Distributed Partial Clustering

Sudipto Guha — Upenn

Yi Li — NTU

Qin Zhang — IUB

SPAA 2017
July 25, 2017

Clustering

- Metric space (X, d)
- n input points A; want to find k centers
- Objective function (k-median):

$$\min_{K \subseteq A : |K| = k} \sum_{p \in A} d(p, K)$$

- k-means: $\sum_{p \in A} d^2(p, K)$
  k-center: $\max_{p \in A} d(p, K)$

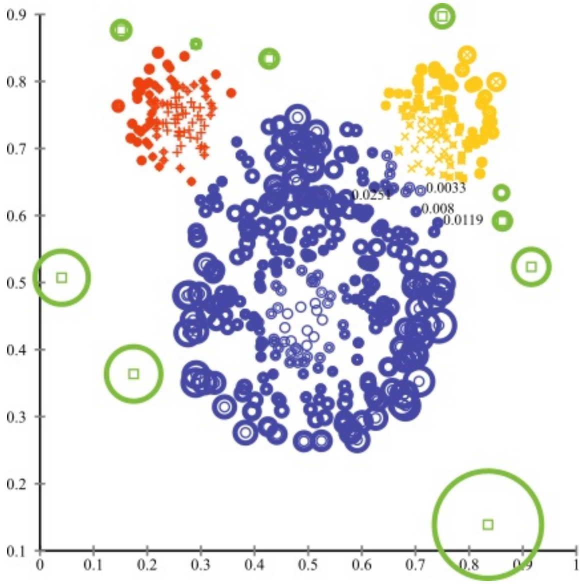
Clustering with outliers

- Metric space (X, d)
- n input points A; want to find k centers, t outliers
- Objective function ((k,t)-median):

$$\min_{K, O \subseteq A : |K| = k, |O| \leq t} \sum_{p \in A \setminus O} d(p, K)$$

- (k,t)-means: $\sum_{p \in A \setminus O} d^2(p, K)$
  (k,t)-center: $\max_{p \in A \setminus O} d(p, K)$

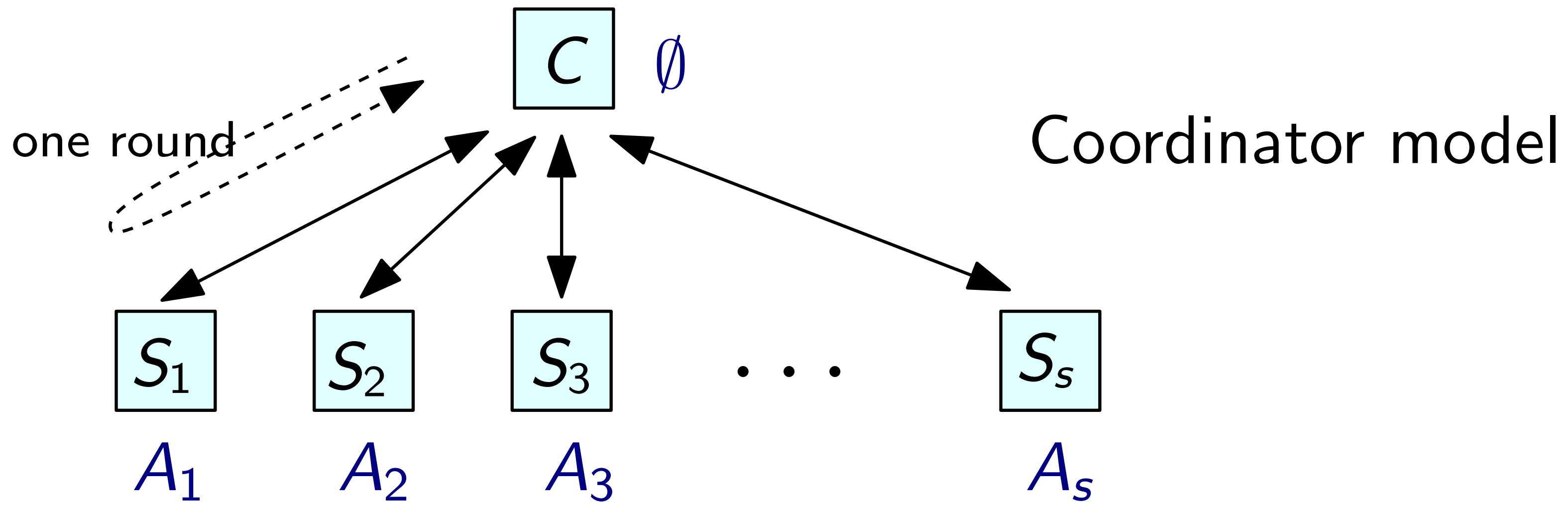
Motivation: partial optimization
gives much better results



Distributed clustering

- s sites, coordinator model
- Site i gets A_i, parties want to cluster $A = A_1 \cup \ldots \cup A_s$
- Want to minimize comm. cost and #comm. rounds
- For simiplicity, assume each point takes $\tilde{O}(1)$ bits

Motivation: data is inherently distributed /
data is big and does not fit one machine



Clustering on uncertain data

- Each data item j is a distribution; call it a *node*.
  Motivation: data is noisy; a subfield in databases
  Let $\sigma(j)$ denote a realization, $\pi(j)$ the center point to which j is attached.

Clustering on uncertain data

- Each data item j is a distribution; call it a *node*.
  Motivation: data is noisy; a subfield in databases
  Let $\sigma(j)$ denote a realization, $\pi(j)$ the center point to which j is attached.

- Objective function (k-median):

$$\min_{K \subseteq A : |K| = k} \sum_{j \in A} \mathrm{E}_\sigma[d(\sigma(j), \pi(j))]$$

- k-means: replace $d(p, K)$ with $d^2(p, K)$.

- k-center has two versions: E and max do not commute.
  - $\max_{j \in A} \mathrm{E}[d(\sigma(j), \pi(j))]$
  - $\mathrm{E}[\max_{j \in A} d(\sigma(j), \pi(j))]$

Clustering with outlier on uncertain data

- Each data item j is a distribution; call it a *node*.
  Let $\sigma(j)$ denote a realization, $\pi(j)$ the center point to which j is attached.

- Objective function ((k, t)-median):

$$\min_{K, O \subseteq A : |K| = k, |O| \leq t} \sum_{p \in A \setminus O} \mathrm{E}_{\sigma}[d(\sigma(j), \pi(j))]$$

- (k, t)-means: replace $d(p, K)$ with $d^2(p, K)$.

- (k, t)-center has two versions: E and max do not commute.
  - $\max_{j \in A \setminus O} \mathrm{E}[d(\sigma(j), \pi(j))]$ (k, t)-center-pp
  - $\mathrm{E}\left[\max_{j \in A \setminus O} d(\sigma(j), \pi(j))\right]$ (k, t)-center-global

Old and New Problems

Problems studied before

- Clustering [??, XXXX]
- Clustering with outliers [CKMN, 2001]
- Clustering on uncertain data [CM, 2008]
- Distributed clustering [??, XXXX]
- Distributed clustering with outliers for k-center
  [MKCWM, 2015]
  Implicitly also in [GMMMO, 2003]

New problems

- Distributed clustering with outliers for k-median/means
- Distributed clustering (with outliers) for uncertain data

Old and New Problems

Problems studied before

- Clustering [??, XXXX]
- Clustering with outliers [CKMN, 2001]
- Clustering on uncertain data [CM, 2008]
- Distributed clustering [??, XXXX]
- Distributed clustering with outliers for k-center This paper
  [MKCWM, 2015]
  Implicitly also in [GMMMO, 2003]

New problems

- Distributed clustering with outliers for k-median/means
  This paper
- Distributed clustering (with outliers) for uncertain data
  This paper

Main results

Bicriteria: (α, β)-approx if the cost of SOL is at most αC while excluding βt points, where C is OPT for excluding t points

Main results

Bicriteria: (α, β)-approx if the cost of SOL is at most αC while excluding βt points, where C is OPT for excluding t points

Main results (all in 2 rounds; under the same framework):

- $(O(1), 1)$-approx with $\tilde{O}(sk + t)$ comm. for (k, t)-median/center

Main results

Bicriteria: (α, β)-approx if the cost of SOL is at most αC while excluding βt points, where C is OPT for excluding t points

Main results (all in 2 rounds; under the same framework):

- $(O(1), 1)$-approx with $\tilde{O}(sk + t)$ comm. for (k, t)-median/center
- $((1 + 1/\epsilon), 1 + \epsilon)$-approx with $\tilde{O}(sk + t)$ comm. for (k, t)-median and (k, t)-means, with quadratic local time

Main results

Bicriteria: (α, β)-approx if the cost of SOL is at most αC while excluding βt points, where C is OPT for excluding t points

Main results (all in 2 rounds; under the same framework):

- $(O(1), 1)$-approx with $\tilde{O}(sk + t)$ comm. for (k, t)-median/center
- $((1 + 1/\epsilon), 1 + \epsilon)$-approx with $\tilde{O}(sk + t)$ comm.
  for (k, t)-median and (k, t)-means, with quadratic local time

  Also leads to subquadratic time $(O(1), O(1))$-approx centralized algorithms (open for many years)

Main results

Bicriteria: (α, β)-approx if the cost of SOL is at most αC while excluding βt points, where C is OPT for excluding t points

Main results (all in 2 rounds; under the same framework):

- $(O(1), 1)$-approx with $\tilde{O}(sk + t)$ comm. for (k, t)-median/center
- $((1 + 1/\epsilon), 1 + \epsilon)$-approx with $\tilde{O}(sk + t)$ comm. for (k, t)-median and (k, t)-means, with quadratic local time

  Also leads to subquadratic time $(O(1), O(1))$-approx centralized algorithms (open for many years)
- $((1 + 1/\epsilon), 1 + \epsilon)$-approx with $\tilde{O}(sk + t)$ comm. for uncertain (k, t)-median/means and (k, t)-center-pp

Main results

Bicriteria: (α, β)-approx if the cost of SOL is at most αC while excluding βt points, where C is OPT for excluding t points

Main results (all in 2 rounds; under the same framework):

- $(O(1), 1)$-approx with $\tilde{O}(sk + t)$ comm. for (k, t)-median/center
- $((1 + 1/\epsilon), 1 + \epsilon)$-approx with $\tilde{O}(sk + t)$ comm.
  for (k, t)-median and (k, t)-means, with quadratic local time

  Also leads to subquadratic time $(O(1), O(1))$-approx centralized algorithms (open for many years)
- $((1 + 1/\epsilon), 1 + \epsilon)$-approx with $\tilde{O}(sk + t)$ comm.
  for uncertain (k, t)-median/means and (k, t)-center-pp
- $((1 + 1/\epsilon), 1 + \epsilon)$-approx with $\tilde{O}(sk + tI + s \log \Delta)$ comm.
  for uncertain (k, t)-center-global, where I is the info to encode the distribution, and Δ is the max-distance/min-distance

Previous results on distributed clustering with outliers

- $\tilde{O}(sk + st)$ bits in 2 rounds for k-center

  (Malkomes, Kusner, Chen, Weinberger, Moseley. 2015)

- $\tilde{O}(sk + st)$ bits in 1 round for k-median/means/center

  Can be derived from (Guha, Meyerson, Mishra, Motwani, O'Callaghan. 2003)

Previous results on distributed clustering with outliers

- $\tilde{O}(sk + st)$ bits in 2 rounds for k-center

  (Malkomes, Kusner, Chen, Weinberger, Moseley. 2015)

- $\tilde{O}(sk + st)$ bits in 1 round for k-median/means/center

  Can be derived from (Guha, Meyerson, Mishra, Motwani, O'Callaghan. 2003)

- Interesting range of parameters: $n \gg t \gg k, s$.

  Consider a modest data set, say $n = 10^8$.

  Suppose that 0.1% is noise, thus $t = 0.001 \times n = 10^5$.

  Say $s = 1000$, and $k = 100$. Then $sk = 10^5$, and $st = 10^8$
  Consequently $sk + st = 10^8$ while $sk + t = 10^5$

Previous results on distributed clustering with outliers

- $\tilde{O}(sk + st)$ bits in 2 rounds for k-center
  (Malkomes, Kusner, Chen, Weinberger, Moseley. 2015)
- $\tilde{O}(sk + st)$ bits in 1 round for k-median/means/center
  Can be derived from (Guha, Meyerson, Mishra, Motwani, O'Callaghan. 2003)
- Interesting range of parameters: $n \gg t \gg k, s$.
  Consider a modest data set, say $n = 10^8$.
  Suppose that 0.1% is noise, thus $t = 0.001 \times n = 10^5$.
  Say $s = 1000$, and $k = 100$. Then $sk = 10^5$, and $st = 10^8$
  Consequently $sk + st = 10^8$ while $sk + t = 10^5$
- Goal: reduce the st term to t, since
  the difference $\Rightarrow$ your data/energy/time bill.

More related work

(Centralized) clustering with outliers

- 3-approx for (k, t)-center, $(O(1), O(1))$-approx for (k, t)-median (Charikar, Khuller, Mount, Narasimhan, 2001)

  $O(1)$-approx for (k, t)-median by Ke Chen (2008)

- (k, t)-median with different loss functions. (Feldman, Schulman, 2012)

More related work

(Centralized) clustering with outliers

- 3-approx for (k, t)-center, $(O(1), O(1))$-approx for (k, t)-median (Charikar, Khuller, Mount, Narasimhan, 2001)

  $O(1)$-approx for (k, t)-median by Ke Chen (2008)
- (k, t)-median with different loss functions. (Feldman, Schulman, 2012)

Uncertain data

- Uncertain k-center/median/means (Cormode, McGregor, 2008)
- Better results for k-center (Guha, Munagala, 2009)

More related work

(Centralized) clustering with outliers

- 3-approx for (k, t)-center, $(O(1), O(1))$-approx for (k, t)-median (Charikar, Khuller, Mount, Narasimhan, 2001)
  $O(1)$-approx for (k, t)-median by Ke Chen (2008)
- (k, t)-median with different loss functions. (Feldman, Schulman, 2012)

Uncertain data

- Uncertain k-center/median/means (Cormode, McGregor, 2008)
- Better results for k-center (Guha, Munagala, 2009)

Distributed clustering (coordinator model)

- $O(1)$-approx with $\tilde{O}(kd + sk)$ for k-median/means in d-dim Euclidean space (Balcan, Ehrlich, Liang, 2013)
- Better results for k-means by (Liang, Balcan, Kanchanapally, Woodruff, 2014), and (Cohen, Elder, Musco, Musco, Persu, 2015).

Distributed (k, t)-median and the Algorithm Framework

Two-level distributed clustering (GMMMO, 2003)

- Consider k-median. Let $A = A_1 \cup \ldots \cup A_s$.
- Site i computes M_i which is the set of centers of the local k-median solution on A_i.

  Weight of each $p \in M_i$ is the number of points assigned to p

  Let $M = M_1 \cup \ldots \cup M_s$

  Let L : sum of costs of local solutions
- Coordinator computes weighted clustering on M

Two-level distributed clustering (GMMMO, 2003)

- Consider k-median. Let $A = A_1 \cup \ldots \cup A_s$.
- Site i computes M_i which is the set of centers of the local k-median solution on A_i.

  Weight of each $p \in M_i$ is the number of points assigned to p

  Let $M = M_1 \cup \ldots \cup M_s$

  Let L : sum of costs of local solutions
- Coordinator computes weighted clustering on M

Any $O(1)$-approx, but can assume optimal for now

Two-level distributed clustering (GMMMO, 2003)

- Consider k-median. Let $A = A_1 \cup \ldots \cup A_s$.
- Site i computes M_i which is the set of centers of the local k-median solution on A_i.

  Weight of each $p \in M_i$ is the number of points assigned to p

  Let $M = M_1 \cup \ldots \cup M_s$

  Let L : sum of costs of local solutions
- Coordinator computes weighted clustering on M
- It holds that $C_{sol}(M, k) \leq O(1) \cdot (L + C_{opt}(A, k)) + L$,

  and $L = O(C_{opt}(A, k))$; we thus get $O(1)$-approx

Two-level distributed clustering (GMMMO, 2003)

- Consider k-median. Let $A = A_1 \cup \ldots \cup A_s$.
- Site i computes M_i which is the set of centers of the local k-median solution on A_i.

  Weight of each $p \in M_i$ is the number of points assigned to p

  Let $M = M_1 \cup \ldots \cup M_s$

  Let L : sum of costs of local solutions
- Coordinator computes weighted clustering on M
- It holds that $C_{sol}(M, k) \leq O(1) \cdot (L + C_{opt}(A, k)) + L$,

  and $L = O(C_{opt}(A, k))$; we thus get $O(1)$-approx
- Similar result holds for (k, t)-median

Two-level distributed clustering (GMMMO, 2003)

- Consider k-median. Let $A = A_1 \cup \ldots \cup A_s$.
- Site i computes M_i which is the set of centers of the local k-median solution on A_i.

  Weight of each $p \in M_i$ is the number of points assigned to p

  Let $M = M_1 \cup \ldots \cup M_s$

  Let L : sum of costs of local solutions
- Coordinator computes weighted clustering on M
- It holds that $C_{sol}(M, k) \leq O(1) \cdot (L + C_{opt}(A, k)) + L$,

  and $L = O(C_{opt}(A, k))$; we thus get $O(1)$-approx
- Similar result holds for (k, t)-median

  – Each site computes a local (k, t)-median and then sends both k centers (and their weights) and t outliers to the coordinator
  – Coordinator performs a second-level clustering
  Comm. cost: $\tilde{O}(sk + st)$

Local solutions

- Let t_i^* be # excluded points in A_i in $\mathrm{OPT}(A, k, t)$
- One can show that

$$\sum_{i \in [s]} C_{opt}(A_i, k, t_i^*) \leq O(1) \cdot C_{opt}(A, k, t)$$

- Thus if we knows t_i^*, then we can apply the result on the previous slide. (but of course we do not know t_i^*)

Local solutions

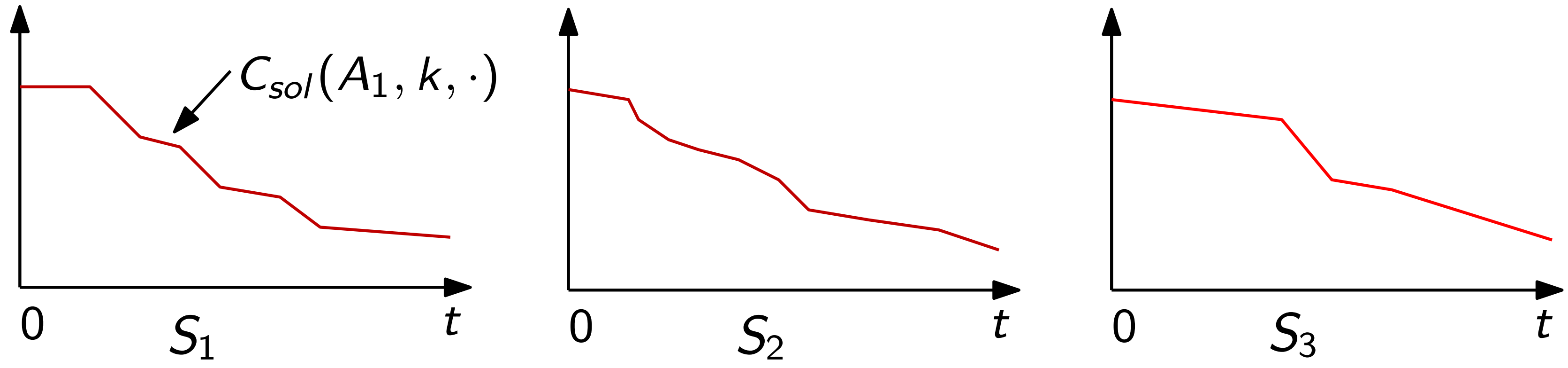
- Let t_i^* be # excluded points in A_i in OPT(A, k, t)
- One can show that

$$\sum_{i \in [s]} C_{opt}(A_i, k, t_i^*) \leq O(1) \cdot C_{opt}(A, k, t)$$

- Thus if we knows t_i^*, then we can apply the result on the previous slide. (but of course we do not know t_i^*)
- We want to miminize $\sum_{i \in [s]} C_{sol}(A_i, k, t_i)$ s.t. $\sum_{i \in [s]} t_i = t$.

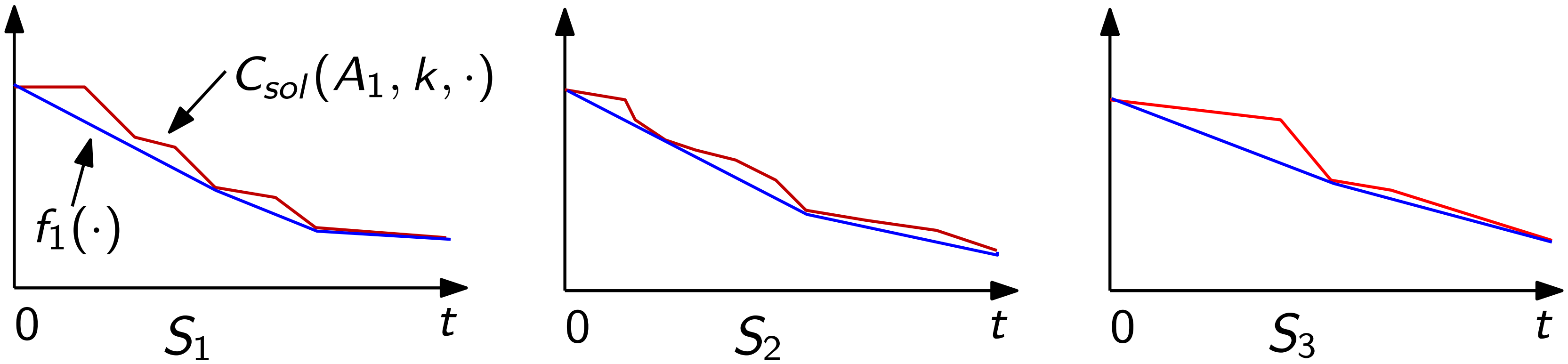
Waterfilling

- Miminize $\sum_{i \in [s]} C_{sol}(A_i, k, t_i)$ s.t. $\sum_{i \in [s]} t_i = t$.
- Convexify the functions



Waterfilling

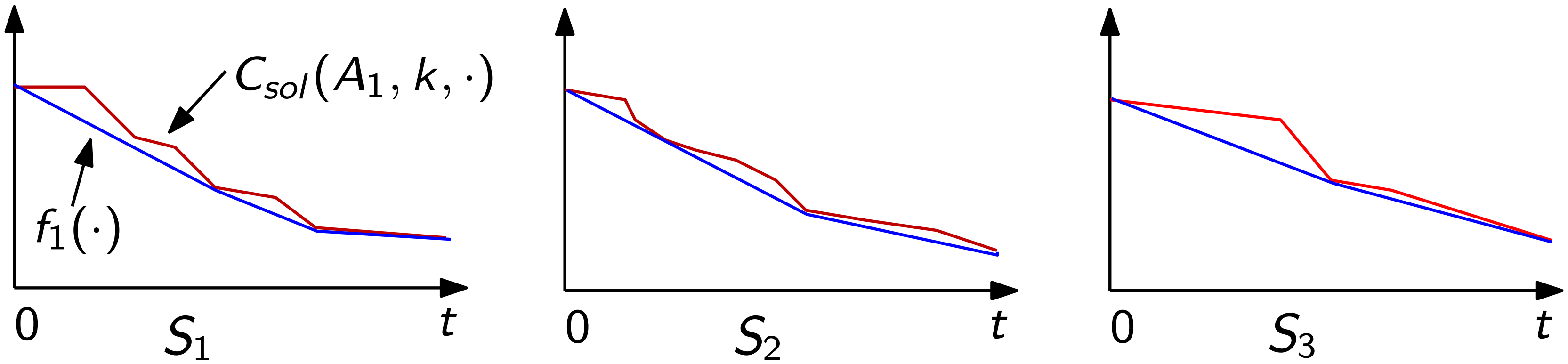
- Miminize $\sum_{i\in[s]} C_{sol}(A_i, k, t_i)$ s.t. $\sum_{i\in[s]} t_i = t$.
- Convexify the functions



Blue curve $f_i(\cdot)$ is the lower convex hull of red curve $C_{sol}(A_i, k, \cdot)$

Waterfilling

- Miminize $\sum_{i \in [s]} C_{sol}(A_i, k, t_i)$ s.t. $\sum_{i \in [s]} t_i = t$.
- Convexify the functions

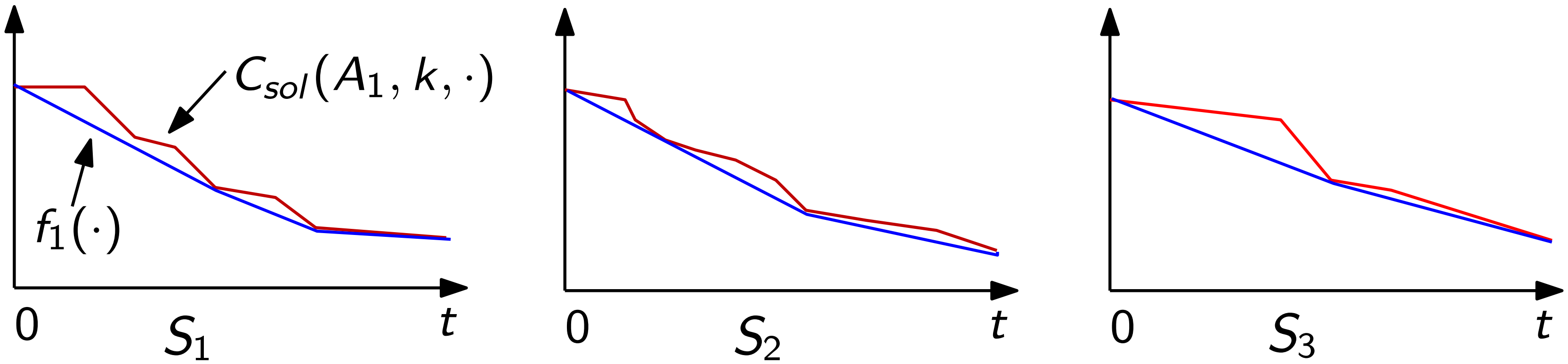


Blue curve $f_i(\cdot)$ is the lower convex hull of red curve $C_{sol}(A_i, k, \cdot)$

Need some work to show that $f_i(\cdot)$ is a good approx of $C_{sol}(A_i, k, \cdot)$

Waterfilling

- Miminize $\sum_{i\in[s]} C_{sol}(A_i, k, t_i)$ s.t. $\sum_{i\in[s]} t_i = t$.
- Convexify the functions



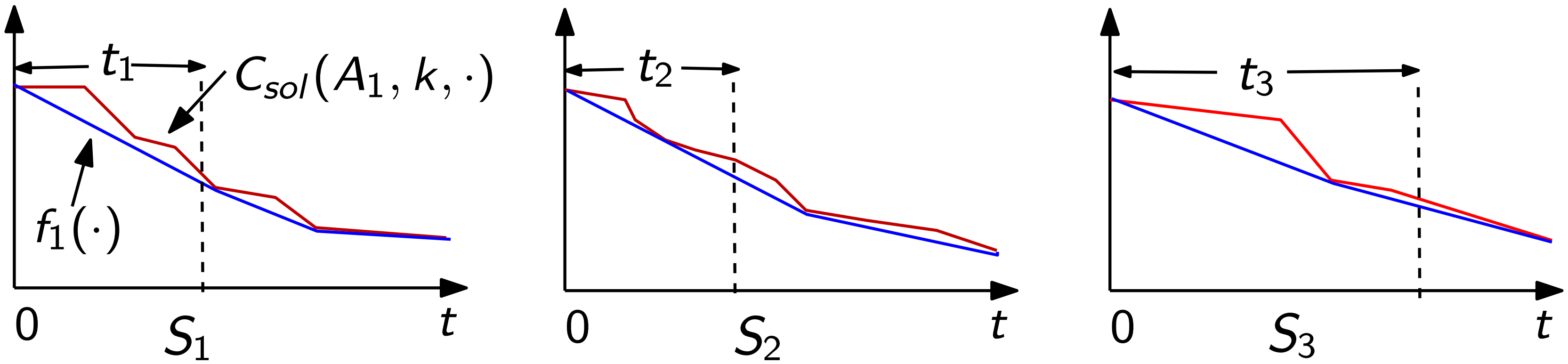
Blue curve $f_i(\cdot)$ is the lower convex hull of red curve $C_{sol}(A_i, k, \cdot)$

Need some work to show that $f_i(\cdot)$ is a good approx of $C_{sol}(A_i, k, \cdot)$

- Sort all "slopes" $\ell(i, q) = f_i(q-1) - f_i(q)$ ($i \in [s]$, $q \in [t]$) and choose the value η of rank t as threshold.
  Note: for a fixed i, $\ell(i, q)$ ($q \in [t]$) are non-increasing

Waterfilling

- Miminize $\sum_{i \in [s]} C_{sol}(A_i, k, t_i)$ s.t. $\sum_{i \in [s]} t_i = t$.
- Convexify the functions



Blue curve $f_i(\cdot)$ is the lower convex hull of red curve $C_{sol}(A_i, k, \cdot)$

Need some work to show that $f_i(\cdot)$ is a good approx of $C_{sol}(A_i, k, \cdot)$

- Sort all "slopes" $\ell(i, q) = f_i(q-1) - f_i(q)$ $(i \in [s],\ q \in [t])$ and choose the value η of rank t as threshold.

  Note: for a fixed i, $\ell(i, q)$ $(q \in [t])$ are non-increasing

t_i is the number of slopes $\ell(i, \cdot)$ at Site i that are at least η.

Two-round algorithm

- Each site i sends $\ell(i, q)$ to coordinator for $q = 1, 2, 3, \ldots, t$
- Coordinator determines the threshold (rank t element) and sends it to sites
- Each site i determines t_i and sends local centers (and their weights) and the t_i outliers.
  Recall that $\sum_{i \in [s]} t_i = t$
- Coordinator solves the (k, t)-median problem on the (weighted) centers and outliers

Two-round algorithm

- Each site i sends $\ell(i, q)$ to coordinator for $q = 1, 2, 3, \ldots, t$
- Coordinator determines the threshold (rank t element) and sends it to sites
- Each site i determines t_i and sends local centers (and their weights) and the t_i outliers.
  Recall that $\sum_{i \in [s]} t_i = t$
- Coordinator solves the (k, t)-median problem on the (weighted) centers and outliers

Comm. cost $\tilde{O}(sk + st)$. No improvement, hmm?

Two-round algorithm (cont.)

- Each site i sends $\ell(i, q)$ to coordinator for $q = 1, 2, 4, 8, \ldots, t$

- Coordinator determines the threshold (rank $2t$ element)
- Each site i determines t_i and sends local centers (and their weights) and the t_i outliers
  Can show that $\sum_{i \in [s]} t_i \leq 3t$
- Coordinator solves the (k, t)-median problem on the (weighted) centers and outliers

Comm. cost $\tilde{O}(sk + t)$.

Two-round algorithm (cont.)

- Each site i sends $\ell(i, q)$ to coordinator for $q = 1, 2, 4, 8, \ldots, t$

  we can do this because $\ell(i, q)$ are non-increasing for a fixed i

- Coordinator determines the threshold (rank $2t$ element)
- Each site i determines t_i and sends local centers (and their weights) and the t_i outliers
  Can show that $\sum_{i \in [s]} t_i \leq 3t$
- Coordinator solves the (k, t)-median problem on the (weighted) centers and outliers

Comm. cost $\tilde{O}(sk + t)$.

Two-round algorithm, bicriteria

- Each site i sends $\ell(i, q)$ to coordinator for $q = 1, 2, 4, 8, \ldots, t$
- Coordinator determines the threshold (rank $2t$ element)
- Each site i determines t_i and sends local centers (with the number of associated points) and the t_i outliers. Can show that $\sum_{i \in [s]} t_i \leq 3t$
- Coordinator solves the $(k, (1+\epsilon)t)$-median problem on the (weighted) centers and ignored points

Comm. cost $\tilde{O}(sk + t)$.

Quadratic time locally

Subquadratic-time centralized algorithm

The reduction: (for (k, t)-median/means)
A $(\gamma, O(1))$-approx centralized algo with time $\tilde{O}(n^{1+\alpha} k^2)$
$\Rightarrow$
A $(O(\gamma), 2)$-approx centralized algo with time $\tilde{O}(t^2 + n^{\frac{2+2\alpha}{2+\alpha}} k^2)$

Subquadratic-time centralized algorithm

The reduction: (for (k,t)-median/means)
A $(\gamma, O(1))$-approx centralized algo with time $\tilde{O}(n^{1+\alpha} k^2)$
$\Rightarrow$
A $(O(\gamma), 2)$-approx centralized algo with time $\tilde{O}(t^2 + n^{\frac{2+2\alpha}{2+\alpha}} k^2)$

- Apply the distributed (k,t)-median/means algo after dividing the set of points arbitrarily into s pieces of size n/s.
- The sequential simulation of the s sites takes time $\tilde{O}(s\,(n/s)^{1+\alpha}\,k^2)$.
- The coordinator requires time $\tilde{O}((sk+t)^2) = \tilde{O}(s^2 k^2) + \tilde{O}(t^2)$.
- Finally balance $n^{1+\alpha} = s^{2+\alpha}$

Subquadratic-time centralized algorithm

The reduction: (for (k,t)-median/means)
A $(\gamma, O(1))$-approx centralized algo with time $\tilde{O}(n^{1+\alpha} k^2)$
$\Rightarrow$
A $(O(\gamma), 2)$-approx centralized algo with time $\tilde{O}(t^2 + n^{\frac{2+2\alpha}{2+\alpha}} k^2)$

- Apply the distributed (k,t)-median/means algo after
  dividing the set of points arbitrarily into s pieces of size n/s.
- The sequential simulation of the s sites takes time $\tilde{O}(s\,(n/s)^{1+\alpha}\,k^2)$.
- The coordinator requires time $\tilde{O}((sk+t)^2) = \tilde{O}(s^2 k^2) + \tilde{O}(t^2)$.
- Finally balance $n^{1+\alpha} = s^{2+\alpha}$

Apply the reduction $O(1)$ times to further reduce the running time to $\tilde{O}(n^{1.01} k^2)$ (assuming $t \leq \sqrt{n}$) at the cost of larger (but still $O(1)$) approx.

Distributed (k, t)-Center

Gonzalez's algorithm

Gonzalez's algorithm for k-center

- Let $S = \{z_1, \ldots, z_n\}$ be a data set
- Choose $z_1 \in S$ arbitrary as the first center. Let $Z_i = \{z_1, \ldots, z_i\}$
- For $i = 2$ to n, set $z_i = \arg\max_{x \in S} d(x, Z_{i-1})$

Get an ordering $z_1, \ldots, z_n$ of S

Two-round algorithm for distributed (k, t)-center

- Site i runs Gonzales's algorithm and obtain a re-ordering $\{a_1, \ldots, a_{n_i}\}$ of the points in A_i
- Site i, for each $1 \leq q \leq t$, computes

$$\ell(i, q) \leftarrow \min_{j < k+q} d(a_j, a_{k+q})$$

- Sites and coordinator sort $\{\ell(i, q)\}$, and then follow the subsequent steps in the previous framework. In the second level clustering we use an algo for k-center with exactly t outliers.

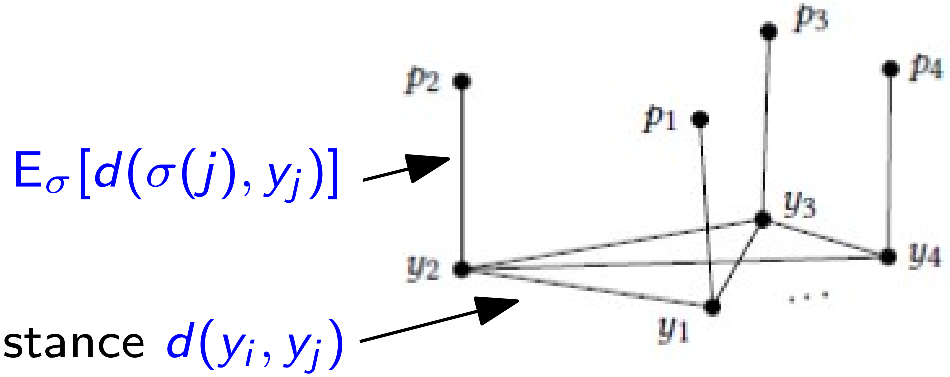
Uncertain Data

Uncertain data – (k, t)-median/means/center-pp

- Reduce the clustering problems to the deterministic case; collapse each node/cloud j to its optimal center.

$$y_j = \arg\min_{y} \mathrm{E}_\sigma[d(\sigma(j), y)]$$

- Fully connect y_j's using the metric distance
- Attach a vertex p_j to y_j, with edge cost $\mathrm{E}_\sigma[d(\sigma(j), y_j)]$
- Apply the previous framework on the compressed graph; $d_G(u, v)$ between $u, v \in G$ is the shortest path distance



Uncertain data – (k, t)-center-global

- Use the idea from [Guha, Munagala, 2009], reduce center to median.
- Use a truncated distance function

  $L_\tau(u, v) = \max\{d(u, v) - \tau, 0\}$
  $\rho_\tau(j, u) = \mathrm{E}_\sigma[L_\tau(\sigma(j), u)]$

- Perform a parametric search on τ, and then apply our previous framework
- Find a τ s.t. $\sum_i C_{sol}(A_i, 2k, t_i(\tau), \rho_\tau) \approx \tau$, where $t_i(\tau)$ is #local outliers at Site i (after applying the previous framework)

Concluding remarks

Summary

- For (k, t)-median/means: $\tilde{O}(sk + t)$ communication, 2 rounds, $O(1)$-approx, ($t(1 + \epsilon)$ outliers for k-means).
- A subquadratic time $(O(1), O(1))$-approx centralized algorithms for (k, t)-median/means
- Can handle uncertain data cases with similar comm. and round costs.

Open problems

- Lower bounds. $\Omega(sk + t)$ for (k, t)-median/means/center if the algo needs to output all the outliers. What if not?
- Better approximation ratios?

Thank you!
Questions?