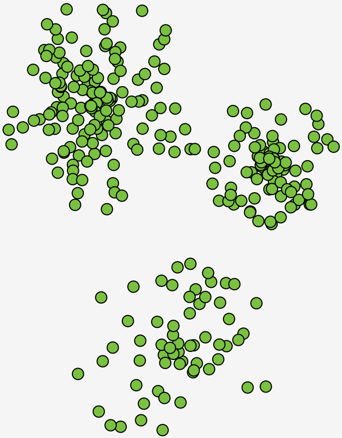
Small coresets and a dimensionality reduction for the k-means problem

Dan Feldman, Christian Sohler, Melanie Schmidt

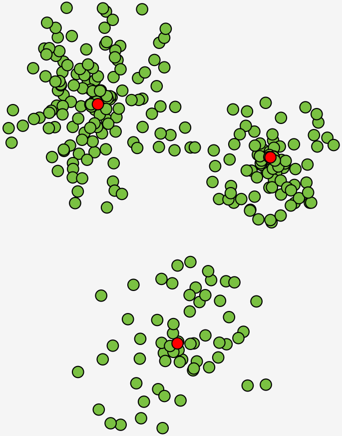
02/18/2015

The k-means problem



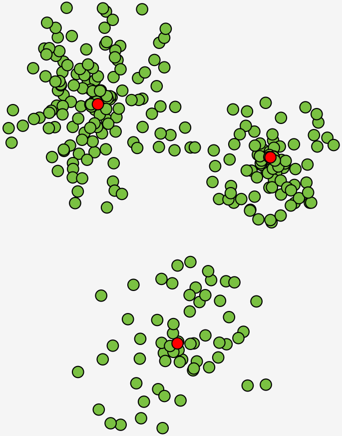
The k-means problem

- Given a point set $P \subseteq \mathbb{R}^n$,



The k-means problem

- Given a point set $P \subseteq \mathbb{R}^n$,
- compute a set $C \subseteq \mathbb{R}^n$ with $|C| = k$ centers

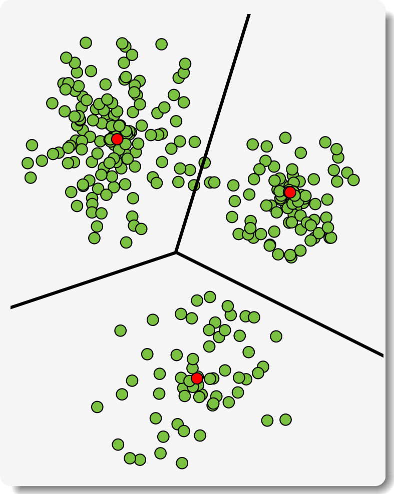


The k-means problem

- Given a point set $P \subseteq \mathbb{R}^n$,
- compute a set $C \subseteq \mathbb{R}^n$ with $|C| = k$ centers
- which minimizes cost(P, C)

$$= \sum_{p \in P} \min_{c \in C} ||p - c||^2,$$

the sum of the squared distances.



The k-means problem

- Given a point set $P \subseteq \mathbb{R}^n$,
- compute a set $C \subseteq \mathbb{R}^n$ with $|C| = k$ centers
- which minimizes cost(P, C)

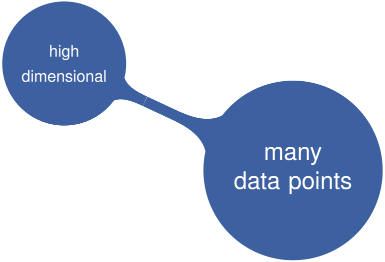
$$= \sum_{p \in P} \min_{c \in C} ||p - c||^2,$$

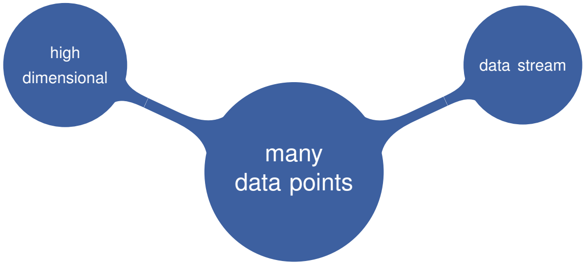
the sum of the squared distances.

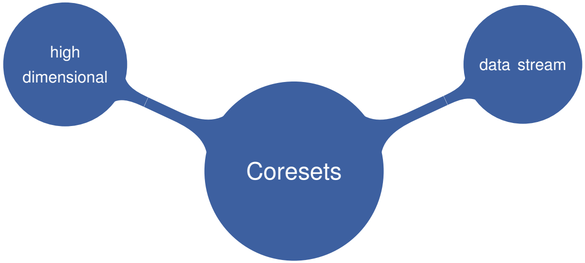
- induces a partitioning of the input point set

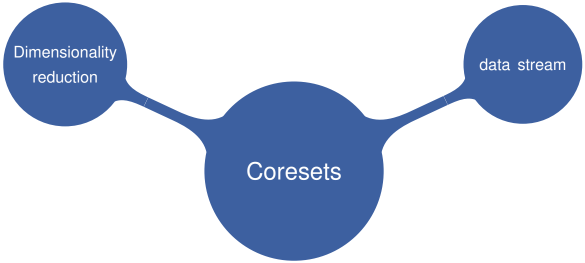

Big
Data

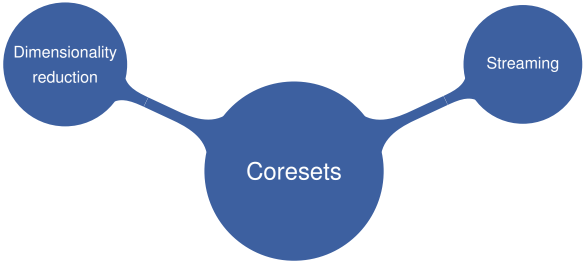

many
data points

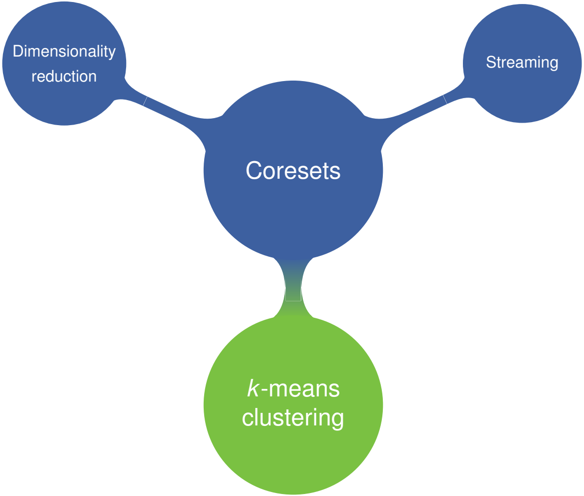

high
dimensional
many
data points

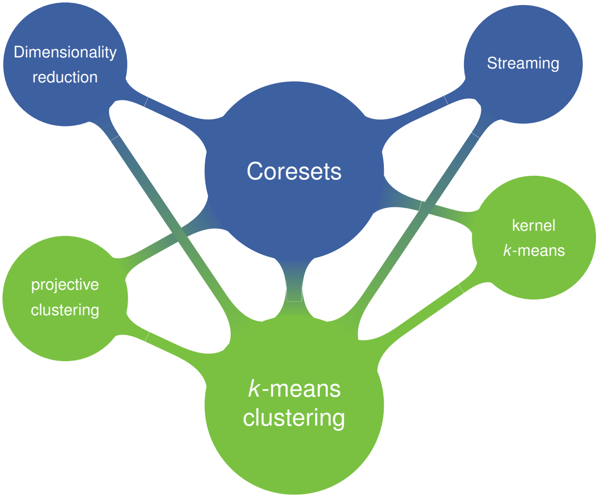

high
dimensional
data stream
many
data points

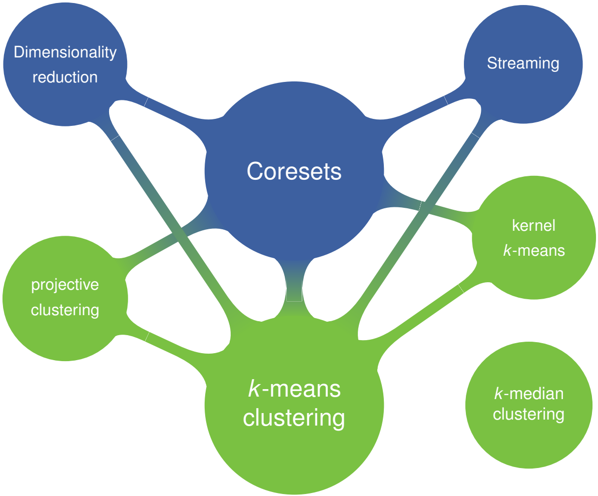

high dimensional
data stream
Coresets


Dimensionality reduction
data stream
Coresets


Dimensionality reduction
Streaming
Coresets


Dimensionality reduction
Streaming
Coresets
k-means clustering


Dimensionality reduction
Streaming
Coresets
kernel k-means
projective clustering
k-means clustering


Dimensionality reduction
Streaming
Coresets
kernel k-means
projective clustering
k-means clustering
k-median clustering

Coreset (idea)

- compute a smaller weighted point set
- that preserves the k-means objective,
- i.e., the sum of the weighted squared distances is similar
- for all sets of k centers

Coreset (idea)

- compute a smaller weighted point set
- that preserves the k-means objective,
- i.e., the sum of the weighted squared distances is similar
- for all sets of k centers

Why for all centers?

- coreset and input should look alike for k-means

Coreset (idea)

- compute a smaller weighted point set
- that preserves the k-means objective,
- i.e., the sum of the weighted squared distances is similar
- for all sets of k centers

Why for all centers?

- coreset and input should look alike for k-means

- assume optimizing over the possible centers
- if the cost is underestimated for certain center sets, then they might be mistakenly assumed to be optimal

Coreset (idea)

- compute a smaller weighted point set
- that preserves the k-means objective,
- i.e., the sum of the weighted squared distances is similar
- for all sets of k centers

Why for all centers?

- coreset and input should look alike for k-means

- assume optimizing over the possible centers
- if the cost is underestimated for certain center sets, then they might be mistakenly assumed to be optimal

Very convenient, e.g. for usage in data streams or distributed settings

Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon\, \mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \, \text{cost}(P, C)$$

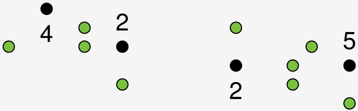
holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon\, \mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

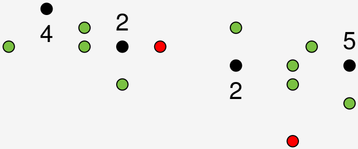


Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon\, \mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

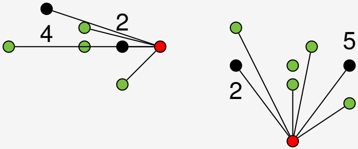


Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \, \text{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

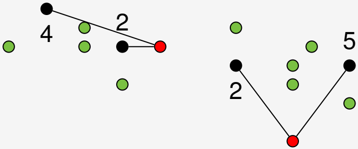


Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon \, \mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

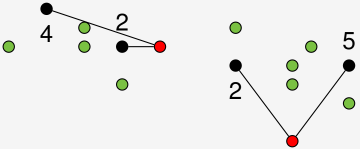


Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \, \text{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.



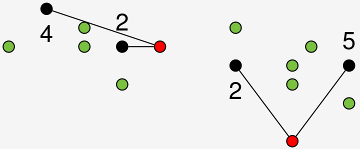
Space reduction: Size of S should be polylogarithmic in n or constant

Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ is a $(1+\varepsilon)$-coreset if

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon \, \mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.



Space reduction: Size of S should be polylogarithmic in n or constant

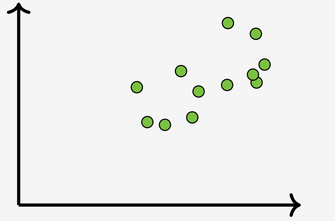
Earlier coreset definitions e.g. in [AHPV04], [BHPI02], [I99], [MOP01]

Dimensionality reduction

Replace P by a point set P' of smaller intrinsic dimension

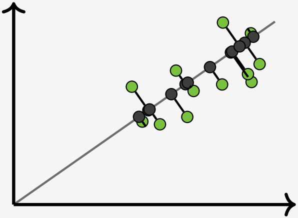
Dimensionality reduction

Replace P by a point set P' of smaller intrinsic dimension



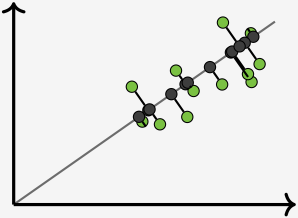
Dimensionality reduction

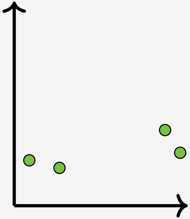
Replace P by a point set P' of smaller intrinsic dimension



Dimensionality reduction

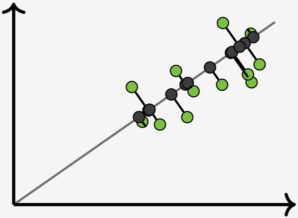
Replace P by a point set P' of smaller intrinsic dimension

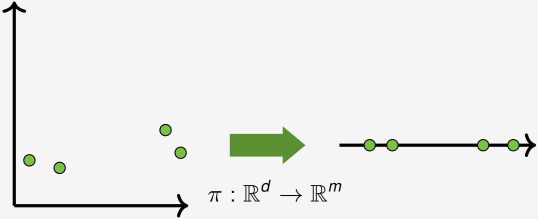




Dimensionality reduction

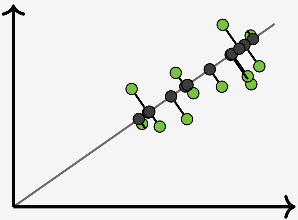
Replace P by a point set P' of smaller intrinsic dimension

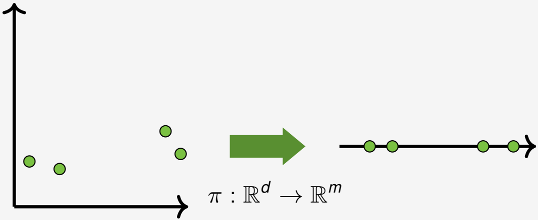




Dimensionality reduction

Replace P by a point set P' of smaller intrinsic dimension



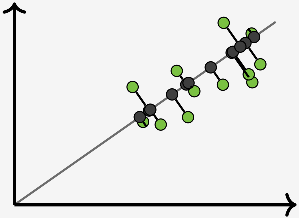


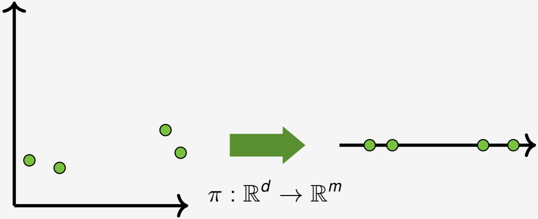
[Drineas et. al., 1999]

- projection to first k principal components
- 2-approximation

Dimensionality reduction

Replace P by a point set P' of smaller intrinsic dimension





[Drineas et. al., 1999]

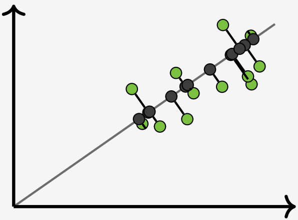
- projection to first k principal components
- 2-approximation

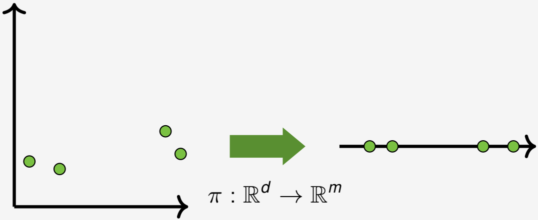
[Johnson, Lindenstrauss, 1984]

- random projection, target dimension $\Theta(\log n/\varepsilon^2)$
- $(1+\varepsilon)$-coreset-type guarantee

Dimensionality reduction

Replace P by a point set P' of smaller intrinsic dimension





[Drineas et. al., 1999]

- projection to first k principal components
- 2-approximation

[Johnson, Lindenstrauss, 1984]

- random projection, target dimension $\Theta(\log n/\varepsilon^2)$
- $(1+\varepsilon)$-coreset-type guarantee

[BMD09] $2+\varepsilon$, $\tilde{\Theta}(k/\varepsilon^2)$

[BZD10] $2+\varepsilon$, $\Theta(k/\varepsilon^2)$

Dimensionality reduction

$P \subset \mathbb{R}^d$ is replaced by $P' \subset \mathbb{R}^d$ of smaller intrinsic dimension such that

$$\left|\text{cost}(P', C) - \text{cost}(P, C)\right| \leq \varepsilon \, \text{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Dimensionality reduction

$P \subset \mathbb{R}^d$ is replaced by $P' \subset \mathbb{R}^d$ of smaller intrinsic dimension such that

$$\left|\mathrm{cost}(P', C) - \mathrm{cost}(P, C)\right| \le \varepsilon\,\mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Strong Coresets [Har-Peled, Mazumdar, 2004]

For a $P \subset \mathbb{R}^d$, a weighted set $S \subset \mathbb{R}^d$ with $|S| < |P|$ is a $(1+\varepsilon)$-coreset if

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \le \varepsilon\,\mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Moving points to reduce their complexity [HPM04,FS05]

Move points in P by using a mapping $\pi : P \to \mathbb{R}^d$ that satisfies

$$\sum_{x \in P} ||x - \pi(x)||^2 \leq \frac{\varepsilon^2}{16} \cdot OPT.$$

Then it holds for every set of k centers $C \subset \mathbb{R}^d$ that

$$|\,\mathrm{cost}(\pi(P), C) - \mathrm{cost}(P, C)| \leq \varepsilon \cdot \mathrm{cost}(P).$$

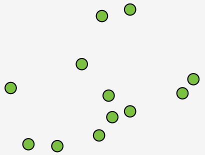
Moving points to reduce their complexity [HPM04,FS05]

Move points in P by using a mapping $\pi : P \to \mathbb{R}^d$ that satisfies

$$\sum_{x \in P} ||x - \pi(x)||^2 \leq \frac{\varepsilon^2}{16} \cdot OPT.$$

Then it holds for every set of k centers $C \subset \mathbb{R}^d$ that

$$|\,\mathrm{cost}(\pi(P), C) - \mathrm{cost}(P, C)| \leq \varepsilon \cdot \mathrm{cost}(P).$$



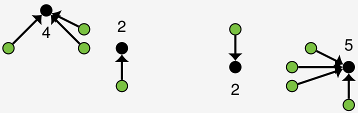
Moving points to reduce their complexity [HPM04,FS05]

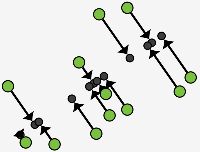
Move points in P by using a mapping $\pi : P \to \mathbb{R}^d$ that satisfies

$$\sum_{x \in P} ||x - \pi(x)||^2 \leq \frac{\varepsilon^2}{16} \cdot OPT.$$

Then it holds for every set of k centers $C \subset \mathbb{R}^d$ that

$$|\operatorname{cost}(\pi(P), C) - \operatorname{cost}(P, C)| \leq \varepsilon \cdot \operatorname{cost}(P).$$





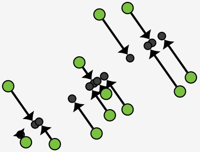
Moving points to reduce their complexity [HPM04,FS05]

Move points in P by using a mapping $\pi : P \to \mathbb{R}^d$ that satisfies

$$\sum_{x \in P} ||x - \pi(x)||^2 \leq \frac{\varepsilon^2}{16} \cdot OPT.$$

Then it holds for every set of k centers $C \subset \mathbb{R}^d$ that

$$|\,\mathrm{cost}(\pi(P), C) - \mathrm{cost}(P, C)| \leq \varepsilon \cdot \mathrm{cost}(P).$$



Used in combination with grids [HPM04], [HPK05], [FS05], [FGSSS13]

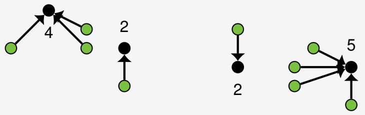
Moving points to reduce their complexity [HPM04,FS05]

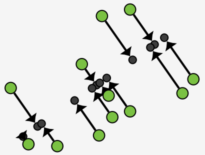
Move points in P by using a mapping $\pi : P \to \mathbb{R}^d$ that satisfies

$$\sum_{x \in P} ||x - \pi(x)||^2 \leq \frac{\varepsilon^2}{16} \cdot OPT.$$

Then it holds for every set of k centers $C \subset \mathbb{R}^d$ that

$$|\operatorname{cost}(\pi(P), C) - \operatorname{cost}(P, C)| \leq \varepsilon \cdot \operatorname{cost}(P).$$





Used in combination with grids [HPM04], [HPK05], [FS05], [FGSSS13]
(Coreset sizes depend exponentially on the dimension d)

Random Sampling

- draw a point $x \in P$ uniformly at random
- $\rightarrow$ unbiased extimator for cost(P, C)
- for any fixed set of k centers $C \subset \mathbb{R}^d$

Random Sampling

- draw a point $x \in P$ uniformly at random
- $\rightarrow$ unbiased extimator for $\text{cost}(P, C)$
- for any fixed set of k centers $C \subset \mathbb{R}^d$

Problem

- high variance
- large
  sample set

Random Sampling

- draw a point $x \in P$ uniformly at random
- $\rightarrow$ unbiased extimator for $\text{cost}(P, C)$
- for any fixed set of k centers $C \subset \mathbb{R}^d$

Problem

- high variance
- large
  sample set

[Hoeffding, 1963], [Haussler, 1992], [MOP, 2001], [Chen, 2006]

$\mathcal{O}(k \cdot \log n \cdot n \cdot diam(P)/(\varepsilon^2 \cdot OPT))$ is a sufficient sample size

Random Sampling

- draw a point $x \in P$ uniformly at random
- $\rightarrow$ unbiased extimator for cost(P, C)
- for any fixed set of k centers $C \subset \mathbb{R}^d$

Problem

- high variance
- large

  sample set

[Hoeffding, 1963], [Haussler, 1992], [MOP, 2001], [Chen, 2006]

$\mathcal{O}(k \cdot \log n \cdot n \cdot diam(P)/(\varepsilon^2 \cdot OPT))$ is a sufficient sample size

Reduce variance by...

- partitioning P into sets with small diameter [C06]
- sampling according to cost based probabilities [FMS07]
- sampling according to sensitivity based probabilities [LS10, FL11]

Random Sampling

- draw a point $x \in P$ uniformly at random
- $\rightarrow$ unbiased extimator for cost(P, C)
- for any fixed set of k centers $C \subset \mathbb{R}^d$

Problem

- high variance
- large sample set

[Hoeffding, 1963], [Haussler, 1992], [MOP, 2001], [Chen, 2006]

$\mathcal{O}(k \cdot \log n \cdot n \cdot diam(P)/(\varepsilon^2 \cdot OPT))$ is a sufficient sample size

Reduce variance by...

- partitioning P into sets with small diameter [C06]
- sampling according to cost based probabilities [FMS07]
- sampling according to sensitivity based probabilities [LS10, FL11]

Feldman, Langberg (2011) get a coreset size of $\tilde{\mathcal{O}}(kd/\varepsilon^{-4})$.

[Zhang, Ramakrishnan, Livny, 1996]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

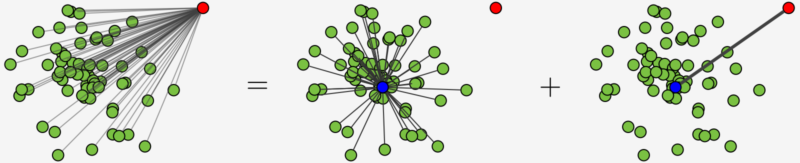
where $\mu(P) = \sum_{x \in P} x / |P|$ is the centroid of P.

[Zhang, Ramakrishnan, Livny, 1996]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.



[Zhang, Ramakrishnan, Livny, 1996]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x / |P|$ is the centroid of P.

Implications

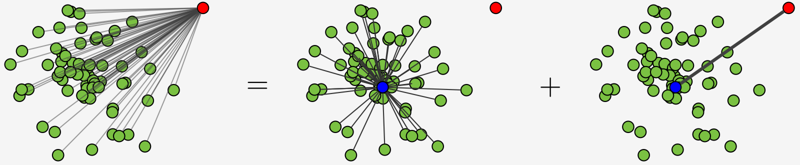
- centroid is always the optimal 1-means solution
- optimal solution consists of centroids of subsets
- centroid (plus constant) is an $(1, \varepsilon)$-coreset with no error

[Zhang, Ramakrishnan, Livny, 1996]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x / |P|$ is the centroid of P.

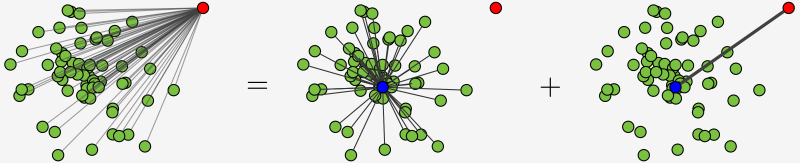


[Zhang, Ramakrishnan, Livny, 1996]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

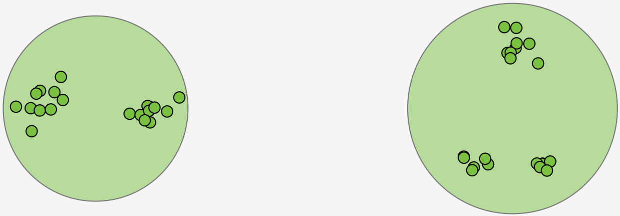
where $\mu(P) = \sum_{x \in P} x / |P|$ is the centroid of P.



Neat exact coreset for $k = 1$: centroid plus constant

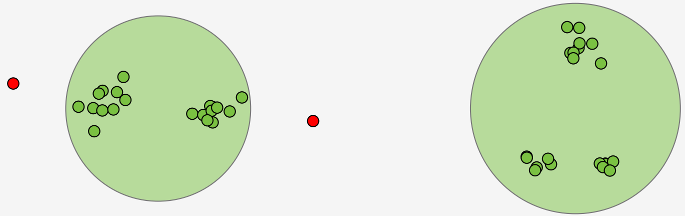
Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost



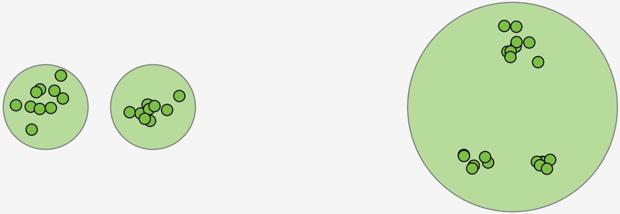
Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost



Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost

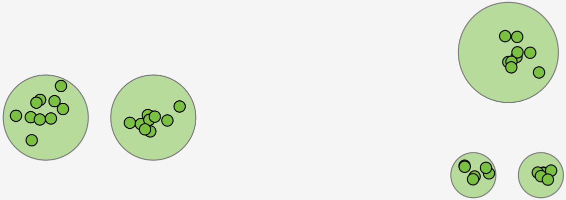


Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost

Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost



Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost

1. start with an (approximately) optimal clustering
2. for each subset in the partitioning, test:
3. optimal k-means cost $\leq$ optimal 1-means cost $/ (1 + \varepsilon)$?
4. If yes, subdivide and recurse on the subsets
5. If not, replace by centroid plus constant

Notice: Stop recursion at level $\mathcal{O}(\log_{1+\varepsilon} \varepsilon^{-2})$ and replace by centroid

Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost

1. start with an (approximately) optimal clustering
2. for each subset in the partitioning, test:
3. optimal k-means cost $\leq$ optimal 1-means cost $/ (1+\varepsilon)$?
4. If yes, subdivide and recurse on the subsets
5. If not, replace by centroid plus constant

Notice: Stop recursion at level $\mathcal{O}(\log_{1+\varepsilon} \varepsilon^{-2})$ and replace by centroid

- Corset has size $\mathcal{O}\left(k^{\mathcal{O}(\log_{1+\varepsilon} \varepsilon^{-2})}\right) = \mathcal{O}(k^{\mathcal{O}(\varepsilon^{-2} \log \varepsilon^{-1})})$
- number of points is independent of n and d

Application for coresets

- Idea: Store fixed costs in an additional constant
- Subset of points with same center pay a fixed basic cost

For all subsets in our partitioning:

- either a we stop dividing at some point

$\rightarrow$ points can pick the same center with not much error

- or 1-means cost falls below threshold

$\rightarrow$ use movement lemma to move points to the centroid

- Corset has size $\mathcal{O}\left(k^{\mathcal{O}(\log_{1+\varepsilon}\varepsilon^{-2})}\right) = \mathcal{O}(k^{\mathcal{O}(\varepsilon^{-2}\log\varepsilon^{-1})})$
- number of points is independent of n and d

Now

- much smaller coreset size
- obtained by reducing the intrinsic dimension first

Now

- much smaller coreset size
- obtained by reducing the intrinsic dimension first

- Recall: Feldman, Langberg obtain coreset with $\tilde{O}(kd/\varepsilon^4)$ points
- reduce dimension, compute coreset
- *d* vanishes from coreset size

Now

- much smaller coreset size
- obtained by reducing the intrinsic dimension first

- Recall: Feldman, Langberg obtain coreset with $\tilde{O}(kd/\varepsilon^4)$ points
- reduce dimension, compute coreset
- d vanishes from coreset size

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil 18k/\varepsilon^2 \rceil$, there exists a P' with intrinsic dimension $\lceil 18k/\varepsilon^2 \rceil$ and a constant Δ such that

$$|\,\mathrm{cost}(P', C) + \Delta - \mathrm{cost}(P, C)| \leq \varepsilon\, \mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Now

- much smaller coreset size
- obtained by reducing the intrinsic dimension first

- Recall: Feldman, Langberg obtain coreset with $\tilde{O}(kd/\varepsilon^4)$ points
- reduce dimension, compute coreset
- d vanishes from coreset size $\rightarrow$ $\tilde{O}(k^2/\varepsilon^6)$ points

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil 18k/\varepsilon^2 \rceil$, there exists a P' with intrinsic dimension $\lceil 18k/\varepsilon^2 \rceil$ and a constant Δ such that

$$|\operatorname{cost}(P', C) + \Delta - \operatorname{cost}(P, C)| \leq \varepsilon \operatorname{cost}(P, C)$$

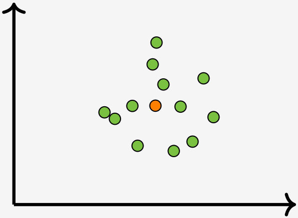
holds for all sets $C \subset \mathbb{R}^d$ of k centers.

[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

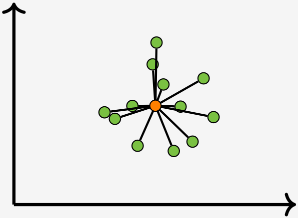


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

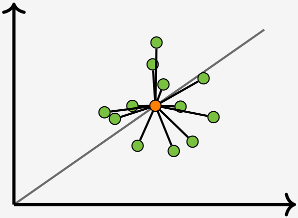


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

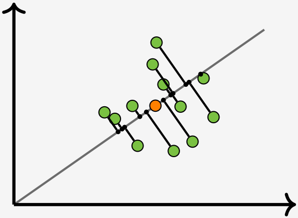


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

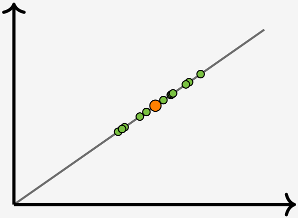


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

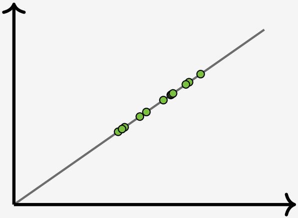


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

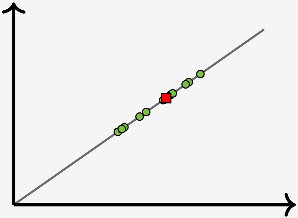


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

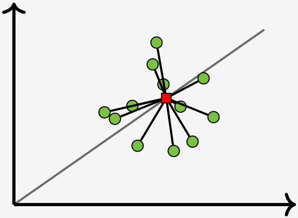


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg\min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.

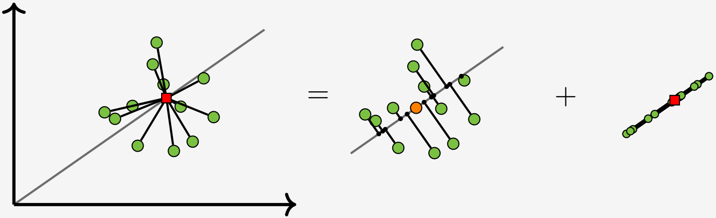


[Drineas, Frieze, Kannan, Vempala, Vinay, 1999]

Let P be a set of n points in $\mathbb{R}^n$. Consider the best fit subspace

$$V_k := \arg \min_{\dim(V)=k} \sum_{p \in P} d(p, V)^2 \subset \mathbb{R}^n.$$

Solving the projected instance in V_k yields a 2-approximation.



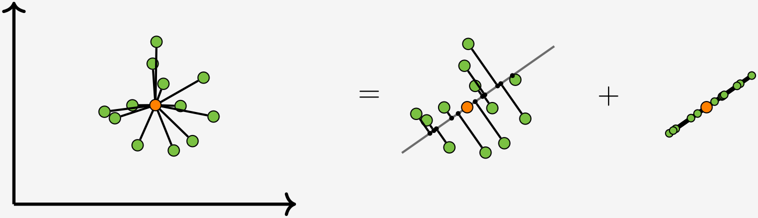
Plan

- $\mathcal{O}(k/\varepsilon^2)$ instead of k dimensions $\rightarrow$ $(1+\varepsilon)$-approximation
- coreset-type guarantee

Plan

- $\mathcal{O}(k/\varepsilon^2)$ instead of k dimensions $\rightarrow$ $(1+\varepsilon)$-approximation
- coreset-type guarantee

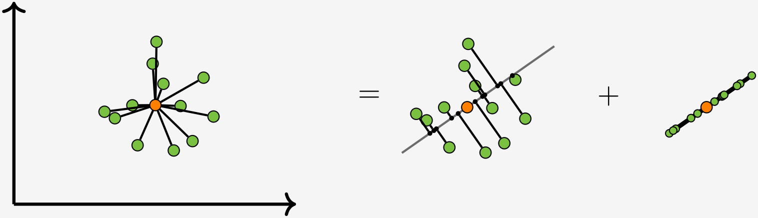
Step 1: Split cost into two terms



Plan

- $\mathcal{O}(k/\varepsilon^2)$ instead of k dimensions $\rightarrow$ $(1+\varepsilon)$-approximation
- coreset-type guarantee

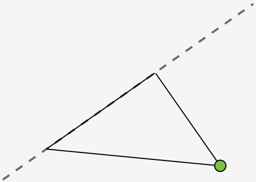
Step 1: Split cost into two terms



For any k-dimensional subspace,
approximate squared distances to and within the subspace!

Step 2: Squared distances to any subspace are correct (approx.)

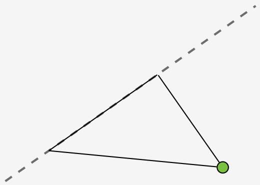
What is the squared distance between a point and a subspace?



$$\text{dist}^2(x, V) = ||x||^2 - ||\phi_V(x)||^2$$

Step 2: Squared distances to any subspace are correct (approx.)

What is the squared distance between a point and a subspace?

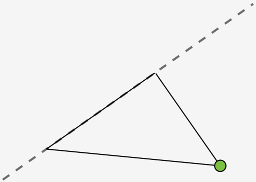


$$\text{dist}^2(x, V) = ||x||^2 - ||\phi_V(x)||^2$$

- gets closer to $||x||^2$ if k is small compared to d
- subspace 'chooses' k directions where the length is disregarded

Step 2: Squared distances to any subspace are correct (approx.)

What is the squared distance between a point and a subspace?



$$\text{dist}^2(x, V) = ||x||^2 - ||\phi_V(x)||^2$$

- gets closer to $||x||^2$ if k is small compared to d
- subspace 'chooses' k directions where the length is disregarded

- First idea: Just say $\sum_{x \in P} ||x||^2$!
- Problem: P lies within k dimensions $\rightarrow$ true answer is 0

- query subspace ‘disregards’ length in *k* directions
- we want to report $\sum ||x||^2 -$ disregarded length

- query subspace ‘disregards’ length in k directions
- we want to report $\sum ||x||^2 -$ disregarded length

Best fit subspace, singular value decomposition (SVD)

Write points in row of a matix A. Then the SVD gives

- singular values $\sigma_1 \geq \ldots \geq \sigma_d$ and vectors $v_1, \ldots, v_d$, form a basis
- $v_1, \ldots, v_m$ span the best fit subspace of P,
- $A = \sum \sigma_i^2 u_i v_i^T$ and projection to V_m is $A_m = \sum_i^m \sigma_i^2 u_i v_i^T$
- $||A||_F^2 = \sum \sigma_i^2$

- query subspace ‘disregards’ length in k directions
- we want to report $\sum ||x||^2 -$ disregarded length

Best fit subspace, singular value decomposition (SVD)

Write points in row of a matix A. Then the SVD gives

- singular values $\sigma_1 \geq \ldots \geq \sigma_d$ and vectors $v_1, \ldots, v_d$, form a basis
- $v_1, \ldots, v_m$ span the best fit subspace of P,
- $A = \sum \sigma_i^2 u_i v_i^T$ and projection to V_m is $A_m = \sum_i^m \sigma_i^2 u_i v_i^T$
- $||A||_F^2 = \sum \sigma_i^2$

Assume that subspace is aligned to singular vectors

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_d^2$$

- we report $\sum_{i=m+1}^d \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

- query subspace 'disregards' length in k directions
- we want to report $\sum ||x||^2 -$ disregarded length

Best fit subspace, singular value decomposition (SVD)

Write points in row of a matix A. Then the SVD gives

- singular values $\sigma_1 \geq \ldots \geq \sigma_d$ and vectors $v_1, \ldots, v_d$, form a basis
- $v_1, \ldots, v_m$ span the best fit subspace of P,
- $A = \sum \sigma_i^2 u_i v_i^T$ and projection to V_m is $A_m = \sum_i^m \sigma_i^2 u_i v_i^T$
- $||A||_F^2 = \sum \sigma_i^2$

Assume that subspace is aligned to singular vectors

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_d^2$$

- we report $\sum_{i=m+1}^d \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

- query subspace 'disregards' length in k directions
- we want to report $\sum ||x||^2 -$ disregarded length

Best fit subspace, singular value decomposition (SVD)

Write points in row of a matix A. Then the SVD gives

- singular values $\sigma_1 \geq \ldots \geq \sigma_d$ and vectors $v_1, \ldots, v_d$, form a basis
- $v_1, \ldots, v_m$ span the best fit subspace of P,
- $A = \sum \sigma_i^2 u_i v_i^T$ and projection to V_m is $A_m = \sum_i^m \sigma_i^2 u_i v_i^T$
- $||A||_F^2 = \sum \sigma_i^2$

Assume that subspace is aligned to singular vectors

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_d^2$$

- we report $\sum_{i=m+1}^d \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

Assume that subspace is aligned to singular vectors

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_d^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

Assume that subspace is aligned to singular vectors

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_d^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

Core idea

Make m large enough such that $\sigma_{m+1}^2 + \ldots + \sigma_{m+k}^2$
is small compared to $\sigma_1^2 + \sigma_2^2 \ldots + \ldots + \sigma_m^2$! $\quad \rightarrow m \geq \lceil k/\varepsilon \rceil$

Assume that subspace is aligned to singular vectors

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_d^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

Core idea

Make m large enough such that $\sigma_{m+1}^2 + \ldots + \sigma_{m+k}^2$
is small compared to $\sigma_1^2 + \sigma_2^2 \ldots + \ldots + \sigma_m^2$! $\rightarrow m \geq \lceil k/\varepsilon \rceil$

Step 3: Squared distances within the subspace

Follows with similar measures, introduces the ε^{-2} and the constant 18

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil 18k/\varepsilon^2 \rceil$, there exists a P' with intrinsic dimension $\lceil 18k/\varepsilon^2 \rceil$ and a constant Δ such that

$$|\operatorname{cost}(P', C) + \Delta - \operatorname{cost}(P, C)| \leq \varepsilon \operatorname{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil 18k/\varepsilon^2 \rceil$, there exists a P' with intrinsic dimension $\lceil 18k/\varepsilon^2 \rceil$ and a constant Δ such that

$$|\operatorname{cost}(P', C) + \Delta - \operatorname{cost}(P, C)| \leq \varepsilon \operatorname{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil ck/\varepsilon^2 \rceil$, there exists a weighted set S with $\tilde{\mathcal{O}}(k^2/\varepsilon^6)$ points and a constant Δ such that

$$|\operatorname{cost}(S, C) + \Delta - \operatorname{cost}(P, C)| \leq \varepsilon \operatorname{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil 18k/\varepsilon^2 \rceil$, there exists a P' with intrinsic dimension $\lceil 18k/\varepsilon^2 \rceil$ and a constant Δ such that

$$|\operatorname{cost}(P', C) + \Delta - \operatorname{cost}(P, C)| \leq \varepsilon \operatorname{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil ck/\varepsilon^2 \rceil$, there exists a weighted set S with $\tilde{\mathcal{O}}(k^2/\varepsilon^6)$ points and a constant Δ such that

$$|\operatorname{cost}(S, C) + \Delta - \operatorname{cost}(P, C)| \leq \varepsilon \operatorname{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Thank you for your attention!