Fast Construction of Nets in Low Dimensional Metrics, and Their Applications

Sariel Har-Peled[1] Manor Mendel[2]

[1]Univ. of Illinois, Urbana-Champaign

[2]California Institute of Technology

June '05

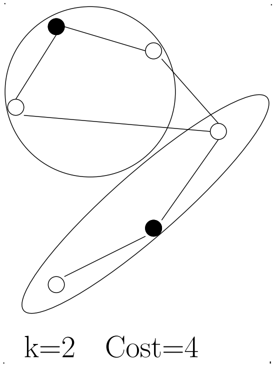
k-Median

- Given: Graph $G = (V, E)$, $k \in \mathbb{N}$.
- Finite metric: $d_G()$ - shortest distance in graph.
- Find: Centers $c_1, \ldots, c_k \in V$, minimizing $\sum_{v \in V} \min_{i \in \{1,\ldots,k\}} d_G(v, c_i)$.
- NP-Hard optimization problem.
- $\exists$ Poly-time $3 + \varepsilon$ approx algorithm **[Arya et al., 2001]**.
- Can do much better in low-dim Euclidean settings...
- Q: $\exists$ cases such that finite metric $\approx$ Euclidean metric?

k=2

k-Median

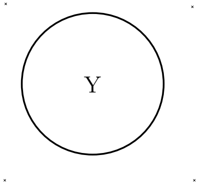
- Given: Graph $G = (V, E)$, $k \in \mathbb{N}$.
- Finite metric: $d_G()$ - shortest distance in graph.
- Find: Centers $c_1, \ldots, c_k \in V$, minimizing $\sum_{v \in V} \min_{i \in \{1,\ldots,k\}} d_G(v, c_i)$.
- NP-Hard optimization problem.
- $\exists$ Poly-time $3 + \varepsilon$ approx algorithm **[Arya et al., 2001]**.
- Can do much better in low-dim Euclidean settings...
- Q: $\exists$ cases such that finite metric $\approx$ Euclidean metric?



Doubling Dimension

Definition (Doubling Dimension ddim(X))

The least ρ, $\forall Y \subset X\ \exists Z_1, \ldots Z_{2^\rho}$, $\mathrm{diam}(Z_i) \leq \frac{\mathrm{diam}(Y)}{2}$ and $\bigcup_i Z_i \supset Y$.

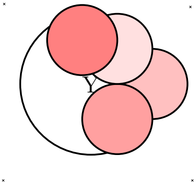


- $\mathrm{ddim}(\mathbb{R}^d) = \Theta(d)$.
- Applicable to finite metrics.
- $Y \subset X \;\Rightarrow\; \dim(Y) \leq \dim(X)$.
- $\dim(X) \leq \log |X|$.

Doubling Dimension

Definition (Doubling Dimension ddim(X))

The least ρ, $\forall Y \subset X\ \exists Z_1, \ldots Z_{2^\rho}$, $\mathrm{diam}(Z_i) \leq \frac{\mathrm{diam}(Y)}{2}$ and $\bigcup_i Z_i \supset Y$.

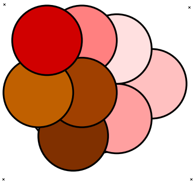


- $\mathrm{ddim}(\mathbb{R}^d) = \Theta(d)$.
- Applicable to finite metrics.
- $Y \subset X \Rightarrow \dim(Y) \leq \dim(X)$.
- $\dim(X) \leq \log |X|$.

Doubling Dimension

Definition (Doubling Dimension ddim(X))

The least ρ, $\forall Y \subset X\ \exists Z_1, \ldots Z_{2^\rho}$, $\mathrm{diam}(Z_i) \leq \frac{\mathrm{diam}(Y)}{2}$ and $\bigcup_i Z_i \supset Y$.



- $\mathrm{ddim}(\mathbb{R}^d) = \Theta(d)$.
- Applicable to finite metrics.
- $Y \subset X \ \Rightarrow\ \dim(Y) \leq \dim(X)$.
- $\dim(X) \leq \log |X|$.

Why Doubling Metrics?

1. Naturally extends Euclidean settings.
2. Truely extended Euclidean metrics.
3. More abstract.
4. Forces conceptually simple solution for metric problems.

Why work on finite metric spaces?

- Strong (geometric) intuition.
- Beautiful topic. Deep math.
- Push CG to high dimensions.
- Hot topic in theory.

Why Doubling Metrics?

1. Naturally extends Euclidean settings.
2. Truely extended Euclidean metrics.
3. More abstract.
4. Forces conceptually simple solution for metric problems.

Why work on finite metric spaces?

- Strong (geometric) intuition.
- Beautiful topic. Deep math.
- Push CG to high dimensions.
- Hot topic in theory.

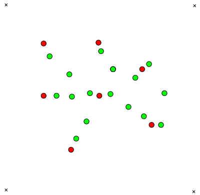
History of Doubling Metric

- Related to Hausdorff-Besicovitch dimension.
- Used in the Geometric Analysis community **[Assouad, 1983]**.
- Introduced to CS by **[Gupta et al., 2003]**. (Also **[Clarkson, 1999]**.)
- Approx. Nearest Neighbor Search (ANNS) **[Krauthgamer and Lee, 2004b, Krauthgamer and Lee, 2004a]**
- Quasi PTAS for TSP **[Talwar, 2004]**.
- Efficient distance labeling **[Talwar, 2004, Slivkins, 2004]**

Nets in Metric Spaces

r-net S:

- Balls of radius r centered at S cover the space.
- Balls of radius $r/2$ centered at S are packed in the space.



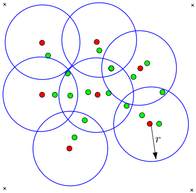
The importance of nets

- r-net captures the geometry at scale $3r$ in the metric.
- In fixed doubling metrics:
  - A ball of radius $2r$ contains $O(1)$ r-net points — allows efficient manipulation of hierarchy of nets.

Nets in Metric Spaces

r-net S:

- Balls of radius r centered at S cover the space.
- Balls of radius $r/2$ centered at S are packed in the space.



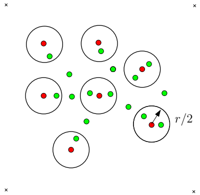
The importance of nets

- r-net captures the geometry at scale $3r$ in the metric.
- In fixed doubling metrics:
  - A ball of radius $2r$ contains $O(1)$ r-net points — allows efficient manipulation of hierarchy of nets.

Nets in Metric Spaces

r-net S:

- Balls of radius r centered at S cover the space.
- Balls of radius $r/2$ centered at S are packed in the space.



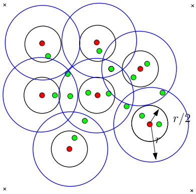
The importance of nets

- r-net captures the geometry at scale $3r$ in the metric.
- In fixed doubling metrics:
  - A ball of radius $2r$ contains $O(1)$ r-net points — allows efficient manipulation of hierarchy of nets.

Nets in Metric Spaces

r-net S:

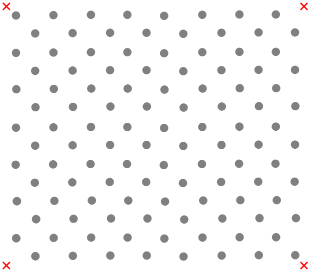
- Balls of radius r centered at S cover the space.
- Balls of radius $r/2$ centered at S are packed in the space.



The importance of nets

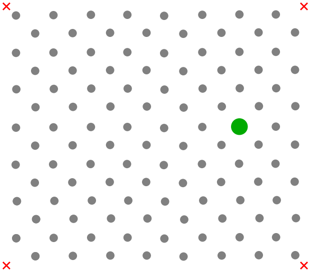
- r-net captures the geometry at scale $3r$ in the metric.
- In fixed doubling metrics:
  - A ball of radius $2r$ contains $O(1)$ r-net points — allows efficient manipulation of hierarchy of nets.

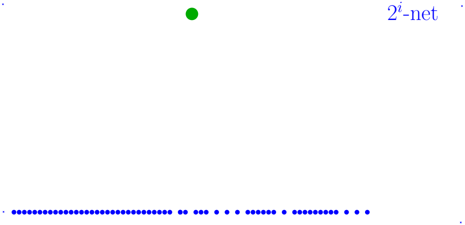
Hierarchy of Nets: Net-tree



This is just quadtree or compressed quadtree in geometric settings...

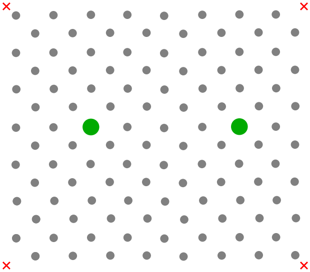
Hierarchy of Nets: Net-tree

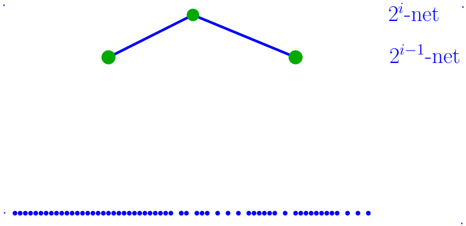




This is just quadtree or compressed quadtree in geometric settings...

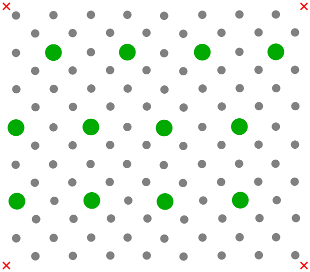
Hierarchy of Nets: Net-tree

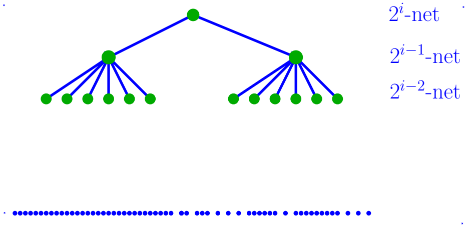




This is just quadtree or compressed quadtree in geometric settings...

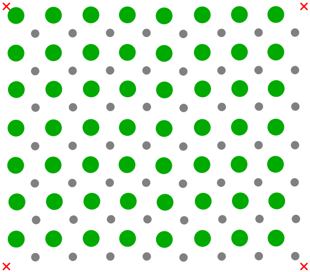
Hierarchy of Nets: Net-tree

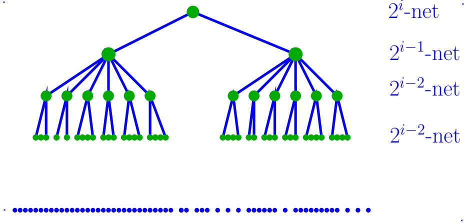




This is just quadtree or compressed quadtree in geometric settings...

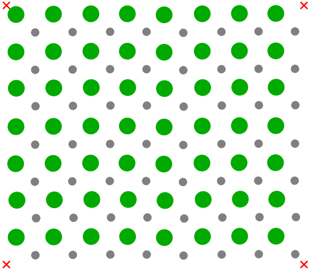
Hierarchy of Nets: Net-tree

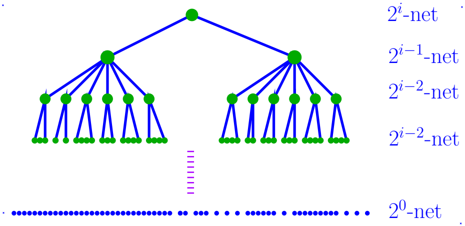




This is just quadtree or compressed quadtree in geometric settings...

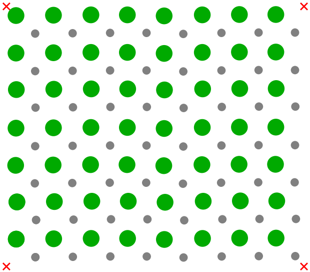
Hierarchy of Nets: Net-tree

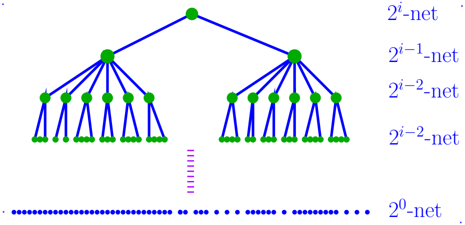




This is just quadtree or compressed quadtree in geometric settings...

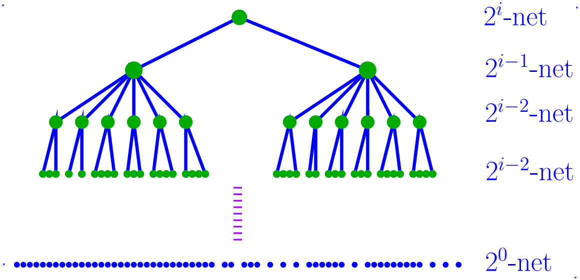
Hierarchy of Nets: Net-tree



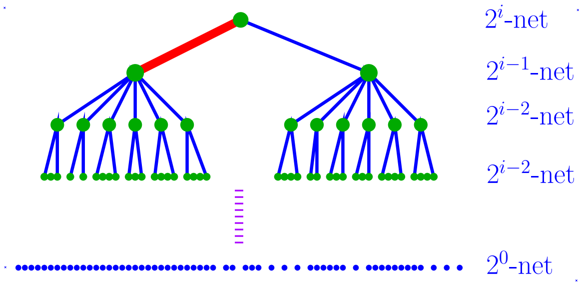


This is just quadtree or compressed quadtree in geometric settings...

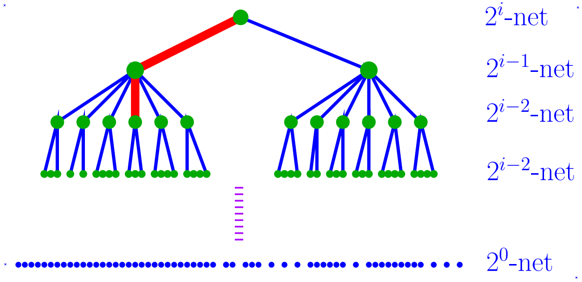
Nets and ANNS



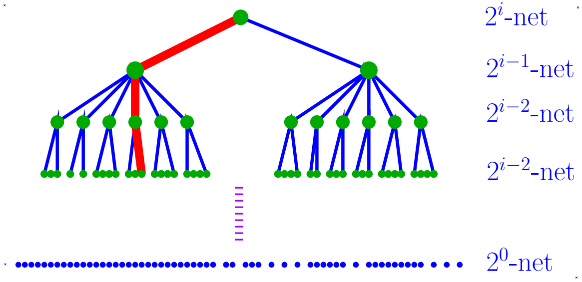
Nets and ANNS



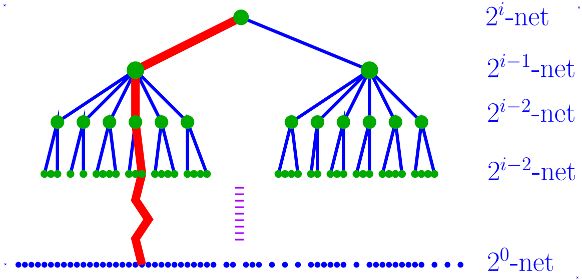
Nets and ANNS



Nets and ANNS



Nets and ANNS



Main Technical Result

Theorem

A hierarchy of nets at all scales can be constructed in $2^{O(ddim)} n \log n$ *time.*

Remark

The input size is $\Theta(n^2)$. The alg' is randomized and uses sampling.

Observation

Ideas leads to simpler construction of compressed quadtrees.

Main Technical Result

Theorem

A hierarchy of nets at all scales can be constructed in $2^{O(ddim)} n \log n$ time.

Remark

The input size is $\Theta(n^2)$. The alg' is randomized and uses sampling.

Observation

Ideas leads to simpler construction of compressed quadtrees.

Main Technical Result

Theorem

A hierarchy of nets at all scales can be constructed in $2^{O(ddim)} n \log n$ *time.*

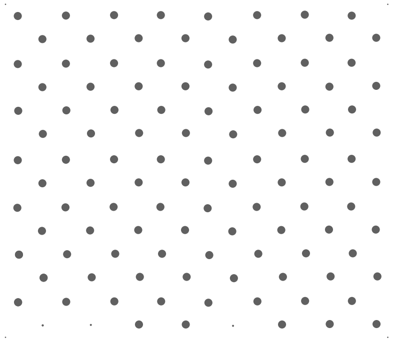
Remark

The input size is $\Theta(n^2)$. The alg' is randomized and uses sampling.

Observation

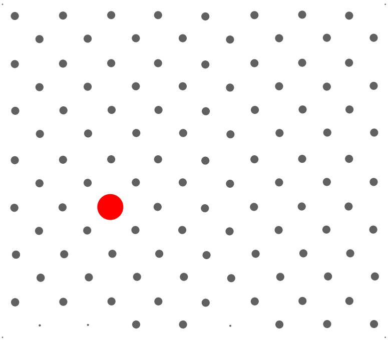
Ideas leads to simpler construction of compressed quadtrees.

Gonzalez Sequence



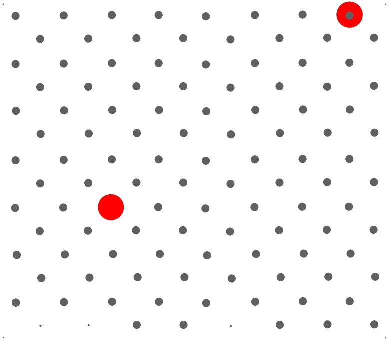
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



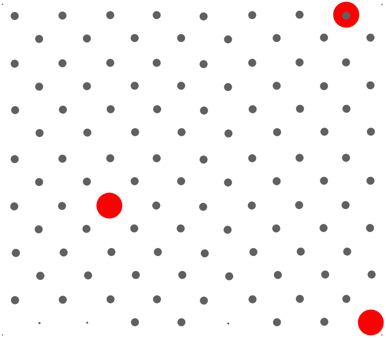
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



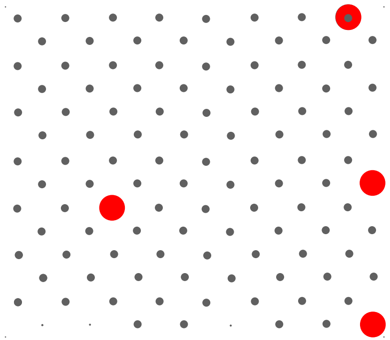
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



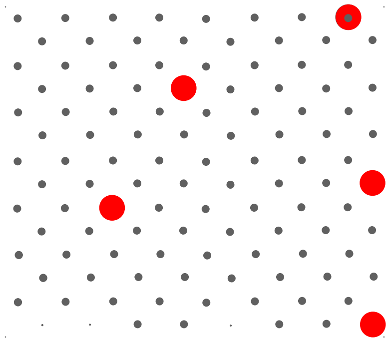
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



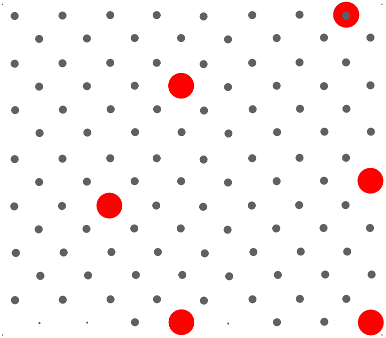
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



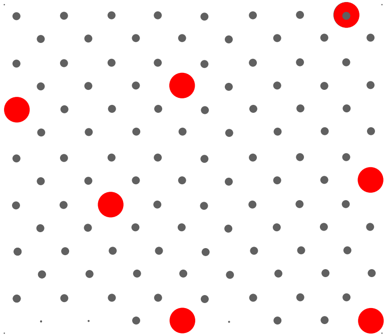
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



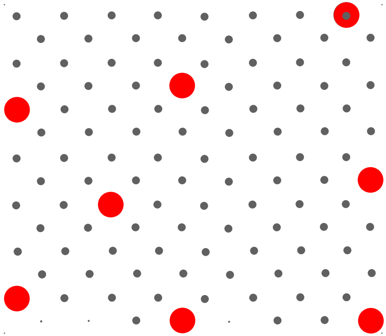
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



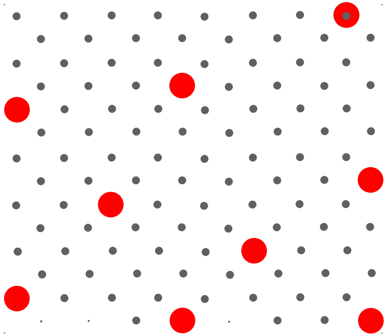
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



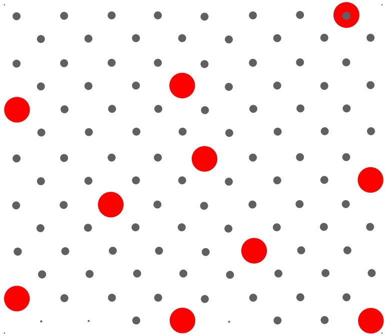
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



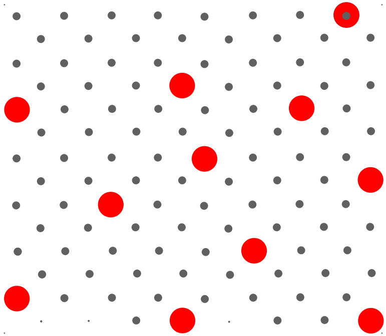
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



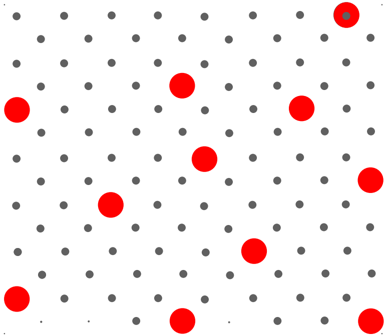
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



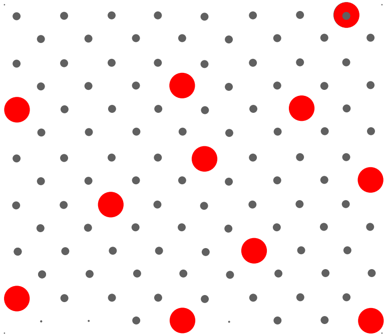
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



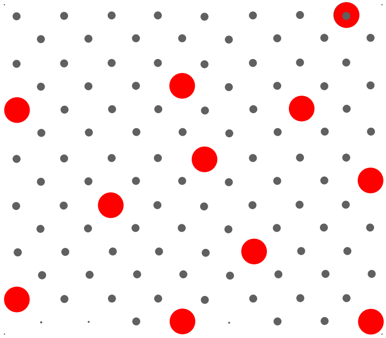
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



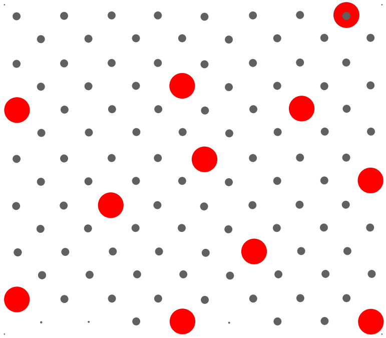
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



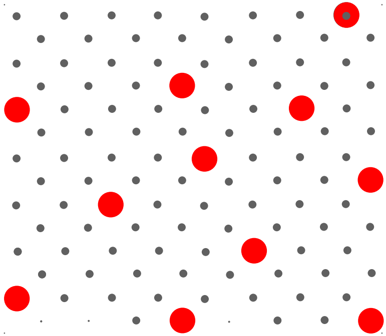
1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by [Gonzalez, 1985] takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Gonzalez Sequence



1. Prefix of the sequence: r-net for some r.
2. Set of prefixes: Nets at all scales.
3. "invented" by **[Gonzalez, 1985]** takes $O(n^2)$ time.
4. New result: compute sequence in $O(n \log n)$ time.
5. Not trivial... Nasty issues with spread.

Aprox. Nearest Neighbor in Doubling Metrics

Theorem

Any n-point metric with fixed doubling dimension admits ANNS with
Approx $= 1 + \varepsilon$, *Query* $= O(\varepsilon^{-\mathrm{ddim}} + \log n)$,
Space $= O(n)$, *Preprocessing time* $= O(n \log n)$.

- Basic idea appears already in [Clarkson, 1999] (without the analysis and fast construction).
- Similar to results known for $\mathbb{R}^d$.
- Improves over:
  - [Krauthgamer and Lee, 2004b] - dependency on spread.
  - [Krauthgamer and Lee, 2004a] - quadratic space.

Aprox. Nearest Neighbor in Doubling Metrics

Theorem

Any n-point metric with fixed doubling dimension admits ANNS with
Approx$= 1 + \varepsilon$, *Query*$=O(\varepsilon^{-\mathrm{ddim}} + \log n)$,
Space$=O(n)$, *Preprocessing time*$= O(n \log n)$.

- Basic idea appears already in **[Clarkson, 1999]** (without the analysis and fast construction).
- Similar to results known for $\mathbb{R}^d$.
- Improves over:
  - **[Krauthgamer and Lee, 2004b]** - dependency on spread.
  - **[Krauthgamer and Lee, 2004a]** - quadratic space.

Aprox. Nearest Neighbor in Doubling Metrics

Theorem

Any n-point metric with fixed doubling dimension admits ANNS with
Approx$= 1 + \varepsilon$, *Query*$=O(\varepsilon^{-\mathrm{ddim}} + \log n)$,
Space$=O(n)$, *Preprocessing time*$= O(n \log n)$.

- Basic idea appears already in **[Clarkson, 1999]** (without the analysis and fast construction).
- Similar to results known for $\mathbb{R}^d$.
- Improves over:
  - **[Krauthgamer and Lee, 2004b]** - dependency on spread.
  - **[Krauthgamer and Lee, 2004a]** - quadratic space.

Aprox. Nearest Neighbor in Doubling Metrics

Theorem

Any n-point metric with fixed doubling dimension admits ANNS with
Approx$= 1 + \varepsilon$, *Query*$=O(\varepsilon^{-\mathrm{ddim}} + \log n)$,
Space$=O(n)$, *Preprocessing time*$= O(n \log n)$.

- Basic idea appears already in **[Clarkson, 1999]** (without the analysis and fast construction).
- Similar to results known for $\mathbb{R}^d$.
- Improves over:
  - **[Krauthgamer and Lee, 2004b]** - dependency on spread.
  - **[Krauthgamer and Lee, 2004a]** - quadratic space.

Well-separated pairs decomposition

- P: set of n points with ddim dim.
- Result: Compute WSPD in near linear time.
- # of pairs is $n\varepsilon^{-O(\mathsf{dim})}$
- Improves **[Talwar, 2004]** work (spread).
- Matches the results of **[Callahan and Kosaraju, 1995]**

Spanners.

- $1 + \varepsilon$-spanners
- number of edges $n\varepsilon^{-O(\mathrm{dim})}$
- Similar result by **[Chan et al., 2005]**.

Compact Representation Scheme (CRS).

- Build a data-structure.
- linear size
- approx. distance queries between pairs of points in constant time.
- “approximate distance oracles” **[Thorup and Zwick, 2001]**
- Extends **[Gudmundsson et al., 2002a, Gudmundsson et al., 2002b]** (geometric settings)
- Improve/unify: **[Talwar, 2004, Slivkins, 2005]**.

A variant of Assouad's embedding.

Theorem

$\mathcal{M}$: *metric space with doubling dim* $\dim$
Can be embedded in ℓ_∞^d.
Where $d \leq \varepsilon^{-O(\dim)}$
$(\mathcal{M}, \sqrt{d_\mathcal{M}})$ *is distorted by* $1+\varepsilon$ *factor.*

Embedding can be computed quickly...
[Indyk and Naor, 2005]
Better dimension reduction for such cases.

A variant of Assouad's embedding.

Theorem

$\mathcal{M}$: *metric space with doubling dim* dim
Can be embedded in ℓ_∞^d.
Where $d \leq \varepsilon^{-O(\mathrm{dim})}$
$(\mathcal{M}, \sqrt{d_{\mathcal{M}}})$ *is distorted by* $1 + \varepsilon$ *factor.*

Embedding can be computed quickly...
[Indyk and Naor, 2005]
Better dimension reduction for such cases.

Doubling Measure.

- A doubling measure μ:
  $x \in P$ and $r > 0$, the ratio $\mu(\mathbf{b}(x, 2r))/\mu(\mathbf{b}(x, r))$ is bounded.
  $\mathbf{b}(x, r) = \{y : \; d(x, y) \le r\}$.
- Another definition of dimension. Weaker.
- **[Vol′berg and Konyagin, 1987]**
  Any doubling-dim metric can be converted to be doubling measure.
  Weights of points exponential.
- Present a near linear time algorithm for computing it.

Lipschitz Constant of a Mapping.

- $f : P \to B$
- Compute Lipschitz constant of f?
- Approx follows from WSPD.
  Near linear time.
- $P \subseteq \mathbb{R}^2$ or $P \subseteq \mathbb{R}$:
  Compute Lipschitz constant exactly in near linear time.

Computing the Doubling Dimension.

- $\mathcal{M}$: Finite metric space.
- Q: What is the doubling dim of $\mathcal{M}$?

New result

One can compute $O(\dim(\mathcal{M}))$ in $2^{O(\dim)} n \log n$ time

No need to know the doubling dim in advance...

Computing the Doubling Dimension.

- $\mathcal{M}$: Finite metric space.
- Q: What is the doubling dim of $\mathcal{M}$?

New result

One can compute $O(\dim(\mathcal{M}))$ in $2^{O(\dim)} n \log n$ time

No need to know the doubling dim in advance...

Computing the Doubling Dimension.

- $\mathcal{M}$: Finite metric space.
- Q: What is the doubling dim of $\mathcal{M}$?

New result

One can compute $O(\dim(\mathcal{M}))$ in $2^{O(\dim)} n \log n$ time

No need to know the doubling dim in advance...

Open problems

- Does doubling metric really exists in nature?
  Temporary answer: yes and no.
- Simplify net-tree construction?
- Good spanners - light and low diam?
- What is dimension?
  (Recent survey by Clarkson.)

Thank You

Arya, V., Garg, N., Khandekar, R., Munagala, K., and Pandit, V. (2001).
Local search heuristic for k-median and facility location problems.
In *Proc. 33rd Annu. ACM Sympos. Theory Comput.*, pages 21–29.

Assouad, P. (1983).
Plongements lipschitziens dans $\mathbf{R}^n$.
Bull. Soc. Math. France, 111(4):429–448.

Callahan, P. B. and Kosaraju, S. R. (1995).
A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields.
J. Assoc. Comput. Mach., 42:67–90.

Chan, H. T.-H., Gupta, A., Maggs, B. M., and Zhou, S. (2005).

On hierarchical routing in bounded growth metrics.
In *Proc. 16th ACM-SIAM Sympos. Discrete Algorithms*.

Clarkson, K. L. (1999).
Nearest neighbor queries in metric spaces.
Discrete Comput. Geom., 22(1):63–93.

Gonzalez, T. (1985).
Clustering to minimize the maximum intercluster distance.
Theoret. Comput. Sci., 38:293–306.

Gudmundsson, J., Levcopoulos, C., Narasimhan, G., and Smid, M. (2002a).
Approximate distance oracles for geometric graphs.
In *Proc. 13th ACM-SIAM Sympos. Discrete Algorithms*, pages 828–837. Society for Industrial and Applied Mathematics.

Gudmundsson, J., Levcopoulos, C., Narasimhan, G., and Smid, M. H. M. (2002b).
Approximate distance oracles revisited.

In *Proc. 13th Annu. Internat. Sympos. Algorithms Comput.*, pages 357–368. Springer-Verlag.

Gupta, A., Krauthgamer, R., and Lee, J. R. (2003).
Bounded geometries, fractals, and low-distortion embeddings.
In *Proc. 44th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 534–543.

Indyk, P. and Naor, A. (2005).
Nearest neighbor preserving embeddings.
unpublished.

Krauthgamer, R. and Lee, J. R. (2004a).
The black-box complexity of nearest neighbor search.
In *Proc. 31st Internat. Colloq. Automata Lang. Prog.*, page to appear.

Krauthgamer, R. and Lee, J. R. (2004b).
Navigating nets: simple algorithms for proximity search.

In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 798–807. Society for Industrial and Applied Mathematics.

Slivkins, A. (2004).
Distributed approached to triangulation and embedding.

Slivkins, A. (2005).
Distributed approaches to triangulation and embedding.
In *Proc. 16th ACM-SIAM Sympos. Discrete Algorithms*.

Talwar, K. (2004).
Bypassing the embedding: algorithms for low dimensional metrics.
In *Proc. 36th Annu. ACM Sympos. Theory Comput.*, pages 281–290. ACM Press.

Thorup, M. and Zwick, U. (2001).
Approximate distance oracles.

In *Proc. 33rd Annu. ACM Sympos. Theory Comput.*, pages 183–192. ACM Press.

Vol′berg, A. L. and Konyagin, S. V. (1987).
On measures with the doubling condition.
Izv. Akad. Nauk SSSR Ser. Mat., 51(3):666–675.