Sampling in Geometric and Combinatorial Set Systems

Nabil H. Mustafa

ESIEE
PARIS

UNIVERSITÉ
PARIS-EST

Outline

Configurations

graphs, sets, k-tuples, random, ...

Approximations

samples, coresets, kernels, sketches, ...

Applications

optimisation, graph algorithms, learning, ...

Approximations

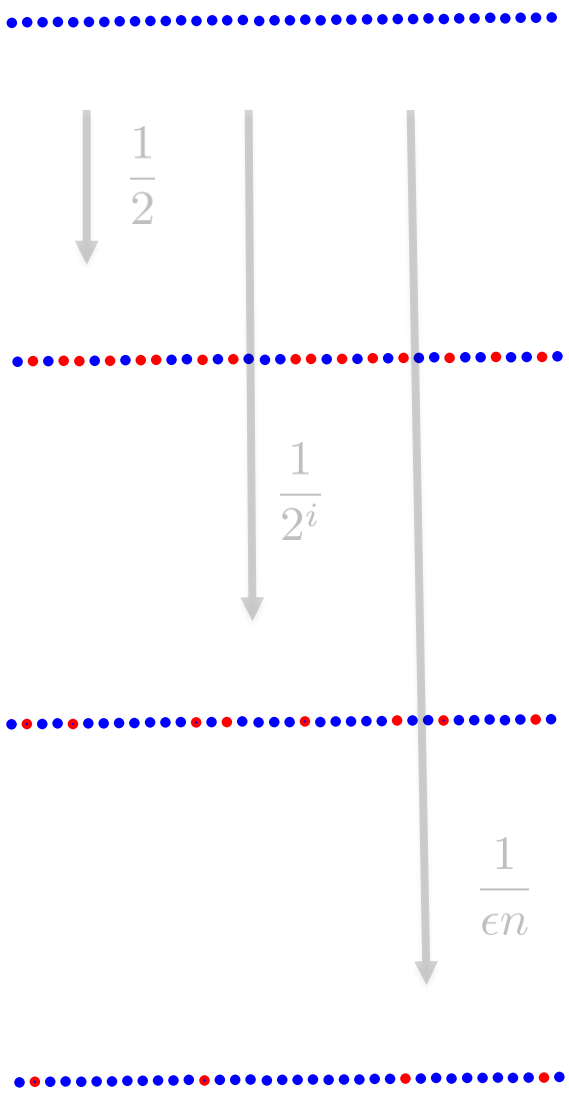
n elements $\quad$ m sets $\quad = O(n^d)$

set system $(X, \mathcal{R})$

$A \subseteq X$ $\quad$ $S \in \mathcal{R}$

$|A|$ $\quad$ $\mathbf{E}[|A \cap S|]$

		discrepancy
$t = \frac{n}{2}$	$\frac{\lvert S\rvert}{2}$	$\frac{\lvert S\rvert}{2} \pm$ error
	i steps	$\frac{\lvert S\rvert}{2^i} \pm \epsilon t$
		ϵ-approximations
$t = \frac{n}{2^i}$	$\frac{\lvert S\rvert}{2^i}$	$\frac{\lvert S\rvert\, t}{n} \pm \epsilon\, t$
	$\log(\epsilon n)$ steps, $\quad \epsilon > 0$ a given parameter	
$t = \frac{1}{\epsilon}$	$\frac{\lvert S\rvert}{\epsilon n}$	



Approximations

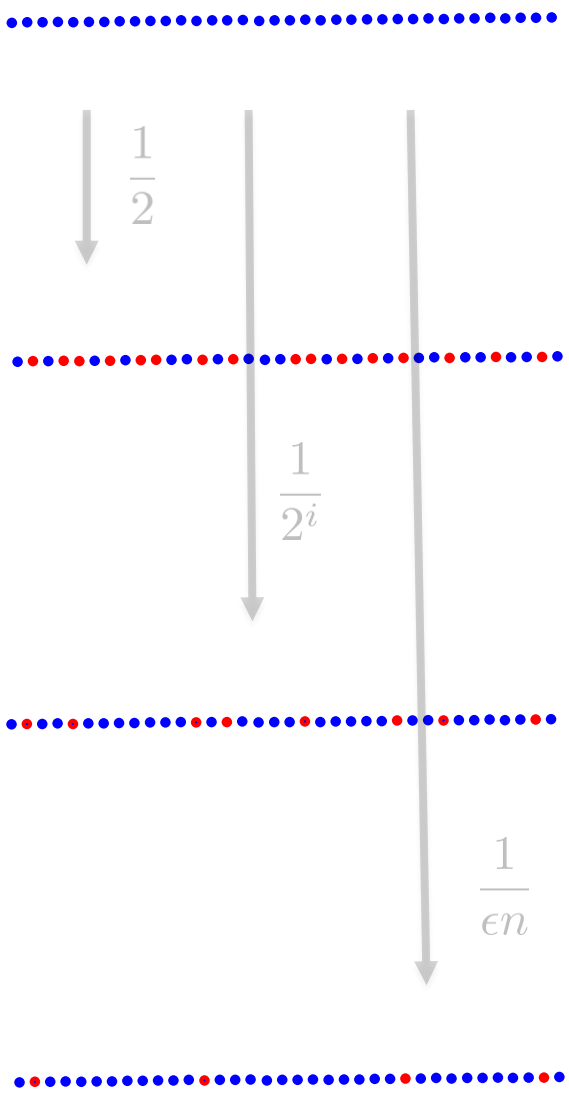
n elements $\quad$ m sets $\quad = O\left(n^d\right)$

set system $(X, \mathcal{R})$

$S \in \mathcal{R}$

$A \subseteq X$

$\lvert A\rvert$	$\mathbf{E}[\lvert A \cap S\rvert]$	
$t = \frac{n}{2}$	$\frac{\lvert S\rvert}{2}$	**discrepancy** $\quad \frac{\lvert S\rvert}{2} \pm$ error
	i steps	
$t = \frac{n}{2^i}$	$\frac{\lvert S\rvert}{2^i}$	**ϵ-approximations** $\quad \frac{\lvert S\rvert\, t}{n} \pm \epsilon\, t$
	$\log(\epsilon\, n)$ steps, $\quad \epsilon > 0$ a given parameter	
$t = \frac{1}{\epsilon}$	$\frac{\lvert S\rvert}{\epsilon n}$	**ϵ-nets** $\quad \geq 1$ if $\lvert S\rvert \geq \epsilon\, n$



Random Sampling for Approximations

n m

set system $(X, \mathcal{R})$

A: a uniform random sample of X of size t.

$$\mathbf{E}\left[|A \cap S|\right] = |S| \cdot \frac{t}{n}$$

Chernoff's bound: For any $\eta > 0$ and $S \in \mathcal{R}$

$$\Pr\left[|A \cap S| \notin \frac{|S|t}{n} \pm \eta\right] \leq 2\exp\left(-\frac{\eta^2 n}{2|S|t + \eta n}\right) \leq 2\exp\left(-\frac{\eta^2}{3t}\right).$$

discrepancy

$$t = \frac{n}{2} \qquad \frac{|S|}{2} \pm \texttt{error}$$

$t = \frac{n}{2}$

probability of discrepancy being at least η

$$\leq \; m \cdot \exp\left(-\Theta\left(\frac{\eta^2}{n}\right)\right)$$

$\eta = \Theta\left(\sqrt{n \ln m}\right)$

$$\leq m \cdot e^{-\ln 2m} \leq \tfrac{1}{2}$$

$$\boldsymbol{O\left(\sqrt{n \ln m}\right)}$$

Random Sampling for Approximations

n m

set system $(X, \mathcal{R})$

A: a uniform random sample of X of size t.

$$\mathbf{E}\left[|A \cap S|\right] = |S| \cdot \frac{t}{n}$$

Chernoff's bound: For any $\eta > 0$ and $S \in \mathcal{R}$

$$\Pr\left[|A \cap S| \notin \frac{|S|t}{n} \pm \eta\right] \leq 2\exp\left(-\frac{\eta^2 n}{2|S|t + \eta n}\right) \leq 2\exp\left(-\frac{\eta^2}{3t}\right).$$

discrepancy

$$t = \frac{n}{2} \qquad \frac{|S|}{2} \pm \texttt{error}$$

$t = \frac{n}{2}$

probability of discrepancy being at least η

$$\leq m \cdot \exp\left(-\Theta\left(\frac{\eta^2}{n}\right)\right)$$

$\eta = \Theta\left(\sqrt{n \ln m}\right)$

$$\leq m \cdot e^{-\ln 2m} \leq \frac{1}{2}$$

$$O\left(\sqrt{n \ln m}\right)$$

ϵ-approximations

$$\frac{|S|\,t}{n} \pm \epsilon\, t$$

$\eta = \epsilon\, t$ $\qquad$ $t = \Theta\left(\frac{1}{\epsilon^2} \ln \frac{m}{\gamma}\right)$

probability of error being at least η for some set $S \in \mathcal{R}$

$$\leq m \cdot \exp\left(-\Theta\left(\epsilon^2 t\right)\right)$$

$$\leq m \cdot e^{-\ln \frac{m}{\gamma}} \leq \gamma$$

$$\boldsymbol{O\left(\frac{1}{\epsilon^2} \ln m\right)}$$

Random Sampling for Approximations

n m

set system $(X, \mathcal{R})$

A: a uniform random sample of X of size t.

$$\mathbf{E}\left[|A \cap S|\right] = |S| \cdot \frac{t}{n}$$

Chernoff's bound: For any $\eta > 0$ and $S \in \mathcal{R}$

$$\Pr\left[|A \cap S| \notin \frac{|S|t}{n} \pm \eta\right] \leq 2\,\exp\left(-\frac{\eta^2 n}{2|S|t + \eta n}\right) \leq 2\exp\left(-\frac{\eta^2}{3t}\right).$$

discrepancy

$$t = \frac{n}{2} \qquad \frac{|S|}{2} \pm \texttt{error}$$

$t = \frac{n}{2}$

probability of discrepancy being at least η

$$\leq\ m \cdot \exp\left(-\Theta\left(\frac{\eta^2}{n}\right)\right)$$

$\eta = \Theta\left(\sqrt{n \ln m}\right)$

$$\leq m \cdot e^{-\ln 2m} \leq \tfrac{1}{2}$$

$$O\left(\sqrt{n \ln m}\right)$$

ϵ-approximations

$$\frac{|S|\,t}{n} \pm \epsilon\, t$$

$\eta = \epsilon\, t \qquad t = \Theta\left(\frac{1}{\epsilon^2} \ln \frac{m}{\gamma}\right)$

probability of error being at least η for some set $S \in \mathcal{R}$

$$\leq\ m \cdot \exp\left(-\Theta\left(\epsilon^2 t\right)\right)$$

$$\leq m \cdot e^{-\ln \frac{m}{\gamma}} \leq \gamma$$

$$O\left(\tfrac{1}{\epsilon^2} \ln m\right)$$

ϵ-nets

$$\geq 1 \quad \text{for} \quad |S| \geq \epsilon\, n$$

$t = \Theta\left(\frac{1}{\epsilon} \ln \frac{m}{\gamma}\right)$

probability of not hitting some set $S \in \mathcal{R}$

$$\leq\ m \cdot (1-\epsilon)^t \leq m \cdot \exp\left(-\epsilon t\right)$$

$$\leq m \cdot e^{-\ln \frac{m}{\gamma}} \leq \gamma$$

$$\boldsymbol{O\left(\frac{1}{\epsilon} \ln m\right)}$$

Random Sampling for Approximations

n m

set system $(X, \mathcal{R})$

A: a uniform random sample of X of size t.

$$\mathbf{E}\left[|A \cap S|\right] = |S| \cdot \frac{t}{n}$$

Random Sampling

Chernoff's bound: For any $\eta > 0$ and $S \in \mathcal{R}$

$$\Pr\left[|A \cap S| \notin \frac{|S|t}{n} \pm \eta\right] \leq 2\exp\left(-\frac{\eta^2 n}{2|S|t + \eta n}\right) \leq 2\exp\left(-\frac{\eta^2}{3t}\right).$$

discrepancy

$$t = \frac{n}{2} \qquad \frac{|S|}{2} \pm \texttt{error}$$

$t = \frac{n}{2}$

probability of discrepancy being at least η

$$\leq m \cdot \exp\left(-\Theta\left(\frac{\eta^2}{n}\right)\right)$$

$\eta = \Theta\left(\sqrt{n \ln m}\right)$

$$\leq m \cdot e^{-\ln 2m} \leq \frac{1}{2}$$

$$\boldsymbol{O\left(\sqrt{n \ln m}\right)}$$

ϵ-approximations

$$\frac{|S|\,t}{n} \pm \epsilon\, t$$

$\eta = \epsilon\, t \qquad t = \Theta\left(\frac{1}{\epsilon^2} \ln \frac{m}{\gamma}\right)$

probability of error being at least η for some set $S \in \mathcal{R}$

$$\leq m \cdot \exp\left(-\Theta\left(\epsilon^2 t\right)\right)$$

$$\leq m \cdot e^{-\ln \frac{m}{\gamma}} \leq \gamma$$

$$\boldsymbol{O\left(\frac{1}{\epsilon^2} \ln m\right)}$$

ϵ-nets

$$\geq 1 \quad \text{for} \quad |S| \geq \epsilon\, n$$

$t = \Theta\left(\frac{1}{\epsilon} \ln \frac{m}{\gamma}\right)$

probability of not hitting some set $S \in \mathcal{R}$

$$\leq m \cdot (1 - \epsilon)^t \leq m \cdot \exp\left(-\epsilon t\right)$$

$$\leq m \cdot e^{-\ln \frac{m}{\gamma}} \leq \gamma$$

$$\boldsymbol{O\left(\frac{1}{\epsilon} \ln m\right)}$$

Application

ϵ-nets

Optimisation

Minimum hitting set problem on input $(P, \mathcal{R})$ [Vazirani 2003]

Algorithm: Greedy

$N = \emptyset$

While N not a hitting set for $\mathcal{R}$

Add $q \in P$ that hits maximum new sets, to N

Return N

$O(\mathrm{OPT} \cdot \log m)$

Algorithm: Linear Program

Solve LP

N : random sample of P w.r.t weights $\boldsymbol{x_p \cdot \log m}$

Return N

$$\text{Minimize} \quad \sum_{p \in P} x_p \qquad \leq \mathrm{OPT}$$

subject to

$$\text{(C1)} \quad \sum_{p \in R} x_p \geq 1 \quad \forall R \in \mathcal{R},$$

$$\text{(C2)} \quad 0 \leq x_p \leq 1 \quad \forall p \in P.$$

$\epsilon = \frac{1}{\mathrm{OPT}} \quad \rightarrow \quad O(\mathrm{OPT} \cdot \log m)$

Claim : If $(P, \mathcal{R})$ has ϵ-nets of size $\frac{1}{\epsilon} \cdot f\left(\frac{1}{\epsilon}\right)$ then $f(\mathrm{OPT})$-approximation

[Long 2001]

Optimisation

Minimum hitting set problem on input $(P, \mathcal{R})$

Algorithm: Randomized Greedy + LP

Solve LP

N : random sample of P w.r.t weights ~~$x_p \cdot \log m$~~ $x_p \cdot S$ (smaller)

$S = 1, 2, \ldots, \log m$

While N not a hitting set for $\mathcal{R}$

R : any set of $\mathcal{R}$ not hit by N

q : an element randomly sampled from R according to weights $\{x_p : p \in R\}$

$N = N \cup \{q\}$

Return N

[M. 2019]

Claim: optimal approximation bounds (within constant factors) for most well-studied geometric systems

runs for expected OPT iterations

Minimize $\sum_{p \in P} x_p$

subject to

(C1) $\sum_{p \in R} x_p \geq 1 \quad \forall R \in \mathcal{R},$

(C2) $0 \leq x_p \leq 1 \quad \forall p \in P.$

Locally nice systems

Theorem : A uniform random sample of X of size $\Theta\left(\frac{1}{\epsilon^2}\ln m\right)$ is an ϵ-approximation with constant probability.

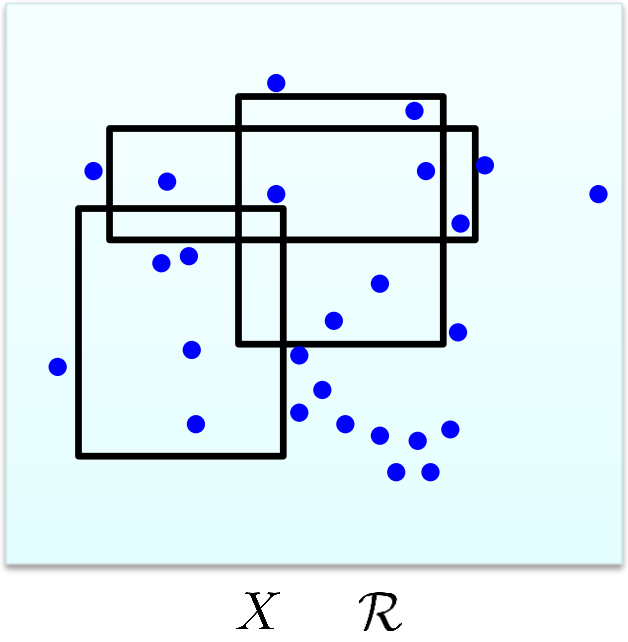
Surprising Theorem: A uniform random sample of X of size $\Theta\left(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon}\right)$ is an ϵ-approximation with constant probability, for a locally nice system

[Vapnik, Chervonenkis 71]

'locally nice' set system

total number of sets $|\mathcal{R}|$: $O\left(n^4\right)$

number of subsets on $Y \subseteq X$: $O\left(|Y|^4\right)$



X $\mathcal{R}$

combinatorially

$$\mathcal{R}|_Y = \{Y \cap R \ : \ R \in \mathcal{R}\}$$

the *projection* of $\mathcal{R}$ onto Y

a constant d such that

$$|\mathcal{R}|_Y| = O\left(|Y|^d\right) \text{ for any } Y \subseteq X$$

Sampling in Nice Systems

ϵ-approximations: A uniform random sample of X of size $\Theta\left(\frac{d}{\epsilon^2}\ln\frac{1}{\epsilon}\right)$ is an ϵ-approximation of $\mathcal{R}$, with constant probability, for nice systems. [Vapnik, Chervonenkis 71]

typically computational learning theory uses
more complicated technique called *symmetrization*

Previous bound : a uniform random sample of size $t = \Theta\left(\frac{1}{\epsilon^2}\ln m\right)$ is an ϵ-approximation of $\mathcal{R}$.

$\boldsymbol{X}$ $\quad |X| = n$

$\frac{\epsilon}{2}$-approximation of $\mathcal{R}$ ↓ $\quad |A'| = O\left(\frac{1}{\epsilon^2}\ln n^d\right)$

Random Sampling

$\boldsymbol{A'}$

$\frac{\epsilon}{2}$-approximation of $\mathcal{R}|_{A'}$ ↓ $\quad |A| = O\left(\frac{1}{\epsilon^2}\ln|A'|^d\right) = O\left(\frac{d}{\epsilon^2}\ln\left(\frac{1}{\epsilon^2}\ln n^d\right)\right) = O\left(\frac{d}{\epsilon^2}\ln\frac{1}{\epsilon^2} + \frac{d}{\epsilon^2}\ln\ln n^d\right)$

$\boldsymbol{A}$

ϵ-approximation of $\mathcal{R}$

$$T(\epsilon) = O\left(\frac{1}{\epsilon^2}\ln T\left(\frac{\epsilon}{2}\right)^d\right) \longrightarrow \Theta\left(\frac{d}{\epsilon^2}\ln\frac{1}{\epsilon}\right)$$

[Csikos, M., 2020]

ϵ-nets: A uniform random sample of X of size $O\left(\frac{d}{\epsilon}\log\frac{1}{\epsilon}\right)$ is an ϵ-net with constant probability

[Haussler, Welzl 87]

VC Dimension

VC Dimension: A classical measure of complexity of set systems $(X, \mathcal{R})$

size of the largest set for which all subsets are possible; that is,

$$\mathcal{R}|_Y = 2^Y \qquad \text{where } \mathcal{R}|_Y = \{Y \cap R : \ R \in \mathcal{R}\}$$

Sauer-Shelah lemma

related to the dimension of the Euclidean space

$d+1$ for half-spaces in $\mathbb{R}^d$ $\qquad$ $d+1$ for balls in $\mathbb{R}^d$ $\qquad$ $|\mathcal{R}|_Y| = O(|Y|^d)$

an important case: VC dimension of union of k balls, half-spaces in $\mathbb{R}^d$

key parameter behind certain high-dimensional clustering algorithms

$O(d \cdot k \log k)$ [Blumer, Ehrenfeucht, Haussler, Warmuth '89]

$\Omega(d \cdot k \log k)$ { probabilistic construction: take a random set system [Eisenstat, Angluin, 2007]
a different construction for half-spaces, balls [Csikos, Kupavskii, M., 2019]

alterations: pay too much if we sample large-enough to eliminate *all* 'bad events'

probability, statistics, learning

the biggest risk is
to not take any risk!

n elements
m subsets
d dimension

	Ideal	Random Arbitrary	VC dimension Sampling	 Sampling + Combinatorics
Discrepancy	0	$\sqrt{n \ln m}$	$\sqrt{dn \ln n}$	
↓				
Approximations	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon^2} \ln m$	$\frac{d}{\epsilon^2} \ln \frac{1}{\epsilon}$	$\frac{d}{\epsilon^2}$
↓				
Nets	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon} \ln m$	$\frac{d}{\epsilon} \ln \frac{1}{\epsilon}$	

random set system
[Komlos, Pach, Woeginger 1992]

chaining
[Talagrand, 1994]
[Li, Long, Srinivasan, 2001]

Epsilon-Nets for Abstract Set Systems

ϵ-nets: A uniform random sample of X of size $O\left(\frac{d}{\epsilon}\log\frac{1}{\epsilon}\right)$ is an ϵ-net with constant probability

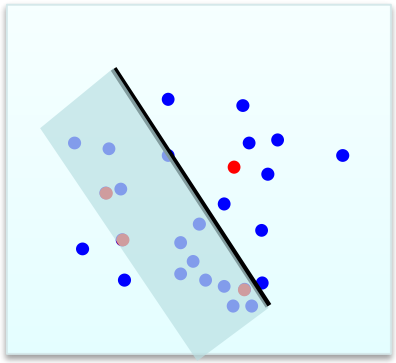
shortest paths in planar graphs

$O\left(\frac{1}{\epsilon}\right)$-sized nets

Non-Optimal

Throughout the 1990s and the 2000s

several new techniques developed for $o(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$ sized nets

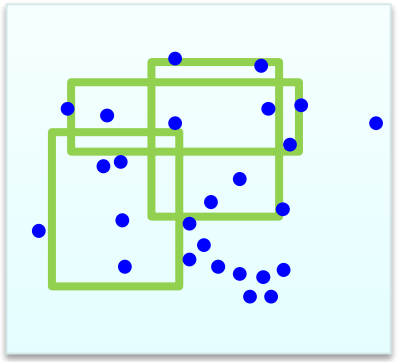


$O\left(\frac{1}{\epsilon}\right)$-sized nets

[Clarkson, Varadarajan 2006] [Pyrga, Ray 2008] [Varadarajan 2008]

[Varadarajan 2009] [Aronov, Ezra, Sharir 2010] [Chan et al. 2012]

[Har-Peled, Kaplan, Sharir, Smorodinsky 2014] [Dutta, Ghosh, M. 2018]



$O\left(\frac{1}{\epsilon}\log\log\frac{1}{\epsilon}\right)$-sized nets

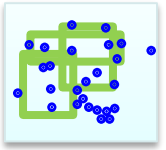
VC-dimension not fine enough

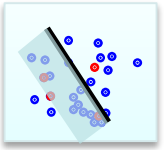
The strengthened ϵ-net theorem

INSIGHT

$(X, \mathcal{R})$ has *shallow-cell complexity* $\varphi(\cdot, \cdot)$ if for any $Y \subseteq X$, and integer k

number of sets in $\mathcal{R}|_Y$ of size at most k is $O\Big(|Y| \cdot \varphi(|Y|, k)\Big)$





Strengthened Epsilon-Net theorem: Let $(X, \mathcal{R})$ be a set system with shallow-cell complexity $\varphi(\cdot, \cdot)$. Then there exists an ϵ-net of size

$$O\left(\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\varphi_{\mathcal{F}}\left(\frac{8d}{\epsilon}, 24d\right)\right)$$

[Varadarajan 2009]
[Chan et al. 2012]
[Dutta, Ghosh, M. 2018]

algorithm requires computing additional structures

optimal: probabilistic construction
[Kupavskii, M., Pach 2017]

An Algorithm

General Net-Finder $\Big((\mathbf{X}, \mathcal{F}), \epsilon > \mathbf{0}\Big)$.

N: pick a uniform random subset of X
while *there exists a set* $S \in \mathcal{F}$ *not hit by* N **do**
  Add $\Theta(1)$ uniformly chosen random elements of S to N.

return N.

chaining + alterations

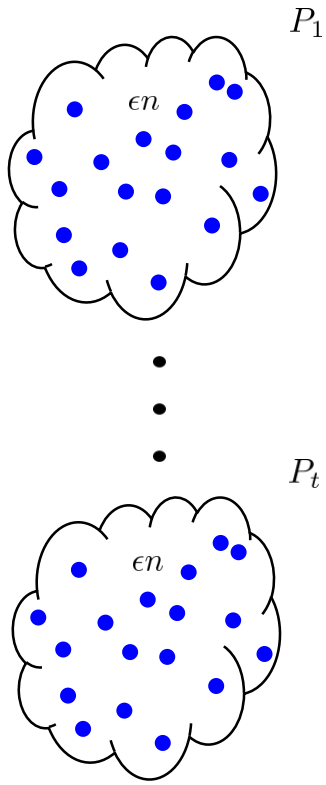
gives all known bounds

An Algorithm

General Net-Finder $\Big((\mathbf{X}, \mathcal{F}), \epsilon > \mathbf{0}\Big)$.

N: pick a uniform random subset of X
while *there exists a set* $S \in \mathcal{F}$ *not hit by* N **do**
  Add $\Theta(1)$ uniformly chosen random elements of S to N.
return N.

[M. 2019]



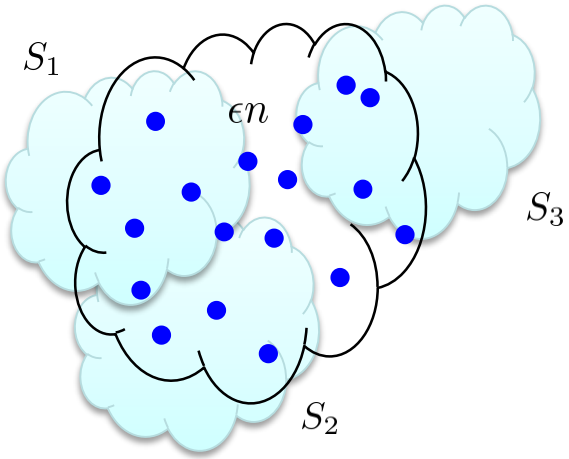
[Haussler 1995]

1: clustering. there are a few key sets of $\mathcal{F}$ $\quad O\left(\frac{1}{\epsilon}\varphi_{\mathcal{F}}\left(\frac{d}{\epsilon}, d\right)\right)$

2. reduction. the initial random sample reduces it even further

3. potential function. each iteration makes progress on one of them

probabilistic charging argument



n elements
m subsets
d dimension

	Ideal	Random	VC dimension	
			Sampling	Sampling + Combinatorics
Discrepancy	0	$\sqrt{n \ln m}$	$\sqrt{n \ln m}$	$n^{\frac{1}{2} - \frac{1}{2d}}$
↓				
Approximations	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon^2} \ln m$	$\frac{d}{\epsilon^2}$	$\frac{1}{\epsilon^{2 - \frac{2}{d+1}}}$
↓				
Nets	$\frac{1}{\epsilon}$	$\frac{1}{\epsilon} \ln m$	$\frac{d}{\epsilon} \ln \frac{1}{\epsilon}$	$O\left(\frac{1}{\epsilon} \log \varphi_{\mathcal{F}}\left(\frac{16d}{\epsilon}, 64d\right)\right)$

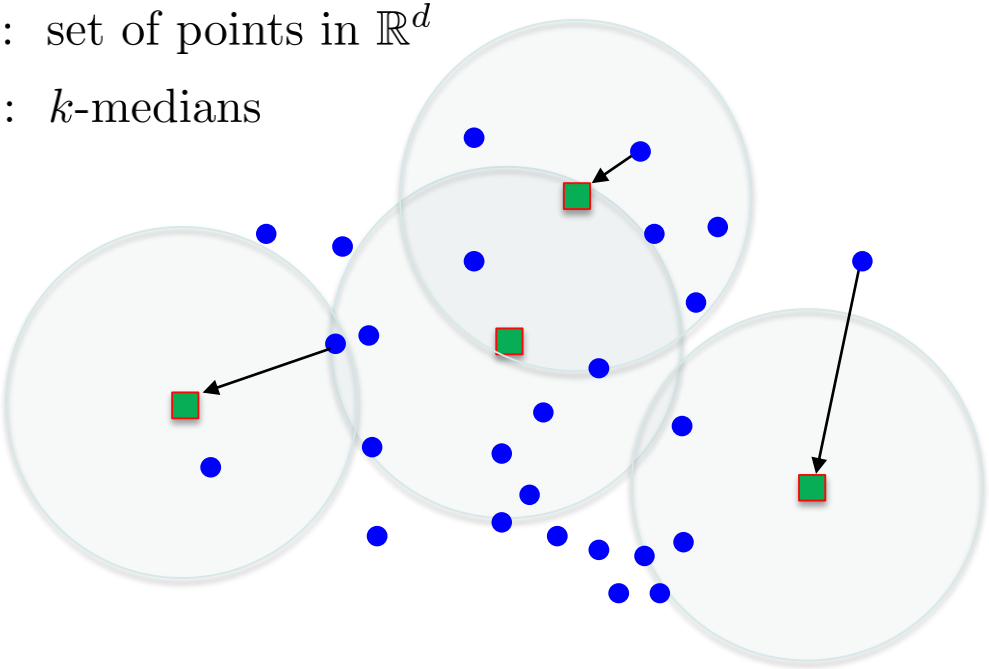
Application

ϵ-approximations

Clustering

P: set of points in $\mathbb{R}^d$

C: k-medians



Goal: compute a small-sized set A such that

for any C, $|C| = k$: $$\sum_{p \in A} \texttt{distance}\,(p, C) \cdot w\,(p) = (1 \pm \epsilon) \cdot \sum_{p \in P} \texttt{distance}\,(p, C)$$

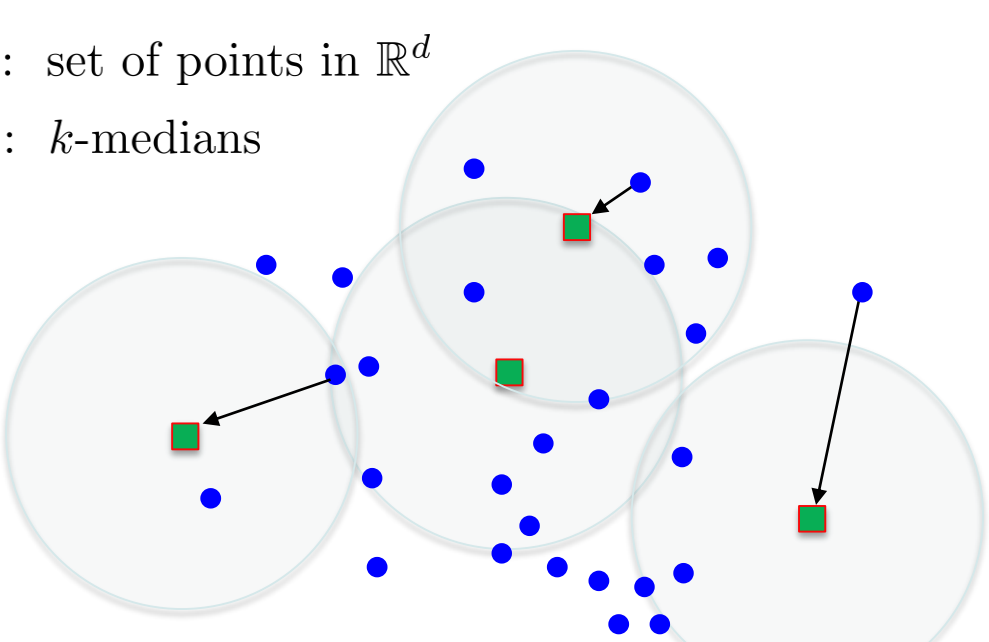
[Feldman, Langberg 2011]

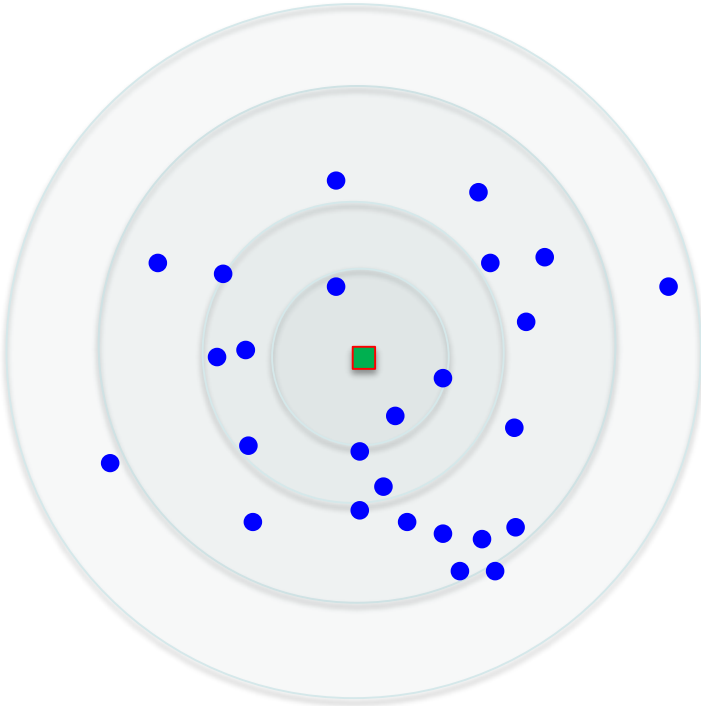
$\Longrightarrow$ **computing optimal clustering on A gives approximate clustering on P**

Clustering

P: set of points in $\mathbb{R}^d$

C: k-medians





Theorem : there exists assignment of weights, $w\colon P \to \mathbb{R}$, such that

any ϵ-approximation A for union of k balls w.r.t. these weights satisfies

$$\text{for any } C,\ |C| = k\colon \qquad \sum_{p \in A} \texttt{distance}\,(p, C) \cdot w\,(p) = (1 \pm \epsilon) \cdot \sum_{p \in P} \texttt{distance}\,(p, C)$$

[Feldman, Langberg 2011]

$\Longrightarrow$ **computing optimal clustering on A gives approximate clustering on P**

Conclusion

random sampling on combinatorial structures

probability, statistics, learning

structures improve sampling bounds and analysis

'altering' samples gives better bounds

combinatorics, geometry

newer algorithms not much more complicated

applications of these compact structures

optimisation, graphs, algorithms

step 1. show the existence of a probability distribution on a set system

step 2. computing approximation under this distribution

LP, FL

Thank you