Smaller Coresets for k-Median and k-Means Clustering

Sariel Har-Peled Akash Kushal
UIUC, Urbana, IL

Coresets for Clustering

Clustering

P: set of **n** points in $\mathbb{R}^{\mathbf{d}}$, compute a set **C** of **k** centers –

- **k-median**: sum dist. points in **P** to nearest center.
  $\nu_{\mathbf{C}}(\mathbf{P}) = \sum_{\mathbf{p}\in\mathbf{P}} \mathbf{d}(\mathbf{p}, \mathbf{C})$
- **k-means**: sum squares dist. points in **P** to nearest center.

Target: Find **C** that minimizes price.

Coreset

Definition (Coreset)

$\mathcal{S} \subseteq \mathbf{P}$ - A weighted subset is $(\mathbf{k}, \varepsilon)$-coreset if

$$\forall \mathbf{C} \subseteq \mathbb{R}^{\mathbf{d}} \quad \nu_{\mathbf{C}}(\mathbf{P}) \approx_{\varepsilon} \nu_{\mathbf{C}}(\mathcal{S}),$$

where $\mathbf{a} \approx_{\varepsilon} \mathbf{b}$ if $(1-\varepsilon)\mathbf{a} \leq \mathbf{b} \leq (1+\varepsilon)\mathbf{a}$.

Coreset = small sketch of the input for clustering.

Question

What is the smallest sketch possible?

Coreset

Definition (Coreset)

$\mathcal{S} \subseteq \mathbf{P}$ - A weighted subset is $(\mathbf{k}, \varepsilon)$-coreset if

$$\forall \mathbf{C} \subseteq \mathbb{R}^{\mathbf{d}} \quad \nu_{\mathbf{C}}(\mathbf{P}) \approx_{\varepsilon} \nu_{\mathbf{C}}(\mathcal{S}),$$

where $\mathbf{a} \approx_{\varepsilon} \mathbf{b}$ if $(1-\varepsilon)\mathbf{a} \leq \mathbf{b} \leq (1+\varepsilon)\mathbf{a}$.

Coreset = small sketch of the input for clustering.

Question

What is the smallest sketch possible?

Coreset

Definition (Coreset)

$\mathcal{S} \subseteq \mathbf{P}$ - A weighted subset is $(\mathbf{k}, \varepsilon)$-coreset if

$$\forall \mathbf{C} \subseteq \mathbb{R}^{\mathbf{d}} \quad \nu_{\mathbf{C}}(\mathbf{P}) \approx_{\varepsilon} \nu_{\mathbf{C}}(\mathcal{S}),$$

where $\mathbf{a} \approx_{\varepsilon} \mathbf{b}$ if $(1-\varepsilon)\mathbf{a} \leq \mathbf{b} \leq (1+\varepsilon)\mathbf{a}$.

Coreset = small sketch of the input for clustering.

Question

What is the smallest sketch possible?

k-median

- **[Arora et al., 1998]**:
  $O\left(n^{O(1/\varepsilon)}\right)$ time $(1+\varepsilon)$-approximation for points in plane
- **[Kolliopoulos and Rao, 1999]** (Discrete version):
  $O(\varrho n \log n \log k)$ time, where $\varrho = \exp\left[O\left(\varepsilon^{-1}\log 1/\varepsilon\right)^{d-1}\right]$.
- **[Har-Peled and Mazumdar, 2004]**:
  Coreset of size $O(\frac{k \log n}{\varepsilon^d})$
  Improve running time to $O\left(n + \varrho k^{O(1)} \log^{O(1)} n\right)$

New Result:

Show coreset of size $O(k^2/\varepsilon^d)$

Independent of n!

k-median

- **[Arora et al., 1998]**:
  $O\left(n^{O(1/\varepsilon)}\right)$ time $(1+\varepsilon)$-approximation for points in plane
- **[Kolliopoulos and Rao, 1999]** (Discrete version):
  $O(\varrho n \log n \log k)$ time, where $\varrho = \exp\left[O\left(\varepsilon^{-1}\log 1/\varepsilon\right)^{d-1}\right]$.
- **[Har-Peled and Mazumdar, 2004]**:
  Coreset of size $O(\frac{k \log n}{\varepsilon^d})$
  Improve running time to $O\left(n + \varrho k^{O(1)} \log^{O(1)} n\right)$

New Result:

Show coreset of size $O(k^2/\varepsilon^d)$

Independent of n!

k-median

- **[Arora et al., 1998]**:
  $\mathbf{O}\left(\mathbf{n}^{\mathbf{O}(1/\varepsilon)}\right)$ time $(1+\varepsilon)$-approximation for points in plane
- **[Kolliopoulos and Rao, 1999]** (Discrete version):
  $\mathbf{O}(\varrho \mathbf{n} \log \mathbf{n} \log \mathbf{k})$ time, where $\varrho = \exp\left[\mathbf{O}\left(\varepsilon^{-1} \log 1/\varepsilon\right)^{\mathbf{d}-1}\right]$.
- **[Har-Peled and Mazumdar, 2004]**:
  Coreset of size $\mathbf{O}(\frac{\mathbf{k} \log \mathbf{n}}{\varepsilon^{\mathbf{d}}})$

  Improve running time to $\mathbf{O}\left(\mathbf{n} + \varrho \mathbf{k}^{\mathbf{O}(1)} \log^{\mathbf{O}(1)} \mathbf{n}\right)$

New Result:

Show coreset of size $\mathbf{O}(\mathbf{k}^2/\varepsilon^{\mathbf{d}})$

Independent of **n**!

k-median

- **[Arora et al., 1998]**:
  $\mathbf{O}\left(\mathbf{n}^{\mathbf{O}(1/\varepsilon)}\right)$ time $(1+\varepsilon)$-approximation for points in plane
- **[Kolliopoulos and Rao, 1999]** (Discrete version):
  $\mathbf{O}(\varrho \mathbf{n} \log \mathbf{n} \log \mathbf{k})$ time, where $\varrho = \exp\left[\mathbf{O}\left(\varepsilon^{-1} \log 1/\varepsilon\right)^{\mathbf{d}-1}\right]$.
- **[Har-Peled and Mazumdar, 2004]**:
  Coreset of size $\mathbf{O}(\frac{\mathbf{k} \log \mathbf{n}}{\varepsilon^{\mathbf{d}}})$
  Improve running time to $\mathbf{O}\left(\mathbf{n} + \varrho \mathbf{k}^{\mathbf{O}(1)} \log^{\mathbf{O}(1)} \mathbf{n}\right)$

New Result:

Show coreset of size $\mathbf{O}(\mathbf{k}^2/\varepsilon^{\mathbf{d}})$

Independent of **n**!

New coreset for k-median of size

- independent of **n**!
- disconnected from **n**!!
- free of **n**!!!
- autunomous of **n**!!!!
- unregimented by **n**!!!!!
- unblemished by **n**!!!!!!
- unmolested by **n**!!!!!!!
- !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

New coreset for k-median of size

- independent of **n**!
- disconnected from **n**!!
- free of **n**!!!
- autunomous of **n**!!!!
- unregimented by **n**!!!!!
- unblemished by **n**!!!!!!
- unmolested by **n**!!!!!!!
- !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

k-means

- **[de la Vega et al., 2003]**: $(1+\varepsilon)$-approx in high dim. Running time $\mathbf{O(g(k,\varepsilon)dn\log^k n)}$, where $\mathbf{g(k,\varepsilon) = \exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon))\ln k]}$
- **[Kumar et al., 2004]**: improved to $\mathbf{O(h(k,\varepsilon)dn)}$, where $\mathbf{h(k,\varepsilon) = 2^{(k/\varepsilon)^{O(1)}}}$
- **[Matoušek, 2000]**: running time $\mathbf{O_{d,k}\left(n\varepsilon^{-2k^2d}\log^k n\right)}$
- **[Har-Peled and Mazumdar, 2004]**: Coresets of size $\mathbf{O\left(\frac{k\log n}{\varepsilon^d}\right)}$ ε-approx in $\mathbf{O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1}(n/\varepsilon))}$ time
- New Result: show coreset of size $\mathbf{O\left(k^3/\varepsilon^{d+1}\right)}$

k-means

- **[de la Vega et al., 2003]**: $(1+\varepsilon)$-approx in high dim. Running time $\mathbf{O(g(k,\varepsilon)dn\log^k n)}$, where $\mathbf{g(k,\varepsilon) = \exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon))\ln k]}$
- **[Kumar et al., 2004]**: improved to $\mathbf{O(h(k,\varepsilon)dn)}$, where $\mathbf{h(k,\varepsilon) = 2^{(k/\varepsilon)^{O(1)}}}$
- **[Matoušek, 2000]**: running time $\mathbf{O_{d,k}\left(n\varepsilon^{-2k^2d}\log^k n\right)}$
- **[Har-Peled and Mazumdar, 2004]**:
  Coresets of size $\mathbf{O\left(\frac{k\log n}{\varepsilon^d}\right)}$
  ε-approx in $\mathbf{O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1}(n/\varepsilon))}$ time
- New Result: show coreset of size $\mathbf{O\left(k^3/\varepsilon^{d+1}\right)}$

k-means

- **[de la Vega et al., 2003]**: $(1+\varepsilon)$-approx in high dim. Running time $\mathbf{O(g(k,\varepsilon)dn\log^k n)}$, where $\mathbf{g(k,\varepsilon) = \exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon))\ln k]}$
- **[Kumar et al., 2004]**: improved to $\mathbf{O(h(k,\varepsilon)dn)}$, where $\mathbf{h(k,\varepsilon) = 2^{(k/\varepsilon)^{O(1)}}}$
- **[Matoušek, 2000]**: running time $\mathbf{O_{d,k}\left(n\varepsilon^{-2k^2d}\log^k n\right)}$
- **[Har-Peled and Mazumdar, 2004]**:
  Coresets of size $\mathbf{O\left(\frac{k\log n}{\varepsilon^d}\right)}$
  ε-approx in $\mathbf{O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1}(n/\varepsilon))}$ time
- New Result: show coreset of size $\mathbf{O\left(k^3/\varepsilon^{d+1}\right)}$

k-means

- **[de la Vega et al., 2003]**: $(1+\varepsilon)$-approx in high dim. Running time $O(g(k,\varepsilon)dn\log^k n)$, where $g(k,\varepsilon) = \exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon))\ln k]$
- **[Kumar et al., 2004]**: improved to $O(h(k,\varepsilon)dn)$, where $h(k,\varepsilon) = 2^{(k/\varepsilon)^{O(1)}}$
- **[Matoušek, 2000]**: running time $O_{d,k}\left(n\varepsilon^{-2k^2d}\log^k n\right)$
- **[Har-Peled and Mazumdar, 2004]**: Coresets of size $O\left(\frac{k\log n}{\varepsilon^d}\right)$ ε-approx in $O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1}(n/\varepsilon))$ time
- New Result: show coreset of size $O\left(k^3/\varepsilon^{d+1}\right)$

k-means

- **[de la Vega et al., 2003]**: $(1+\varepsilon)$-approx in high dim. Running time $O(g(k,\varepsilon)dn\log^k n)$, where $g(k,\varepsilon) = \exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon))\ln k]$
- **[Kumar et al., 2004]**: improved to $O(h(k,\varepsilon)dn)$, where $h(k,\varepsilon) = 2^{(k/\varepsilon)^{O(1)}}$
- **[Matoušek, 2000]**: running time $O_{d,k}\left(n\varepsilon^{-2k^2d}\log^k n\right)$
- **[Har-Peled and Mazumdar, 2004]**: Coresets of size $O\left(\frac{k\log n}{\varepsilon^d}\right)$ ε-approx in $O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1}(n/\varepsilon))$ time
- New Result: show coreset of size $O\left(k^3/\varepsilon^{d+1}\right)$

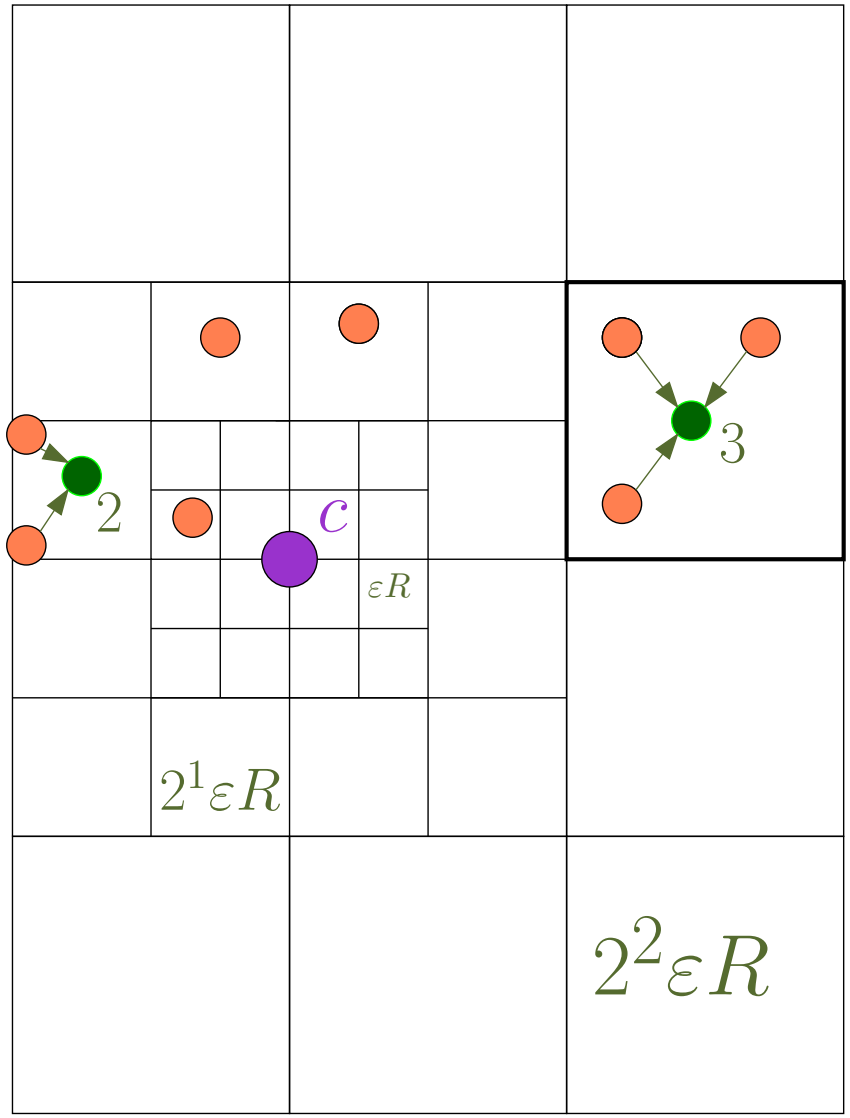
k-means - Effros & Schulman result

- Deterministic clustering with data nets
  [Effros and Schulman, 2003]
- For **k** means only.
- Show the existence of centroid set of size independent of **n**.
- Complicated.
- Inspiration to conjecture that small coresets exists.
- New result implies their result (better parameters).
- New result much simpler.
- Effros & Schulman recently showed existance of small coresets for **k**-means.

Old Coreset Construction

Construction

- **D** - a set of **k** centers $\nu_{\mathbf{D}}(\mathbf{P}) = \mathbf{O}(\nu_{\mathbf{opt}})$.
- build exponential grid of $\mathbf{O(\log n)}$ levels around each center,
- snap the points of **P** to this grid.
- *price of snapping is smaller than* $\varepsilon\nu_{\mathbf{opt}}(\mathbf{P}, \mathbf{k})$



Old Coreset Construction – limitations

Lemma $\Rightarrow$ Problem

If snapping is "cheap" ($\leq \varepsilon \nu_{\mathrm{opt}}(\mathbf{P}, \mathbf{k})$), then $|\mathcal{S}| = \Omega(\log n)$.

Conclusions

For $\mathcal{S}$ to be small

$\Rightarrow$ Must pick $\mathcal{S}$ such that snapping errors *cancel each other out*.

Old Coreset Construction – limitations

Lemma $\Rightarrow$ Problem

If snapping is "cheap" ($\leq \varepsilon \nu_{\mathrm{opt}}(\mathbf{P}, \mathbf{k})$), then $|\mathcal{S}| = \Omega(\log n)$.

Conclusions

For $\mathcal{S}$ to be small

$\Rightarrow$ Must pick $\mathcal{S}$ such that snapping errors *cancel each other out*.

Warmup exercise - Coreset in One Dimension

Problem

$\mathbf{P} \subseteq \mathbb{R}$ - $\mathbf{n}$ *points on real line.*
Pick $(\mathbf{k}, \varepsilon)$*-coreset for* $\mathbf{P}$ *for centers on the line.*

Basic Idea:

Break the point set into batches, and use the mean point of batch, as the representative for the coreset.

Warmup exercise - Coreset in One Dimension

Problem

$\mathbf{P} \subseteq \mathbb{R}$ - $\mathbf{n}$ *points on real line.*
Pick $(\mathbf{k}, \varepsilon)$*-coreset for* $\mathbf{P}$ *for centers on the line.*

Basic Idea:

Break the point set into batches, and use the mean point of batch, as the representative for the coreset.

Warmup exercise - Coreset in One Dimension

Problem

$\mathbf{P} \subseteq \mathbb{R}$ - $\mathbf{n}$ *points on real line.*
Pick $(\mathbf{k}, \varepsilon)$*-coreset for* $\mathbf{P}$ *for centers on the line.*

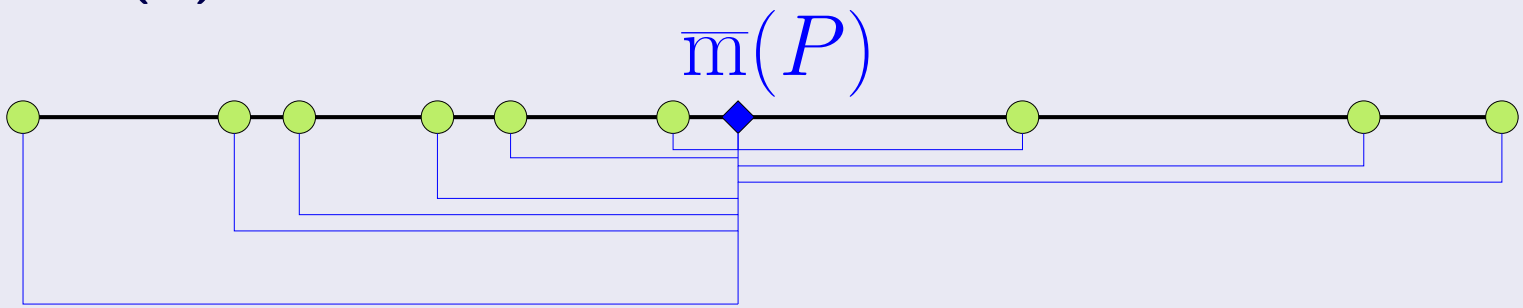
Basic Idea:

Break the point set into batches, and use the mean point of batch, as the representative for the coreset.

Warmup exercise - Coreset in One Dimension - cont'd

Definition (Cumulative Error)

$\mathcal{E}_\nu(\mathbf{P}) = \nu_{\overline{\mathbf{m}}}(\mathbf{P}) = \sum_{\mathbf{p}\in\mathbf{P}} \mathbf{w_p}\|\mathbf{p}\overline{\mathbf{m}}\|$
$\overline{\mathbf{m}} = \overline{\mathbf{m}}(\mathbf{P})$ is mean of $\mathbf{P}$.



$\mathcal{E}_\nu(\mathbf{P})$ = Total length of blue segments.

Warmup exercise - Coreset in One Dimension - cont'd

- $\mathbf{V}$: $\nu_{\mathrm{opt}}(\mathbf{P}, \mathbf{k}) \leq \mathbf{V} \leq \mathbf{c} \cdot \nu_{\mathrm{opt}}(\mathbf{P}, \mathbf{k})$
- Sweep $\mathbf{P}$ from left to right.
- Break into batches $\mathbf{B_i}$ with $\mathcal{E}_\nu(\mathbf{B_i}) = \frac{\varepsilon}{\mathbf{10ck}}\mathbf{V}$ (allow fractional points)
- # batches is $\alpha = \mathbf{O(k/\varepsilon)}$
- Coreset point: $(\overline{\mathbf{m}}(\mathbf{B_i}), \mathbf{w}(\mathbf{B_i}))$
- Coreset: $\left\{ (\overline{\mathbf{m}}(\mathbf{B_i}), \mathbf{w}(\mathbf{B_i})) \;\middle|\; \mathbf{i} = 1, \ldots, \alpha \right\}$.

Correctness

- $\mathbf{B}$ - a batch of points
- $\mathcal{I}(\mathbf{B})$ - interval containing $\mathbf{B}$.
- $\forall \mathbf{c}$ outside interval $\mathcal{I}(\mathbf{B})$ then $\nu_{\mathbf{c}}(\mathbf{P}) = \nu_{\mathbf{c}}(\mathcal{S})$.
- *problematic* batches = contribute to the error
  - Batches served by two or more centers ($\leq \mathbf{k}-\mathbf{1}$ batches)
  - batches with center point inside them ($\leq \mathbf{k}$ such batches).
- Error of problematic batches $\leq \mathbf{L} = \mathcal{E}_{\nu}(\mathbf{B_i}) = \frac{\varepsilon}{\mathbf{10ck}}\mathbf{V}$
- Total error $\leq (\mathbf{2k}-\mathbf{1}) \cdot \mathbf{L} \leq \varepsilon \nu_{\mathrm{opt}}(\mathbf{P}, \mathbf{k})$.

Construction can be extended to:

- Points on line $\ell \subseteq \mathbb{R}^{\mathbf{d}}$.
- Centers somewhere in $\mathbb{R}^{\mathbf{d}}$
- Extension not trivial (but doable - details in paper)

Correctness

- $\mathbf{B}$ - a batch of points
- $\mathcal{I}(\mathbf{B})$ - interval containing $\mathbf{B}$.
- $\forall \mathbf{c}$ outside interval $\mathcal{I}(\mathbf{B})$ then $\nu_{\mathbf{c}}(\mathbf{P}) = \nu_{\mathbf{c}}(\mathcal{S})$.
- *problematic* batches $=$ contribute to the error
  - Batches served by two or more centers ($\leq \mathbf{k-1}$ batches)
  - batches with center point inside them ($\leq \mathbf{k}$ such batches).
- Error of problematic batches $\leq \mathbf{L} = \mathcal{E}_{\nu}(\mathbf{B_i}) = \frac{\varepsilon}{\mathbf{10ck}}\mathbf{V}$
- Total error $\leq (\mathbf{2k-1}) \cdot \mathbf{L} \leq \varepsilon \nu_{\mathrm{opt}}(\mathbf{P}, \mathbf{k})$.

Construction can be extended to:

- Points on line $\ell \subseteq \mathbb{R}^{\mathbf{d}}$.
- Centers somewhere in $\mathbb{R}^{\mathbf{d}}$
- Extension not trivial (but doable - details in paper)

Correctness

- $\mathbf{B}$ - a batch of points
- $\mathcal{I}(\mathbf{B})$ - interval containing $\mathbf{B}$.
- $\forall \mathbf{c}$ outside interval $\mathcal{I}(\mathbf{B})$ then $\nu_{\mathbf{c}}(\mathbf{P}) = \nu_{\mathbf{c}}(\mathcal{S})$.
- *problematic* batches $=$ contribute to the error
  - Batches served by two or more centers ($\leq \mathbf{k}-\mathbf{1}$ batches)
  - batches with center point inside them ($\leq \mathbf{k}$ such batches).
- Error of problematic batches $\leq \mathbf{L} = \mathcal{E}_{\nu}(\mathbf{B_i}) = \frac{\varepsilon}{\mathbf{10ck}}\mathbf{V}$
- Total error $\leq (\mathbf{2k}-\mathbf{1}) \cdot \mathbf{L} \leq \varepsilon \nu_{\mathbf{opt}}(\mathbf{P}, \mathbf{k})$.

Construction can be extended to:

- Points on line $\ell \subseteq \mathbb{R}^{\mathbf{d}}$.
- Centers somewhere in $\mathbb{R}^{\mathbf{d}}$
- Extension not trivial (but doable - details in paper)

Correctness

- $\mathbf{B}$ - a batch of points
- $\mathcal{I}(\mathbf{B})$ - interval containing $\mathbf{B}$.
- $\forall \mathbf{c}$ outside interval $\mathcal{I}(\mathbf{B})$ then $\nu_{\mathbf{c}}(\mathbf{P}) = \nu_{\mathbf{c}}(\mathcal{S})$.
- *problematic* batches $=$ contribute to the error
  - Batches served by two or more centers ($\leq \mathbf{k}-\mathbf{1}$ batches)
  - batches with center point inside them ($\leq \mathbf{k}$ such batches).
- Error of problematic batches $\leq \mathbf{L} = \mathcal{E}_{\nu}(\mathbf{B_i}) = \frac{\varepsilon}{\mathbf{10ck}}\mathbf{V}$
- Total error $\leq (\mathbf{2k}-\mathbf{1}) \cdot \mathbf{L} \leq \varepsilon \nu_{\mathbf{opt}}(\mathbf{P}, \mathbf{k})$.

Construction can be extended to:

- Points on line $\ell \subseteq \mathbb{R}^{\mathbf{d}}$.
- Centers somewhere in $\mathbb{R}^{\mathbf{d}}$
- Extension not trivial (but doable - details in paper)

What we currently have...

Lemma

- $\mathbf{P}$ *a set of points on a line* ℓ
- ℓ *- line in* $\mathbb{R}^{\mathbf{d}}$
- *Coreset of size* $\mathbf{O(k/\varepsilon)}$ *for* $\mathbf{P}$
- *For* **k***-median clustering.*

Problem

General point-set do not lie on a single line.

What we currently have...

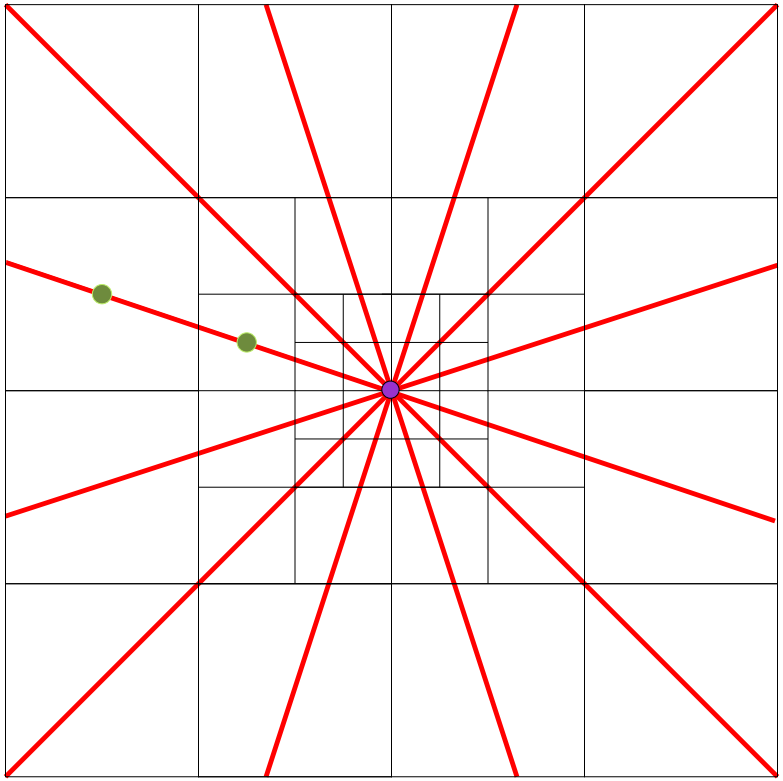
Lemma

- $\mathbf{P}$ *a set of points on a line* ℓ
- ℓ *- line in* $\mathbb{R}^{\mathbf{d}}$
- *Coreset of size* $\mathbf{O(k/\varepsilon)}$ *for* $\mathbf{P}$
- *For* **k***-median clustering.*

Problem

General point-set do not lie on a single line.

Snapping to lines instead of grid



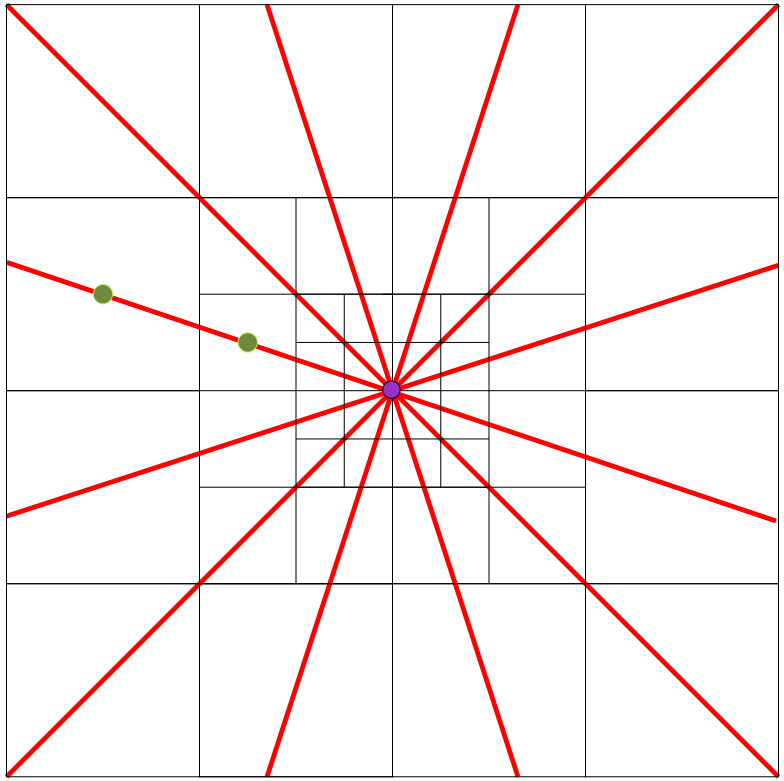
# of grid cell centers: $\mathbf{O(\varepsilon^{-d} \log n)}$.

# of lines: $\mathbf{O(1/\varepsilon^{d})}$.

Conclusion

We can snap points to a small number of lines.

Snapping to lines instead of grid



# of grid cell centers: $\mathbf{O(\varepsilon^{-d} \log n)}$.
# of lines: $\mathbf{O(1/\varepsilon^{d})}$.

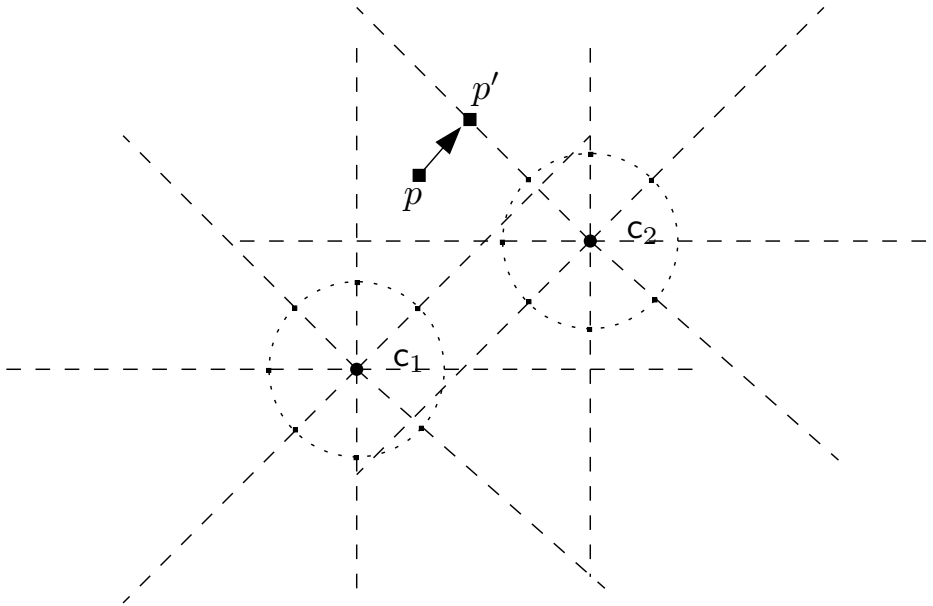
Conclusion

We can snap points to a small number of lines.

Extending to Higher Dimensions

- comp. $\mathbf{C} = \{\mathbf{c_1}, \ldots, \mathbf{c_k}\}$ s.t. $\nu_{\mathbf{C}}(\mathbf{P}) = \mathbf{O}(\nu_{\mathbf{opt}}(\mathbf{P}, \mathbf{k}))$
- construct a fan of lines around points of $\mathbf{C}$
- snap points of $\mathbf{P}$ to the closest line
- build coreset for each set on a line.

Correctness



- Number of lines is $\mathbf{O(\frac{k}{\varepsilon^{d-1}})} \Rightarrow$ coreset size is $\mathbf{O(\frac{k^2}{\varepsilon^d})}$
- snapping error is bounded by $\frac{\varepsilon}{3}\nu_{\mathbf{opt}}(\mathbf{P}, \mathbf{k})$
- error by construction $\frac{\varepsilon}{3}\nu_{\mathbf{opt}}(\mathbf{P'}, \mathbf{k})$ ($\mathbf{P'}$ are snapped points)
- Total error $\leq \varepsilon\nu_{\mathbf{opt}}(\mathbf{P}, \mathbf{k})$

k-means: coreset construction

- Similar scheme works for **k**-means.
- Analysis slightly simpler.
- $\Rightarrow$ Coreset size $\mathbf{O(k^3/\varepsilon^{d+1})}$
- Full details in the paper.

k-means: results

Definition (Centroid Set)

$\mathcal{D} \subseteq \mathbb{R}^{\mathbf{d}}$: $(\mathbf{k}, \varepsilon)$-*approximate centroid set* for $\mathbf{P}$, if $\exists \mathbf{C} \subseteq \mathcal{D}$ such that $\mu_{\mathbf{C}}(\mathbf{P}) \leq (1+\varepsilon)\mu_{\mathbf{opt}}(\mathbf{P}, \mathbf{k})$

- Matoušek **[Matoušek, 2000]**: $\exists$ an ε-approx. centroid set. Size $\mathbf{O}(\mathbf{n}\varepsilon^{-\mathbf{d}} \log(1/\varepsilon))$
- **[Effros and Schulman, 2003]** Centroid set of size $\mathbf{O}\left(\varepsilon^{-\mathbf{d}-\mathbf{1}}(\mathbf{k}^{\mathbf{4}} + \mathbf{k}^{\mathbf{2}}\varepsilon^{-\mathbf{2}})\right)$.

New Result:

A $(\mathbf{k}, \varepsilon)$-centroid set for $\mathbf{k}$-means clustering with size $\mathbf{O}\left((\mathbf{k}^{\mathbf{3}}/\varepsilon^{\mathbf{2d+1}}) \log(1/\varepsilon)\right)$.
Fast running time...

Minor improvement in RT for $\mathbf{k}$-means clustering (over **[Har-Peled and Mazumdar, 2004]**)

Open Problems

- Improve running time of approx. **k**-means clustering?
- FPTAS for **k**-median and **k**-means (in both **k** and $\mathbf{1}/\varepsilon$)?
- A coreset with only polynomial dependency on the dimension and no dependency on **n**?
  Relevant results **[Bădoiu et al., 2002, Chen, 2004]**.
- Improve dependency on **k** and ε in coresets size?


Arora, S., Raghavan, P., and Rao, S. (1998).
Approximation schemes for Euclidean **k**-median and related problems.
In *Proc. 30th Annu. ACM Sympos. Theory Comput.*, pages 106–113.


Bădoiu, M., Har-Peled, S., and Indyk, P. (2002).
Approximate clustering via coresets.
In *Proc. 34th Annu. ACM Sympos. Theory Comput.*, pages 250–257.


Chen, K. (2004).
Clustering algorithms using adaptive sampling.
manuscript.


de la Vega, W. F., Karpinski, M., Kenyon, C., and Rabani, Y. (2003).
Approximation schemes for clustering problems.
In *Proc. 35th Annu. ACM Sympos. Theory Comput.*, pages 50–58.


Effros, M. and Schulman, L. J. (2003).
Deterministic clustering with data nets.
Technical Report TR04-085, Elec. Colloq. Comp. Complexity.
http://www.eccc.uni-trier.de/eccc-reports/2004/TR04-085/.


Har-Peled, S. and Mazumdar, S. (2004).
Coresets for **k**-means and **k**-median clustering and their applications.
In *Proc. 36th Annu. ACM Sympos. Theory Comput.*, pages 291–300.


Kolliopoulos, S. G. and Rao, S. (1999).
A nearly linear-time approximation scheme for the euclidean κ-median problem.
In *Proc. 7th Annu. European Sympos. Algorithms*, pages 378–389.


Kumar, A., Sabharwal, Y., and Sen, S. (2004).

Linear time algorithms for clustering problems in any dimension.
manuscript.

Matoušek, J. (2000).
On approximate geometric **k**-clustering.
Discrete Comput. Geom., 24:61–84.