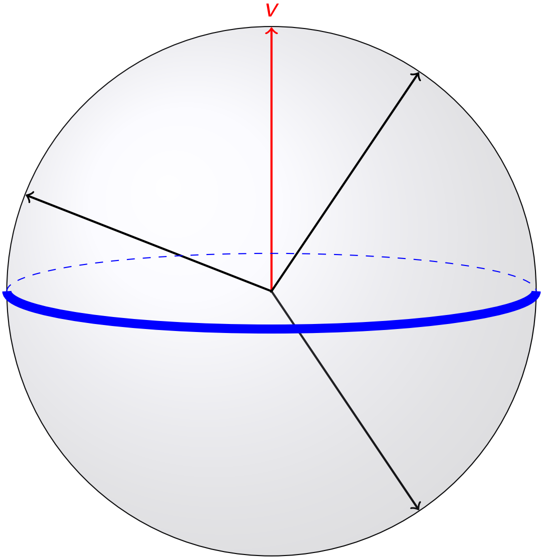
Random Instance

- Input: $X = [x_1, \ldots, x_n]$, $x_i \in \{-1, 1\}^d$, $\rho \in (0, 1]$.
- For $i = 1, \ldots, n$:
  - either $\exists j \neq i$, $|\langle x_i, x_j \rangle| \geq \rho \cdot d$, ($\rho$-correlated)
  - or x_i is chosen uniformly at random.
- Objective:
  - Packing: any two vectors in the net are not ρ-correlated,
  - Covering: any vector is ρ-correlated with some net vector.

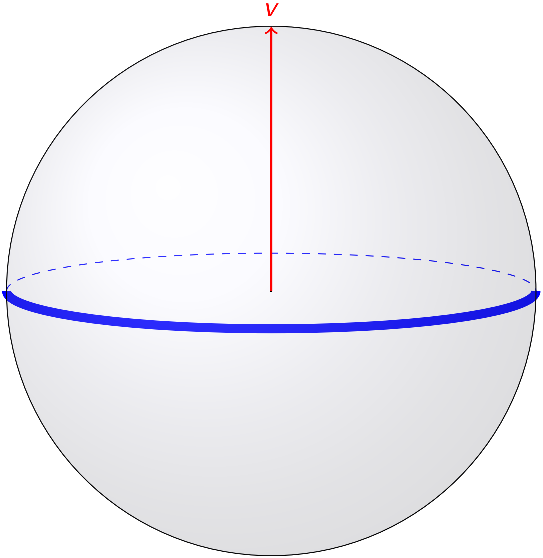
Random instance



Random instance



Random instance



Random instance with tight concentration

Observation

Let x be non-correlated with vectors in $S \subseteq X \subseteq \{-1,1\}^d$, where $|S| = n^\alpha$, $\alpha \in (0,1)$. If $d \approx n^{2\alpha}/\rho^2$, then with high probability,

$$|\langle x, \sum_{y \in S} y \rangle| = |\sum_{y \in S} \langle x, y \rangle| < \rho \cdot d.$$

RandomInstanceNet

Input: $X = [x_1, \ldots, x_n]$, $x_i \in \{-1, 1\}^d$, $d \approx n^{2\alpha}/\rho^2$, $\alpha, \rho \in (0, 1)$
Output: ρ-net $N \subseteq \{x_1, \ldots, x_n\}$

- Repeat $\sqrt{n}$ times: //Decrease the number of correlations
  - Choose a column x_i uniformly at random.
  - $N \leftarrow N \cup \{x_i\}$; Delete x_i from X.
  - Delete each x_j from X s.t. $|\langle x_i, x_j \rangle| \geq \rho$.
- $n \leftarrow$ #remaining columns.
- Randomly partition vectors into disjoint subsets $S_1, \ldots, S_{n^{1-\alpha}}$.
- Set $d \times n^{1-\alpha}$ matrix Z : column $Z_k = \sum_{x_j \in S_k} x_j$. //Compress
- Compressed Gram matrix: $W = X^T Z$, size $n \times n^{1-\alpha}$.
- For W rows/vectors $i = 1, \ldots$: //Search for correlations
  - $N \leftarrow N \cup \{x_i\}$
  - For each $|w_{ik}| \geq \rho$: For each ρ-correlated $x_j \in S_k$, delete row j.

$\alpha = 1/3 \implies$ time: $O(dn^{1.94})$

Nets under inner product

Definition (Approximate inner product nets)

For any $X \subset \mathbb{S}^{d-1}$, an approximate ρ-net for $(X, \langle \cdot, \cdot \rangle)$, with additive approximation parameter $\epsilon > 0$, is a subset $N \subseteq X$ which satisfies the following properties:

- *for any two $p \neq q \in N$, $\langle p, q \rangle < \rho$, and*
- *for any $x \in X$, there exists $p \in N$ s.t. $\langle x, p \rangle \geq \rho - \epsilon$.*

Sphere to Hypercube

MakeUniform [Charikar '02]

There exists an algorithm running in $O(\frac{dn \log n}{\delta^2})$ with the following properties.

Input: $X = [x_1, \ldots, x_n]$ s.t. $x_i \in \mathbb{S}^{d-1}$.
Output: $Y = [y_1, \ldots, y_n] \in \{-1, 1\}^{n \times d'}$, $d' = O(\log n / \delta^2)$.

With probability $1 - o(1/n^2)$, for all pairs $i, j \in [n]$,

$$\left| \frac{\langle y_i, y_j \rangle}{d'} - \left(1 - 2 \cdot \frac{\arccos(\langle x_i, x_j \rangle)}{\pi}\right) \right| \leq \delta.$$

Simulate tight concentration

ChebyshevEmbedding [Valiant '12]

There exists an algorithm with the following properties.

Input: $X = [x_1, \ldots, x_n]$ s.t. $x_i \in \{-1,1\}^d$, $\rho \in [-1,1]$.
Output: $Y, Y' \in \{-1,1\}^{n \times d'}$, $d' = n^{0.2}$.

With probability $1 - o(1/n)$, for all $i, j \in [n]$,

- $\langle x_i, x_j \rangle \leq \rho \cdot d \implies |\langle y_i, y'_j \rangle| \leq 3n^{0.16}$,
- $\langle x_i, x_j \rangle \geq (\rho + \delta) \cdot d \implies |\langle y_i, y'_j \rangle| \geq 3n^{0.16+\sqrt{\delta}/100}$.

Gap amplification simulates tight concentration in random instance.

`InnerProductApprxNet`

Input: $X = [x_1, \ldots, x_n]$ with $x_i \in \mathbb{S}^{d-1}$, $\rho \in [-1, 1]$, $\epsilon \in (0, 1/2]$.
Output: ρ-net $N \subseteq [n]$.

- $(Y, \rho') \leftarrow$`MakeUniform`$(X,\ \delta = \epsilon/2\pi)$.
- $(Z, Z', \rho'') \leftarrow$`ChebyshevEmbedding`$(Y,\ \rho')$.
- $N \leftarrow$Simulate `RandomInstanceNet`(ρ'', Z, Z').

Theorem

The algorithm `InnerProductApprxNet`*, on input $X = [x_1, \ldots, x_n]$ with each $x_i \in \mathbb{S}^{d-1}$, $\rho \in [-1, 1]$ and $\epsilon \in (0, 1/2]$, computes an approximate ρ-net with additive error ϵ. The algorithm runs in time $\tilde{O}(dn + n^{2-\sqrt{\epsilon}/600})$ and succeeds with probability $1 - O(1/n^{0.2})$.*

1. Previous Work

2. High dimensional approximate r-nets
   - Random instance
   - Nets under inner product
   - Nets under Euclidean distance

Nets under Euclidean distance

We can reduce to the inner-product net problem.

Theorem (Avarikioti, Emiris, Kavouras, P '17)

Given n points in $\mathbb{R}^d$, *a parameter* $r > 0$ *and an approximation parameter* $\epsilon \in (0, 1/2]$, *with probability* $1 - o(1/n^{0.04})$, `ApprxNet` *will return a* $(1+\epsilon)r$*-net, in* $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$ *time.*

Future work

- ANN for other norms or general methods for classes of norms.
- Recently, [Alman, Chan, and Williams '16] improved upon [Valiant '12] for the approximate closest pair problem. Similar improvement for r-nets?

Thank you!