Finding Haystacks (and Similar Structures) in Geometry

Sariel Har-Peled

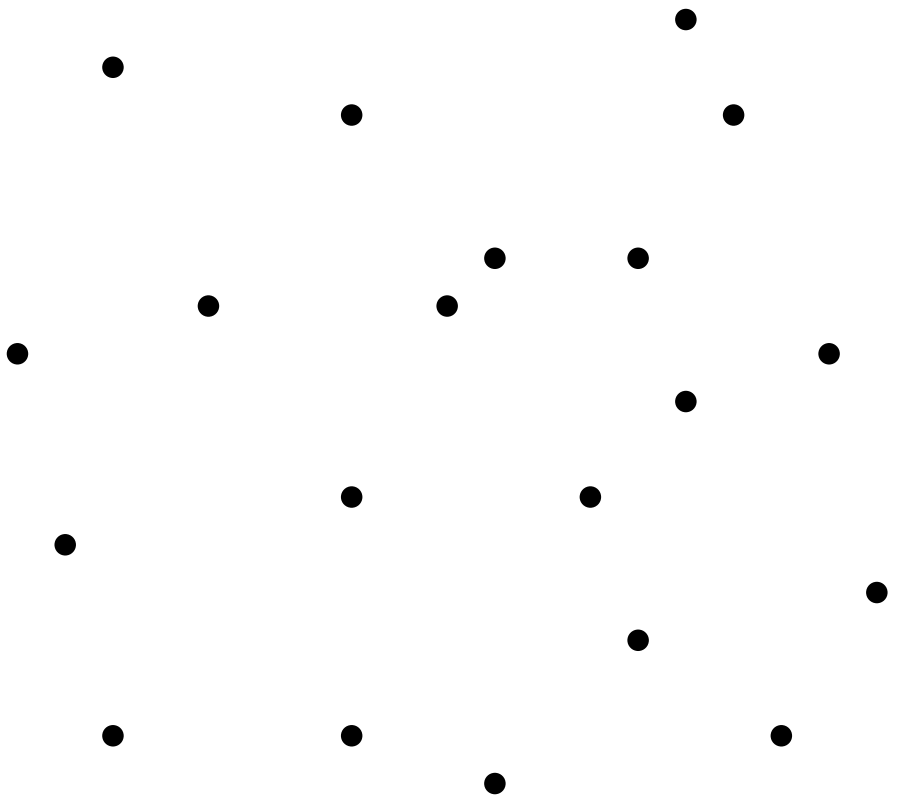
[2]UIUC, Illinois, USA

September 2, 2010

2: Motivating problem

Placing an antenna.

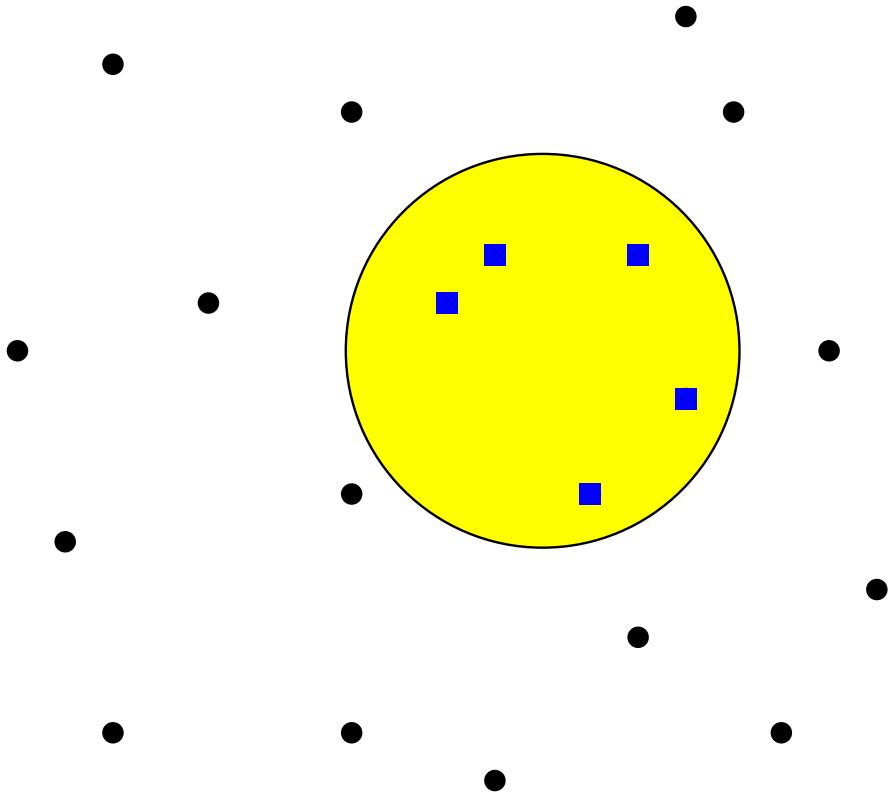
P: Set of **n** points (customers)
Q: Find location of antenna that serves maximum number of customers.



2: Motivating problem

Placing an antenna.

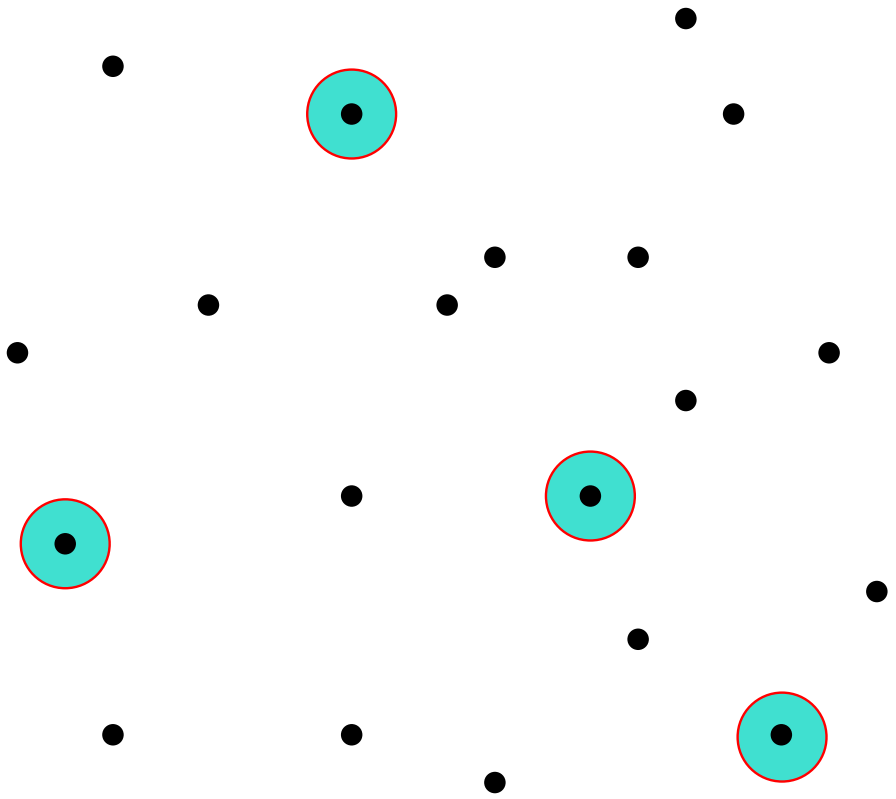
P: Set of **n** points (customers)
Q: Find location of antenna that serves maximum number of customers.



2: Motivating problem

Placing an antenna.

P: Set of **n** points (customers)
Q: Find location of antenna that serves maximum number of customers.



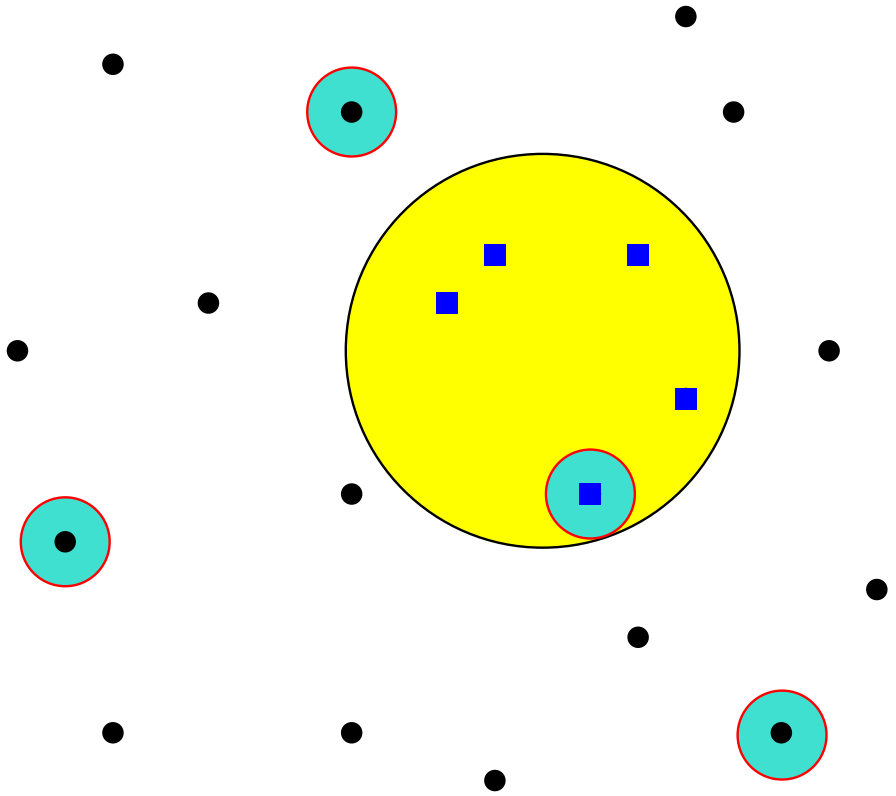
Approach...

- Pick a random sample.
- Count sample points.

2: Motivating problem

Placing an antenna.

P: Set of **n** points (customers)
Q: Find location of antenna that serves maximum number of customers.



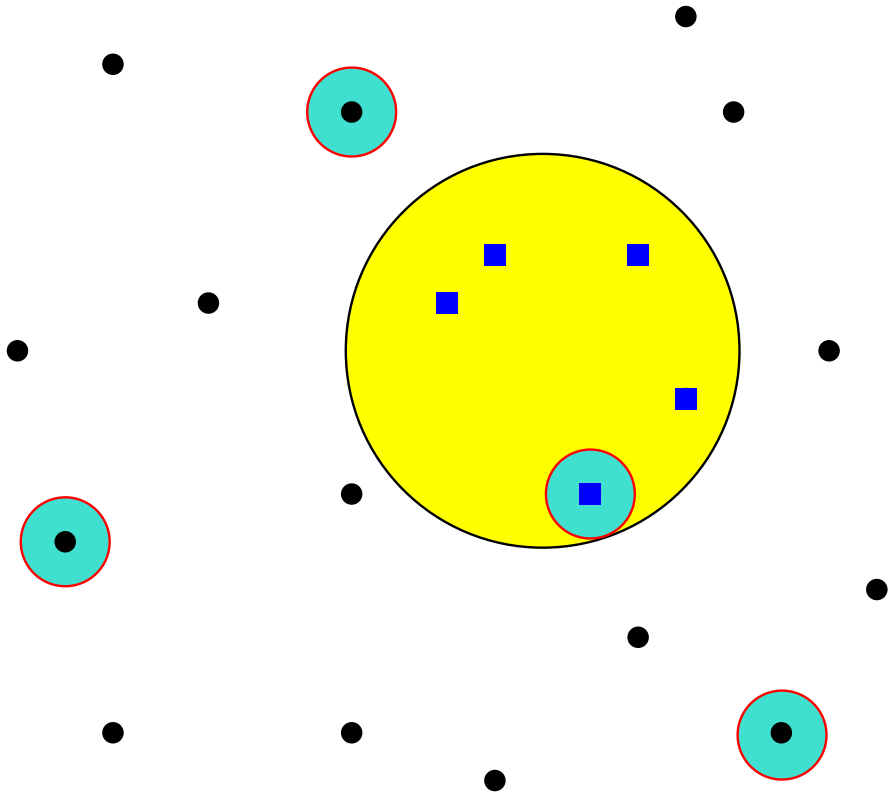
Approach...

- Pick a random sample.
- Count sample points.

2: Motivating problem

Placing an antenna.

P: Set of **n** points (customers)
Q: Find location of antenna that serves maximum number of customers.



Approach...

- Pick a random sample.
- Count sample points.

Challenge...

- Good for all disks.
- ... Infinite number of disks.

3: What we want

Definitions

$\mathbf{P}$: Set of points in the plane.

Measure

For any disk $\mathbf{D}$ its **measure** is $\overline{\mathbf{P}}(\mathbf{D}) = \dfrac{|\mathbf{D} \cap \mathbf{P}|}{|\mathbf{P}|}$.

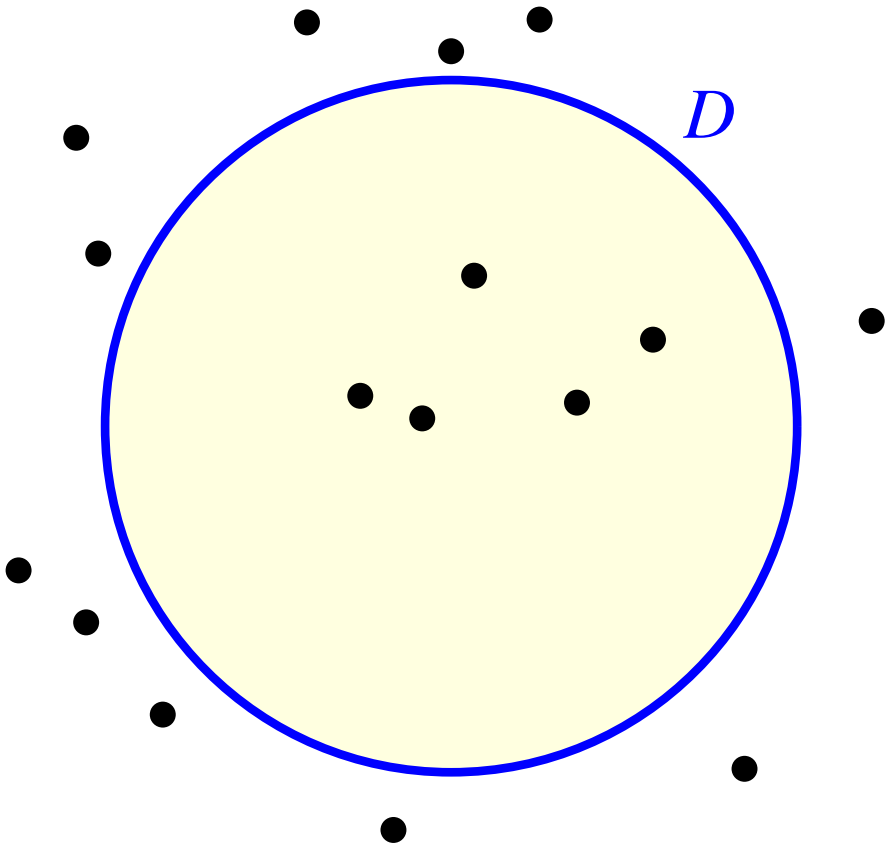
ε-Sample

$\varepsilon > \mathbf{0}$: Approx parameter.
$\mathbf{S} \subseteq \mathbf{P}$ is **ε-sample** if

$$\forall \mathbf{D} \quad \left|\overline{\mathbf{P}}(\mathbf{D}) - \overline{\mathbf{S}}(\mathbf{D})\right| < \varepsilon$$

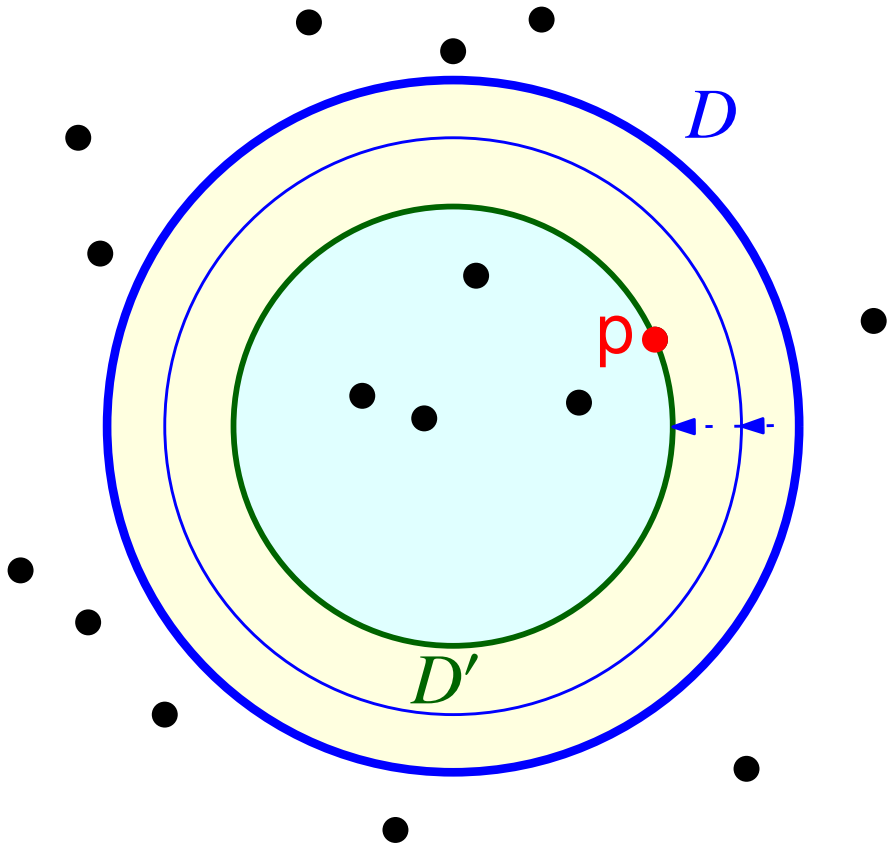
4: Bounding infinity...

...or, it just looks infinite.



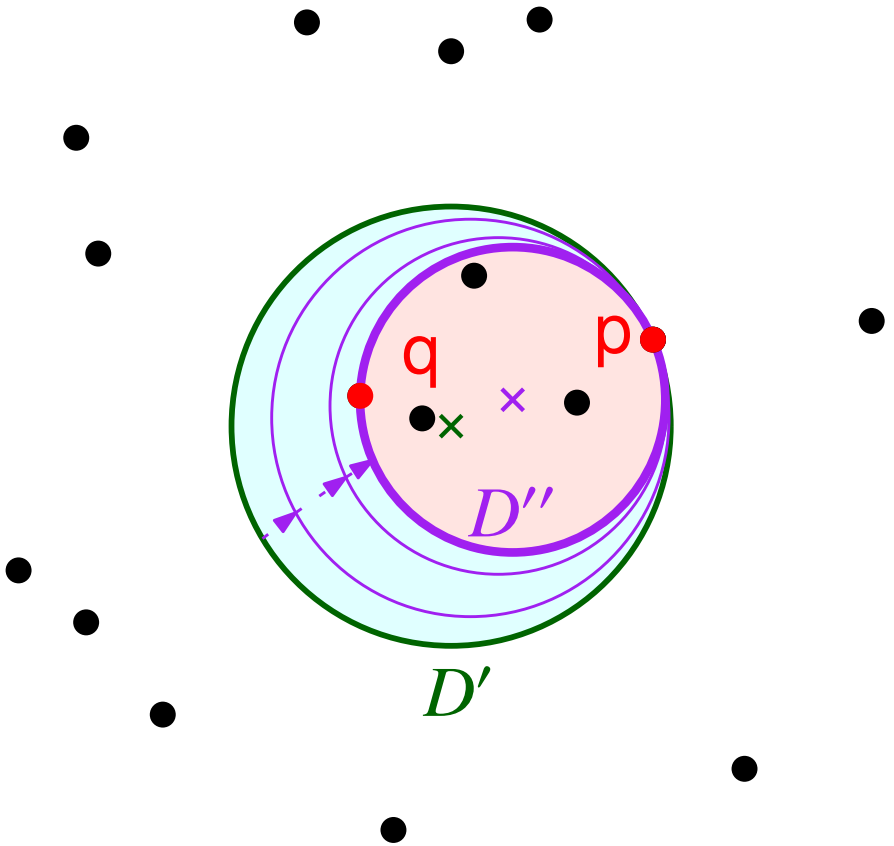
4: Bounding infinity...

...or, it just looks infinite.



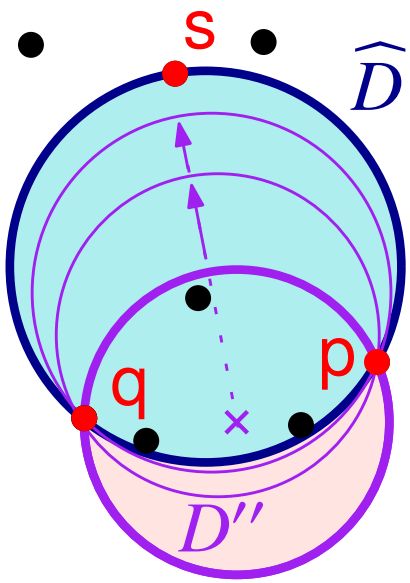
4: Bounding infinity...

...or, it just looks infinite.



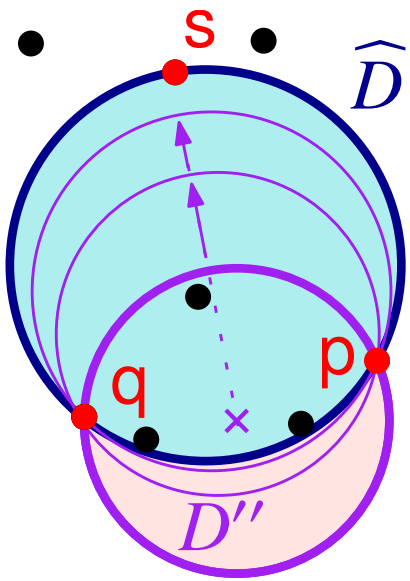
4: Bounding infinity...

...or, it just looks infinite.



4: Bounding infinity...

...or, it just looks infinite.



As such...

n: number of points.
Only $\mathbf{O(n^3)}$ different subsets.

Conclusion: ε-Sample

Std. random sampling...
$\mathbf{O}\left(\frac{\log n}{\varepsilon^2}\right)$.

Known: ε-Sample

[Vapnik and Chervonenkis, 1971]
[Li et al., 2001] $\mathbf{O}\left(\frac{1}{\varepsilon^2}\right)$.

5: Application

...or why size matters.

Problem

P: **n** points in the plane

Compute: Smallest disk containing half the points.

min_disk$(\mathbf{P}, \mathbf{n}/2)$: $\mathbf{O(n^2)}$ time exact algorithm.

Approximation algorithm

$\varepsilon > \mathbf{0}$: Approx parameter.

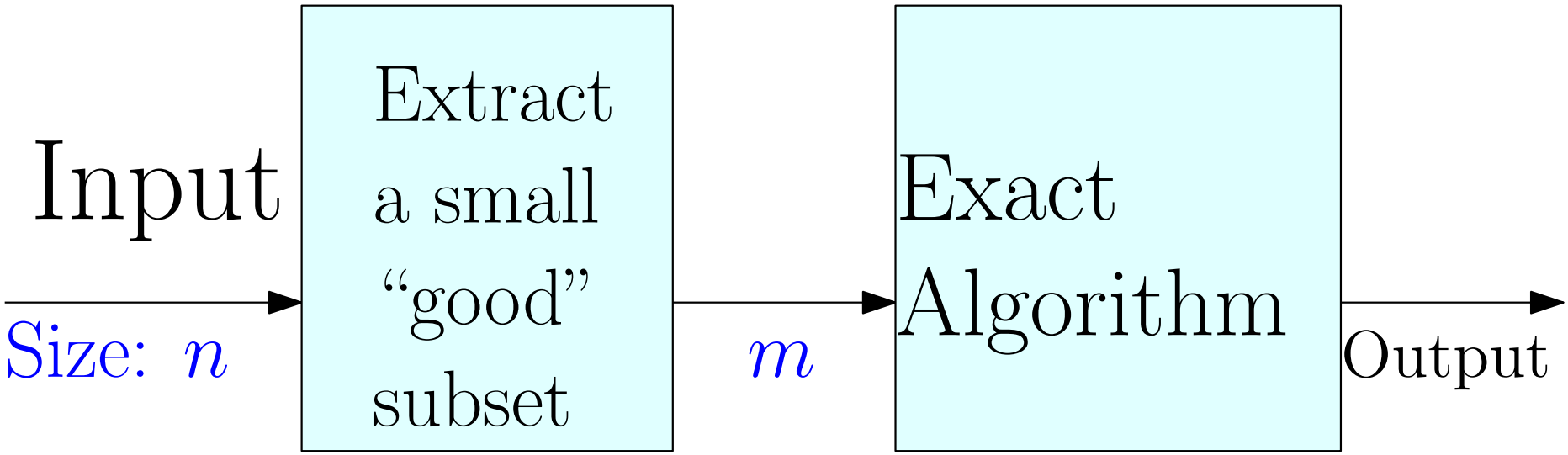
$\mathbf{S} \subseteq \mathbf{P}$: ε-Sample of size **m**.

Return **min_disk**$(\mathbf{S}, \mathbf{(1-\varepsilon)m/2})$.

Ret disk has $\geq \mathbf{(1/2 - 2\varepsilon)n}$ points and smaller than exact answer.

6: Application

...or why size matters.

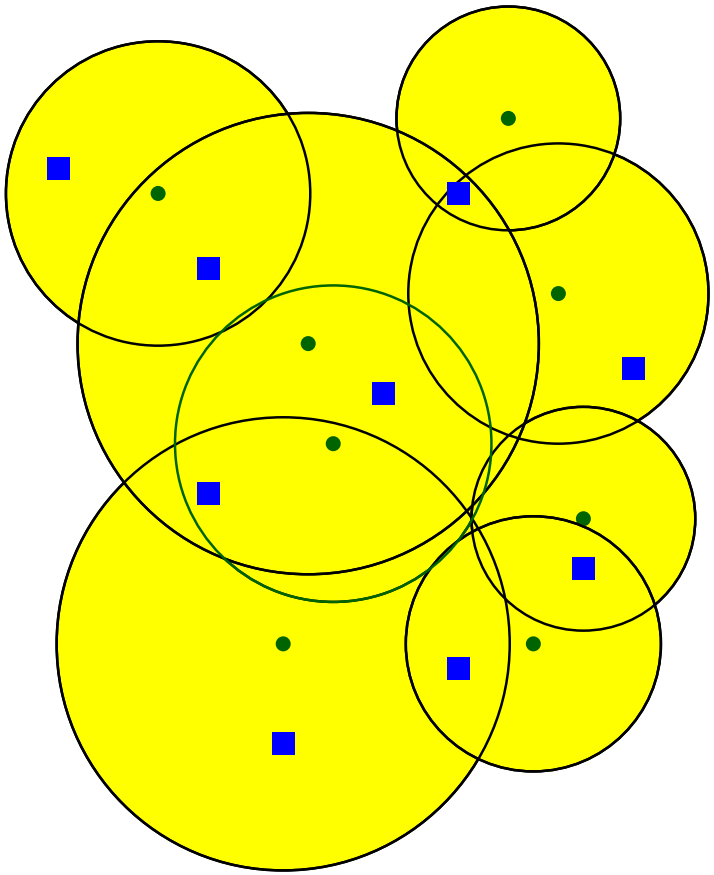


Exact Running time: $\mathbf{T_{exact}(n)}$
New : $\mathbf{T_{extract}(n) + T_{exact}(m)}$

Example: Min disk containing $\mathbf{n/2}$ points

$\mathbf{T_{extract}(n) = O(m)}$, $\mathbf{T_{exact}(n) = O(n^2)}$, $\mathbf{m = 1/\varepsilon^2}$.
Running time: $\mathbf{O(1/\varepsilon^4)}$.

7: Geometric set systems



- $\mathcal{F}$: $\mathbf{m}$ of disks in the plane.
- $\mathbf{P}$: set of points
- Every disk $\mathbf{D} \in \mathcal{F}$ corresponds to the subset $\mathbf{P} \cap \mathbf{D}$.
- Induced **set system**:

$$\left(\mathbf{P}, \left\{ \mathbf{P} \cap \mathbf{D} \;\middle|\; \mathbf{D} \in \mathcal{F} \right\}\right).$$

8: Sketch of a sketch is a sketch

...or how many sketches would a sketch sketch, if a sketch could sketch sketches?

Lemma (Sketch property.)

$\mathbf{X} \subset \mathbf{Y}$: δ-sample.
$\mathbf{Y} \subset \mathbf{Z}$: δ'-sample.
$\Longrightarrow$ $\mathbf{X}$ is a $(\delta + \delta')$-sample of $\mathbf{Z}$.

$\mathbf{Z} = \mathbf{P}$: Set of $\mathbf{n}$ points
$\exists \mathbf{Y} \subset \mathbf{Z}$: ε-Sample $|\mathbf{Y}| = \mathbf{O}\left(\frac{\log n}{\varepsilon^2}\right)$.
$\exists \mathbf{X} \subset \mathbf{Y}$: ε-Sample $|\mathbf{X}| = \mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.
$\Longrightarrow$
$\mathbf{X}$ is 2ε-sample of $\mathbf{P}$ of size $\mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.

Theorem [Vapnik and Chervonenkis, 1971]

Exists ε-sample of size $\mathbf{O}\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$.

8: Sketch of a sketch is a sketch

...or how many sketches would a sketch sketch, if a sketch could sketch sketches?

Lemma (Sketch property.)

$\mathbf{X} \subset \mathbf{Y}$: δ-sample.
$\mathbf{Y} \subset \mathbf{Z}$: δ'-sample.
$\Longrightarrow$ $\mathbf{X}$ is a $(\delta + \delta')$-sample of $\mathbf{Z}$.

$\mathbf{Z} = \mathbf{P}$: Set of $\mathbf{n}$ points
$\exists \mathbf{Y} \subset \mathbf{Z}$: ε-Sample $|\mathbf{Y}| = \mathbf{O}\left(\frac{\log n}{\varepsilon^2}\right)$.
$\exists \mathbf{X} \subset \mathbf{Y}$: ε-Sample $|\mathbf{X}| = \mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.

$\Longrightarrow$

$\mathbf{X}$ is 2ε-sample of $\mathbf{P}$ of size $\mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.

Theorem [Vapnik and Chervonenkis, 1971]

Exists ε-sample of size $\mathbf{O}\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$.

8: Sketch of a sketch is a sketch

...or how many sketches would a sketch sketch, if a sketch could sketch sketches?

Lemma (Sketch property.)

$\mathbf{X} \subset \mathbf{Y}$: δ-sample.
$\mathbf{Y} \subset \mathbf{Z}$: δ'-sample.
$\Longrightarrow$ $\mathbf{X}$ is a $(\delta + \delta')$-sample of $\mathbf{Z}$.

$\mathbf{Z} = \mathbf{P}$: Set of $\mathbf{n}$ points
$\exists \mathbf{Y} \subset \mathbf{Z}$: ε-Sample $|\mathbf{Y}| = \mathbf{O}\left(\frac{\log n}{\varepsilon^2}\right)$.
$\exists \mathbf{X} \subset \mathbf{Y}$: ε-Sample $|\mathbf{X}| = \mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.
$\Longrightarrow$
$\mathbf{X}$ is 2ε-sample of $\mathbf{P}$ of size $\mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.

Theorem [Vapnik and Chervonenkis, 1971]

Exists ε-sample of size $\mathbf{O}\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$.

8: Sketch of a sketch is a sketch

...or how many sketches would a sketch sketch, if a sketch could sketch sketches?

Lemma (Sketch property.)

$\mathbf{X} \subset \mathbf{Y}$: δ-sample.
$\mathbf{Y} \subset \mathbf{Z}$: δ'-sample.
$\Longrightarrow$ $\mathbf{X}$ is a $(\delta + \delta')$-sample of $\mathbf{Z}$.

$\mathbf{Z} = \mathbf{P}$: Set of $\mathbf{n}$ points
$\exists \mathbf{Y} \subset \mathbf{Z}$: ε-Sample $|\mathbf{Y}| = \mathbf{O}\left(\frac{\log n}{\varepsilon^2}\right)$.
$\exists \mathbf{X} \subset \mathbf{Y}$: ε-Sample $|\mathbf{X}| = \mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.
$\Longrightarrow$
$\mathbf{X}$ is 2ε-sample of $\mathbf{P}$ of size $\mathbf{O}\left(\frac{\log \log n}{\varepsilon^2}\right)$.

Theorem [Vapnik and Chervonenkis, 1971]

Exists ε-sample of size $\mathbf{O}\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$.

9: Netting a lot with a little

or, getting less with even less

Smaller ε-samples are better.

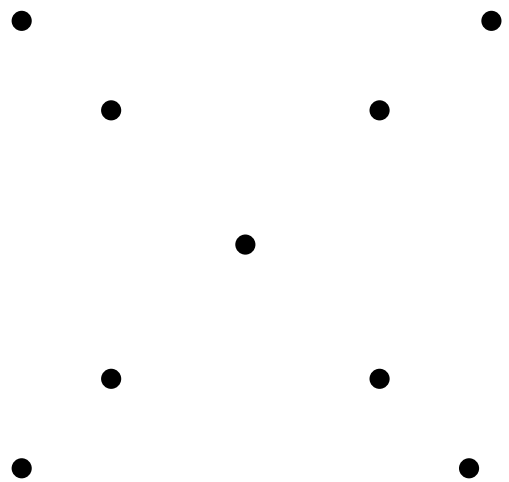
$\mathbf{P}$: $\mathbf{n}$ points in the plane.

ε-Net

$\varepsilon > \mathbf{0}$: Approx parameter.

$\mathbf{S} \subseteq \mathbf{P}$ is ε-**net** if

$$\forall \mathbf{D} \quad \left|\overline{\mathbf{P}}(\mathbf{D})\right| \geq \varepsilon \implies \mathbf{S} \cap \mathbf{D} \neq \emptyset.$$



9: Netting a lot with a little

or, getting less with even less

Smaller ε-samples are better.

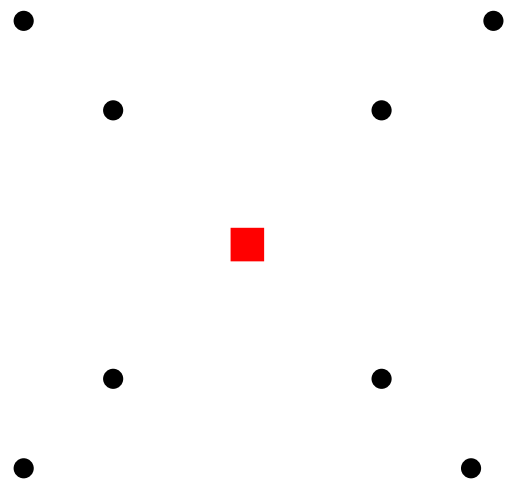
P: **n** points in the plane.

ε-Net

$\varepsilon > 0$: Approx parameter.

S $\subseteq$ **P** is ε-**net** if

$$\forall \mathbf{D} \quad \left|\overline{\mathbf{P}}(\mathbf{D})\right| \geq \varepsilon \implies \mathbf{S} \cap \mathbf{D} \neq \emptyset.$$



9: Netting a lot with a little

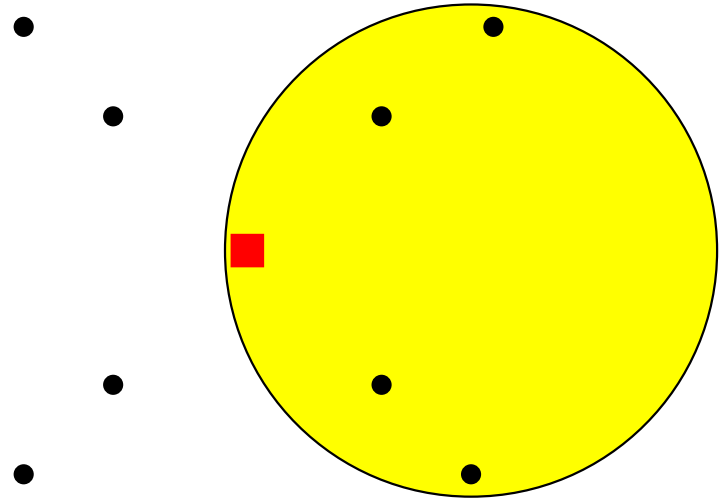
or, getting less with even less

Smaller ε-samples are better.
P: **n** points in the plane.

ε-Net

$\varepsilon > 0$: Approx parameter.
$\mathbf{S} \subseteq \mathbf{P}$ is ε-**net** if

$$\forall \mathbf{D} \quad \left|\overline{\mathbf{P}}(\mathbf{D})\right| \geq \varepsilon \implies \mathbf{S} \cap \mathbf{D} \neq \emptyset.$$



10: ε-Sample vs. ε-net

P: **n** points in the plane.
$\varepsilon > 0$: Approx parameter.

ε-Net

S $\subseteq$ **P** is ε-**net** if $\forall$**D** $\left|\overline{\mathbf{P}}(\mathbf{D})\right| \geq \varepsilon \implies \mathbf{S} \cap \mathbf{D} \neq \emptyset$.

ε-Sample

S $\subseteq$ **P** is ε-**sample** if $\forall$**D** $\left|\overline{\mathbf{P}}(\mathbf{D}) - \overline{\mathbf{S}}(\mathbf{D})\right| < \varepsilon$.

ε-sample is an ε-net.

Theorem (ε-net theorem)

[Haussler and Welzl, 1987]
Random sample of size $\mathbf{O}\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)$ *is an* ε*-net.*

10: ε-Sample vs. ε-net

P: **n** points in the plane.
$\varepsilon > 0$: Approx parameter.

ε-Net

S $\subseteq$ **P** is ε-**net** if $\forall$**D** $\left|\overline{\mathbf{P}}(\mathbf{D})\right| \geq \varepsilon \implies \mathbf{S} \cap \mathbf{D} \neq \emptyset$.

ε-Sample

S $\subseteq$ **P** is ε-**sample** if $\forall$**D** $\left|\overline{\mathbf{P}}(\mathbf{D}) - \overline{\mathbf{S}}(\mathbf{D})\right| < \varepsilon$.

ε-sample is an ε-net.

Theorem (ε-net theorem)

[Haussler and Welzl, 1987]
Random sample of size $\mathbf{O}\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)$ *is an* ε*-net.*

10: ε-Sample vs. ε-net

P: **n** points in the plane.
$\varepsilon > 0$: Approx parameter.

ε-Net

S $\subseteq$ **P** is ε-**net** if $\quad \forall \mathbf{D} \quad \left|\overline{\mathbf{P}}(\mathbf{D})\right| \geq \varepsilon \implies \mathbf{S} \cap \mathbf{D} \neq \emptyset.$

ε-Sample

S $\subseteq$ **P** is ε-**sample** if $\quad \forall \mathbf{D} \quad \left|\overline{\mathbf{P}}(\mathbf{D}) - \overline{\mathbf{S}}(\mathbf{D})\right| < \varepsilon.$

ε-sample is an ε-net.

Theorem (ε-net theorem)

[Haussler and Welzl, 1987]
Random sample of size $\mathbf{O}\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)$ *is an* ε*-net.*

11: Hitting set via LP relaxation

ε-nets and Integrality gap.

LP relax' of hitting set.

$$\min \quad \mathrm{Opt} = \sum_{i=1}^{n} x_i$$

$$\text{s.t.} \sum_{p_i \in D_j} x_i \geq 1 \qquad \forall j$$

$$x_i \geq 0 \qquad \forall i$$

Lemma [Long, 2001]

In geometric settings there is an ε-net of size $O(K/\varepsilon)$ iff the integrality gap is K.

ε-net of size...

$O(1/\varepsilon \log 1/\varepsilon) \implies O(\log \mathrm{Opt})$-approximation hitting-set

$O(1/\varepsilon) \implies O(1)$-approximation

12: On small ε-nets

Upper bounds

- $\mathbf{O}\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)$: ε-net theorem.
- $\mathbf{O(1/\varepsilon)}$: halfplanes, halfspaces, pseudo-disks.
- $\mathbf{O}\left(\frac{1}{\varepsilon}\log\frac{\mathbf{U}(1/\varepsilon)}{1/\varepsilon}\right)$
  [Aronov, Ezra and Sharir, 2009]
  $\mathbf{U(n)}$: Union complexity of $\mathbf{n}$ shapes.

Lower bounds

- $\mathbf{\Omega}\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)$: Lower bound **[Komlós et al., 1992]**.
- $\mathbf{\Omega}\left(\frac{1}{\varepsilon}\mathbf{w}\left(\frac{1}{\varepsilon}\right)\right)$: points and lines **[Alon, 2010]**.

13: Other notions of samples

Name	Property $\forall \mathbf{D} \in \mathcal{F}$	Sample size
ε-net	$\overline{\mathbf{m}} = \overline{\mathbf{m}}(\mathbf{r}), \bar{\mathbf{s}} = \bar{\mathbf{s}}(\mathbf{r})$ $\overline{\mathbf{m}} \geq \varepsilon \quad \Rightarrow \quad \bar{\mathbf{s}} > 0$	$\mathbf{O}\left(\frac{\delta}{\varepsilon} \log \frac{1}{\varepsilon}\right)$
ε-sample	$\lvert\overline{\mathbf{m}} - \bar{\mathbf{s}}\rvert \leq \varepsilon$	$\mathbf{O}\left(\frac{\delta}{\varepsilon^2}\right)$
Sensitive ε-approx.	$\lvert\overline{\mathbf{m}} - \bar{\mathbf{s}}\rvert \leq \frac{\varepsilon}{2}\left(\sqrt{\overline{\mathbf{m}}} + \varepsilon\right)$	$\mathbf{O}\left(\frac{\delta}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$
Relative $(\varepsilon, \mathbf{p})$-approx.	$\overline{\mathbf{m}} \geq \mathbf{p} \quad \Rightarrow$ $(1-\varepsilon)\overline{\mathbf{m}} \leq \bar{\mathbf{s}} \leq (1+\varepsilon)\overline{\mathbf{m}}$	$\mathbf{O}\left(\frac{\delta}{\varepsilon^2 \mathbf{p}} \log \frac{1}{\mathbf{p}}\right)$

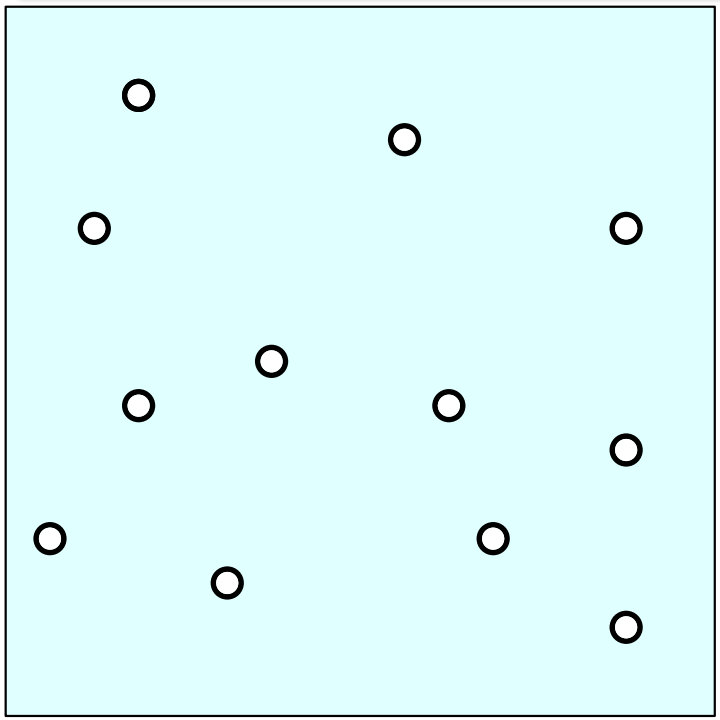
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(1/\sqrt{\mathbf{n}})$-sample.

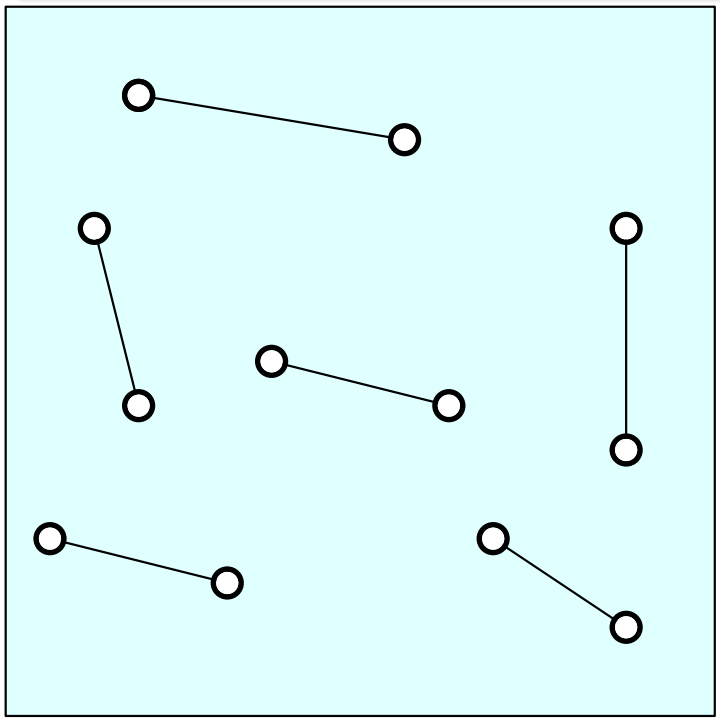
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(1/\sqrt{\mathbf{n}})$-sample.

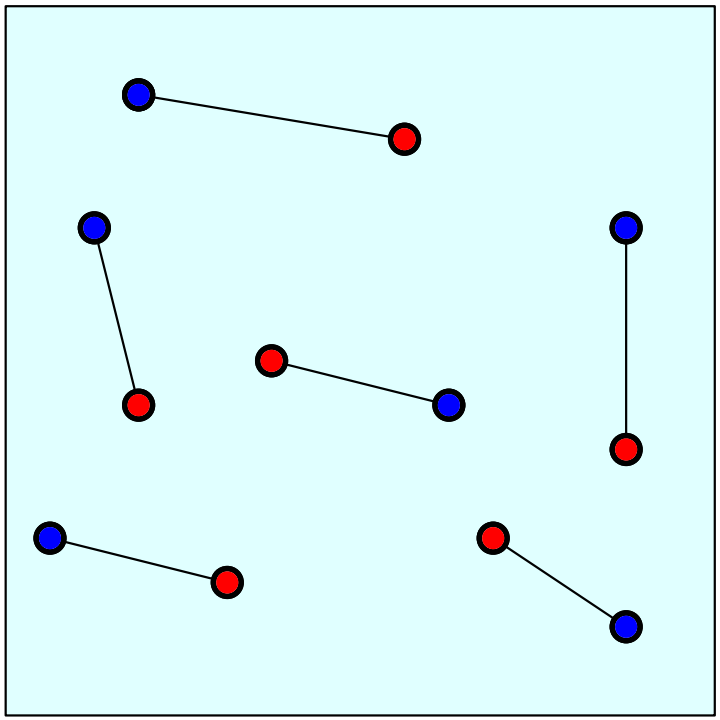
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(1/\sqrt{\mathbf{n}})$-sample.

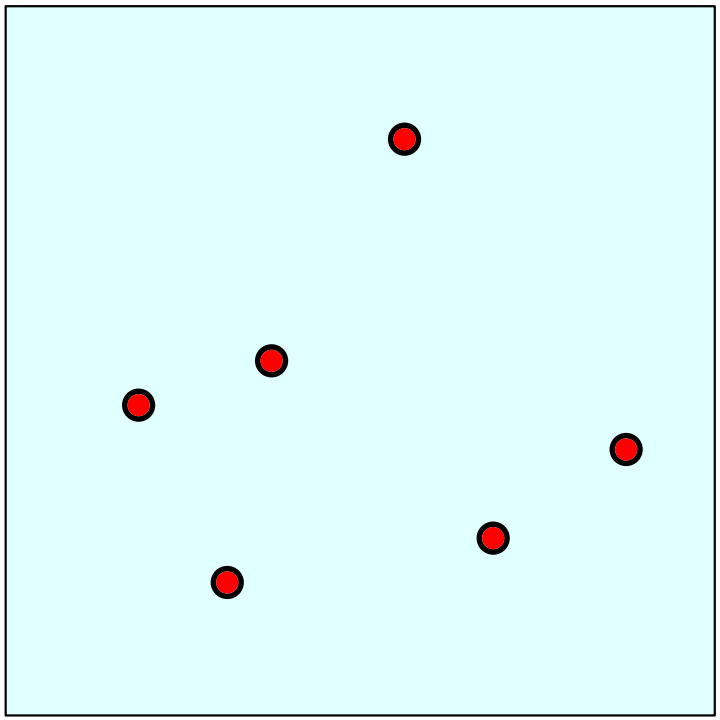
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(1/\sqrt{\mathbf{n}})$-sample.

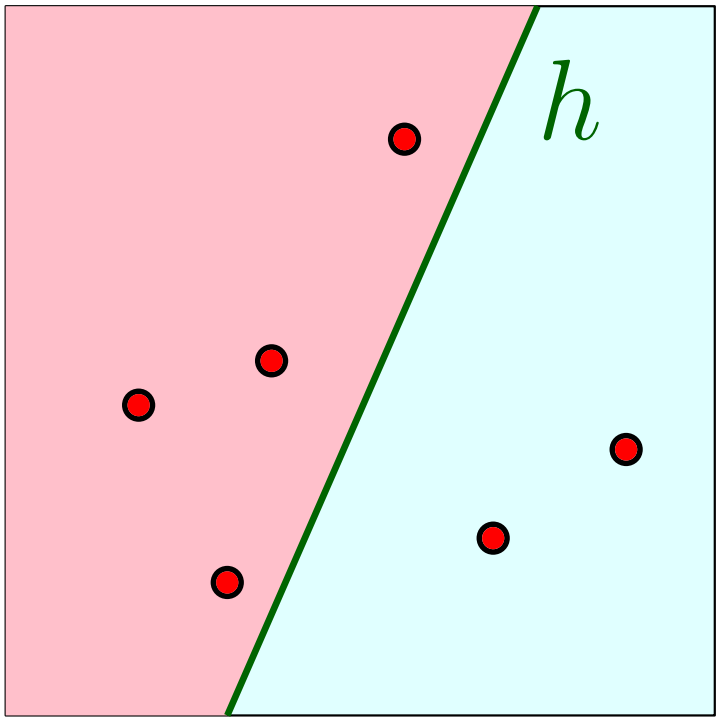
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(1/\sqrt{\mathbf{n}})$-sample.

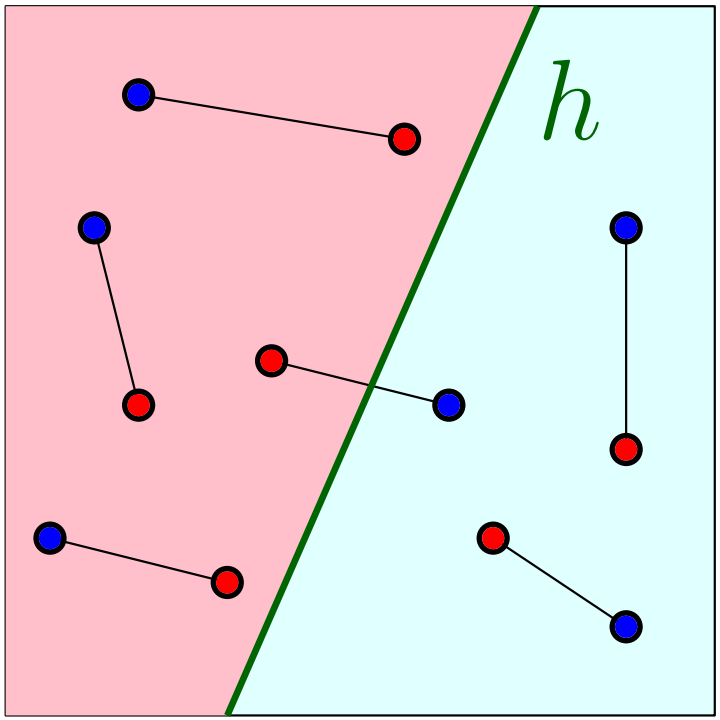
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(1/\sqrt{\mathbf{n}})$-sample.

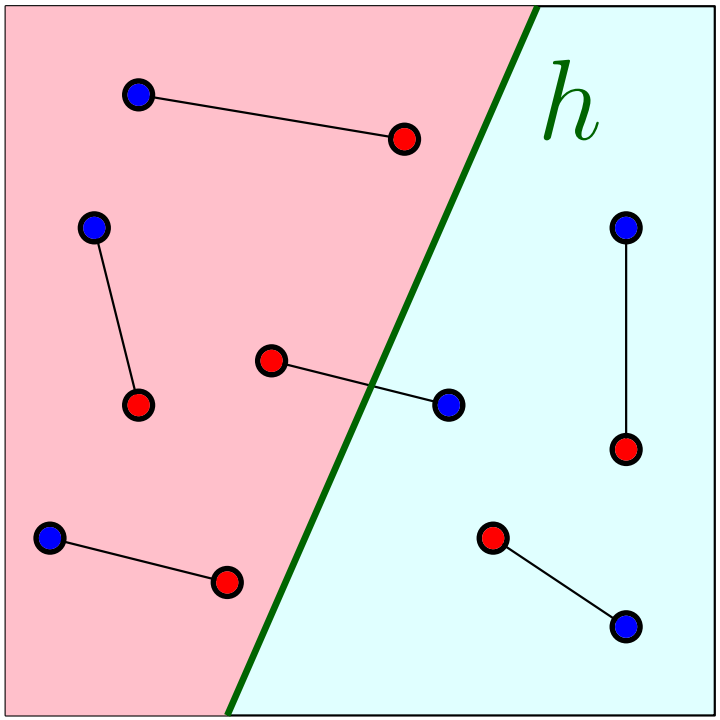
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(1/\sqrt{\mathbf{n}})$-sample.

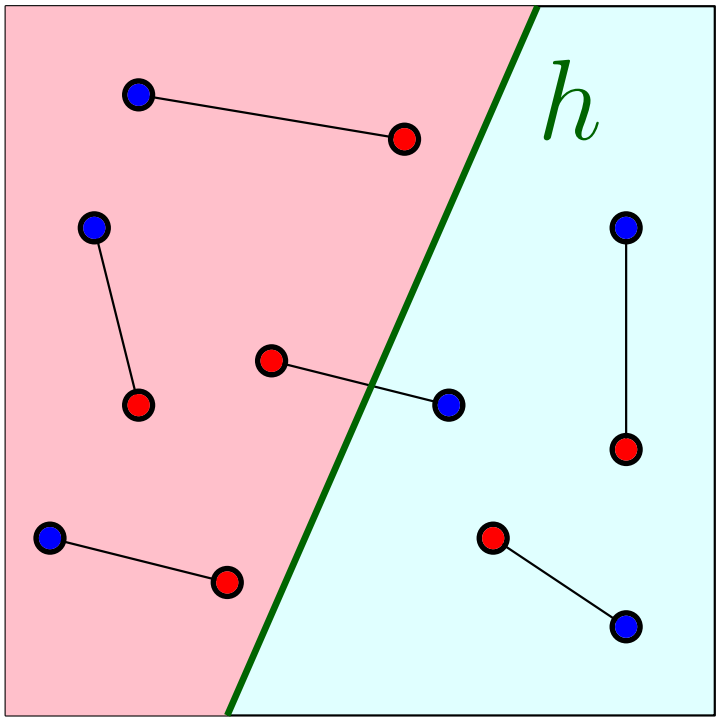
14: Getting a good sketch via discrepancy

Best sample using half the points?

Q: What is best approximation to **P** using half the points?

Discrepancy

Color points by red/blue and take the red subset as approximation.



E: # of edges of matching crossing **h**

Result: Discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E}\log \mathbf{n}}\right)$

Lemma

Red points form $\widetilde{\mathbf{O}}(\mathbf{1}/\sqrt{\mathbf{n}})$-sample.

15: Spanning tree with low crossing number

[Welzl, 1992]

P: **n** points

T: spanning tree for **P**

Such that any line ℓ crosses

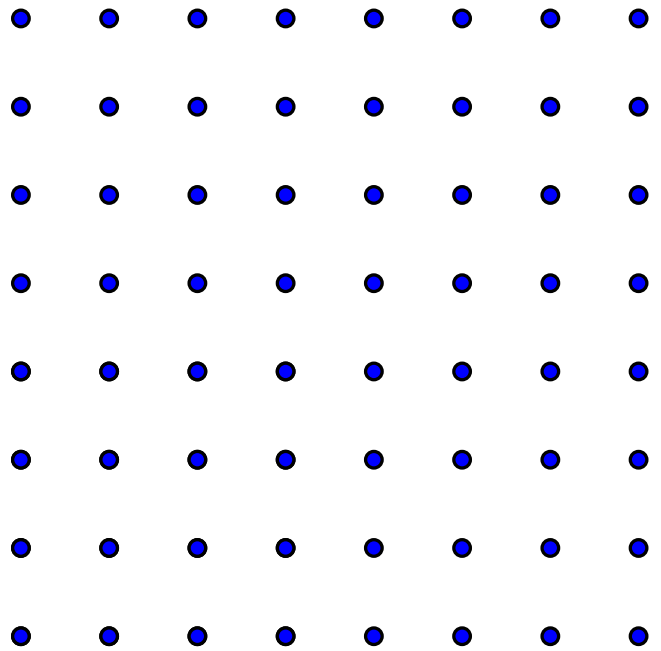
$\mathbf{O(\sqrt{n})}$ edges of **T**

Proof: Uses reweighting.

15: Spanning tree with low crossing number

[Welzl, 1992]

P: **n** points
T: spanning tree for **P**
Such that any line ℓ crosses
$\mathbf{O}(\sqrt{\mathbf{n}})$ edges of **T**
Proof: Uses reweighting.



15: Spanning tree with low crossing number

[Welzl, 1992]

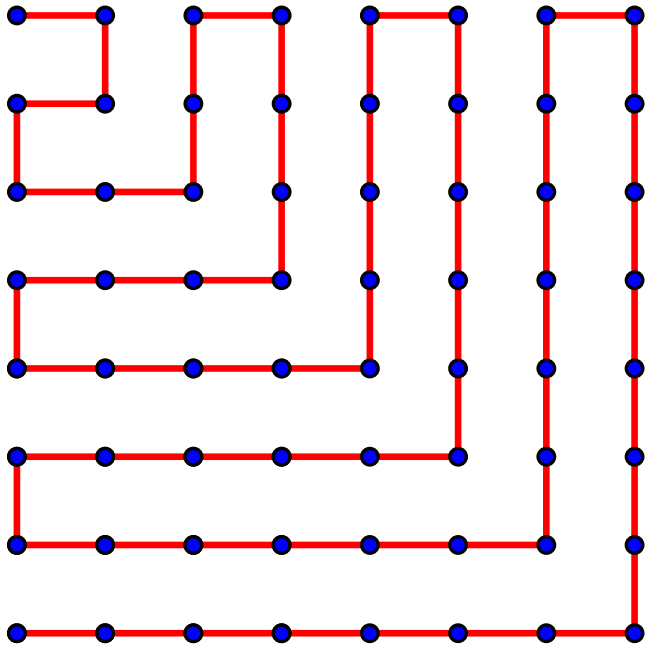
P: **n** points

T: spanning tree for **P**

Such that any line ℓ crosses

$\mathbf{O(\sqrt{n})}$ edges of **T**

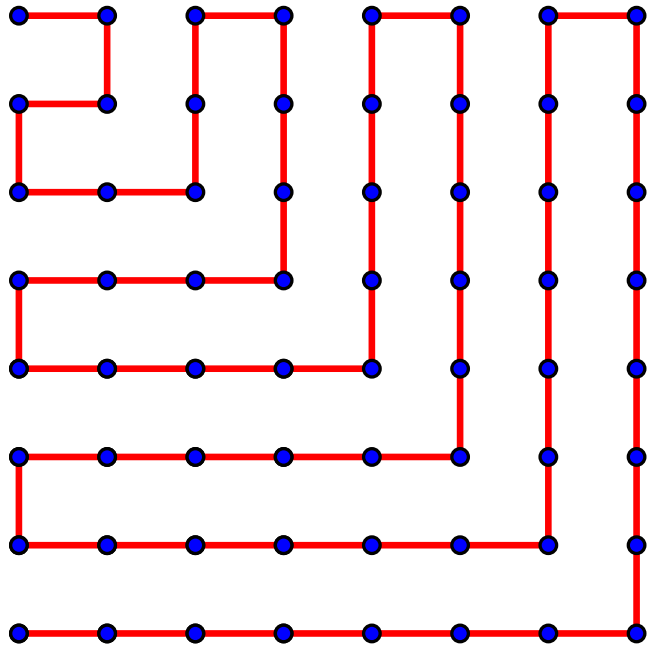
Proof: Uses reweighting.



15: Spanning tree with low crossing number

[Welzl, 1992]

P: **n** points
T: spanning tree for **P**
Such that any line ℓ crosses
$\mathbf{O(\sqrt{n})}$ edges of **T**
Proof: Uses reweighting.



Any point set behaves like a grid.
$\Longrightarrow$ tour of **P** with
$\mathbf{cr(P) = O(\sqrt{n})}$.
$\Longrightarrow$ matching of **P** with
$\mathbf{cr(P) = O(\sqrt{n})}$.

16: Smaller ε-sample for halfplanes

Matching with low crossing number

$\mathbf{n = |P|}$

Plugging matching into disc. construction:

Theorem

$\mathbf{P}$: *Set of* $\mathbf{n}$ *points in the plane.*

One can compute coloring with discrepancy $\mathbf{O}\left(\sqrt{\mathbf{E} \log \mathbf{n}}\right) = \widetilde{\mathbf{O}}\left(\mathbf{n}^{1/4}\right)$.

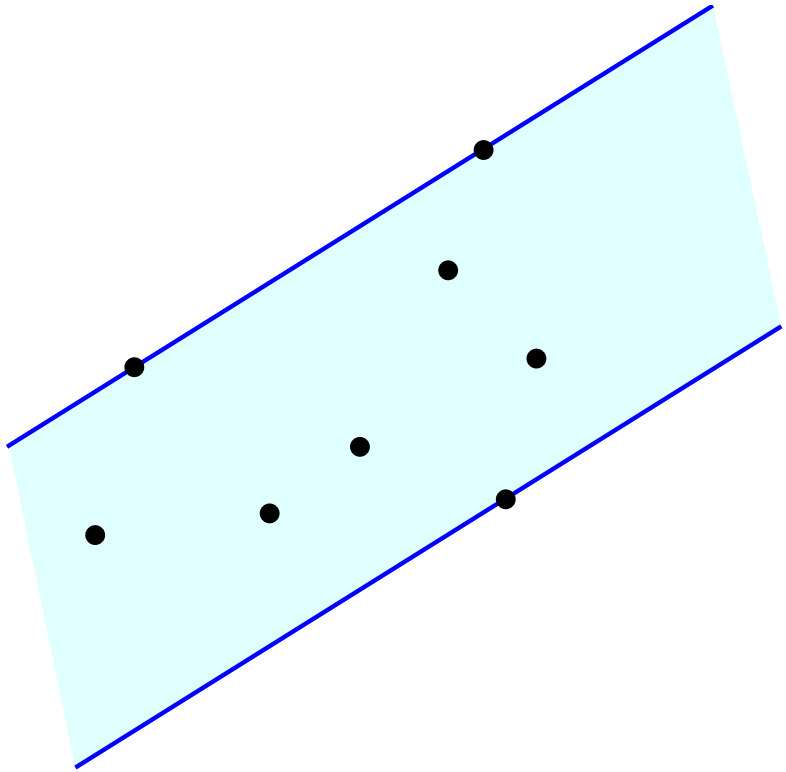
$\Longrightarrow$ $\widetilde{\mathbf{O}}(\mathbf{1}/\mathbf{n}^{3/4})$-sample using $\mathbf{n}/\mathbf{2}$ points.

Theorem

For (*points*, *halfplanes*) *there is* ε*-sample of size* $\widetilde{\mathbf{O}}(\mathbf{1}/\varepsilon^{4/3})$.

Compute width of a point set

Input: $\mathbf{P}$ - set points in $\mathbb{R}^{\mathbf{d}}$.
Q: Find the width of $\mathbf{P}$.
Find two parallel planes of min dist. that encloses $\mathbf{P}$.

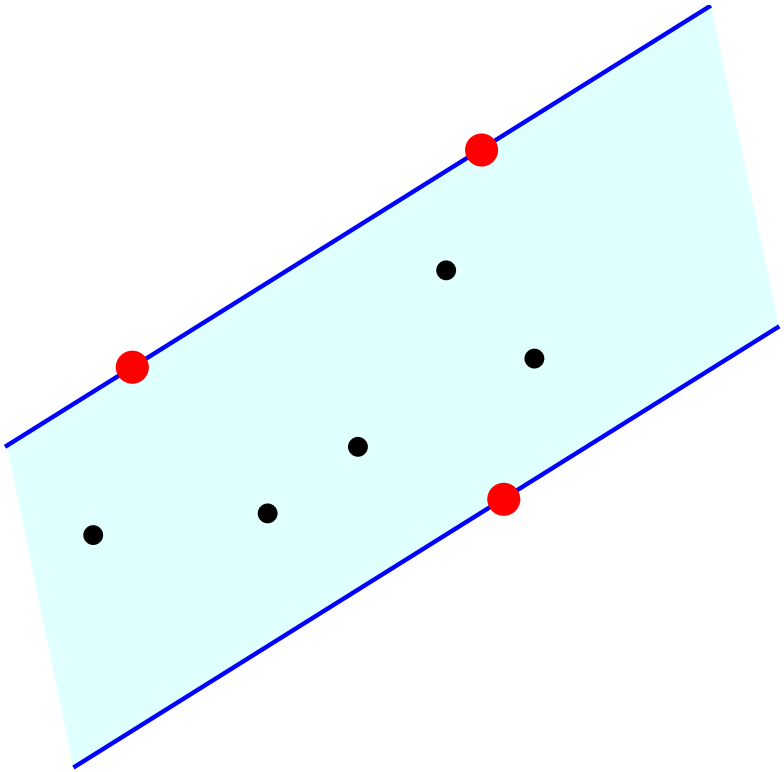


- Nasty problem for $\mathbf{d} > \mathbf{2}$. (Exact algorithms are slow.)
- Approximation?

17: Motivation: Another notion of a sketch

Compute width of a point set

Input: $\mathbf{P}$ - set points in $\mathbb{R}^{\mathbf{d}}$.
Q: Find the width of $\mathbf{P}$.
Find two parallel planes of min dist. that encloses $\mathbf{P}$.



- Nasty problem for $\mathbf{d} > \mathbf{2}$. (Exact algorithms are slow.)
- Approximation?

18: Motivation:

Geometric Optimization

- Find "quickly" a subset $\mathbf{Q} \subseteq \mathbf{P}$, s.t.,

$$\mathbf{width(Q)} \geq \mathbf{(1-\varepsilon)width(P).}$$

  ε - approx parameter.
- $\mathbf{Q}$ - is a ε-**coreset** for $\mathbf{P}$
  (for the width function)
- Compute width of $\mathbf{Q}$ using naive algorithm.
- Running time $\mathbf{O\left(n + 1/\varepsilon^{O(1)}\right)}$.

19: Coresets

- $\mathbf{f}$: A monotone function defined over subsets of $\mathbb{R}^{\mathbf{d}}$

$$\mathbf{S \subseteq T \quad \Rightarrow \quad f(S) \leq f(T).}$$

- $\mathbf{P} \subseteq \mathbb{R}^{\mathbf{d}}$: input
- **Q:** Compute $\mathbf{f(P)}$.
- $\mathcal{S} \subseteq \mathbf{P}$ is a ε-**coreset** for $\mathbf{f}$ if

$$\mathbf{f}(\mathcal{S}) \geq (1 - \varepsilon)\mathbf{f}(\mathbf{P}).$$

20: Coresets - more general definition

- $\mathbf{f(P, v)}$: A monotone function defined over subsets of $\mathbb{R}^{\mathbf{d}}$

$$\mathbf{S \subseteq T \quad \Rightarrow \quad f(S, v) \leq f(T, v).}$$

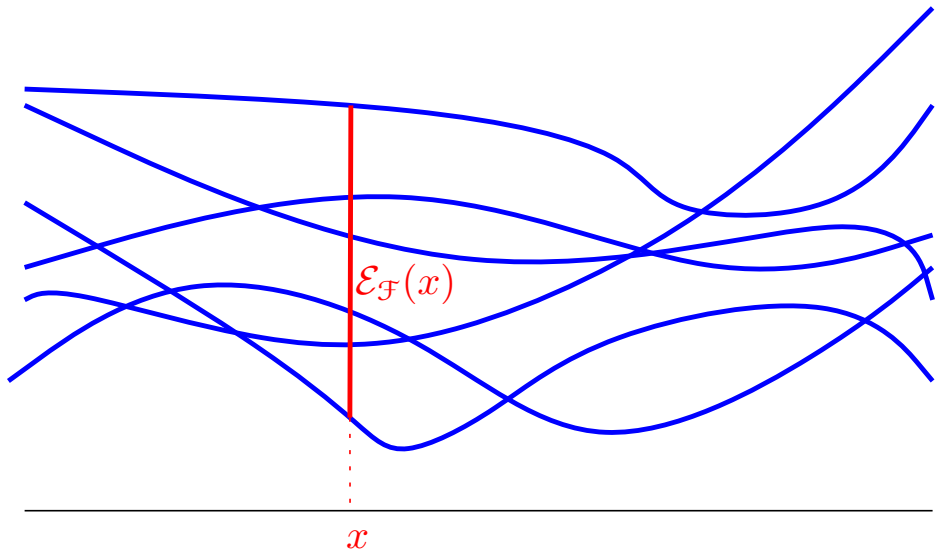
- $\mathbf{P \subseteq \mathbb{R}^d}$: input
- $\mathcal{S} \subseteq \mathbf{P}$ is a ε-**coreset** for $\mathbf{f}$ if

$$\forall \mathbf{v} \quad \mathbf{f}(\mathcal{S}, \mathbf{v}) \geq (1 - \varepsilon)\mathbf{f}(\mathbf{P}, \mathbf{v}).$$

21: Coresets - For family of functions

Vertical extent

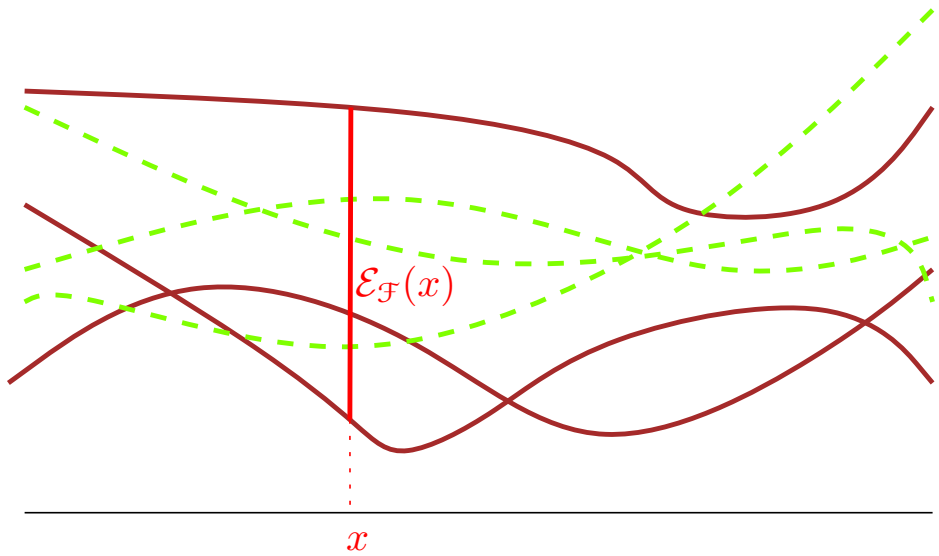
- $\mathcal{F}$ - set of functions for $\mathbb{R}^{d-1}$ to $\mathbb{R}$.
  (i.e., surfaces in $\mathbb{R}^d$).
- $\mathcal{E}(\mathbf{x}) = \left(\max_{\mathbf{f}\in\mathcal{F}} \mathbf{f}(\mathbf{x})\right) - \left(\min_{\mathbf{f}\in\mathcal{F}} \mathbf{f}(\mathbf{x})\right)$
  (Dist. upper/lower envelope of $\mathcal{F}$ at $\mathbf{x}$).



21: Coresets - For family of functions

Vertical extent

- $\mathcal{F}$ - set of functions for $\mathbb{R}^{d-1}$ to $\mathbb{R}$.
  (i.e., surfaces in $\mathbb{R}^d$).
- $\mathcal{E}(x) = \left(\max_{f \in \mathcal{F}} f(x)\right) - \left(\min_{f \in \mathcal{F}} f(x)\right)$
  (Dist. upper/lower envelope of $\mathcal{F}$ at x).



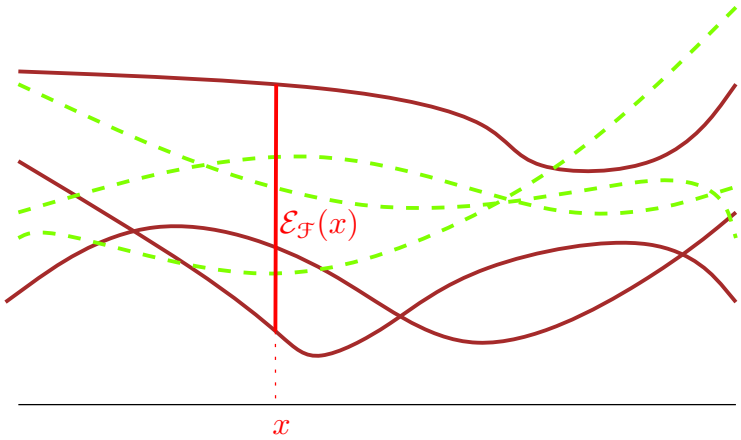
22: Coresets - For family of funcs

The result

[Agarwal et al., 2004]

$\mathcal{F}$: set of $\mathbf{n}$ functions
$\exists \mathcal{G} \subseteq \mathcal{F}$ s.t.

- $\mathcal{G}$ is ε-coreset of $\mathcal{F}$:
  $\forall \mathbf{x} \quad \boldsymbol{\mathcal{E}_{\mathcal{G}}(x) \geq (1-\varepsilon)\mathcal{E}_{\mathcal{F}}(x)}$
- $|\mathcal{G}| = \mathbf{1/\varepsilon^{O(1)}}$
- Construction time:
  $\mathbf{O(n + 1/\varepsilon^{O(1)})}$.



Functions in $\mathcal{F}$ either:

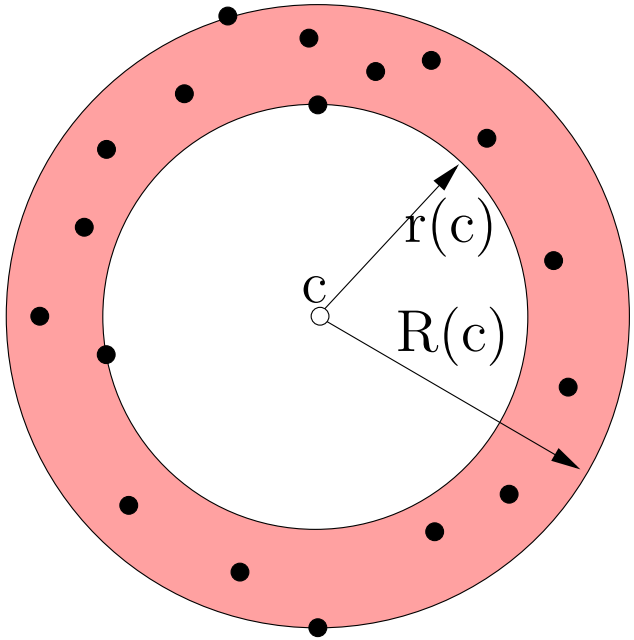
- polynomials
- square root of polynomial.

...c constant number of variables.

23: Application

Min-width annulus

- **P** - set of **n** points in the plane.
- Find smallest annulus that covers **P**.
- Motivation: meterology.



Quality: **R(c) − r(c)**

24: Min-width annulus

Reduction to extent

- $p_i = (x_i, y_i) \in P$ defines distance function:

$$f_i(c) = f_i((x, y)) = \|p_i c\| = \sqrt{(x_i - x)^2 + (y_i - y)^2}$$

$\mathcal{F} = \{f_1, \ldots, f_n\}$

- Width of annulus at c: $\mathcal{E}_{\mathcal{F}}(c) = (\max_i f_i(c)) - (\min_j f_j(c))$.
- Optimal annulus center: $\arg\min_{c \in \mathbb{R}^2} \mathcal{E}_{\mathcal{F}}(c)$
- There is a coreset $\exists \mathcal{S} \subseteq \mathcal{F}$
  - (A) $\forall c \quad \mathcal{E}_{\mathcal{S}}(c) \geq (1 - \varepsilon)\mathcal{E}_{\mathcal{F}}(c)$
  - (B) $|\mathcal{S}| = O(1/\varepsilon^2)$
- Compute shortest extent on coreset (brute force)
- Running time: $O(n + 1/\varepsilon^{O(1)})$.

25: Other applications

- A generic/general algorithm.
- Other $(1+\varepsilon)$-approximations:
  - Min radius cylinder.
  - Min width cylindrical shell.
  - Diameter.
  - Min enclosing ball.
  - Min volume bounding box.

  All of the above for moving points.
- Handling outliers.
  (A) $\mu(\mathbf{P}, \mathbf{k})$: min price of shape in $\mathcal{F}$ cover all but $\mathbf{k}$ points $\mathbf{P}$.
  (B) $\exists \mathcal{S} \subseteq \mathbf{P}$ s.t. $\mu(\mathcal{S}, \mathbf{k}) \geq (1-\varepsilon)\mu(\mathbf{P}, \mathbf{k})$.
  $|\mathcal{S}| = \mathbf{O}\left(\mathbf{k}/\varepsilon^{\mathbf{d}}\right)$.
  (C) Useful # of outliers small.

26: Sketches: Samples/Coresets/etc

Merging & sketching

Merge

- $\mathcal{S} \subseteq \mathbf{P}$ is ε-sketch
  $\mathcal{T} \subseteq \mathbf{Q}$ is ε-sketch

$\Rightarrow$ $\mathcal{S} \cup \mathcal{T}$ is ε-sketch for $\mathbf{P} \cup \mathbf{Q}$.

Sketch

- $\mathbf{A} \subseteq \mathbf{B}$ is ε-sketch
  $\mathbf{B} \subseteq \mathbf{C}$ is δ-sketch

  $\Rightarrow$ $\mathbf{A}$ is $(\varepsilon + \delta)$-sketch for $\mathbf{C}$.

27: Sketches - streaming

- Points comes in a stream.
- Only small space is available.
- Maintain approx. to stream of points.

Result

If $\exists \varepsilon$-sketch for a problem of size $\mathbf{f}(\varepsilon)$, then one can solve problem on a stream using

$$\mathbf{O}\left(\mathbf{f}\left(\frac{\varepsilon}{\log \mathbf{n}}\right) \log \mathbf{n}\right)$$

space.

$\mathbf{n}$ = number of elements in the stream.

28: Clustering

- **Clustering** = cover points with several shapes.
- Clustering price
  - $\mathbf{L}_\infty$: Price of worst shape.
  - $\mathbf{L}_1$: Sum of prices.
- Example: **k** center problem
  Cover point set with **k** balls of min max radius.
- Almost all clustering problems are NP-Complete.
- Extract sketch $\Rightarrow$ cluster sketch.

29: Clustering - results

Fast algs. using coresets for all the following problems:

- **k**-center clustering (easy).
- **k**-center clustering for moving points.
- **k**-line center clustering
  Cover points with **k** lines that min max dist of point to a line.
- **k**-median clustering
- **k**-means clustering.

30: Coresets in high dimensions

- $\mathbf{P}$ - set of points in $\mathbb{R}^{\mathbf{d}}$
  $\mathcal{S} \subseteq \mathbf{P}$ is ε-coreset for min-ball if:
  - $\mathbf{b}$: ball in $\mathbb{R}^{\mathbf{d}}$ covering $\mathcal{S}$
  - $(\mathbf{1}+\varepsilon)\mathbf{b}$ covers $\mathbf{P}$.
  - Holds **for all** possible balls.
- ... Any coreset for min-ball has to be of size $\Omega\left(\mathbf{1}/\varepsilon^{(\mathbf{d}-\mathbf{1})/\mathbf{2}}\right)$.

31: Weak Sketches in high dimensions

- We still want to compute min enclosing ball.
- $\mathbf{P}$ - set of points in $\mathbb{R}^{\mathbf{d}}$
  $\mathcal{S} \subseteq \mathbf{P}$ is ε-**shellset** for min-ball if:
  - $\mathbf{b}$: smallest ball in $\mathbb{R}^{\mathbf{d}}$ covering $\mathcal{S}$
  - $(\mathbf{1}+\varepsilon)\mathbf{b}$ covers $\mathbf{P}$.

[Bădoiu and Clarkson, 2003]

$\exists$ ε-shellset of size $\mathbf{O(1/\varepsilon)}$ of min-enclosing ball.
Computed in $\mathbf{O}\left(\mathbf{dn}/\varepsilon + \mathbf{d}/\varepsilon^{\mathbf{O(1)}}\right)$ time.

32: Weak-sketches in high dimensions

Applications

"Fast" approximation to the following problems:

- Min enclosing ball with outliers
- Min enclosing cylinder (linear running time!)
- Min **k**-flat fitting
- Projective clustering with outliers.

The meta question...

To which optimization problems $\exists$ short certificate of approximate optimal solution?

33: Conclusions

- Many problems in geometry posses small sketches.
- ... Far from having a unified theory.
- Open problems:

Approximate ε-sample

Given that $\exists$ ε-sample of size $\mathbf{t}$, compute a sample of size $\mathbf{O(t \log(1/\varepsilon))}$.

... How to compute/approximate most compact sketch?


Agarwal, P. K., Har-Peled, S., and Varadarajan, K. R. (2004).
Approximating extent measures of points.
J. Assoc. Comput. Mach., 51(4):606–635.


Alon, N. (2010).
A non-linear lower bound for planar epsilon-nets.
In *Proc. 51st Annu. IEEE Sympos. Found. Comput. Sci.*


Bădoiu, M. and Clarkson, K. L. (2003).
Optimal coresets for balls.
`http://cm.bell-labs.com/who/clarkson/coresets2.pdf`.


Haussler, D. and Welzl, E. (1987).
ε-nets and simplex range queries.
Discrete Comput. Geom., 2:127–151.


Komlós, J., Pach, J., and Woeginger, G. (1992).
Almost tight bounds for ε-nets.
Discrete Comput. Geom., 7:163–173.


Li, Y., Long, P. M., and Srinivasan, A. (2001).
Improved bounds on the sample complexity of learning.
J. Comput. Syst. Sci., 62(3):516–527.


Long, P. M. (2001).
Using the pseudo-dimension to analyze approximation algorithms for integer programming.
In *Proc. 7th Workshop Algorithms Data Struct.*, volume 2125 of *Lecture Notes Comput. Sci.*, pages 26–37.


Vapnik, V. N. and Chervonenkis, A. Y. (1971).
On the uniform convergence of relative frequencies of events to their probabilities.
Theory Probab. Appl., 16:264–280.


Welzl, E. (1992).
On spanning trees with low crossing numbers.

In *Data Structures and Efficient Algorithms, Final Report on the DFG Special Joint Initiative*, volume 594 of *Lect. Notes in Comp. Sci.*, pages 233–249. Springer-Verlag.